Sparsity-Based Generalization Bounds for Predictive Sparse Coding

Nishant A. Mehta NICHE@CC.GATECH.EDU
Alexander G. Gray AGRAY@CC.GATECH.EDU
College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract

The goal of predictive sparse coding is to learn a representation of examples as sparse linear combinations of elements from a dictionary, such that a learned hypothesis linear in the new representation performs well on a predictive task. Predictive sparse coding has demonstrated impressive performance on a variety of supervised tasks, but its generalization properties have not been studied. We establish the first generalization error bounds for predictive sparse coding, in the overcomplete setting, where the number of features k exceeds the original dimensionality d. The learning bound decays as $\tilde{O}(\sqrt{dk/m})$ with respect to d, k, and the size m of the training sample. It depends intimately on stability properties of the learned sparse encoder, as measured on the training sample. Consequently, we also present a fundamental stability result for the LASSO, a result that characterizes the stability of the sparse codes with respect to dictionary perturbations.

1. Introduction

Learning architectures such as the support vector machine and other linear predictors enjoy strong theoretical properties (Steinwart & Christmann, 2008; Kakade et al., 2009), but a learning-theoretic view of many more complex learning architectures is lacking. Predictive methods based on *sparse coding* recently have emerged which simultaneously learn a data representation via a nonlinear encoding scheme and an estimator linear in that representation (Bradley & Bagnell, 2009; Mairal et al., 2009; 2012). A sparse coding representation $z \in \mathbb{R}^k$ of a data point $x \in \mathbb{R}^d$ is learned by representing x as a sparse linear combination of k *atoms* $D_j \in \mathbb{R}^d$ of a *dictionary* $D = (D_1, \ldots, D_k) \in \mathbb{R}^{d \times k}$. In the coding $x \approx \sum_{j=1}^k z_j D_j$, all but a few z_j are zero.

Predictive sparse coding methods such as Mairal et al.'s (2012) *task-driven dictionary learning* recently have achieved state-of-the-art results on many tasks, including the MNIST digits task. Whereas standard sparse coding minimizes an unsupervised, reconstructive ℓ_2 loss, predictive sparse coding seeks to minimize a supervised loss by learning a dictionary and a linear predictor in the space of codes induced by that dictionary. There is much empirical evidence that sparse coding can provide good abstraction by finding higher-level representations which are useful in predictive tasks (Yu et al., 2009). Intuitively, the power of prediction-driven dictionaries is that they pack more atoms in parts of the representational space where the prediction task is more difficult. However, despite the empirical successes of predictive sparse coding, it is unknown how well it generalizes in a theoretical sense.

In this work, we develop what to our knowledge are the first generalization error bounds for predictive sparse coding algorithms; in particular, we focus on ℓ_1-*regularized sparse coding*. Maurer & Pontil (2010) and Vainsencher et al. (2011) previously established generalization bounds for the classical, reconstructive sparse coding setting. Extending their analysis to the predictive setting introduces certain difficulties related to the complexity of the class of sparse encoders. Whereas in the reconstructive setting, this complexity can be controlled directly by exploiting the stability of the *reconstruction error* to dictionary perturbations, in the predictive setting it appears that the complexity hinges upon the stability of the *sparse codes themselves* to dictionary perturbations. This latter notion of stability is much harder to prove; moreover, it can be realized only with additional assumptions which depend on the dictionary, the data, and their interaction (see Theorem 4). Furthermore, when the assumptions hold for the learned dictionary and data, we also need to guarantee that the assumptions hold on a newly drawn sample.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

Contributions We provide a learning bound for the *overcomplete setting* in predictive sparse coding, where the dictionary size, or number of learned features, k exceeds the ambient dimension d. The bound holds provided the size m of the training sample is large enough, where the critical size for the bound to kick in depends on a certain notion of stability of the learned representation. This work's core contributions are:

- Under mild conditions, a stability bound for the LASSO (Tibshirani, 1996) under dictionary perturbations (Theorem 4).
- In the overcomplete setting, a learning bound that is essentially of order $\sqrt{\frac{dk}{m}} + \frac{\sqrt{s}}{\lambda \mu_s(D)}$, where each sparse code has at most s non-zero components (Theorem 5). The term $\frac{1}{\mu_s(D)}$ is the inverse s-incoherence (see Definition 1) and is roughly the worst condition number among all linear systems induced by taking s columns of D.

The stability of the sparse codes are absolutely crucial to this work. Proving that the notion of stability of contribution 1 holds is highly nontrivial because the LASSO objective (see (1) below) is not strongly convex in general. Consequently, much of the technical difficulty of this work is owed to finding conditions under which the LASSO is stable under dictionary perturbations and proving that when these conditions hold with respect to the learned hypothesis and the training sample, they also hold with respect to a future sample.

1.1. The predictive sparse coding problem

Let P be a probability measure over $B_{\mathbb{R}^d} \times \mathcal{Y}$, the product of an input space $B_{\mathbb{R}^d}$ (the unit ball of $\mathbb{R}^d$) and a space $\mathcal{Y}$ of univariate labels; examples of $\mathcal{Y}$ include a bounded subset of $\mathbb{R}$ for regression and $\{-1, 1\}$ for classification. Let $\mathbf{z} = (z_1, \ldots, z_m)$ be a sample of m points drawn iid from P, where each labeled point z_i equals (x_i, y_i) for $x_i \in B_{\mathbb{R}^d}$ and $y_i \in \mathcal{Y}$. In the reconstructive setting, labels are not of interest and we can just as well consider an unlabeled sample $\mathbf{x}$ of m points drawn iid from the marginal probability measure Π on $B_{\mathbb{R}^d}$.

