Adversarial Robustness of Supervised Sparse Coding

Jeremias Sulam
Johns Hopkins University
jsulam1@jhu.edu

Ramchandran Muthukumar
Johns Hopkins University
rmuthuk1@jhu.edu

Raman Arora
Johns Hopkins University
arora@cs.jhu.edu

Abstract

Several recent results provide theoretical insights into the phenomena of adversarial examples. Existing results, however, are often limited due to a gap between the simplicity of the models studied and the complexity of those deployed in practice. In this work, we strike a better balance by considering a model that involves learning a representation while at the same time giving a precise generalization bound and a robustness certificate. We focus on the hypothesis class obtained by combining a sparsity-promoting encoder coupled with a linear classifier, and show an interesting interplay between the expressivity and stability of the (supervised) representation map and a notion of margin in the feature space. We bound the robust risk (to ℓ_2-bounded perturbations) of hypotheses parameterized by dictionaries that achieve a mild encoder gap on training data. Furthermore, we provide a robustness certificate for end-to-end classification. We demonstrate the applicability of our analysis by computing certified accuracy on real data, and compare with other alternatives for certified robustness.

1 Introduction

With machine learning applications becoming ubiquitous in modern-day life, there exists an increasing concern about the robustness of the deployed models. Since first reported in [Szegedy et al., 2013, Goodfellow et al., 2014, Biggio et al., 2013], these *adversarial attacks* are small perturbations of the input, imperceptible to the human eye, which can nonetheless completely fluster otherwise well-performing systems. Because of clear security implications [DARPA, 2019], this phenomenon has sparked an increasing amount of work dedicated to devising defense strategies [Metzen et al., 2017, Gu and Rigazio, 2014, Madry et al., 2017] and correspondingly more sophisticated attacks [Carlini and Wagner, 2017, Athalye et al., 2018, Tramer et al., 2020], with each group trying to triumph over the other in an arms-race of sorts.

A different line of research attempts to understand adversarial examples from a theoretical standpoint. Some works have focused on giving robustness certificates, thus providing a guarantee to withstand the attack of an adversary under certain assumptions [Cohen et al., 2019, Raghunathan et al., 2018, Wong and Kolter, 2017]. Other works address questions of learnabiltiy [Shafahi et al., 2018, Cullina et al., 2018, Bubeck et al., 2018, Tsipras et al., 2018] or sample complexity [Schmidt et al., 2018, Yin et al., 2018, Tu et al., 2019], in the hope of better characterizing the increased difficulty of learning hypotheses that are robust to adversarial attacks. While many of these results are promising, the analysis is often limited to simple models.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Here, we strike a better balance by considering a model that involves learning a representation while at the same time giving a precise generalization bound and a robustness certificate. In particular, we focus our attention on the adversarial robustness of the supervised sparse coding model [Mairal et al., 2011], or task-driven dictionary learning, consisting of a linear classifier acting on the representation computed via a supervised sparse encoder. We show an interesting interplay between the expressivity and stability of a (supervised) representation map and a notion of margin in the feature space. The idea of employing sparse representations as data-driven features for supervised learning goes back to the early days of deep learning [Coates and Ng, 2011, Kavukcuoglu et al., 2010, Zeiler et al., 2010, Ranzato et al., 2007], and has had a significant impact on applications in computer vision and machine learning [Wright et al., 2010, Henaff et al., 2011, Mairal et al., 2008, 2007, Gu et al., 2014]. More recently, new connections between deep networks and sparse representations were formalized by Papyan et al. [2018], which further helped deriving stability guarantees [Papyan et al., 2017b], providing architecture search strategies and analysis [Tolooshams et al., 2019, Murdock and Lucey, 2020, Sulam et al., 2019], and other theoretical insights [Xin et al., 2016, Aberdam et al., 2019, Aghasi et al., 2020, Aberdam et al., 2020, Moreau and Bruna, 2016]. While some recent work has leveraged the stability properties of these latent representations to provide robustness guarantees against adversarial attacks [Romano et al., 2019], these rely on rather stringent generative model assumptions that are difficult to be satisfied and verified in practice. In contrast, our assumptions rely on the existence of a positive *gap* in the encoded features, as proposed originally by Mehta and Gray [2013]. This distributional assumption is significantly milder – it is directly satisfied by making traditional sparse generative model assumptions – and can be directly quantified from data.

This work makes two main contributions: The first is a bound on the robust risk of hypotheses that achieve a mild encoder gap assumption, where the adversarial corruptions are bounded in ℓ_2-norm. Our proof technique follows a standard argument based on a minimal ϵ-cover of the parameter space, dating back to Vapnik and Chervonenkis [1971] and adapted for matrix factorization and dictionary learning problems in Gribonval et al. [2015]. However, the analysis of the Lipschitz continuity of the adversarial loss with respect to the model parameters is considerably more involved. The increase in the sample complexity is mild with adversarial corruptions of size ν manifesting as an additional term of order $\mathcal{O}\left((1+\nu)^2/m\right)$ in the bound, where m is the number of samples, and a minimal encoder gap of $\mathcal{O}(\nu)$ is necessary. Much of our results extend directly to other supervised learning problems (e.g. regression). Our second contribution is a robustness certificate that holds for every hypothesis in the function class for ℓ_2 perturbations for multiclass classification. In a nutshell, this result guarantees that the label produced by the hypothesis will not change if the encoder gap is *large enough* relative to the energy of the adversary, the classifier margin, and properties of the model (e.g. dictionary incoherence).

2 Preliminaries and Background

In this section, we first describe our notation and the learning problem, and then proceed to situate our contribution in relation to prior work.

Consider the spaces of inputs, $\mathcal{X} \subseteq B_{\mathbb{R}^d}$, i.e. the unit ball in $\mathbb{R}^d$, and labels, $\mathcal{Y}$. Much of our analysis is applicable to a broad class of label spaces, but we will focus on binary and multi-class classification setting in particular. We assume that the data is sampled according to some unknown distribution P over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y}'\}$ denote a hypothesis class mapping inputs into some output space $\mathcal{Y}' \subseteq \mathbb{R}$. Of particular interest to us are norm-bounded linear predictors, $f(\cdot) = \langle \mathbf{w}, \cdot \rangle$, parametrized by d-dimensional vectors $\mathbf{w} \in \mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq B\}$.

From a learning perspective, we have a considerable understanding of the linear hypothesis class, both in a stochastic non-adversarial setting as well as in an adversarial context [Charles et al., 2019, Li et al., 2019]. However, from an application standpoint, linear predictors are often too limited, and rarely applied directly on input features. Instead, most state-of-the-art systems involve learning a representation. In general, an *encoder* map $\varphi : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^p$, parameterized by parameters θ, is composed with a linear function so that $f(\mathbf{x}) = \langle \mathbf{w}, \varphi_\theta(\mathbf{x}) \rangle$, for $\mathbf{w} \in \mathbb{R}^p$. This description applies to a large variety of popular models, including kernel-methods, multilayer perceptrons and deep convolutional neural networks. Herein we focus on an encoder given as the solution to a Lasso problem [Tibshirani, 1996]. More precisely, we consider $\varphi_{\mathbf{D}}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^p$, defined by

$$\varphi_{\mathbf{D}}(\mathbf{x}) \coloneqq \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1. \tag{1}$$

Note that when $\mathbf{D}$ is overcomplete, i.e. $p > d$, this problem is not strongly convex. Nonetheless, we will assume that that solution to Problem 1 is unique[1], and study the hypothesis class given by $\mathcal{H} = \{f_{\mathbf{D},\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \varphi_{\mathbf{D}}(\mathbf{x}) \rangle : \mathbf{w} \in \mathcal{W}, \mathbf{D} \in \mathcal{D}\}$, where $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq B\}$, and $\mathcal{D}$ is the oblique manifold of all matrices with unit-norm columns (or *atoms*); i.e. $\mathcal{D} = \{\mathbf{D} \in \mathbb{R}^{d \times p} : \|\mathbf{D}_i\|_2 = 1 \; \forall i \in [p]\}$. While not explicit in our notation, $\varphi_{\mathbf{D}}(\mathbf{x})$ depends on the value of λ. For notational simplicity, we also suppress subscripts $(\mathbf{D}, \mathbf{w})$ in $f_{\mathbf{D},\mathbf{w}}(\cdot)$ and simply write $f(\cdot)$.

