Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Compute Less to Get More: Using ORC to Improve Sparse Filtering

Johannes Lederer
Department of Statistical Science
Cornell University
Ithaca, NY 14853
johanneslederer@cornell.edu

Sergio Guadarrama
EECS Department
University of California
Berkeley, CA 94720
sergio.guadarrama@berkeley.edu

Abstract

Sparse Filtering is a popular feature learning algorithm for image classification pipelines. In this paper, we connect the performance of Sparse Filtering with spectral properties of the corresponding feature matrices. This connection provides new insights into Sparse Filtering; in particular, it suggests early stopping of Sparse Filtering. We therefore introduce the Optimal Roundness Criterion (ORC), a novel stopping criterion for Sparse Filtering. We show that this stopping criterion is related with pre-processing procedures such as Statistical Whitening and demonstrate that it can make image classification with Sparse Filtering considerably faster and more accurate.

Introduction

Standard ways to improve image classification are to collect more samples or to change the representation and the processing of the data. In practice, the number of samples is typically limited, so that the second approach becomes relevant. An important tool for this second approach are feature learning algorithms, which aim at easing the classification task by transforming the data. Recently proposed deep learning methods intend to jointly learn learn a feature transformation and the classification (Krizhevsky, Sutskever, and Hinton 2012). In this work, however, we focus on unsupervised feature learning, especially on Sparse Filtering, because of their simplicity and scalability.

Feature learning algorithms for image classification pipelines typically consists of three steps: pre-processing, (un)supervised dictionary learning, and encoding. An abundance of procedures is available for each of these steps, but for accurate image classification, we need procedures that are effective and interact beneficially with each other (Agarwal and Triggs 2006; Coates and Ng 2011; Coates, Ng, and Lee 2011; Jia, Huang, and Darrell 2012; Le 2013; LeCun, Huang, and Bottou 2004). Therefore, a profound understanding of these procedures is crucial to ensure accurate results and efficient computations.

In this paper, we study the performance of Sparse Filtering (Ngiam et al. 2011) for image classification. Our main contributions are:

- we show that Sparse Filtering can strongly benefit from early stopping;
- we show that the performance of Sparse Filtering is correlated with spectral properties of feature matrices on tests sets;
- we introduce the Optimal Roundness Criterion (ORC), a stopping criterion for Sparse Filtering based on the above correlation, and demonstrate that the ORC can considerably improve image classification.

Feature Learning for Image Classification

Feature learning algorithms often consist of two steps: In a first step, a dictionary is learned, and in a second step, the samples are encoded based on this dictionary. A typical dictionary learning step for image classification is sketched in Figure 1: First, random patches (samples) are extracted from the training images. These patches are then pre-processed using, for example, Statistical Whitening or Contrast Normalization. Finally, an unsupervised learning algorithm is applied to learn a dictionary from the pre-processed patches. Once a dictionary is learnt, several further steps need to be applied to finally train an image classifier, see, for example, (Coates and Ng 2011; Coates, Ng, and Lee 2011; Jia, Huang, and Darrell 2012; Le 2013). Our pipeline is similar to the one in (Coates and Ng 2011): We extract square patches comprising 9×9 pixels, pre-process them with Contrast Normalization[1] and/or Statistical Whitening, and finally pass them to Random Patches or Sparse Filtering. (Note that our outcomes differ slightly from those in (Coates and Ng 2011) because we use square patches comprising 9×9 pixels instead of 6×6 pixels.) Subsequently, we apply soft-thresholding for encoding, 4×4 spatial max pooling for extracting features from the training data images, and finally L2 SVM classification (cf. (Coates and Ng 2011)).

Numerous examples show that feature learning can considerably improve classification. Therefore, insight in the underlying principles of feature learning algorithms such as Statistical Whitening and Sparse Filtering is of great interest.

[1] Contrast normalization consists of subtracting the mean and dividing by the standard deviation of the pixel values.

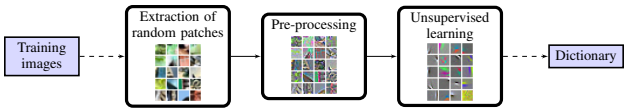

Figure 1: A typical dictionary learning step. Statistical Whitening and Contrast Normalization are examples for pre-processing procedures; Random Patches and Sparse Filtering are examples for unsupervised learning procedures.