The sparse coding problem is to represent each point x_i as a sparse linear combination of k basis vectors, or *atoms* $D_1, \ldots, D_k$. The atoms form the columns of a *dictionary* D living in a space of dictionaries $\mathcal{D} := (B_{\mathbb{R}^d})^k$, for $D_i = (D_i^1, \ldots, D_i^d)^T$ in the unit ℓ_2 ball. An encoder φ_D can be used to express ℓ_1 sparse coding:

$$\varphi_D(x) := \arg\min_z \|x - Dz\|_2^2 + \lambda \|z\|_1; \qquad (1)$$

hence, encoding x as $\varphi_D(x)$ amounts to solving a LASSO problem. The reconstructive ℓ_1 sparse coding objective is then

$$\min_{D \in \mathcal{D}} \mathsf{E}_{x \sim \Pi} \|x - D\varphi_D(x)\|_2^2 + \lambda \|\varphi_D(x)\|_1.$$

Generalization bounds for the empirical risk minimization (ERM) variant of this objective have been established. In the infinite-dimensional setting, Maurer & Pontil (2010) showed[1] that with probability $1 - \delta$ over the training sample $\mathbf{x} = (x_1, \ldots, x_m)$:

$$\sup_{D \in \mathcal{D}} \mathsf{E}_{x \sim \Pi} f_D(x) - \frac{1}{m} \sum_{i=1}^m f_D(x_i) \leq \frac{k}{\sqrt{m}} \left(\frac{14}{\lambda} + \frac{1}{2} \sqrt{\log\left(16m/\lambda^2\right)} \right) + \sqrt{\frac{\log(1/\delta)}{2m}}, \qquad (2)$$

where $f_D(x) := \min_{z \in \mathbb{R}^k} \|x - Dz\|_2^2 + \lambda \|z\|_1$. This bound is *independent* of the dimension d and hence useful when $d \gg k$, as in general Hilbert spaces. They also showed a similar bound in the overcomplete setting where the k is replaced by $\sqrt{dk}$. Vainsencher et al. (2011) handled the overcomplete setting, producing a bound that is $O\left(\sqrt{dk/m}\right)$ as well as fast rates of $O(dk/m)$, with only logarithmic dependence on $\frac{1}{\lambda}$.

Predictive sparse coding (Mairal et al., 2012), minimizes a supervised loss with respect to a representation and an estimator linear in the representation. Let $\mathcal{W}$ be a space of linear hypotheses with $\mathcal{W} := rB_{\mathbb{R}^k}$, the ball in $\mathbb{R}^k$ scaled to radius r. A predictive sparse coding hypothesis function f is identified by $f = (D, w) \in \mathcal{D} \times \mathcal{W}$ and defined as $f(x) = \langle w, \varphi_D(x) \rangle$. The function class $\mathcal{F}$ is the set of such hypotheses. The loss will be measured via $l : \mathcal{Y} \times \mathbb{R} \to [0, b]$, $b > 0$, a bounded loss function that is L-Lipschitz in its second argument.

The predictive sparse coding objective is[2]

$$\min_{D \in \mathcal{D}, w \in \mathcal{W}} \mathsf{E}_{(x,y) \sim P} l(y, \langle w, \varphi_D(x) \rangle) + \frac{1}{r} \|w\|_2^2; \qquad (3)$$

In this work, we analyze the ERM variant of (3):

$$\min_{D \in \mathcal{D}, w \in \mathcal{W}} \frac{1}{m} \sum_{i=1}^m l(y_i, \langle w, \varphi_D(x_i) \rangle) + \frac{1}{r} \|w\|_2^2. \qquad (4)$$

This objective is not convex, and it is unclear how to find global minima, so *a priori* we cannot say whether an optimal or nearly optimal hypothesis will be returned by any learning algorithm. However, we can

[1] To see this, take Theorem 1.2 of Maurer & Pontil (2010) with $Y = \{y \in \mathbb{R}^k : \|y\|_1 < \frac{1}{\lambda}\}$ and $\mathcal{T} = \{T : \mathbb{R}^k \to \mathbb{R}^d : \|Te_j\| \leq 1, j \in [k]\}$, so that $\|\mathcal{T}\|_Y \leq \frac{1}{\lambda}$.

[2] While the focus of this work is (3), formally the predictive sparse coding framework admits swapping out the squared ℓ_2 norm regularizer on w for any other regularizer.

and will bet on certain sparsity-related stability properties holding with respect to the learned hypothesis and the training sample. Consequently, the presented learning bound will hold uniformly *not over the set of all hypotheses but rather potentially much smaller random subclasses of hypotheses.* The presented bound will be algorithm-independent[3], but algorithm design can influence the learned hypothesis's stability and hence the best *a posteriori* learning bound.

Encoder stability Defining the encoder (1) via the ℓ_1 sparsity-inducing regularizer (or *sparsifier*) is just one choice for the encoder. The choice of sparsifier seems to be pivotal both from an empirical perspective and a theoretical one. Bradley & Bagnell (2009) used a differentiable *approximate* sparsifier based on the Kullback-Leibler divergence (true sparsity may not result). The ℓ_1 sparsifier $\|\cdot\|_1$ is the most popular and notably is the tightest convex lower bound for the ℓ_0 "norm": $\|x\|_0 := |\{i : x_i \neq 0\}|$ (Fazel, 2002). Regrettably, from a stability perspective the ℓ_1 sparsifier is not well-behaved in general. Indeed, due to the lack of strict convexity, each x need not have a unique image under φ_D. It also is unclear how to analyze the class of mappings φ_D, parameterized by D, if the map changes drastically under small perturbations to D. Hence, we will begin by establishing sufficient conditions under which φ_D is stable under perturbations to D.

2. Conditions and main result

In this section, we develop several quantities that are central to the statement of the main result. Throughout this paper, let $[n] := \{1, \ldots, n\}$ for $n \in \mathbb{N}$. Also, for $t \in \mathbb{R}^k$, define $\text{supp}(t) := \{i \in [k] : t_i \neq 0\}$.

Definition 1 (s-incoherence) *For $s \in [k]$ and $D \in \mathcal{D}$, the s-*incoherence *$\mu_s(D)$ is defined as the square of the minimum singular value among s-atom subdictionaries of D. Formally,*

$$\mu_s(D) = \left(\min\left\{\varsigma_s(D_\Lambda) : \Lambda \subseteq [k], |\Lambda| = s\right\}\right)^2,$$

where $\varsigma_s(A)$ is the $s^{\,th}$ singular value of A.

The s-incoherence can used to guarantee that sparse codes are stable in a certain sense. We also introduce some key parameter-and-data-dependent properties. The first property regards the sparsity of the encoder on a sample $\mathbf{x} = (x_1, \ldots, x_m)$.

Definition 2 (s-sparsity) *If every point x_i in the set of points $\mathbf{x}$ satisfies $\|\varphi_D(x_i)\|_0 \leq s$, then φ_D is s-sparse on $\mathbf{x}$. More concisely, the boolean expression s-sparse$(\varphi_D(\mathbf{x}))$ is true.*

This property is critical as the learning bound will exploit the observed sparsity level over the training sample. Finally, we require some margin properties.