We consider a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y}' \to [0, b]$, with Lipschitz constant L_ℓ. The goal of learning is to find an $f \in \mathcal{H}$ with minimal risk, or expected loss, $R(f) = \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\ell(y, f(\mathbf{x}))\right]$. Given a sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, drawn i.i.d. from P, a popular learning algorithm is empirical risk minimization (ERM) which involves finding $f_{\mathbf{D},\mathbf{w}}$ that solves the following problem:

$$\min_{\mathbf{D},\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, f_{\mathbf{D},\mathbf{w}}(\mathbf{x}_i)).$$

Adversarial Learning. In an adversarial setting, we are interested in hypotheses that are robust to adversarial perturbations of inputs. We focus on *evasion attacks*, in which an attack is deployed at test time (while the training samples are not tampered with). As a result, a more appropriate loss that incorporates the robustness to such contamination is the robust loss [Madry et al., 2017], $\tilde{\ell}_\nu(y, f(\mathbf{x})) \coloneqq \max_{\mathbf{v}\in\Delta_\nu} \; \ell(y, f(\mathbf{x} + \mathbf{v}))$, where Δ is some subset of $\mathbb{R}^d$ that restricts the power of the adversary. Herein we focus on ℓ_2 norm-bounded corruptions, $\Delta_\nu = \{\mathbf{v} \in \mathbb{R}^d \; : \; \|\mathbf{v}\|_2 \leq \nu\}$, and denote by $\tilde{R}_S(f) = \frac{1}{m}\sum_{i=1}^m \tilde{\ell}_\nu(y_i, f(\mathbf{x}_i))$ the empirical robust risk of f and $\tilde{R}(f) = \mathbb{E}_{(\mathbf{x},y)\sim P}[\tilde{\ell}_\nu(y, f(\mathbf{x}))]$ its population robust risk w.r.t. distribution P.

Main Assumptions. We make two general assumptions throughout this work. First, we assume that the dictionaries in $\mathcal{D}$ are s-incoherent, i.e, they satisfy a restricted isometry property (RIP). More precisely, for any s-sparse vector, $\mathbf{z} \in \mathbb{R}^p$ with $\|\mathbf{z}\|_0 = s$, there exists a minimal constant $\eta_s < 1$ so that $\mathbf{D}$ is close to an isometry, i.e. $(1 - \eta_s)\|\mathbf{z}\|_2^2 \leq \|\mathbf{D}\mathbf{z}\|_2^2 \leq (1 + \eta_s)\|\mathbf{z}\|_2^2$. Broad classes of matrices are known to satisfy this property (e.g. sub-Gaussian matrices [Foucart and Rauhut, 2017]), although empirically computing this constant for a fixed (deterministic) matrix is generally intractable. Nonetheless, this quantity can be upper bounded by the correlation between columns of $\mathbf{D}$, either via mutual coherence [Donoho and Elad, 2003] or the Babel function [Tropp et al., 2003], both easily computed in practice.

Second, we assume that the map $\varphi_{\mathbf{D}}$ induces a positive *encoder gap* on the computed features. Given a sample $\mathbf{x} \in \mathcal{X}$ and its encoding, $\varphi_{\mathbf{D}}(\mathbf{x})$, we denote by Λ^{p-s} the set of atoms of cardinality $(p - s)$, i.e., $\Lambda^{p-s} = \{\mathcal{I} \subseteq \{1, \ldots, p\} : |\mathcal{I}| = p - s\}$. The encoder gap $\tau_s(\cdot)$ induced by $\varphi_{\mathbf{D}}$ on any sample $\mathbf{x}$ is defined [Mehta and Gray, 2013] as

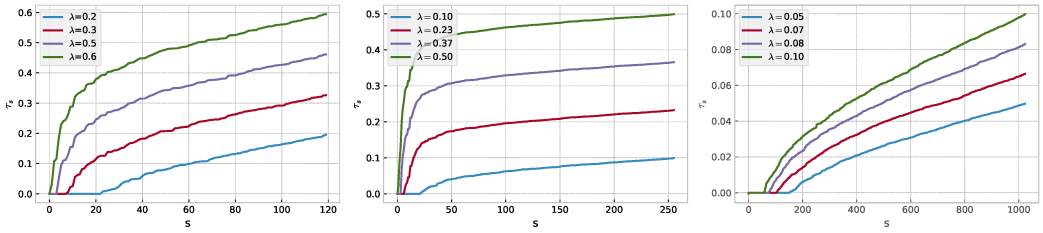
$$\tau_s(\mathbf{x}) \coloneqq \max_{\mathcal{I}\in\Lambda^{p-s}} \min_{i\in\mathcal{I}} \; \left(\lambda - |\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\rangle|\right).$$

An equivalent and conceptually simpler definition for $\tau_s(\mathbf{x})$ is the $(s+1)^{th}$ smallest entry in the vector $\lambda\mathbf{1} - |\langle \mathbf{D}, \mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\rangle|$. Intuitively, this quantity can be viewed as a measure of maximal energy along any dictionary atom that is not in the support of an input vector. More precisely, recall from the optimality conditions of Problem (1) that $|\mathbf{D}_i^T(\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}))| = \lambda$ if $[\varphi_{\mathbf{D}}(\mathbf{x})]_i \neq 0$, and $|\mathbf{D}_i^T(\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}))| \leq \lambda$ otherwise. Therefore, if τ_s is large, this indicates that there exist a set $\mathcal{I}$ of $(p - s)$ atoms that are *far* from entering the support of $\varphi_{\mathbf{D}}(\mathbf{x})$. If $\varphi_{\mathbf{D}}(\mathbf{x})$ has exactly k non-zero entries, we may choose some $s > k$ to obtain $\tau_s(\mathbf{x})$. In general, $\tau_s(\cdot)$ depends on the energy of the residual, $\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})$, the correlation between the atoms, the parameter λ, and the cardinality s. In a nutshell, if a dictionary $\mathbf{D}$ provides a quickly decaying approximation error as a function of the cardinality s, then a positive encoder gap exists for some s.

We consider dictionaries that induce a positive encoder gap in every input sample from a dataset, and define the minimum such margin as $\tau_s^* := \min_{i\in[m]} \tau_s(\mathbf{x}_i) > 0$. Such a positive encoder exist for quite general distributions, such as s-sparse and approximately sparse signals. However, this definition is more general and it will allow us to avoid making any other stronger distributional assumptions. We now illustrate such the encoder gap with both analytic and numerical examples[2].

[1]The solution is unique under mild assumptions [Tibshirani et al., 2013], and otherwise our results hold for any solution returned by a deterministic solver.

[2]Code to reproduce all of our experiments is made available at our github repository.



(a) (b) (c)

Figure 1: Encoder gap, τ_s^*, for synthetic approximately sparse signals (a) MNIST digits (b) and CIFAR10 images (c).

Approximate k-sparse signals Consider signals $\mathbf{x}$ obtained as $\mathbf{x} = \mathbf{D}\mathbf{z} + \mathbf{v}$, where $\mathbf{D} \in \mathcal{D}$, $\|\mathbf{v}\|_2 \leq \nu$ and $\mathbf{z}$ is sampled from a distribution of sparse vectors with up to k non-zeros, with $k < \frac{1}{3}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$, where $\mu(\mathbf{D}) = \max_{i \neq j} \langle \mathbf{D}_i, \mathbf{D}_j \rangle$ is the mutual coherence of $\mathbf{D}$. Then, for a particular choice of λ, we have that $\tau_s(\mathbf{x}) > \lambda - \frac{15\mu\nu}{2}, \forall s > k$. This can be shown using standard results in [Tropp, 2006]; we defer the proof to the Appendix A. Different values of λ provide different values of $\tau_s(\mathbf{x})$. To illustrate this trade-off, we generate synthetic approximately k-sparse signals ($k = 15$) from a dictionary with 120 atoms in 100 dimensions and contaminate them with Gaussian noise. We then numerically compute the value of τ_s^* as a function of s for different values of λ, and present the results in Figure 1a.

Image data We now demonstrate that a positive encoder exist for natural images as well. In Figure 1b we similarly depict the value of $\tau_s(\cdot)$, as a function of s, for an encoder computed on MNIST digits and CIFAR images (from a validation set) with learned dictionaries (further details in Section 6).