In mathematical terms, a feature learning algorithm provides a transformation

$$\begin{aligned} \mathcal{F} : \mathbb{R}^{l\times p} &\to \mathbb{R}^{n\times p} \\ X &\mapsto \mathcal{F}(X) \end{aligned} \tag{1}$$

of an original feature matrix $X \in \mathbb{R}^{l\times p}$ to a new feature matrix $\mathcal{F}(X) \in \mathbb{R}^{n\times p}$. We adopt the convention that the rows of the matrices correspond to the features, the columns to the samples; this convention implies in particular that $l \in \mathbb{N}$ is the number of original features, $p \in \mathbb{N}$ the number of samples, and $n \in \mathbb{N}$ the number of new features.

The Optimal Roundness Criterion

Roundness of Feature Matrices

Feature learning can be seen as trade-off between reducing the correlations of the feature representation and preservation of relevant information. This trade-off can be readily understood looking at Statistical Whitening. For this, recall that pre-processing with Statistical Whitening transforms a set of image patches into a new set of patches by changing the local correlation structure. More precisely, Statistical Whitening transforms patches $X_{\text{Patch}} \in \mathbb{R}^{n'\times p}$ ($n' < n$), that is, subsets of the entire feature matrix, into new patches $\mathcal{F}_{\text{Patch}}(X_{\text{Patch}})$ such that

$$\mathcal{F}_{\text{Patch}}(X_{\text{Patch}})^T \mathcal{F}_{\text{Patch}}(X_{\text{Patch}}) = n' \, \mathrm{I}_{n'} \, .$$

Statistical Whitening therefore acts locally: while the correlation structures of the single patches are directly and radically changed, the structure of the entire matrix is affected only indirectly. However, these indirect effects on the entire matrix are important for the following. To capture these effects, we therefore introduce the roundness of a feature matrix $F := \mathcal{F}(X)$ given an original feature matrix X. On a high level, we say that the new feature matrix F is round if the spectrum of the associated Gram matrix $FF^T \in \mathbb{R}^{n\times n}$ is narrow. To specify this notion, we denote the ordered eigenvalues of FF^T by $\sigma_1(F) \geq \cdots \geq \sigma_n(F) \geq 0$ and their mean by $\overline{\sigma}(F) := \frac{1}{n}\sum_{i=1}^{n} \sigma_i(F)$ and define roundness as follows:

Definition 1. *For any matrix $F \neq 0$, we define its roundness as*

$$r(F) := \frac{\overline{\sigma}(F)}{\sigma_1(F)} \in [0, 1].$$

The largest eigenvalue σ_1 measures the width of the spectrum of the Gram matrix; alternative measures of the width such as the standard deviation of the eigenvalues would serve the same purpose. The mean of the eigenvalues $\overline{\sigma}$, on the other hand, is basically a normalization as the following result illustrates (the proof is found in the supplementary material):

Theorem 1. *Denote the columns of F by $F^1, \ldots, F^p$. Then, the mean $\overline{\sigma}(F)$ of the eigenvalues of the Gram matrix FF^T is constant on $\mathcal{S} := \{F \in \mathbb{R}^{n\times p} : \|F^1\|_2 = \cdots = \|F^p\|_2 = 1\}$:*

$$\overline{\sigma}(F) := \frac{p}{n} \quad \textit{for all } F \in \mathcal{S}.$$

Definition 1 therefore states that the larger is r, the narrower is the spectrum of the eigenvalues of the Gram matrix of F, and therefore, the rounder is the matrix F. With this notion of roundness at hand, we can now understand the effects of Statistical Whitening: On the one hand, Definition 1 indicates that Statistical Whitening renders single patches perfectly round, that is, $r(\mathcal{F}_{\text{Patch}}) = 1$. On the other hand, Statistical Whitening preserves global structures in the feature matrix. In particular, the entire feature matrix is made rounder but *not* rendered perfectly round, that is, $r(\mathcal{F}(X)) < 1$. In this sense, Statistical Whitening can be seen as trade-off between increasing of roundness and preservation of global structures. It therefore remains to connect roundness and randomization.