Definition 3 (s-margin) *Given a dictionary D and a point $x_i \in B_{\mathbb{R}^d}$, the s-*margin of D on x_i *is*

$$\text{margin}_s(D, x_i) := \max_{\substack{\mathcal{I} \subseteq [k] \\ |\mathcal{I}| = k-s}} \min_{j \in \mathcal{I}} \left\{\lambda - \left|\langle D_j, x_i - D\varphi_D(x_i)\rangle\right|\right\}.$$

*The sample s-margin is the maximum s-margin that holds for all points in $\mathbf{x}$, or the s-*margin of D on $\mathbf{x}$*:*

$$\text{margin}_s(D, \mathbf{x}) := \min_{x_i \in \mathbf{x}} \text{margin}_s(D, x_i).$$

The importance of the s-margin properties flows directly from the upcoming Sparse Coding Stability Theorem (Theorem 4). Intuitively, if the s-margin of D on x is high, there is a set of $(k-s)$ inactive atoms that correlate poorly with the optimal residual $x - D\varphi_D(x)$; hence these $k-s$ atoms are far from being included in the set of active atoms. More formally, $\text{margin}_s(D, x_i)$ is equal to the $(s+1)^{\,\text{th}}$ smallest element of the set of k elements $\{\lambda - |\langle D_j, x_i - D\varphi_D(x_i)\rangle|\}_{j \in [k]}$. Note that if $\|\varphi_D(x_i)\|_0 = s$, we can use the $(s+\rho)$-margin for any integer $\rho \geq 0$. Indeed, $\rho > 0$ is justified when $\varphi_D(x_i)$ has only s non-zero dimensions but for precisely one index j^* outside the support set $|\langle D_{j^*}, x_i - D\varphi_D(x_i)\rangle|$ is arbitrarily close to λ. In this scenario, the s-margin of D on x_i is trivially small; however, the $(s+1)$-margin is non-trivial because the max in the definition of the margin will remove j^* from the min's choices $\mathcal{I}$. Empirical evidence shown in Section 5 suggests that even when ρ is small, the $(s+\rho)$-margin is not too small.

Sparse coding stability Our first result is a fundamental stability result for the LASSO. In addition to being critical in motivating the presented conditions, the result may be of interest in its own right.

Theorem 4 (Sparse Coding Stability) *Let $D, \tilde{D} \in \mathcal{D}$ satisfy $\mu_s(D), \mu_s(\tilde{D}) \geq \mu$ and $\|D - \tilde{D}\|_2 \leq \varepsilon$, and let $x \in B_{\mathbb{R}^d}$. Suppose that there exists an index set $\mathcal{I} \subseteq [k]$ of $k-s$ indices such that for all $i \in \mathcal{I}$:*

$$|\langle D_i, x - D\varphi_D(x)\rangle| < \lambda - \tau, \tag{5}$$

with

$$\varepsilon \leq \frac{\tau^2 \lambda}{27}. \tag{6}$$

Then the following stability bound holds:

$$\|\varphi_D(x) - \varphi_{\tilde{D}}(x)\|_2 \leq \frac{3}{2}\frac{\varepsilon\sqrt{s}}{\lambda\mu}.$$

[3] Empirically, stochastic gradient approaches such as the one of Mairal et al. (2012) perform quite well.

Moreover, if $\varepsilon = \frac{\tau'^2\lambda}{27}$ *for* $\tau' < \tau$*, then for all* $i \in \mathcal{I}$*:*

$$\left|\langle \tilde{D}_i, x - \tilde{D}\varphi_{\tilde{D}}(x)\rangle\right| \leq \lambda - (\tau - \tau').$$

Thus, some margin, and hence sparsity, is retained after perturbation.

Condition (5) means that at least $k-s$ inactive atoms in the coding $\varphi_D(x)$ do not have too high absolute correlation with the residual $x - D\varphi_D(x)$. We refer to the right-hand side of (6) as the permissible radius of perturbation (PRP) because it is the largest perturbation for which the theorem can guarantee encoder stability. In short, the theorem says that if problem (1) admits a stable sparse solution, then a small perturbation to the dictionary will not change the fact that a certain set of $k-s$ atoms remains inactive in the new solution.

The proof of Theorem 4 is quite long; we leave all but the following high-level sketch to Appendix A.

Proof sketch First, we show that the solution $\varphi_{\tilde{D}}(x)$ is s-sparse and, in particular, has support contained in the complement of $\mathcal{I}$. Second, we reframe the LASSO as a quadratic program (QP). By exploiting the convexity of the QP and the fact that both solutions have their support contained in a set of s atoms, simple linear algebra yields the desired stability bound. The first step appears much more difficult than the second. The quartet below is our strategy for the first step:

1. OPTIMAL VALUE STABILITY: The two problems' optimal objective values are close; this is an easy consequence of the closeness of D and $\tilde{D}$.
2. STABILITY OF NORM OF RECONSTRUCTOR: The *norms* of the optimal reconstructors ($D\varphi_D(x)$ and $\tilde{D}\varphi_{\tilde{D}}(x)$) of the two problems are close. We show this using OPTIMAL VALUE STABILITY and
$$(x - D\varphi_D(x))^T D\varphi_D(x) = \lambda\|\varphi_D(x)\|_1, \quad (7)$$
the latter of which holds due to the subgradient of (1) with respect to z (Osborne et al., 2000).
3. RECONSTRUCTOR STABILITY: The optimal reconstructors of the two problems are close. This fact is a consequence of STABILITY OF NORM OF RECONSTRUCTOR, using the ℓ_1 norm's convexity and the equality (7).
4. PRESERVATION OF SPARSITY: The solution to the perturbed problem also is supported on the complement of $\mathcal{I}$. To show this, it is sufficient to show that the absolute correlation of each atom $\tilde{D}_i$ ($i \in \mathcal{I}$) with the residual in the perturbed problem is less than λ. This last claim is a relatively easy consequence of RECONSTRUCTOR STABILITY. ■

2.1. Main result

Some notation will aid the result below and the subsequent analysis. Recall that the loss l is bounded by b and L-Lipschitz in its second argument. Also recall that $\mathcal{F}$ is the set of predictive sparse coding hypothesis functions $f(x) = \langle w, \varphi_D(x)\rangle$ indexed by $D \in \mathcal{D}$ and $w \in \mathcal{W}$. For $f \in \mathcal{F}$, define $l(\cdot, f) : \mathcal{Y} \times \mathbb{R}^d \to [0, b]$ as the loss-composed function $(y, x) \mapsto l(y, f(x))$. Let $l \circ \mathcal{F}$ be the class of such functions induced by the choice of $\mathcal{F}$ and l. A probability measure P operates on functions and loss-composed functions as:

$$Pf = \mathsf{E}_{(x,y)\sim P} f(x) \qquad Pl(\cdot, f) = \mathsf{E}_{(x,y)\sim P} l(y, f(x)).$$