In summary, the encoder gap is a measure of the ability of a dictionary to sparsely represent data, and one can induce a larger encoder gap by increasing the regularization parameter or the cardinality s. As we will shortly see, this will provide us with a a controllable knob in our generalization bound.

3 Prior Work

Many results exist on the approximation power and stability of Lasso (see [Foucart and Rauhut, 2017]), which most commonly rely on assuming data is (approximately) k-sparse under a given dictionary. As explained in the previous section, we instead follow an analysis inspired by Mehta and Gray [2013], which relies on the encoder gap. Mehta and Gray [2013] leverage encoder gap to derive a generalization bound for the supervised sparse coding model in a stochastic (non-adversarial) setting. Their result, which follows a symmetrization technique [Mendelson and Philips, 2004], scales as $\tilde{\mathcal{O}}(\sqrt{(dp + \log(1/\delta))/m}$, and requires a minimal number of samples that is $\mathcal{O}(1/(\tau_s \lambda))$. In contrast, we study an generalization in the adversarially robust setting, detailed above. Our analysis is based on an ϵ-cover of the parameter space and on analyzing a local-Lipschitz property of the adversarial loss. The proof of our generalization bound is simpler, and shows a mild deterioration of the upper bound on the generalization gap due to adversarial corruption.

Our work is also inspired by the line of work initiated by Papyan et al. [2017a] who regard the representations computed by neural networks as approximations for those computed by a Lasso encoder across different layers. In fact, a first analysis of adversarial robustness for such a model is presented by Romano et al. [2019]; however, they make strong generative model assumptions and thus their results are not applicable to real-data practical scenarios. Our robustness certificate mirrors the analysis from the former work, though leveraging a more general and new stability bound (Lemma 5.2) relying instead on the existence of positive encoder gap. In a related work, and in the context of neural networks, Cisse et al. [2017] propose a regularization term inspired by Parseval frames, with the empirical motivation of improving adversarial robustness. Their regularization term can in fact be related to minimizing the (average) mutual coherence of the dictionaries, which naturally arises as a control for the generalization gap in our analysis.

Lastly, several works have employed sparsity as a beneficial property in adversarial learning [Marzi et al., 2018, Demontis et al., 2016], with little or no theoretical analysis, or in different frameworks

(e.g. sparse weights in deep networks [Guo et al., 2018, Balda et al., 2019], or on different domains [Bafna et al., 2018]). Our setting is markedly different from that of Chen et al. [2013] who study adversarial robustness of Lasso as a sparse predictor directly on input features. In contrast, the model we study here employs Lasso as an encoder with a data-dependent dictionary, on which a linear hypothesis is applied. A few works have recently begun to analyze the effect of learned representations in an adversarial learning setting [Ilyas et al., 2019, Allen-Zhu and Li, 2020]. Adding to that line of work, our analysis demonstrates that benefits can be provided by exploiting a trade-off between expressivity and stability of the computed representations, and the classifier margin.

4 Generalization bound for robust risk

In this section, we present a bound on the robust risk for models satisfying a positive encoder gap. Recall that given a b-bounded loss ℓ with Lipschitz constant L_ℓ, $\tilde{R}_S(f) = \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}_\nu(y_i, f(\mathbf{x}_i))$ is the empirical robust risk, and $\tilde{R}(f) = \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\tilde{\ell}_\nu(y, f(\mathbf{x}))\right]$ is the population robust risk w.r.t. distribution P. Adversarial perturbations are bounded in ℓ_2 norm by ν. Our main result below guarantees that if a hypothesis $f_{\mathbf{D},\mathbf{w}}$ is found with a sufficiently large encoder gap, and a large enough training set, its generalization gap is bounded as $\tilde{\mathcal{O}}\left(b\sqrt{\frac{(d+1)p}{m}}\right)$, where $\tilde{\mathcal{O}}$ ignores poly-logarithmic factors.

Theorem 4.1. *Let $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq B\}$, and $\mathcal{D}$ be the set of column-normalized dictionaries with p columns and with RIP at most η_s^*. Let $\mathcal{H} = \{f_{\mathbf{D},\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \varphi_{\mathbf{D}}(\mathbf{x})\rangle : \mathbf{w} \in \mathcal{W}, \mathbf{D} \in \mathcal{D}\}$. Denote τ_s^* the minimal encoder gap over the m samples. Then, with probability at least $1-\delta$ over the draw of the m samples, the generalization gap for any hypothesis $f \in \mathcal{H}$ that achieves an encoder gap on the samples of $\tau_s^* > 2\nu$, satisfies*

$$\left|\tilde{R}_S(f) - \tilde{R}(f)\right| \leq \frac{b}{\sqrt{m}}\left((d+1)p\log\left(\frac{3m}{2\lambda(1-\eta_s^*)}\right) + p\log(B) + \log\frac{4}{\delta}\right)^{\frac{1}{2}} + b\sqrt{\frac{2\log(m/2)+2\log(2/\delta)}{m}} + 12\frac{(1+\nu)^2 L_\ell B\sqrt{s}}{m},$$

as long as $m > \frac{\lambda(1-\eta_s)}{(\tau_s^*-2\nu)^2}K_\lambda$, where $K_\lambda = \left(2\left(1+\frac{1+\nu}{2\lambda}\right) + \frac{5(1+\nu)}{\sqrt{\lambda}}\right)^2$.

A few remarks are in order. First, note that adversarial generalization incurs a polynomial dependence on the adversarial perturbation ν. This is mild, especially since it only affects the fast $\mathcal{O}(1/m)$ term. Second, the bound requires a minimal number of samples. Such a requirement is intrinsic to the stability of Lasso (see Lemma 4.2 below) and it exists also in the non-adversarial setting [Mehta and Gray, 2013]. In the adversarial case, this requirement becomes more demanding, as reflected by the term $(\tau_s^* - 2\nu)$ in the denominator. Moreover, a minimal encoder gap $\tau_s^* > 2\nu$ is needed as well.

Theorem 4.1 suggests an interesting trade-off. One can obtain a large τ_s^* by increasing λ and s – as demonstrated in in Figure 1. But increasing λ may come at an expense of hurting the empirical error, while increasing s makes the term $1-\eta_s$ smaller. Therefore, if one obtains a model with small training error, along with large τ_s^* over the training samples for an appropriate choice of λ and s while ensuring that η_s is bounded away from 1, then $f_{\mathbf{D},\mathbf{w}}$ is guaranteed to generalize well. Furthermore, note that the excess error depends mildly (poly logarithmically) on λ and η_s.

Our proof technique is based on a minimal ϵ-cover of the parameter space, and the full proof is included in the Appendix B. Special care is needed to ensure that the encoder gap of the dictionary holds for a sample drawn from the population, as we can only measure this gap on the provided m samples. To address this, we split the data equally into a training set and a development set: the former is used to learn the dictionary, and the latter to provide a high probability bound on the event that $\tau_s(\mathbf{x}) > \tau_s^*$. This is to ensure that the random samples of the encoder margin are i.i.d. for measure concentration. Ideally, we would like to utilize the entire dataset for learning the predictor; we leave that for future work.

While most of the techniques we use are standard [3], the Lipschitz continuity of the robust loss function requires a more delicate analysis. For that, we have the following result.

[3] See [Seibert, 2019] for a comprehensive review on these tools in matrix factorization problems.

Lemma 4.2 (Parameter adversarial stability). *Let* $\mathbf{D}, \mathbf{D}' \in \mathcal{D}$. *If* $\|\mathbf{D} - \mathbf{D}'\|_2 \leq \epsilon \leq 2\lambda/(1+\nu)^2$, *then*

$$\max_{\mathbf{v}\in\Delta} \|\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) - \varphi_{\mathbf{D}'}(\mathbf{x}+\mathbf{v})\|_2 \leq \gamma(1+\nu)^2\epsilon, \tag{2}$$

with $\gamma = \frac{3}{2}\frac{\sqrt{s}}{\lambda(1-\eta_s)}$, *as long as* $\tau_s(\mathbf{x}) \geq 2\nu + \sqrt{\epsilon}\left(\sqrt{\frac{25}{\lambda}}(1+\nu) + 2\left(\frac{(1+\nu)}{\lambda}+1\right)\right)$.