Roundness and Randomness

A connection between roundness and randomization is provided by random matrix theory. To illustrate this connection, we first recall Gordon's theorem for Gaussian random matrices (see (Eldar and Kutyniok 2012, Chapter 5) for a recent introduction to random matrix theory):

Theorem 2 (Gordon). *Let $F \in \mathbb{R}^{n\times p}$ be a random matrix with independent standard normal entries. Then,*

$$\begin{aligned} 1 - \sqrt{n/p} \leq \mathbb{E}\left[\sqrt{\sigma_n(F)/p}\right] \\ \leq \mathbb{E}\left[\sqrt{\sigma_1(F)/p}\right] \leq 1 + \sqrt{n/p}. \end{aligned}$$

Such exact bounds are available only for matrices with independent standard normal entries, but sharp bounds in probability are available also for other random matrices. For our purposes, the common message of all these bounds is that random matrices with sufficiently many columns (number of samples) have a small spectrum. This means in particular that such matrices are round as the following asymptotic result illustrates (the proof is based on well-known results from random matrix theory and therefore omitted):

Lemma 1. *Let the number of features $n \equiv n(p)$ be a function of the number of samples p such that $n/p \to 0$. Moreover, for all $p \in \{1, 2, \ldots\}$, let $F \equiv F(p)$ be a random matrix with independent standard normal entries. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left(|r(F) - 1| > \epsilon\right) \to 0 \quad \textit{for } p \to \infty,$$

that is, $r(F)$ converges in probability to 1.

Similar results can be derived for non-Gaussian or correlated entries, indicating that random matrices are typically round.

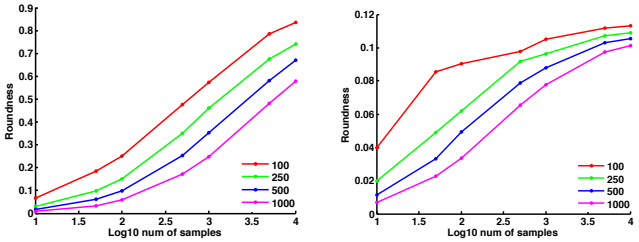

Figure 2: Roundness of random feature matrices with columns drawn from a normal distribution with mean zero and a Toeplitz covariance matrix. For the left plot, $\rho = 0$ (uncorrelated entries). For the right plot, $\rho = 0.8$ (correlations that relate to natural images). The roundness is plotted as a function of the number of samples $p \in [10, 10^4]$ for four different numbers of features $n \in \{100, 250, 500, 1000\}$.

Besides the connection between roundness and randomization, the above results for random matrices also provide a link between roundness and sample sizes. Indeed, we observe that the above results indicate that large samples sizes lead to round matrices. To make this link more tangible, we conduct simulations with Toeplitz matrices, which can model local correlations that are typical for nearby pixels in natural images (Girshick and Malik 2013). To this end, we first recall that for any fixed parameter $\rho \in [0, 1)$, the entries of a Toeplitz matrix T_ρ are defined as[2] $(T_\rho)_{ij} := \rho^{|i-j|}$. We now construct a feature matrix U_ρ by drawing each of its columns, that is, the samples, from the normal distribution with mean zero and covariance matrix $T_\rho \in \mathbb{R}^{n \times n}$. Toeplitz matrices with $\rho = 0$ lead to feature matrices with independent entries; Toeplitz matrices with $\rho = 0.8$ lead to feature matrices with dependence structures that are more similar to dependence structures found in natural images. In Figure 2, we report the roundness of U_ρ for $\rho = 0$ (plot on the left) and $\rho = 0.8$ (plot on the right) as a function of the numbers of samples p for different numbers of features n. The results are commensurate with the theoretical findings above: First, both plots illustrate that the roundness of matrices increases if the number of samples is increased but decreases if the number of features is increased (cf. Theorem 2 and Lemma 1). Second, a comparison of the two plots illustrate that the roundness is larger for $\rho = 0$ than for $\rho = 0.8$ (since T_0 is perfectly round while $T_{0.8}$ is not).