Similarly, an empirical measure $P_{\mathbf{z}}$ associated with sample $\mathbf{z}$ operates on functions and loss-composed functions as:

$$P_{\mathbf{z}} f = \frac{1}{m}\sum_{i=1}^{m} f(x_i) \qquad P_{\mathbf{z}} l(\cdot, f) = \frac{1}{m}\sum_{i=1}^{m} l(y_i, f(x_i)).$$

Classically speaking, the overcomplete setting is the *modus operandi* in sparse coding. In this setting, an overcomplete basis is learned which will be used parsimoniously in coding individual points. The next result bounds the generalization error in the overcomplete setting. The $\tilde{O}(\cdot)$ notation hides $\log(\log(\cdot))$ terms and assumes that $r \leq m^{\min\{d,k\}}$.

Theorem 5 *With probability at least* $1-\delta$ *over* $\mathbf{z} \sim P^m$*, for any* $s \in [k]$ *and any* $f = (D, w) \in \mathcal{F}$ *satisfying* s-sparse$(\varphi_D(\mathbf{x}))$ *and* $m > \frac{243}{\text{margin}_s(D,\mathbf{x})^2\lambda}$*, the generalization error* $(P - P_{\mathbf{z}})l(\cdot, f)$ *is*

$$\tilde{O}\left(\begin{array}{l} b\sqrt{\frac{dk\log m + \log\frac{1}{\delta}}{m}} \\ +\frac{b}{m}\left(dk\log\frac{1}{\text{margin}_s^2(D,\mathbf{x})\cdot\lambda}\right) + \frac{L}{m}\left(\frac{r\sqrt{s}}{\lambda\mu_s(D)}\right)\end{array}\right). \quad (8)$$

Note that this bound also applies to the particular hypothesis learned from the training sample.

Discussion of Theorem 5 The theorem highlights the central role of the stability of the sparse encoder. The bound is data-dependent and exploits properties related to the training sample and the learned hypothesis. Since $k \geq d$ in the overcomplete setting, an ideal learning bound has minimal dependence on k. The $\frac{1}{m}$ term of the learning bound (8) exhibits square root dependence on both the size of the dictionary k and the ambient dimension d. It is unclear whether further improvement is possible, even in the reconstructive setting. The two known results in the reconstructive setting were established by Maurer & Pontil (2010) and later by Vainsencher et al. (2011).

Contrasting the predictive setting with the reconstructive setting, the first term of (8) matches the slower of the rates shown by Vainsencher et al. (2011) for the unsupervised case. Vainsencher et al. also showed fast rates of $\frac{dk}{m}$ (plus a small fraction of the observed empirical risk), but in the predictive setting it is an open question whether similar fast rates are possible. The second term of (8) represents the error in approximating the estimator via an ($\varepsilon = \frac{1}{m}$)-cover of the space of dictionaries. This term reflects the stability of the sparse codes with respect to dictionary perturbations, as quantified by the Sparse Coding Stability Theorem (Theorem 4). The reason for the lower bound on m is that the ε-net used to approximate the space of dictionaries needs to be fine enough to satisfy the PRP condition (6) of the Sparse Coding Stability Theorem. Hence, both this lower bound and the second term are determined primarily by the Sparse Coding Stability Theorem, and so with this proof strategy the extent to which the Sparse Coding Stability Theorem cannot be improved also indicates the extent to which Theorem 5 cannot be improved.

Critically, encoder stability is not necessary in the reconstructive setting because stability in loss (reconstruction error) requires only stability in the *norm of the residual* of the LASSO problem rather than stability in the *value of the solution* to the problem. Stability of the norm of the residual is readily obtainable without any of the incoherence, sparsity, and margin conditions used here.

Remarks on conditions One may wonder about typical values for the various hypothesis-and-data-dependent properties in Theorem 5. In practical applications of reconstructive and predictive sparse coding, the regularization parameter λ is set to ensure that s is small relative to the dimension d. As a result, the incoherence $\mu_s(D)$ of the learned dictionary can be expected to be bounded away from zero. A sufficiently large s-incoherence certainly is necessary if one hopes for any amount of stability of the class of sparse coders with respect to dictionary perturbations. Since our path to reaching Theorem 5 passes through the Sparse Coding Stability Theorem (Theorem 4), it seems that a drastically different strategy needs to be used if it is possible to avoid dependence on $\mu_s(D)$ in the learning bounds.

A curious aspect of the learning bound is its dependence on the s-margin $\text{margin}_s(D, \mathbf{x})$. Suppose a dictionary is learned which is s-sparse on the training sample $\mathbf{x}$, and s is the lowest such integer for which this holds. It may not always be the case that the s-margin is bounded away from zero because for some points a small collection of inactive atoms may be very close to being brought into the optimal solution (the code); however, we can instead use the $(s+\rho)$-margin for some small positive integer ρ for which the $(s+\rho)$-margin is non-trivial. In Section 5 we show empirical evidence that such a non-trivial $(s+\rho)$-margin does exist, with ρ small, when learning predictive sparse codes on real data. Hence, there is evidence that predictive sparse coding learns a dictionary with high s-incoherence $\mu_s(D)$ and non-trivial s-margin $\text{margin}_s(D, \mathbf{x})$ on the training sample for low s.

3. Tools

As before, let $\mathbf{z}$ be a labeled sample of m points (an m-sample) drawn iid from P and $\mathbf{z}'$ be a second (ghost) labeled m-sample drawn iid from P. All epsilon-nets of spaces of dictionaries will use the metric induced by the operator norm $\|\cdot\|_2$.

The next result is essentially due to Mendelson & Philips (2004); it applies symmetrization by a ghost sample for random subclasses.

Lemma 6 (Symmetrization by Ghost Sample) *Let $\mathcal{F}(\mathbf{z}) \subset \mathcal{F}$ be a random subclass which can depend on a labeled sample $\mathbf{z}$. Recall that $\mathbf{z}'$ is a ghost sample of m points. If $m \geq \left(\frac{b}{t}\right)^2$, then*

$$\Pr_{\mathbf{z}}\left\{\exists f \in \mathcal{F}(\mathbf{z}),\ (P - P_{\mathbf{z}})l(\cdot, f) \geq t\right\} \leq 2\Pr_{\mathbf{z}\,\mathbf{z}'}\left\{\exists f \in \mathcal{F}(\mathbf{z}),\ (P_{\mathbf{z}'} - P_{\mathbf{z}})l(\cdot, f) \geq \frac{t}{2}\right\}.$$

For completeness, this lemma is proved in Appendix B. This symmetrization lemma will shift the analysis of the next section from large deviations of the empirical risk from the expected risk to large deviations of two independent empirical risks.