Lemma 4.2 is central to our proof, as it provides a bound on difference between the features computed by the encoder under model deviations. Note that the condition on the minimal encoder gap, $\tau_s(\mathbf{x})$, puts an upper bound on the distance between models $\mathbf{D}$ and $\mathbf{D}'$. This in turn results in the condition imposed on the minimal samples in Theorem 4.1. It is worth stressing that the lower bound on $\tau_s(\mathbf{x})$ is on the *unperturbed* encoder gap – that which can be evaluated on the samples from the dataset, without the need of the adversary. We defer the proof of this Lemma to Appendix B.1.

5 Robustness Certificate

Next, we turn to address an equally important question about robust adversarial learning, that of giving a formal certification of robustness. Formally, we would like to guarantee that the output of the trained model, $f_{\mathbf{D},\mathbf{W}}(\mathbf{x})$, does not change for norm-bounded adversarial perturbations of a certain size. Our second main result provides such a certificate for the supervised sparse coding model.

Here, we consider a multiclass classification setting with $y \in \{1, \ldots, K\}$; simplified results for binary classification are included in Appendix C. The hypothesis class is parameterized as $f_{\mathbf{D},\mathbf{W}}(\mathbf{x}) = \mathbf{W}^T\varphi_{\mathbf{D}}(\mathbf{x})$, with $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K] \in \mathbb{R}^{p\times K}$. The multiclass margin is defined as follows:

$$\rho_{\mathbf{x}} = \mathbf{W}_{y_i}^T\varphi_{\mathbf{D}}(\mathbf{x}) - \max_{j\neq y_i} \mathbf{W}_j^T\varphi_{\mathbf{D}}(\mathbf{x}).$$

We show the following result.

Theorem 5.1 (Robustness certificate for multiclass supervised sparse coding). *Let* $\rho_x > 0$ *be the multiclass classifier margin of* $f_{\mathbf{D},\mathbf{W}}(\mathbf{x})$ *composed of an encoder with a gap of* $\tau_s(\mathbf{x})$ *and a dictionary,* $\mathbf{D}$, *with RIP constant* $\eta_s < 1$. *Let* $c_{\mathbf{W}} := \max_{i\neq j} \|\mathbf{W}_i - \mathbf{W}_j\|_2$. *Then,*

$$\arg\max_{j\in[K]} [\mathbf{W}^T\varphi_{\mathbf{D}}(\mathbf{x})]_j = \arg\max_{j\in[K]} [\mathbf{W}^T\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})]_j, \quad \forall\, \mathbf{v} : \|\mathbf{v}\|_2 \leq \nu, \tag{3}$$

so long as $\nu \leq \min\{\tau_s(\mathbf{x})/2, \rho_{\mathbf{x}}\sqrt{1-\eta_s}/c_{\mathbf{W}}\}$.

Theorem 5.1 clearly captures the potential contamination on two flanks: robustness can no longer be guaranteed as soon as the energy of the perturbation is enough to either significantly modify the computed representation *or* to induce a perturbation larger than the classifier margin on the feature space. Proof of Theorem 5.1, detailed in Appendix C, relies on the following lemma showing that under an encoder gap assumption, the computed features are moderately affected despite adversarial corruptions of the input vector.

Lemma 5.2 (Stability of representations under adversarial perturbations). *Let* $\mathbf{D}$ *be a dictionary with RIP constant* η_s. *Then, for any* $\mathbf{x} \in \mathcal{X}$ *and its additive perturbation* $\mathbf{x}+\mathbf{v}$, *for any* $\|\mathbf{v}\|_2 \leq \nu$, *if* $\tau_s(\mathbf{x}) > 2\nu$, *then we have that*

$$\|\varphi_{\mathbf{D}}(\mathbf{x}) - \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})\|_2 \leq \frac{\nu}{\sqrt{1-\eta_s}}. \tag{4}$$

An extensive set of results exist for the stability of the solution provided by Lasso relying generative model assumptions [Foucart and Rauhut, 2017, Elad, 2010]. The novelty of Lemma 5.2 is in replacing such an assumption with the existence of a positive encoder gap on $\varphi_{\mathbf{D}}(\mathbf{x})$.

Going back to Theorem 5.1, note that the upper bound on ν depends on the RIP constant η_s, which is not computable for a given (deterministic) matrix $\mathbf{D}$. Yet, this result can be naturally relaxed by upper bounding η_s with measures of correlation between the atoms, such as the mutual coherence. This quantity provides a measure of the worst correlation between two atoms in the dictionary $\mathbf{D}$, and it is defined as $\mu(\mathbf{D}) = \max_{i\neq j} |\langle \mathbf{D}_i, \mathbf{D}_j\rangle|$ (for $\mathbf{D}$ with normalized columns). For general (overcomplete and full rank) dictionaries, clearly $0 < \mu(\mathbf{D}) \leq 1$.

While conceptually simple, results that use $\mu(\mathbf{D})$ tend to be too conservative. Tighter bounds on η_s can be provided by the Babel function[4], $\mu_{(s)}$, which quantifies the maximum correlation between an atom and *any other* collection of s atoms in $\mathbf{D}$. It can be shown [Tropp et al., 2003, Elad, 2010, Chapter 2] that $\eta_s \leq \mu_{(s-1)} \leq (s-1)\mu(\mathbf{D})$. Therefore, we have the following:

Corollary 5.3. *Under the same assumptions as those in Theorem 5.1,*

$$\arg\max_{j\in[K]} [\mathbf{W}^T\varphi_{\mathbf{D}}(\mathbf{x})]_j = \arg\max_{j\in[K]} [\mathbf{W}^T\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})]_j, \quad \forall\, \mathbf{v} : \|\mathbf{v}\|_2 \leq \nu \tag{5}$$

so long as $\nu \leq \min\{\tau_s(\mathbf{x})/2, \rho_{\mathbf{x}}\sqrt{1-\mu_{(s-1)}}/c_{\mathbf{W}}\}$.

Although the condition on ν in the corollary above is stricter (and the bound looser), the quantities involved can easily be computed numerically leading to practical useful bounds, as we see next.

6 Experiments

In this section, we illustrate the robustness certificate guarantees both in synthetic and real data, as well as the trade-offs between constants in our sample complexity result. First, we construct samples from a separable binary distribution of k-sparse signals. To this end, we employ a dictionary with 120 atoms in 100 dimensions with a mutual coherence of 0.054. Sparse representations $\mathbf{z}$ are constructed by first drawing their support (with cardinality k) uniformly at random, and drawing its non-zero entries from a uniform distribution away from zero. Samples are obtained as $\mathbf{x} = \mathbf{D}\mathbf{z}$, and normalized to unit norm. We finally enforce separability by drawing $\mathbf{w}$ at random from the unit ball, determining the labels as $y = \text{sign}(\mathbf{w}^T\varphi_{\mathbf{D}}(\mathbf{x}))$, and discarding samples with a margin ρ smaller than a pre-specified amount (0.05 in this case). Because of the separable construction, the accuracy of the resulting classifier is 1.

We then attack the obtained model employing the projected gradient descent method [Madry et al., 2017], and analyze the degradation in accuracy as a function of the energy budget ν. We compare this empirical performance with the bound in Corollary 5.3: given the obtained margin, ρ, and the dictionary's μ_s, we can compute the maximal certified radius for a sample $\mathbf{x}$ as

$$\nu(\mathbf{x}) = \max_s \min\{\tau_s(\mathbf{x})/2, \rho_{\mathbf{x}}\sqrt{1-\mu_{(s-1)}}/c_{\mathbf{W}}\}. \tag{6}$$

For a given dataset, we can compute the minimal certified radius over the samples, $\nu^* = \min_{i\in[n]} \nu(\mathbf{x}_i)$. This is the bound depicted in the vertical line in Figure 2a. As can be seen, despite being somewhat loose, the attacks do not change the label of the samples, thus preserving the accuracy.