Optimal Roundness Criterion (ORC)

The above discussion suggest that optimal feature learning is the result of a trade-off between increasing the roundness of the feature matrix and preserving global structures in the data. In this part, we want to exploit this insight to understand and improve iterative feature learning algorithms. Common feature learning algorithms consist of transformations that are defined as minimizers of a functional. These functionals are then often computed iteratively via a sequence of gradient based operations. In this paper, we therefore focus on feature learning algorithms where the transformation $\mathcal{F}$ as in (1) is the limit of a sequence of transformations $(\mathcal{G}_k)_{k \in \mathbb{N}}$, that is,

$$\mathcal{F} = \lim_{k \to \infty} \mathcal{G}_k,$$

where for all $k \in \mathbb{N}$,

$$\begin{aligned} \mathcal{G}_k : \mathbb{R}^{l \times p} &\to \mathbb{R}^{n \times p} \\ X &\mapsto \mathcal{G}_k(X). \end{aligned}$$

A prominent representative of such iterative algorithms is Sparse Filtering. Sparse Filtering consists of normalizations and the minimization of an ℓ_1-criterion (see next Section). It is reasonable to assume that these operations - similar to the local changes by Statistical Whitening - preserve certain global structures of the feature matrix. In view of a trade-off between roundness and preservation of global structures, we are therefore interested in stopping the iterations as soon as the roundness is maximized. More formally, we introduce the ORC, which serves as stopping criterion to maximize the roundness:

Definition 2. *Let r be the roundness introduced in Definition 1. The Optimal Roundness Criterion (ORC) replaces the transformation $\mathcal{F}$ by*

$$\widehat{\mathcal{G}} := \mathcal{G}_{\widehat{k}}$$

for

$$\widehat{k} := \arg\max_{k \in \mathbb{N}} \{ r(\mathcal{G}_{k'}) < r(\mathcal{G}_k) \text{ for all } k' < k \}$$

if the arg-maximum is finite and $\widehat{\mathcal{G}} := \mathcal{F}$ otherwise.

The ORC assures that the computations continue only as long as the roundness increases. Assuming that certain global structures are preserved by the transformations, the ORC provides an optimization scheme for the performance of iterative feature learning algorithms. One could also think of modifications of the ORC that include an additive constant or a factor to force larger increases or to allow for temporary decreases of the roundness.

Image Classification on CIFAR-10

For our all experiments, we use the CIFAR-10 dataset (Krizhevsky and Hinton 2009)[3]. This dataset consists of 60 000 color images partitioned into 10 classes, each containing 6 000 images. Each of the images comprises 32×32 pixels. The dataset is split into a training set with 50 000 images and a test set with 10 000 images. From the training set, we randomly select 10 000 patches for the unsupervised feature learning. These patches are also used to determine the parameters of Contrast Normalization and Statistical Whitening (if applied).

[2] We set $0^k := 1$ for $k = 0$ and $0^k := 0$ for $k \neq 0$.

[3] http://www.cs.toronto.edu/~kriz/cifar.html

Num	Norm.	White.	Round.	Acc.
243	No	No	0.0041	32.50%
243	Yes	No	0.0131	63.65%
243	No	Yes	0.2080	65.01%
243	Yes	Yes	0.1548	64.34%
486	No	No	0.0021	31.67%
486	Yes	No	0.0062	66.14%
486	No	Yes	0.1134	67.08%
486	Yes	Yes	0.0965	67.84%

Table 1: Roundness and accuracy of Random Patches with and without Contrast Normalization and Statistical Whitening.

Num	ORC	Round.	Acc.
243	No	0.0519	57.66%
243	Yes	0.1425	62.47%
486	No	0.0495	58.19%
486	Yes	0.0908	63.80%

Table 2: Roundness and accuracy of Sparse Filtering with and without early stopping based on the ORC.

Random Patches

For the dictionary learning step, it was shown that simple randomized procedures combined with Statistical Whitening work surprisingly well (Coates and Ng 2011; Jarrett et al. 2009; Saxe et al. 2011). A popular example is Random Patches, which creates a dictionary matrix by simply stacking up randomly selected samples. In Table 1, we report the influence of Contrast Normalization and Statistical Whitening on Random Patches (cf. (Coates and Ng 2011)). We see that Statistical Whitening is very beneficial for Random Patches and increases the roundness of the transformed feature matrix. This suggests that the roundness can be used as an indicator for the performance of feature learning. (Note that the roundness is on different scales for different numbers of features and can therefore not be compared for different numbers of features.)

Sparse Filtering

Sparse Filtering (Ngiam et al. 2011) is an unsupervised feature learning algorithm that computationally scales particularly well with the dimensions. To recall the definition of Sparse Filtering, we denote by $\mathcal{N} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ the function that first normalizes[4] the rows of a matrix in $\mathbb{R}^{n\times p}$ to unit Euclidean norm and then normalizes the columns of the resulting matrix to unit Euclidean norm. For any fixed matrix $X \in \mathbb{R}^{l\times p}$, we then define a matrix $W_X \in \mathbb{R}^{n\times l}$

[4]We set $0/0 := 0 * \infty := \infty$ in the corresponding operations ensure that (2) is well defined.