For a Banach space E of dimension d, the ε-covering numbers of the radius r ball of E are bounded as $\mathcal{N}(rB_E, \varepsilon) \leq (4r/\varepsilon)^d$ (Carl & Stephani, 1990, equation (1.1.10)). For spaces of dictionaries obeying some deterministic property, such as $\mathcal{D}_\mu = \{D \in \mathcal{D} : \mu_s(D) \geq \mu\}$, one must be careful to use a *proper* ε-cover so that the representative elements of the cover also obey the desired property. The following bound relates proper covering numbers to covering numbers (a simple proof is in Vidyasagar 2002, Lemma 2.1):
If E is a Banach space and $T \subseteq E$ is a bounded subset, then $\mathcal{N}(E, \varepsilon, T) \leq \mathcal{N}_{\text{proper}}(E, \varepsilon/2, T)$.

Let $d, k \in \mathbb{N}$. Define $E_\mu := \{E \in (B_{\mathbb{R}^d})^k : \mu_s(D) \geq \mu\}$ and $\mathcal{W} := rB_{\mathbb{R}^d}$. From the above, we have:

Proposition 7 *The proper ε-covering number of E_μ is bounded by $(8/\varepsilon)^{dk}$.*

Proposition 8 *The product of the proper ε-covering number of E_μ and the ε-covering number of $\mathcal{W}$ is bounded by* $\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k}$.

4. Proof of the learning bound

At a high level, our strategy for proving Theorem 5 is to construct an epsilon-net over a subclass of the space of functions $\mathcal{F} := \{f = (D, w) : D \in \mathcal{D}, w \in \mathcal{W}\}$ and to show that the metric entropy of this subclass is of order dk. The main difficulty is that an epsilon-net over $\mathcal{D}$ need not approximate $\mathcal{F}$ to any degree, *unless* one has a notion of encoder stability. Our analysis effectively will be concerned with only a training sample and a ghost sample, and it is similar in style to the luckiness framework of Shawe-Taylor et al. (1998). If we observe that the sufficient conditions for encoder stability hold true on the training sample, then it is enough to guarantee that most points in a ghost sample also satisfy these conditions (at a weaker level).

4.1. Useful conditions and subclasses

Let $\tilde{\mathbf{x}} \subseteq_\eta \mathbf{x}$ indicate that $\tilde{\mathbf{x}}$ is a subset of $\mathbf{x}$ with at most η elements of $\mathbf{x}$ removed. This notation is identical to Shawe-Taylor et al. (1998)'s notation from the luckiness framework.

Our bound uses a PRP-based condition depending on both the learned dictionary and the training sample:

$$\text{margin}_s(D, \mathbf{x}) \geq \iota(\lambda, \varepsilon) \quad \text{for} \quad \iota(\lambda, \varepsilon) = \sqrt{\frac{243\varepsilon}{\lambda}}.$$

For brevity we will refer to ι with its parameters implicit; the dependence on ε, λ, and μ will not be an issue because we first develop bounds with these quantities fixed *a priori*. Lastly, for $\mu > 0$ define a restricted dictionary class $\mathcal{D}_\mu := \{D \in \mathcal{D} : \mu_s(D) \geq \mu\}$ and a function class $\mathcal{F}_\mu := \{f = (D, w) \in \mathcal{F} : D \in \mathcal{D}_\mu\}$.

4.2. Proof of the learning bound

The following proposition is a specialization of Lemma 6 with $\mathcal{F}(\mathbf{z}) := \{f \in \mathcal{F}_\mu : \left[\text{margin}_s(D, \mathbf{x}) > \iota\right]\}$.

Proposition 9 *If $m \geq \left(\frac{b}{t}\right)^2$, then*

$$\Pr_{\mathbf{z}}\left\{ \begin{array}{ll} \exists f \in \mathcal{F}_\mu, & \left[\text{margin}_s(D, \mathbf{x})) > \iota\right] \\ & \textbf{and } ((P - P_{\mathbf{z}})l(\cdot, f) > t) \end{array} \right\}$$
$$\leq 2\Pr_{\mathbf{z}\,\mathbf{z}'}\left\{ \begin{array}{ll} \exists f \in \mathcal{F}_\mu, & \left[\text{margin}_s(D, \mathbf{x}) > \iota\right] \\ & \textbf{and } ((P_{\mathbf{z}'} - P_{\mathbf{z}})l(\cdot, f) > t/2) \end{array} \right\}.$$

In the RHS of the above, let the event whose probability is being measured be

$$J := \left\{ \mathbf{z}\,\mathbf{z}' : \exists f \in \mathcal{F}_\mu, \quad \begin{array}{l} \left[\text{margin}_s(D, \mathbf{x}) > \iota\right] \\ \textbf{and } (P_{\mathbf{z}'} - P_{\mathbf{z}})l(\cdot, f) > t/2) \end{array} \right\}.$$

Define Z as the event that there exists a hypothesis with stable codes on the original sample, in the sense of the Sparse Coding Stability Theorem (Theorem 4), but more than $\eta = \eta(m, d, k, D, \mathbf{x}, \delta)$ points[4] of the ghost sample have codes that are not guaranteed stable by the Sparse Coding Stability Theorem:

$$Z := \left\{ \mathbf{z}\,\mathbf{z}' : \begin{array}{l} \exists f \in \mathcal{F}_\mu, \ \left[\text{margin}_s(D, \mathbf{x}) > \iota\right] \textbf{ and} \\ \left(\nexists\, \tilde{\mathbf{x}} \subseteq_\eta \mathbf{x}' \ \left[\text{margin}_s(D, \tilde{\mathbf{x}}) > \tfrac{1}{3}\text{margin}_s(D, \mathbf{x})\right]\right) \end{array} \right\}.$$

Our strategy will be to show that $\Pr(J)$ is small by use of the fact that[5]

$$\Pr(J) = \Pr(J \cap \bar{Z}) + \Pr(J \cap Z) \leq \Pr(J \cap \bar{Z}) + \Pr(Z).$$

We show that $\Pr(Z)$ and $\Pr(J \cap \bar{Z})$ are small in turn.