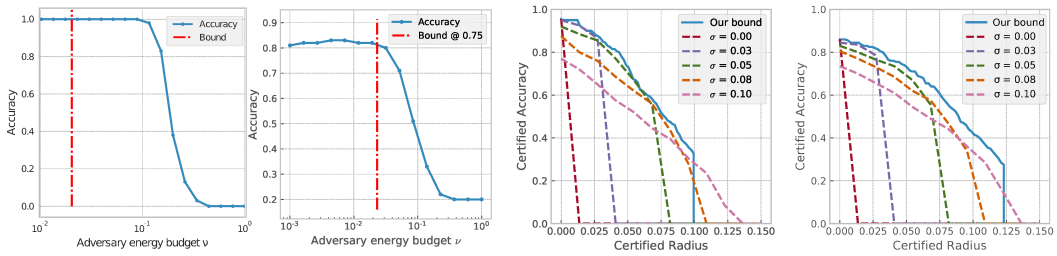
In non-separable distributions, one may study how the accuracy depends on the *soft margin* of the classifier. In this way, one can determine a target margin that results in, say, 75% accuracy on a validation set. One can obtain a corresponding certified radius of ν^* as before, which will guarantee that the accuracy will not drop below 75% as long as $\nu < \nu^*$. This is illustrated in Figure 2b.

An alternative way of employing our results from Section 5 is by studying the *certified accuracy* achieved by the resulting hypothesis. The certified accuracy quantifies the percentage of samples in a test set that are classified correctly while being *certifiable*. In our context, this implies that a sample $\mathbf{x}$ achieves a margin of $\rho_{\mathbf{x}}$, for which a certified radius of ν^* can be obtained with (6). In this way, one may study how the certified accuracy decreases with increasing ν^*.

This analysis lets us compare our bounds with those of other popular certification techniques, such as randomized smoothing [Cohen et al., 2019]. Randomized smoothing provides high probability robustness guarantees against ℓ_2 attacks for *any* classifier by composing them with a Gaussian distribution (though other distributions have been recently explored as well for other l_p norms [Salman et al., 2020]). In a nutshell, the larger the variance of the Gaussian, the larger the certifiable radius becomes, albeit at the expense of a drop in accuracy.

We use the MNIST dataset for this analysis. We train a model with 256 atoms by minimizing the following regularized empirical risk using stochastic gradient descent (employing Adam [Kingma

[4]Let Λ denote subsets (supports) of $\{1, 2, \ldots, p\}$. Then, the Babel function is defined as $\mu_{(s)} = \max_{\Lambda:|\Lambda|=s} \max_{j\notin\Lambda} \sum_{i\in\Lambda} |\langle \mathbf{D}_i, \mathbf{D}_j\rangle|$.



(a) (b) (c) (d)

Figure 2: Numerical demonstrations of our results. (a) synthetic separable distribution. (b) synthetic non-separable distribution. (c-d) certified accuracy on MNIST with $\lambda = 0.2$ and $\lambda = 0.3$, respectively, comparing with Randomized Smoothing with different variance levels.

and Ba, 2014]; the implementation details are deferred to Appendix D)

$$\min_{\mathbf{W},\mathbf{D}} \frac{1}{m}\sum_{i=1}^{m} \ell\left(y_i, \langle \mathbf{W}, \varphi_{\mathbf{D}}(\mathbf{x}_i)\rangle\right) + \alpha\|\mathbf{I} - \mathbf{D}^T\mathbf{D}\|_F^2 + \beta\|\mathbf{W}\|_F^2, \tag{7}$$

where ℓ is the cross entropy loss. Recall that $\varphi_{\mathbf{D}}(\mathbf{x})$ depends on λ, and we train two different models with two values for this parameter ($\lambda = 0.2$ and $\lambda = 0.3$).

Figure 2c and 2d illustrate the certified accuracy on 200 test samples obtained by different degrees of randomized smoothing and by our result. While the certified accuracy resulting from our bound is comparable to that by randomized smoothing, the latter provides a certificate by *defending* (i.e. composing it with a Gaussian distribution). In other words, different *smoothed* models have to be constructed to provide different levels of certified accuracy. In contrast, our model is not defended or modified in any way, and the certificate relies solely on our careful characterization of the function class. Since randomized smoothing makes no assumptions about the model, the bounds provided by this strategy rely on the estimation of the output probabilities. This results in a heavy computational burden to provide a high-probability result (a failure probability of 0.01% was used for these experiments). In contrast, our bound is deterministic and trivial to compute.

Lastly, comparing the results in Figure 2c (where $\lambda = 0.2$) and Figure 2d (where $\lambda = 0.3$), we see the trade-off that we alluded to in Section 4: larger values of λ allow for larger encoder gaps, resulting in overall larger possible certified radius. In fact, λ determines a hard bound on the possible achieved certified radius, given by $\lambda/2$, as per (6). This, however, comes at the expense of reducing the complexity of the representations computed by the encoder $\varphi_{\mathbf{D}}(\mathbf{x})$, which impacts the risk attained.

7 Conclusion

In this paper we study the adversarial robustness of the supervised sparse coding model from two main perspectives: we provide a bound for the robust risk of any hypothesis that achieves a minimum encoder gap over the samples, as well as a robustness certificate for the resulting end-to-end classifier. Our results describe guarantees relying on the interplay between the computed representations, or features, and the classifier margin.

While the model studied is still relatively simple, we envision several ways in which our analysis can be extended to more complex models. First, high dimensional data with shift-invariant properties (such as images) often benefit from convolutional features. Our results do hold for convolutional dictionaries, but the conditions on the mutual coherence may become prohibitive in this setting. An analogous definition of encoder gap in terms of convolutional sparsity [Papyan et al., 2017b] may provide a solution to this limitation. Furthermore, this analysis could also be extended to sparse models with multiple layers, as in [Papyan et al., 2017a, Sulam et al., 2019]. On the other hand, while our result does not provide a uniform learning bound over the hypothesis class, we have found empirically that regularized ERM does indeed return hypotheses satisfying non-trivial encoder gaps. The theoretical underpinning of this phenomenon needs further research. More generally, even though this work focuses on sparse encoders, we believe similar principles could be generalized to other forms of representations in a supervised learning setting, providing a framework for the principled analysis of adversarial robustness of machine learning models.

Broader Impact

This work contributes to the theoretical understanding of the limitations and achievable robustness guarantees for supervised learning models. Our results can therefore provide tools that could be deployed in sensitive settings where these types of guarantees are a priority. On a broader note, this work advocates for the precise analysis and characterization of the data-driven features computed by modern machine learning models, and we hope our results facilitate their generalization to other more complex models.

Acknowledgements

This research was supported, in part, by DARPA GARD award HR00112020004, NSF BIGDATA award IIS-1546482, NSF CAREER award IIS-1943251 and NSF TRIPODS award CCF-1934979. Jeremias Sulam kindly thanks Aviad Aberdam for motivating and inspiring discussions. Raman Arora acknowledges support from the Simons Institute as part of the program on the Foundations of Deep Learning and the Institute for Advanced Study (IAS), Princeton, NJ, as part of the special year on Optimization, Statistics, and Theoretical Machine Learning.

References

Aviad Aberdam, Jeremias Sulam, and Michael Elad. Multi-layer sparse coding: the holistic way. *SIAM Journal on Mathematics of Data Science*, 1(1):46–77, 2019.

Aviad Aberdam, Alona Golts, and Michael Elad. Ada-lista: Learned solvers adaptive to varying models. *arXiv preprint arXiv:2001.08456*, 2020.

Alireza Aghasi, Afshin Abdi, and Justin Romberg. Fast convex pruning of deep neural networks. *SIAM Journal on Mathematics of Data Science*, 2(1):158–188, 2020.

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190*, 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Mitali Bafna, Jack Murtagh, and Nikhil Vyas. Thwarting adversarial examples: An ℓ_0-robust sparse fourier transform. In *Advances in Neural Information Processing Systems*, pages 10075–10085, 2018.

Emilio Rafael Balda, Arash Behboodi, Niklas Koep, and Rudolf Mathar. Adversarial risk bounds for neural networks through sparsity based compression. *arXiv preprint arXiv:1906.00698*, 2019.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

Zachary Charles, Shashank Rajput, Stephen Wright, and Dimitris Papailiopoulos. Convergence and margin of adversarial training on separable data. *arXiv preprint arXiv:1905.09209*, 2019.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782, 2013.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 854–863. JMLR. org, 2017.

Adam Coates and Andrew Y Ng. The importance of encoding versus training with sparse coding and vector quantization. 2011.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pages 230–241, 2018.