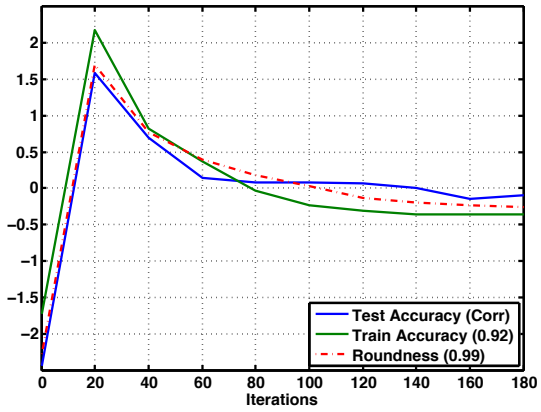

Figure 3: Intermediate outcomes of Sparse Filtering at every 20 iterations. Test accuracy (blue, solid curve), training accuracy (green, solid curve), roundness on the training set (red, dashed curve), and correlations with the test accuracy (numbers in brackets). To enhance visibility, all curves are normalized to have zero mean and unit standard deviation.

such that

$$W_X \in \operatorname*{arg\,min}_{W\in\mathbb{R}^{n\times l}} \|\mathcal{N}(WX)\|_1 \tag{2}$$

if the minimum is finite and $W_X := 0$ otherwise. Sparse Filtering is then the transformation

$$\begin{aligned} \mathcal{F}_{SF} : \mathbb{R}^{l\times p} &\to \mathbb{R}^{n\times p} \\ X &\mapsto \mathcal{F}_{SF}(X) := W_X X. \end{aligned} \qquad \text{(Sparse Filtering)}$$

However, we now show by making the normalizations explicit that these normalizations make Sparse Filtering intricate. For this, we define the rank one matrices $E_1, \ldots, E_n \in \mathbb{R}^{n\times n}$ via

$$(E_i)_{kl} := \delta_{kl}\delta_{ik} \qquad \forall i,k,l \in \{1,\ldots,n\}$$

and $G_1, \ldots, G_p \in \mathbb{R}^{p\times p}$ via

$$(G_i)_{kl} := \delta_{kl}\delta_{ik} \qquad \forall i,k,l \in \{1,\ldots,p\}$$

where δ is the usual Kronecker delta. This then yields the following form of Definition (2).

Theorem 3. *The matrix W_X in (2) is the minimizer of*

$$\left\| \left[\sum_{i=1}^{n} E_i WX(WX)^T E_i\right]^{-\frac{1}{2}} WX \times \right.$$

$$\left. \times \left[\sum_{i=1}^{p} G_i (WX)^T \left[\sum_{i=1}^{n} E_i WX(WX)^T E_i\right]^{-1} WXG_i\right]^{-\frac{1}{2}} \right\|_1$$

over all matrices $W \in \mathbb{R}^{l\times n}$.

Although Sparse Filtering is sometimes claimed to have sparsity properties due to the involvement of the ℓ_1-norm (similar as the Lasso (Tibshirani 1996), for example), the above reformulation demonstrates that this is far from obvious and needs further clarification.

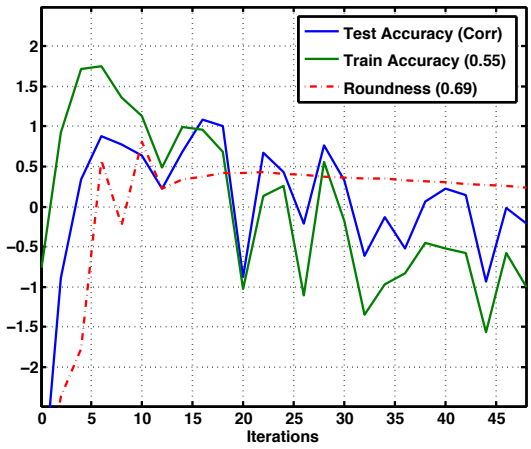

Figure 4: Intermediate outcomes of Sparse Filtering at every 2 iterations. Test accuracy (blue, solid curve), training accuracy (green, solid curve), roundness on the training set (red, dashed curve), and correlations with the test accuracy (numbers in brackets). To enhance visibility, all curves are normalized to have zero mean and unit standard deviation.