The imminent Good Ghost Lemma shadows Shawe-Taylor et al.'s (1998) notion of probable smoothness and provides a bound on $\Pr(Z)$.

Lemma 10 (Good Ghost) *Fix $\mu, \lambda > 0$ and $s \in [k]$. With probability at least $1 - \delta$ over an m-sample $\mathbf{x} \sim P^m$ and a second m-sample $\mathbf{x}' \sim P^m$, for any $D \in \mathcal{D}_\mu$ for which φ_D is s-sparse on $\mathbf{x}$, at least $m - \eta(m, d, k, D, \mathbf{x}, \delta)$ points $\tilde{\mathbf{x}} \subseteq \mathbf{x}'$ satisfy $\left[\text{margin}_s(D, \tilde{\mathbf{x}}) > \frac{1}{3}\text{margin}_s(D, \mathbf{x})\right]$, for*

$$\eta := dk \log \frac{1944}{\text{margin}_s^2(D, \mathbf{x}) \cdot \lambda} + \log(2m+1) + \log \frac{1}{\delta}.$$

Proof By the assumptions of the lemma, consider an arbitrary dictionary D satisfying $\mu_s(D) \geq \mu$ and s-sparse($\varphi_D(\mathbf{x})$). The goal is to guarantee with high probability that all but η points of the ghost sample are coded by φ_D with s-margin of at least $\frac{1}{3}\text{margin}_s(D, \mathbf{x})$.

Let $\varepsilon = \frac{(\frac{1}{3}\text{margin}_s(D, \mathbf{x}))^2 \cdot \lambda}{27}$, and consider a minimum-cardinality ε-proper cover $\mathcal{D}'$ of $\mathcal{D}_\mu$. Let D' be a candidate element of $\mathcal{D}'$ satisfying $\|D - D'\|_2 \leq \varepsilon$. Then the Sparse Coding Stability Theorem (Theorem 4) implies that the coding margin of D' on $\mathbf{x}$ retains over two-thirds the coding margin of D on $\mathbf{x}$; that is, $\left[\text{margin}_s(D', \mathbf{x}) > \frac{2}{3}\text{margin}_s(D, \mathbf{x})\right]$.

Furthermore, most points from the *ghost* sample satisfy $\left[\text{margin}_s(D', \cdot) > \frac{2}{3}\text{margin}_s(D, \mathbf{x})\right]$. To see this,

[4] We use the shorthand $\eta = \eta(m, d, k, D, \mathbf{x}, \delta)$.

[5] Our strategy thus far is similar to the beginning of Shawe-Taylor et al.'s proof of the main luckiness framework learning bound (Shawe-Taylor et al., 1998, Theorem 5.22).

let $\mathcal{F}_D^{\text{marg}} := \{f_{D,\tau}^{\text{marg}} | \tau \in \mathbb{R}_+\}$ be the class of threshold functions defined via

$$f_{D,\tau}^{\text{marg}}(x) := \begin{cases} 1; & \text{if } \text{margin}_s(D,x) > \tau, \\ 0; & \text{otherwise.} \end{cases}$$

The VC dimension of the one-dimensional threshold functions is 1, and so $\text{VC}(\mathcal{F}_D^{\text{marg}}) = 1$. By using the VC dimension of $\mathcal{F}_D^{\text{marg}}$ and the standard permutation argument of Vapnik & Chervonenkis (1968, Proof of Theorem 2), it follows that for a single, *fixed* element of $\mathcal{D}'$, with probability at least $1-\delta$ at most $\log(2m+1) + \log\frac{1}{\delta}$ points from a ghost sample will violate the margin inequality in question. Hence, by the bound on the proper covering numbers provided by Proposition 7, we can we can guarantee for all candidate members $D' \in \mathcal{D}'$ that with probability $1-\delta$ at most

$$\eta = dk \log \frac{1944}{\text{margin}_s^2(D,\mathbf{x}) \cdot \lambda} + \log(2m+1) + \log\frac{1}{\delta}$$

points from the ghost sample violate the s-margin inequality. Thus, for arbitrary $D' \in \mathcal{D}'$ satisfying the conditions of the lemma, with probability $1-\delta$ at most $\eta(m,d,k,D,\mathbf{x},\delta)$ points from the ghost sample violate $\left[\text{margin}_s(D',\cdot) > \frac{2}{3}\text{margin}_s(D,\mathbf{x})\right]$.

Finally, consider the at least $m-\eta$ points in the ghost sample satisfying $\left[\text{margin}_s(D',\cdot) > \frac{2}{3}\text{margin}_s(D,\mathbf{x})\right]$. Since $\|D'-D\|_2 \leq \frac{(\frac{1}{3}\text{margin}_s(D,\mathbf{x}))^2 \cdot \lambda}{27}$, the Sparse Coding Stability Theorem (Theorem 4) implies that these points satisfy $\left[\text{margin}_s(D,\cdot) > \frac{1}{3}\text{margin}_s(D,\mathbf{x})\right]$. ■

It remains to bound $\Pr(J \cap \bar{Z})$.

Lemma 11 (Large Deviation on Good Ghost) *Let* $\varpi := t/2 - \left(2L\beta + \frac{b\eta}{m}\right)$, $\beta := \frac{\varepsilon}{2\lambda}\left(1 + \frac{3r\sqrt{s}}{\mu}\right)$. *Then*

$$\Pr(J \cap \bar{Z}) \leq \left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)).$$

Proof First, note that the event $J \cap \bar{Z}$ is a subset of

$$R := \left\{ \mathbf{z}\mathbf{z}' : \begin{array}{c} \exists f \in \mathcal{F}_\mu,\ \left[\text{margin}_s(D,\mathbf{x}) > \iota\right] \textbf{ and} \\ (\exists \tilde{\mathbf{x}} \subseteq_\eta \mathbf{x}',\ \left[\text{margin}_s(D,\tilde{\mathbf{x}}) > \frac{1}{3}\text{margin}_s(D,\mathbf{x})\right]) \\ \textbf{and } ((P_{\mathbf{z}'} - P_{\mathbf{z}})l(\cdot,f) > t/2) \end{array} \right\}.$$