DARPA. https://www.darpa.mil/news-events/2019-02-06, 2019.

Ambra Demontis, Paolo Russu, Battista Biggio, Giorgio Fumera, and Fabio Roli. On security and sparsity of linear classifiers for adversarial settings. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 322–332. Springer, 2016.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, 2003.

Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

Simon Foucart and Holger Rauhut. A mathematical introduction to compressive sensing. *Bull. Am. Math*, 54:151–165, 2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015.

Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Projective dictionary pair learning for pattern classification. In *Advances in neural information processing systems*, pages 793–801, 2014.

Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. Sparse DNNs with improved adversarial robustness. In *Advances in neural information processing systems*, pages 242–251, 2018.

Mikael Henaff, Kevin Jarrett, Koray Kavukcuoglu, and Yann LeCun. Unsupervised learning of sparse features for scalable audio classification. In *ISMIR*, volume 11, page 2011, 2011.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. Fast inference in sparse coding algorithms with applications to object recognition. *arXiv preprint arXiv:1010.3467*, 2010.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yan Li, Ethan X Fang, Huan Xu, and Tuo Zhao. Inductive bias of gradient descent based adversarial training on separable data. *arXiv preprint arXiv:1906.02931*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2007.

Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Discriminative learned dictionaries for local image analysis. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):791–804, 2011.

Zhinus Marzi, Soorya Gopalakrishnan, Upamanyu Madhow, and Ramtin Pedarsani. Sparsity-based defense against adversarial attacks on linear classifiers. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 31–35. IEEE, 2018.

Nishant Mehta and Alexander Gray. Sparsity-based generalization bounds for predictive sparse coding. In *International Conference on Machine Learning*, pages 36–44, 2013.

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

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Thomas Moreau and Joan Bruna. Understanding trainable sparse coding via matrix factorization. *arXiv preprint arXiv:1609.00285*, 2016.

Calvin Murdock and Simon Lucey. Dataless model selection with the deep frame potential. *arXiv preprint arXiv:2003.13866*, 2020.

Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017a.

Vardan Papyan, Jeremias Sulam, and Michael Elad. Working locally thinking globally: Theoretical guarantees for convolutional sparse coding. *IEEE Transactions on Signal Processing*, 65(21): 5687–5701, 2017b.

Vardan Papyan, Yaniv Romano, Jeremias Sulam, and Michael Elad. Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks. *IEEE Signal Processing Magazine*, 35(4):72–89, 2018.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Marc'Aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

Yaniv Romano, Aviad Aberdam, Jeremias Sulam, and Michael Elad. Adversarial noise attacks of deep learning architectures: Stability analysis via sparse-modeled signals. *Journal of Mathematical Imaging and Vision*, pages 1–15, 2019.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Black-box smoothing: A provable defense for pretrained classifiers. *arXiv preprint arXiv:2003.01908*, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

Matthias Seibert. *Sample Complexity of Representation Learning for Sparse and Related Data Models*. PhD thesis, Technische Universität München, 2019.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

Jeremias Sulam, Aviad Aberdam, Amir Beck, and Michael Elad. On multi-layer basis pursuit, efficient algorithms and convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7: 1456–1490, 2013.

Bahareh Tolooshams, Sourav Dey, and Demba Ba. Deep residual auto-encoders for expectation maximization-based dictionary learning. *arXiv preprint arXiv:1904.08827*, 2019.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.

Joel A Tropp, Anna C Gilbert, Sambavi Muthukrishnan, and Martin J Strauss. Improved sparse approximation over quasiincoherent dictionaries. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–37. IEEE, 2003.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Zhuozhuo Tu, Jingwei Zhang, and Dacheng Tao. Theoretical analysis of adversarial learning: A minimax approach. In *Advances in Neural Information Processing Systems*, pages 12259–12269, 2019.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Theory of Probability and its Applications*, page 264–280. 1971.

Eric Wong and J Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017.

John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6): 1031–1044, 2010.

Bo Xin, Yizhou Wang, Wen Gao, David Wipf, and Baoyuan Wang. Maximal sparsity with deep networks? In *Advances in Neural Information Processing Systems*, pages 4340–4348, 2016.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914*, 2018.

Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE, 2010.

Supplementary Material for Adversarial Robustness of Supervised Sparse Coding

A Encoder Gap for k-sparse signals

Herein we show that a positive encoder gap exists for signals that are (approximately) k-sparse. Consider signals $\mathbf{x}$ obtained as $\mathbf{x} = \mathbf{D}\mathbf{z} + \mathbf{v}$, where $\mathbf{D} \in \mathcal{D}$, $\|\mathbf{v}\|_2 \leq \nu$ and $\mathbf{z}$ is sampled from a distribution of sparse vectors with up to k non-zeros, with $k < \frac{1}{3}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$, where $\mu(\mathbf{D}) = \max_{i \neq j} \langle \mathbf{D}_i, \mathbf{D}_j \rangle$ is the mutual coherence of $\mathbf{D}$. Then, from [Tropp, 2006], if $\lambda = 4\nu$, the (unique) solution recovered by $\boldsymbol{\alpha} = \varphi_{\mathbf{D}}(\mathbf{x} + \mathbf{v})$ satisfies $\|\boldsymbol{\alpha} - \mathbf{z}\|_\infty \leq \frac{15}{2}\nu$, and $Supp(\boldsymbol{\alpha}) \subseteq Supp(\mathbf{z})$. Recall the definition of encoder gap:

$$\tau_s(\mathbf{x}) \coloneqq \max_{\mathcal{I} \in \Lambda^{p-s}} \min_{i \in \mathcal{I}} \ (\lambda - |\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}) \rangle|)$$

and pick $s > k$. Let $S = Supp(\mathbf{z})$. Thus, the maximization over I is achieved by a subset $\mathcal{I}$ which does not contain any of the active atoms in $\mathbf{z}$ (for which $|\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}) \rangle| = \lambda$, by optimality).

Now, define $\Delta = \boldsymbol{\alpha} - \mathbf{z}$, and let Δ_S denote the vector Δ restricted to the support S and $\mathbf{D}_S$ the sub-dictionary obtained $\mathbf{D}$ by restricting it to the same set of atoms. We can then write

$$\begin{aligned}
\max_i \left|\mathbf{D}_i^T(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha})\right| &= \max_i \left|\mathbf{D}_i^T(\mathbf{x} - \mathbf{D}\mathbf{z}) - \mathbf{D}_i^T\mathbf{D}\Delta\right| && (8)\\
&= \max_i \left|\mathbf{D}_i^T\mathbf{D}_S\Delta_S\right| && (9)\\
&\leq \max_i |\mathbf{D}_i^T\mathbf{D}_S| \|\Delta_S\|_\infty && (10)\\
&= \frac{15}{2}\mu(\mathbf{D})\nu, && (11)
\end{aligned}$$

because $i \notin S$. Thus,

$$\min_i \lambda - |\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}) \rangle| \geq \lambda - \frac{15}{2}\mu(\mathbf{D})\nu. \tag{12}$$

In fact, recalling that $\lambda = 4\nu$, we have that $\tau_s \geq \nu(4 - \mu(\mathbf{D})\frac{15}{2})$.

B Robust Generalization Bound

Herein we prove our generalization bound, but first re-state it for completeness.

Theorem 4.1. *Let $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 \leq B\}$, and $\mathcal{D}$ be the set of column-normalized dictionaries with p columns and with RIP at most η_s^*. Let $\mathcal{H} = \{f_{\mathbf{D},\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \varphi_{\mathbf{D}}(\mathbf{x}) \rangle : \mathbf{w} \in \mathcal{W}, \mathbf{D} \in \mathcal{D}\}$. Denote τ_s^* the minimal encoder gap over the m samples. Then, with probability at least $1 - \delta$ over the draw of the m samples, the generalization gap for any hypothesis $f \in \mathcal{H}$ that achieves an encoder gap of $\tau_s^* > 2\nu$, satisfies*

$$\begin{aligned}
\left|\tilde{R}_S(f) - \tilde{R}(f)\right| \leq \frac{b}{\sqrt{m}} \Bigg((d+1)p \log\left(\frac{3m}{2\lambda(1-\eta_s^*)}\right) + p\log(B) + \log\frac{4}{\delta} \Bigg)^{\frac{1}{2}} \\
+ b\sqrt{\frac{2\log(m/2) + 2\log(2/\delta)}{m}} + 12\frac{(1+\nu)^2 L_\ell B\sqrt{s}}{m},
\end{aligned}$$

as long as $m > \frac{\lambda(1-\eta_s)}{(\tau_s^* - 2\nu)^2} K_\lambda$, where $K_\lambda = \left(2\left(1 + \frac{1+\nu}{2\lambda}\right) + \frac{5(1+\nu)}{\sqrt{\lambda}}\right)^2$.