It is apparent that the choice of the number of features n influences the performance of Sparse Filtering. As can be seen in Figure /refWe will see below, however, that the choice of the number of iterations surprisingly can have an even larger influence. We are therefore interested in choosing an appropriate number of iterations. A standard approach would involve l-fold cross-validation schemes, but this requires training of l models and is therefore computationally costly. The ORC, on the other hand, can be a computationally feasible alternative to cross-validation. To illustrate this, we compare in Table 2 the outcomes of Sparse Filtering on the CIFAR-10 dataset with and without application of the ORC. We have also computed the intermediate outcomes of Sparse Filtering at every 20 iterations and report in Figure 3 the corresponding test accuracy, training accuracy, roundness on the training set, and correlations with the test accuracy. The roundness on the test set is basically indistinguishable from the roundness on the training set and is therefore not shown. We make three crucial observations: *(i)* the test accuracy of Sparse Filtering peaks at around 20 iterations and then decreases monotonically; *(ii)* the roundness on the training set is highly correlated with the test accuracy; in particular, the locations of the peaks of these curves coincide; *(iii)* the roundness on the training set is highly correlated with the roundness on the test set. These observations suggest that *(i)* Sparse Filtering should be stopped early; *(ii)* the ORC can optimize the performance of Sparse Filtering; *(iii)* it is sufficient to compute the roundness on the training set. To further support these claims, we have also computed the intermediate outcomes of Sparse Filtering at every 2 iterations in the region around the peaks, that is, we have computed a zoomed-in version of Figure 3. We report the results in Figure 4. We observe that training accuracy, test accuracy, and roundness are highly correlated, which corroborates the above claims and therefore confirms the potential of the ORC. We finally note that the curves in the zoomed-in version are wiggly not only because of the randomness involved but also because computations of gradients over a small number of iterations involve numerical imprecisions.

Conclusions and Outlook

The spectral analysis of feature matrices is a novel and promising approach to feature learning. In particular, our results show that this "geometric" approach can provide new interpretations and substantial improvements of wide-spread feature learning tools such as Statistical Whitening, Random Patches, and Sparse Filtering. For example, we have revealed that Sparse Filtering can, quite surprisingly, deteriorate with increasing number of iterations and can be made considerably faster and more accurate by early stopping according to the spectrum of the intermediate feature matrices.

Regarding the theory, it would be of interest to obtain, for specific procedures, predictions on how the roundness changes with the iterations and to what it converges in the limit.

In an extended version of this paper, we are planning to include an analysis of Roundness in Convolutional Neural Networks (CNNs) (Fukushima 1980). After being neglected for many years, CNNs have received an enormous deal of attention recently, see (Krizhevsky, Sutskever, and Hinton 2012; Girshick et al. 2013) and many others. We therefore expect that the application of our approach to CNNs can be of substantial interest.

Acknowledgments

We thank the reviewers for their insightful comments.

Appendix: Proofs

We denote in the following the columns and rows of any matrix M by $M^1, \dots, M^p$ and $M_1, \dots, M_n$, respectively.

Proof of Theorem 1. The matrix FF^T is symmetric and can therefore be diagonalized. This implies that there is an orthogonal orthogonal matrix $A \in \mathbb{R}^{n\times n}$ such that the diagonal entries of $AFF^TA^T \in \mathbb{R}^{n\times n}$ are $\sigma_1(F), \dots, \sigma_n(F)$. For this matrix A, it then holds

$$\begin{aligned}\overline{\sigma}(F) =& \frac{1}{n}\sum_{i=1}^{n}\sigma_i(F)\\ =& \frac{1}{n}\sum_{i=1}^{n}(AFF^TA^T)_{ii}\\ =& \frac{1}{n}\operatorname{trace}(AFF^TA^T).\end{aligned}$$

Next, we invoke the cyclic property of the trace and the orthogonality of the matrix A to obtain

$$\begin{aligned}\operatorname{trace}(AFF^TA^T) =& \operatorname{trace}(F^TA^TAF)\\ =& \operatorname{trace}(F^TF)\\ =& \sum_{i=1}^{p}(F^TF)_{ii}.\end{aligned}$$