Bounding the probability of the event R is equivalent to bounding the probability of a large deviation (i.e. $((P_{\mathbf{z}'} - P_{\mathbf{z}})l(\cdot,f) > t/2)$) for the random subclass:

$$\tilde{\mathcal{F}}(\mathbf{x},\mathbf{x}') := \left\{ f \in \mathcal{F}_\mu : \begin{array}{l} \left[\text{margin}_s(D,\mathbf{x}) > \iota\right] \textbf{ and} \\ (\exists \tilde{\mathbf{x}} \subseteq_\eta \mathbf{x}',\ \left[\text{margin}_s(D,\tilde{\mathbf{x}}) > \frac{1}{3}\text{margin}_s(D,\mathbf{x})\right]) \end{array} \right\}.$$

Let $\mathcal{F}_\varepsilon = \mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon$, where $\mathcal{D}_\varepsilon$ is a minimum-cardinality proper ε-cover of $\mathcal{D}_\mu$ and $\mathcal{W}_\varepsilon$ is a minimum-cardinality ε-cover of $\mathcal{W}$. It is sufficient to bound the probability of a large deviation for all of $\mathcal{F}_\varepsilon$ and to then consider the maximum difference between an element of $\tilde{\mathcal{F}}(\mathbf{x},\mathbf{x}')$ and its closest representative in $\mathcal{F}_\varepsilon$. Clearly, for each $f = (D,w) \in \tilde{\mathcal{F}}(\mathbf{x},\mathbf{x}')$, there is a $f' = (D',w') \in \mathcal{F}_\varepsilon$ satisfying $\|D - D'\|_2 \leq \varepsilon$ and $\|w - w'\|_2 \leq \varepsilon$. If ε is sufficiently small, then for all but η of the points x_i in the ghost sample (and for all points x_i of the original sample) it is guaranteed that

$$\begin{aligned} &|\langle w, \varphi_D(x_i)\rangle - \langle w', \varphi_{D'}(x_i)\rangle| \\ &\leq \left|\langle w - w', \varphi_D(x_i)\rangle\right| + \left|\langle w', \varphi_D(x_i) - \varphi_{D'}(x_i)\rangle\right| \\ &\leq \frac{\varepsilon}{2\lambda} + r\frac{3}{2}\frac{\varepsilon\sqrt{s}}{\lambda\mu} \;=\; \frac{\varepsilon}{2\lambda}\left(1 + \frac{3r\sqrt{s}}{\mu}\right) \;=\; \beta, \end{aligned}$$

where the second inequality follows from the Sparse Coding Stability Theorem (Theorem 4). Trivially, for the rest of the points x_i in the ghost sample each loss is bounded by b. Hence, on the original sample:

$$\frac{1}{m}\sum_{i=1}^m \left|l(y_i, \langle w, \varphi_D(x_i)\rangle) - l(y_i, \langle w', \varphi_{D'}(x_i)\rangle)\right| \leq L\beta,$$

and on the ghost sample:

$$\begin{aligned} &\frac{1}{m}\sum_{i=1}^m \left|l(y_i', \langle w, \varphi_D(x_i')\rangle) - l(y_i', \langle w', \varphi_{D'}(x_i')\rangle)\right| \\ &\leq \frac{L}{m}\sum_{i\ \text{GOOD}} \left|\langle w, \varphi_D(x_i)\rangle - \langle w', \varphi_{D'}(x_i)\rangle\right| \\ &\quad + \frac{1}{m}\sum_{i\ \text{BAD}} \left|l(y_i', \langle w, \varphi_D(x_i')\rangle) - l(y_i', \langle w', \varphi_{D'}(x_i')\rangle)\right| \\ &\leq L\beta + \frac{b\eta}{m}, \end{aligned}$$

where GOOD denotes the (at least $m-\eta$) points of the ghost sample for which the Sparse Coding Stability Theorem applies, and BAD is the complement thereof.

Concluding the above argument, the difference between the losses of f and f' on the double sample is at most $2L\beta + \frac{b\eta}{m}$. Consequently, if $(P_{\mathbf{z}'} - P_{\mathbf{z}})l(\cdot,f) > t/2$, then the deviation between the loss of f' on the original sample and the loss of f' on the ghost sample must be at least $t/2 - \left(2L\beta + \frac{b\eta}{m}\right)$. To bound the probability of R it therefore is sufficient to control

$$\Pr_{\mathbf{z}\,\mathbf{z}'}\left\{ \begin{array}{c} \exists f = (D',w') \in \mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon \\ (P_{\mathbf{z}'} - P_{\mathbf{z}})l(\cdot,f) > t/2 - \left(2L\beta + \frac{b\eta}{m}\right) \end{array} \right\}.$$

For the case of a fixed $f = (D',w') \in \mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon$, applying Hoeffding's inequality to the random variable $l(y_i, f(x_i)) - l(y_i', f(x_i'))$, with range in $[-b,b]$, yields:

$$\Pr_{\mathbf{z}\,\mathbf{z}'}\left\{(P_{\mathbf{z}'} - P_{\mathbf{z}})l(\cdot,f) > \varpi\right\} \leq \exp(-m\varpi^2/(2b^2)),$$

for $\varpi := t/2 - \left(2L\beta + \frac{b\eta}{m}\right)$. Via a proper covering number bound of $\mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon$ (Proposition 8) and the union bound, this result extends over all of $\mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon$:

$$\Pr_{\mathbf{z}\,\mathbf{z}'}\left\{\exists f=(D',w')\in\mathcal{D}_\varepsilon\times\mathcal{W}_\varepsilon,\ (P_{\mathbf{z}'}-P_{\mathbf{z}})l(\cdot,f)>\varpi\right\}$$
$$\leq\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k}\exp(-m\varpi^2/(2b^2)).$$

The bound on $\Pr(J\cap\bar{Z})$ now follows. ■

We now prove Theorem 5 (full proof in Appendix C).