Proof. Fix $\epsilon > 0$, and consider a minimal ϵ-cover for the parameter space $(\mathcal{D}, \mathcal{W})$ with respect to a metric d and with the elements $(\mathbf{D}_j, \mathbf{w}_j)$, $j \in \{1, \ldots, N^{cov}((\mathcal{D}, \mathcal{W}), \epsilon)\}$. The metric we consider is the max over the operator norm and ℓ_2 norm on $\mathcal{D}$ and $\mathcal{W}$, respectively, i.e. $d\left((\mathbf{D}, \mathbf{w}), (\mathbf{D}', \mathbf{w}')\right) = \max\{\|\mathbf{D} - \mathbf{D}'\|_2, \|\mathbf{w} - \mathbf{w}'\|_2\}$. Now, fixing $(\mathbf{D}, \mathbf{w})$, by the definition of the ϵ-cover, there exist an

index j so that $d((\mathbf{D}_j, \mathbf{w}_j), (\mathbf{D}, \mathbf{w})) \leq \epsilon$. We thus expand the generalization gap into three terms, as follows:

$$\left|\tilde{R}_S(f) - \tilde{R}(f)\right| = \left| \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}_\nu(y_i, f(\mathbf{x}_i)) \;\; - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim P}\left[\tilde{\ell}_\nu(y, f(\mathbf{x}))\right]\right| \tag{13}$$

$$\leq \sup_{k\in[N^{\text{cov}}]} \left|\frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}_\nu(y_i, f_{\mathbf{D}_k,\mathbf{w}_k}(\mathbf{x}_i)) - \mathop{\mathbb{E}}_{(\mathbf{x},y)}\left[\tilde{\ell}_\nu(y, f_{\mathbf{D}_k,\mathbf{w}_k}(\mathbf{x}))\right]\right| \tag{14}$$

$$+ \left|\frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}_\nu(y_i, f_{\mathbf{D},\mathbf{w}}(\mathbf{x}_i)) - \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}_\nu(y_i, f_{\mathbf{D}_j,\mathbf{w}_j}(\mathbf{x}_i))\right| \tag{15}$$

$$+ \left|\mathop{\mathbb{E}}_{(\mathbf{x},y)}\left[\tilde{\ell}_\nu(y, f_{\mathbf{D}_j,\mathbf{w}_j}(\mathbf{x}))\right] - \mathop{\mathbb{E}}_{(\mathbf{x},y)}\left[\tilde{\ell}_\nu(y, f_{\mathbf{D},\mathbf{w}}(\mathbf{x}))\right]\right|. \tag{16}$$

Let us bound the first of these terms. Let $\mathbf{z}_i$ denote the random tuple $(y_i, \mathbf{x}_i)$, and $\mathbf{Z} = [(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)]$. Let $g(\mathbf{Z}) = \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}_f(y_i, \mathbf{x}_i)$. Furthermore, consider $\mathbf{Z}'$ as the set of m random variables $\mathbf{z}$ that only differs from $\mathbf{Z}$ in its i^{th} variable, $\mathbf{z}'_i = (y'_i, \mathbf{x}'_i)$. Then, for any $i \in [m]$,

$$|g(\mathbf{Z}) - g(\mathbf{Z}')| = \left|\frac{1}{m}\left(\tilde{\ell}_{\nu_f}(y_i, \mathbf{x}_i) - \tilde{\ell}_{\nu_f}(y'_i, \mathbf{x}'_i)\right)\right| \leq \frac{b}{m}, \tag{17}$$

since $\ell(y, f(\mathbf{x}))$, and thus $\tilde{\ell}(y, f(\mathbf{x}))$, is bounded.

$$\Pr\left[\, |g(\mathbf{Z}) - \mathbb{E}[g(\mathbf{Z})]| \geq t \,\right] \leq 2\exp\left(\frac{-2mt^2}{b^2}\right). \tag{18}$$

Furthermore, note that $\mathbb{E}[g(\mathbf{Z})] = \mathbb{E}[\tilde{\ell}_{\nu_f}(y, \mathbf{x})]$ (linearity of expectation), and thus we have that

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}_\nu(y_i, f_{\mathbf{D},\mathbf{w}}(\mathbf{x}_i)) - \mathop{\mathbb{E}}_{(\mathbf{x},y)} \tilde{\ell}_\nu(y, f_{\mathbf{D},\mathbf{w}}(\mathbf{x}))\right| > t\right] \leq 2\exp\left(\frac{-2mt^2}{b^2}\right). \tag{19}$$

Next, using a union bound argument, we can bound the probability over the supremum:

$$\Pr\left[\sup_j \left|\frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(y_i, f_{\mathbf{D}_j,\mathbf{w}_j}(\mathbf{x}_i)) - \mathop{\mathbb{E}}_{(\mathbf{x},y)} \tilde{\ell}(y, f_{\mathbf{D}_j,\mathbf{w}_j}(\mathbf{x}))\right| > t\right] \leq$$

$$\sum_{j=1}^{N^{cov}} \Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(y_i, f_{\mathbf{D}_j,\mathbf{w}_j}(\mathbf{x}_i)) - \mathop{\mathbb{E}}_{(\mathbf{x},y)} \tilde{\ell}(y, f_{\mathbf{D}_j,\mathbf{w}_j}(\mathbf{x}))\right| > t\right] \leq 2N^{cov}\exp\left(\frac{-2mt^2}{b^2}\right). \tag{20}$$

Denote this failure probability as $\delta/2$. Thus, with probability at least $1 - \delta/2$, we get

$$\sup_j \left|\frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(y_i, f_{\mathbf{D}_j,\mathbf{w}_j}(\mathbf{x}_i)) - \mathop{\mathbb{E}}_{(\mathbf{x},y)}\left[\tilde{\ell}(y, f_{\mathbf{D}_j,\mathbf{w}_j}(\mathbf{x}))\right]\right| \leq b\sqrt{\frac{\log(N^{cov}) + \log(4/\delta)}{2m}}. \tag{21}$$

Let us now focus on the second and third terms in Eq. (14). In particular, we will upper bound them by analyzing the Lipschitz continuity of the loss function with respect to the parameters, $\mathbf{D}$ and $\mathbf{w}$. We assume that ℓ is L_ℓ-Lipschitz, and we analyze the Lipschitz continuity of $\tilde{\ell}$ w.r.t $\mathbf{D}$ through $f_{\mathbf{D}}(\mathbf{x})$. Noting that the difference of the maxima is upper-bounded by the maximum of the difference, we can write

$$\left|\tilde{\ell}(y, f_{\mathbf{D}}(\mathbf{x})) - \tilde{\ell}(y, f_{\mathbf{D}'}(\mathbf{x}))\right| = \left|\max_{\mathbf{v}\in\Delta} \ell(y, f_{\mathbf{D}}(\mathbf{x}+\mathbf{v})) - \max_{\mathbf{v}\in\Delta} \ell(y, f_{\mathbf{D}'}(\mathbf{x}+\mathbf{v}))\right| \tag{22}$$

$$\leq \max_{\mathbf{v}\in\Delta} \left|\ell(y, f_{\mathbf{D}}(\mathbf{x}+\mathbf{v})) - \ell(y, f_{\mathbf{D}'}(\mathbf{x}+\mathbf{v}))\right| \tag{23}$$

$$\leq L_\ell \max_{\mathbf{v}\in\Delta} |\langle \mathbf{w}^T, \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})\rangle - \langle \mathbf{w}^T, \varphi_{\mathbf{D}'}(\mathbf{x}+\mathbf{v})\rangle| \tag{24}$$

$$\leq L_\ell \|\mathbf{w}\|_2 \max_{\mathbf{v}\in\Delta} \|\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) - \varphi_{\mathbf{D}'}(\mathbf{x}+\mathbf{v})\|_2. \tag{25}$$

We will now bound the term $\max_{\mathbf{v}\in\Delta} \|\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) - \varphi_{\mathbf{D}'}(\mathbf{x}+\mathbf{v})\|_2$. Notice that if the dictionary $\mathbf{D}$ has an encoder gap of at least τ_s^* for an input sample $\mathbf{x}$, then we can use Lemma 4.2 to obtain

$$\max_{\mathbf{v}\in\Delta} \|\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) - \varphi_{\mathbf{D}'}(\mathbf{x}+\mathbf{v})\|_2 \leq \frac{3}{2}\frac{\sqrt{s}(1+\nu)^2}{\lambda(1-\eta_s)}\epsilon.$$

Denote the probability of this event (that $\tau_s(\mathbf{x}) > \tau_s^*$) by $1-\rho$. Note that $\epsilon \leq 2\lambda/(1+\nu)^2$ is required in order to apply Lemma 4.2, but this condition is mild and we will later show that this holds under the condition of minimal number of samples.