Finally, we note that the normalization of the columns of F yields

$$(F^TF)_{ii} = (F^i)^TF^i = \|F^i\|_2^2 = 1$$

for all $i \in \{1,\dots,p\}$. The desired result

$$\overline{\sigma}(F) = \frac{1}{n}\sum_{i=1}^{p} 1 = \frac{p}{n}$$

can now derived combining the three displays. □

Proof of Theorem 3. We first show that for a matrix $A \in \mathbb{R}^{n\times p}$, the corresponding matrix $A^R \in \mathbb{R}^{n\times p}$ with normalized rows can be written as

$$A^R = \left[\sum_{i=1}^{n} E_i A A^T E_i\right]^{\text{-}1/2} A. \qquad \text{(Claim 1)}$$

To this end, we observe that the normalization of the rows of the matrix A corresponds to the matrix multiplication

$$A^R = D^R A,$$

where $D^R \in \mathbb{R}^{n\times n}$ is the diagonal matrix with nonzero entries

$$(D^R)_{ii} = 1/\sqrt{\sum_{j=1}^{p}(A_{ij})^2} \qquad \forall i \in \{1,\dots,n\}.$$

Next, we note that

$$\sum_{j=1}^{p}(A_{ij})^2 = (AA^T)_{ii} \qquad \forall i \in \{1,\dots,n\}$$

and therefore

$$((D^R)^{\text{-}1})_{ii} = 1/(D^R)_{ii} = \sqrt{(AA^T)_{ii}} \qquad \forall i \in \{1,\dots,n\}.$$

This yields the matrix equation

$$(D^R)^{\text{-}2} = \sum_{i=1}^{n} E_i A A^T E_i$$

and therefore

$$D^R = \left[\sum_{i=1}^{n} E_i A A^T E_i\right]^{\text{-}1/2}.$$

This proves the first claim.

We now show that for a matrix $B \in \mathbb{R}^{n\times p}$, the corresponding matrix $B^C \in \mathbb{R}^{n\times p}$ with normalized columns is given by

$$B^C = B\left[\sum_{i=1}^{p} G_i B^T B G_i\right]^{\text{-}1/2}. \qquad \text{(Claim 2)}$$

We first note that we can write the normalization step - this time for the columns - as the matrix multiplication

$$B^C = BD^C,$$

where $D^C \in \mathbb{R}^{p\times p}$ is the diagonal matrix with entries

$$(D^C)_{ii} = 1/\sqrt{\sum_{j=1}^{n}(B_{ji})^2} \qquad \forall i \in \{1,\dots,p\}.$$

Next, we note that

$$\sum_{j=1}^{n}(B_{ji})^2 = (B^TB)_{ii} \qquad \forall i \in \{1,\dots,p\},$$

and therefore for the inverse diagonal matrix

$$((D^C)^{\text{-}1})_{ii} = 1/(D^C)_{ii} = \sqrt{(B^TB)_{ii}} \qquad \forall i \in \{1,\dots,p\}.$$

This yields the matrix equation

$$(D^C)^{\text{-}2} = \sum_{i=1}^{p} G_i B^T B G_i$$

and therefore

$$D^C = \left[\sum_{i=1}^{p} G_i B^T B G_i\right]^{\text{-}1/2}.$$

This proves the second claim.

We now consider $F_W := WX \in \mathbb{R}^{n\times p}$ for an arbitrary matrix $W \in \mathbb{R}^{n\times p}$ and apply Claim 1 and Claim 2: Setting $A = F_W$, we obtain from Claim 1 that normalizing the rows of the matrix F_W yields the matrix $F_W^R \in \mathbb{R}^{n\times p}$ given by

$$F_W^R := \left[\sum_{i=1}^{n} E_i F_W F_W^T E_i\right]^{\text{-}1/2} F_W.$$

This implies in particular

$$(F_W^R)^T F_W^R = F_W^T\left[\sum_{i=1}^{n} E_i F_W F_W^T E_i\right]^{\text{-}1} F_W.$$

Setting then $B = F_W^R$, we obtain from Claim 2 and the two previous displays that the matrix F_W becomes after normalizing its rows and then its columns the matrix

$$\left[\sum_{i=1}^{n} E_i F_W F_W^T E_i\right]^{\text{-}1/2} F_W \times$$
$$\times \left[\sum_{i=1}^{p} G_i F_W^T \left[\sum_{i=1}^{n} E_i F_W F_W^T E_i\right]^{\text{-}1} F_W G_i\right]^{\text{-}1/2}.$$

The desired result can then be deduced from the definition of W_X in (2). □

Appendix

We also present numerical outcomes for a different random splitting of the CIFAR-10 dataset. In particular, we recompute Figures 3 and 4 for a different splitting and give the results in Figures 5 and 6 below. The conclusions are virtually the same as above, which further corroborates our findings.