Proof sketch (of Theorem 5) Proposition 9 and Lemmas 10 and 11 imply that

$$\Pr_{\mathbf{z}}\left\{\begin{array}{ll}\exists f\in\mathcal{F}_\mu, & \left[\text{margin}_s(D,\mathbf{x})>\iota\right]\\ & \textbf{and }((P-P_{\mathbf{z}})l(\cdot,f)>t)\end{array}\right\}$$
$$\leq 2\left(\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k}\exp(-m\varpi^2/(2b^2))+\delta\right).$$

Fix $s\in[k]$ and $\mu>0$ *a priori*. Let $\varepsilon=\frac{1}{m}$; elementary manipulations show that provided $m>\frac{243}{\text{margin}_s(D,\mathbf{x})^2\lambda}$, then with probability at least $1-\delta$ over $\mathbf{z}\sim P^m$, for any $f=(D,w)\in\mathcal{F}$ satisfying $\mu_s(D)\geq\mu$ and $\left[\text{margin}_s(D,\mathbf{x})>\iota\right]$, the generalization error $(P-P_{\mathbf{z}})l(\cdot,f)$ is bounded by:

$$2b\sqrt{\frac{2((d+1)k\log(8m)+k\log\frac{r}{2}+\log\frac{4}{\delta})}{m}}$$
$$+\frac{2L}{m}\left(\frac{1}{\lambda}\left(1+\frac{3r\sqrt{s}}{\mu}\right)\right)$$
$$+\frac{2b}{m}\left(dk\log\frac{1944}{\text{margin}_s^2(D,\mathbf{x})\cdot\lambda}+\log(2m+1)+\log\frac{4}{\delta}\right).$$

The theorem follows after suitably distributing a prior across the bounds for each choice of s and μ. ■

5. An empirical study of the s-margin

Empirical evidence suggests that the s-margin is well above zero even when s is only slightly larger than the observed sparsity level. We performed experiments on the MNIST digit classification task (LeCun et al., 1998), specifically the single binary task of the digit 4 vs all the other digits. All the training data was used, and each data point was normalized to unit norm. The results in Figure 1 show that when the minimum sparsity level is s (indicated by the colored dots on the s-axis of the plots), there is a non-trivial $(s+\rho)$-margin for ρ a small positive integer. Using the $2s$-margin when s-sparsity holds may ensure that there is a moderate margin for only a constant factor increase to s.

6. Discussion and open problems

We have shown the first generalization error bound for predictive sparse coding. The learning bound in Theorem 5 is intimately related to the stability of the sparse encoder, and the bound consequently depends on properties of both the learned dictionary and the training sample. The PRP condition in the Sparse Coding Stability Theorem (Theorem 4) appears to be much stronger than necessary; we conjecture that the PRP actually is $O(\varepsilon)$ rather than $O(\sqrt{\varepsilon})$. If the conjecture is true, the number of samples required before Theorem 5 kicks in would be greatly reduced, as would be the size of many of the constants in the results.

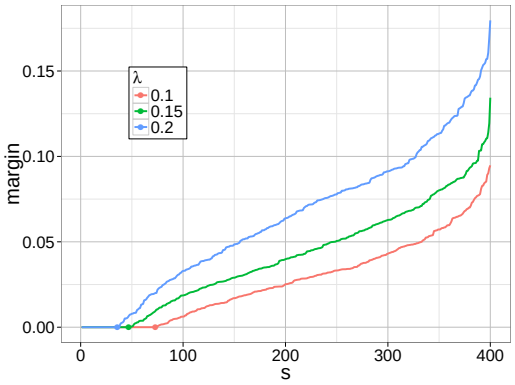

Figure 1. The s-margin for predictive sparse coding with 400 atoms trained on the MNIST training set, digit 4 versus all, for three settings of λ. The sparsity level (maximum number of non-zeros per code, taken across all codes of the training points) is indicated by the dots on the s-axis. Predictive sparse coding was trained as per the stochastic gradient descent approach of Mairal et al. (2012).

In machine learning, we often first map the data implicitly to a space of very high or even infinite dimension and use kernels for computability. In these cases where $d\gg k$ or d is infinite, any learning bound must be independent of d. We in fact have obtained a bound in the infinite-dimensional setting using considerably more sophisticated techniques (Rademacher complexities over "mostly good" random subclasses), but for space we leave this result to the long version.

Though we established an upper bound on the generalization error for predictive sparse coding, two things remain unclear. How close to optimal is the bound of Theorem 5, and what lower bounds can be established? If the conditions on which the bound relies are of fundamental importance, then the presented data-dependent bound provides motivation for an algorithm to prefer dictionaries for which small subdictionaries are well-conditioned and to additionally encourage large coding margin on the training sample.

Acknowledgments

We thank the anonymous reviewers for immeasurably useful feedback on this work. We also thank Dongryeol Lee for many formative conversations and comments.

References

Bradley, David M. and Bagnell, J. Andrew. Differentiable sparse coding. In *Advances in Neural Information Processing Systems 21*, pp. 113–120. MIT Press, 2009.

Carl, B. and Stephani, I. *Entropy, compactness, and the approximation of operators*, volume 98. Cambridge University Press, 1990.

Fazel, Maryam. Matrix rank minimization with applications. *Elec Eng Dept Stanford University*, 54: 1–130, 2002.

Kakade, Sham M., Sridharan, Karthik, and Tewari, Ambuj. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In Koller, Daphne, Schuurmans, Dale, Bengio, Yoshua, and Bottou, Léon (eds.), *Advances in Neural Information Processing Systems 21*, pp. 793–800. MIT Press, 2009.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Mairal, Julien, Bach, Francis, Ponce, Jean, Sapiro, Guillermo, and Zisserman, Andrew. Supervised dictionary learning. In Koller, Daphne, Schuurmans, Dale, Bengio, Yoshua, and Bottou, Léon (eds.), *Advances in Neural Information Processing Systems 21*, pp. 1033–1040. MIT Press, 2009.

Mairal, Julien, Bach, Francis, and Ponce, Jean. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4): 791–804, 2012.

Maurer, A. and Pontil, M. K-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

Mendelson, Shahar and Philips, Petra. On the importance of small coordinate projections. *Journal of Machine Learning Research*, 5:219–238, 2004.

Osborne, Michael R., Presnell, Brett, and Turlach, Berwin A. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, pp. 319–337, 2000.

Shawe-Taylor, John, L., Peter, Williamson, Robert C., and Anthony, Martin. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

Steinwart, Ingo and Christmann, Andreas. *Support vector machines*. Springer, 2008.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Vainsencher, Daniel, Mannor, Shie, and Bruckstein, Alfred M. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12:3259–3281, 2011.

Vapnik, Vladimir N. and Chervonenkis, Alexey Ya. Uniform convergence of frequencies of occurence of events to their probabilities. In *Dokl. Akad. Nauk SSSR*, volume 181, pp. 915–918, 1968.

Vidyasagar, Mathukumalli. *Learning and Generalization with Applications to Neural Networks*. Springer, 2002.

Yu, Kai, Zhang, Tong, and Gong, Yihong. Nonlinear learning using local coordinate coding. In Bengio, Yoshua, Schuurmans, Dale, Lafferty, John, Williams, Christopher K. I., and Culotta, Aron (eds.), *Advances in Neural Information Processing Systems 22*, pp. 2223–2231. MIT Press, 2009.