References

Agarwal, A., and Triggs, B. 2006. Hyperfeatures–multilevel local coding for visual recognition. In *European Conference for Computer Vision (ECCV)*. 30–43.

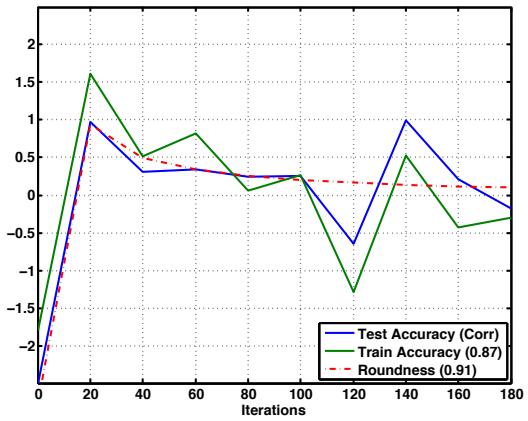

Figure 5: Intermediate outcomes of Sparse Filtering at every 20 iterations. Test accuracy (blue, solid curve), training accuracy (green, solid curve), roundness on the training set (red, dashed curve), and correlations with the test accuracy (numbers in brackets). To enhance visibility, all curves are normalized to have zero mean and unit standard deviation.

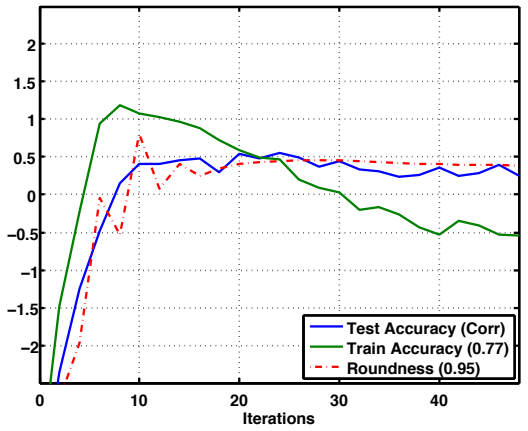

Figure 6: Intermediate outcomes of Sparse Filtering at every 2 iterations. Test accuracy (blue, solid curve), training accuracy (green, solid curve), roundness on the training set (red, dashed curve), and correlations with the test accuracy (numbers in brackets). To enhance visibility, all curves are normalized to have zero mean and unit standard deviation.

Coates, A., and Ng, A. 2011. The importance of encoding versus training with sparse coding and vector quantization. In *28th International Conference on Machine Learning (ICML '11)*, 921–928.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 215–223.

Eldar, Y., and Kutyniok, G., eds. 2012. *Compressed Sensing: Theory and Applications*. Cambridge University Press.

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36(4):193–202.

Girshick, R., and Malik, J. 2013. Training deformable part models with decorrelated features. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *preprint* arxiv:1311.2524.

Jarrett, K.; Kavukcuoglu, K.; Ranzato, M.; and LeCun, Y. 2009. What is the Best Multi-Stage Architecture for Object Recognition? In *IEEE International Conference on Computer Vision (ICCV)*, 2146–2153.

Jia, Y.; Huang, C.; and Darrell, T. 2012. Beyond spatial pyramids: Receptive field learning for pooled image features. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 3370–3377.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Tech. Rep. Computer Science Department, University of Toronto*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. 1097–1105.

Le, Q. 2013. Building high-level features using large scale unsupervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8595–8598. IEEE.

LeCun, Y.; Huang, F.; and Bottou, L. 2004. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition (CVPR)*, II–97.

Ngiam, J.; Koh, P. W.; Chen, Z.; Bhaskar, S.; and Ng, A. 2011. Sparse filtering. In *Advances in Neural Information Processing Systems (NIPS)*, 1125–1133.

Saxe, A.; Koh, P.; Chen, Z.; Bhand, M.; Suresh, B.; and Ng, A. 2011. On random weights and unsupervised feature learning. In *28th International Conference on Machine Learning (ICML '11)*, 1089–1096.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58(1):267–288.