Likewise, $\tilde{\ell}$ is Lipschitz continuous w.r.t $\mathbf{w}$,

$$\left|\tilde{\ell}(y, f_{\mathbf{D},\mathbf{w}}(\mathbf{x})) - \tilde{\ell}(y, f_{\mathbf{D},\mathbf{w}'}(\mathbf{x}))\right| \leq L_\ell \max_{\mathbf{v}\in\Delta} |\langle \mathbf{w}^T, \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})\rangle - \langle \mathbf{w}'^T, \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})\rangle| \tag{26}$$

$$\leq \frac{L_\ell(1+\nu)^2}{\lambda}\|\mathbf{w}-\mathbf{w}'\|_2, \tag{27}$$

since $\max_{\mathbf{v}\in\Delta} \|\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})\|_2 = (1+\nu)^2/\lambda$ (Remark B.2). Furthermore, $\|\mathbf{w}-\mathbf{w}'\|_2 \leq \epsilon$, as follows from our definition of ϵ-cover.

On the other hand, if $\mathbf{D}$ does not achieve this encoder gap ($\tau_s(\mathbf{x}) < \tau^*$), which happens with probability ρ, then we can simply upper bound the worst possible loss, i.e. $\left|\tilde{\ell}(y, f_{\mathbf{D}}(\mathbf{x})) - \tilde{\ell}(y, f_{\mathbf{D}'}(\mathbf{x}))\right| \leq b$.

Let us now analyze this probability, ρ. For simplicity, assume that $\tau_s(\mathbf{x}_i)$ are i.i.d. random variables. e.g. by computing $\tau_s(\mathbf{x}_i)$ on a held-out set with m_2 samples, independent from the m_1 samples that are used to train the dictionary. In particular, we split training and development samples m_1 and m_2 equally $m_1 = m_2 = m/2$. Let $F_{m_2}(\tau) := \frac{1}{m_2}\sum_{i=1}^{m_2} 1_{\{\tau_s(\mathbf{x}_i)<\tau\}}$ denote the fraction of training points that achieve the encoder margin smaller than τ. Let $F(\tau) := \Pr(\tau_s(\mathbf{x}) < \tau)$. Then, uniform convergence [Mohri et al., 2018] yields that for any $\delta/2 > 0$, with probability at least $1-\delta/2$, we have that $\sup_{\tau\in\mathbb{R}} |F_{m_2}(\tau) - F(\tau)| \leq c\sqrt{\frac{\log(m_2)+\log(2/\delta)}{m_2}}$, for some constant c. Since this holds uniformly for any τ, it holds in particular for $\tau = \tau_s^*$. This implies then that $F(\tau) = \Pr(\tau_s(\mathbf{x}) \leq \tau^*) \leq c\sqrt{\frac{\log(m_2)+\log(1/\delta_2)}{m_2}} = \rho$.

Note that the third term in Eq. (14) involves the expectation over the population, and so we can upper bound that term by

$$\frac{L_\ell(1+\nu)^2}{\lambda}\left(1+\frac{3}{2}\frac{B\sqrt{s}}{(1-\eta_s)}\right)\epsilon + \frac{b}{\sqrt{m_2}}c\left(\sqrt{\log(m_2)+\log(2/\delta)}\right),$$

with probability at least $1-\delta/2$. The second term, on the other hand, is the average loss over the training samples. For this, it suffices to note that the uniform bound $\sup_{\tau\in\mathbb{R}} |F_{m_2}(\tau) - F(\tau)|$ holds for *any* choice of τ. In particular, it holds for τ_s^* defined over both training and development samples. As a result, the dictionary satisfies the encoder gap on those samples, and so the second term can be simply upper bounded by

$$\frac{L_\ell(1+\nu)^2}{\lambda}\left(1+\frac{3}{2}\frac{B\sqrt{s}}{(1-\eta_s)}\right)\epsilon.$$

We finally get expressions for the covering number as a function of ϵ. For oblique manifolds (matrices of size $n\times p$ with unit norm columns), $N^{cov}(\mathcal{D},\epsilon) \leq (3/\epsilon)^{dp}$ [Seibert, 2019], while for B-bounded vectors $N^{cov}(\mathcal{W},\epsilon) \leq (3B/\epsilon)^p$. Thus, the covering number of the direct product of the two constraint sets can be bounded by $N^{cov}(\mathcal{D},\mathcal{W},\epsilon) \leq (3/\epsilon)^{(d+1)p}\, B^p$.

Gathering everything together, we can bound the generalization error by

$$\left|\tilde{R}_S(f) - \tilde{R}(f)\right| \leq b\sqrt{\frac{(d+1)p\log(3/\epsilon) + p\log(B) + \log(4/\delta)}{m}} + bc\sqrt{2\frac{\log(m/2)+\log(2/\delta)}{m}} + \frac{2L_\ell(1+\nu)^2}{\lambda}\left(1+\frac{3}{2}\frac{B\sqrt{s}}{(1-\eta_s)}\right)\epsilon. \tag{28}$$

All that remains is to set ϵ appropriately. Set $\epsilon = \lambda(1-\eta_s)/m$, and so

$$\left|\tilde{R}_S(f) - \tilde{R}(f)\right| \leq b\sqrt{\frac{(d+1)p\log(3m/(2\lambda(1-\eta_s))) + p\log(B) + \log 4/\delta}{m}} + bc\sqrt{2\frac{\log(m/2)+\log(2/\delta)}{m}} + 12\frac{B\sqrt{s}L_\ell(1+\nu)^2}{m}. \quad (29)$$

Lastly, the results above holds for ϵ small enough. Due to Lemma Lemma B.6, one needs

$$\tau_s^* > 2\nu + \sqrt{\epsilon}\left(\sqrt{\frac{25}{\lambda}}(1+\nu) + 2\left(\frac{(1+\nu)}{2\lambda}+1\right)\right),$$

implying that

$$\sqrt{\frac{\lambda(1-\eta_s)}{m}} < \frac{\tau_s - 2\nu}{2\left(1+\frac{1+\nu}{2\lambda}\right) + (1+\nu)\sqrt{\frac{25}{\lambda}}}.$$

Recalling that, naturally, $0 < \sqrt{1/m}$, it is enough to require that $\tau_s > 2\nu$ and that

$$m > \frac{\lambda(1-\eta_s)}{(\tau_s - 2\nu)^2}\left(2\left(1+\frac{1+\nu}{2\lambda}\right) + (1+\nu)\sqrt{\frac{25}{\lambda}}\right)^2. \quad (30)$$

Lastly, we need to show that this condition guarantees the assumption $\epsilon \leq 2\lambda/(1+\nu)^2$ is satisfied, in order to apply Lemma 4.2 above. Note that $\epsilon = \lambda(1-\eta_s)/m \leq \lambda/m$. Thus, we need $\lambda/m \leq 2\lambda/(1+\nu)^2$, which is satisfied as long as $m \geq 2 \geq (1+\nu)^2/2$, which is satisfied in all relevant scenarios.

□

B.1 Parameter adversarial stability

In this section we prove the key result in Lemma 4.2, guaranteeing that the perturbation in the encoded features under model deviations and adversarial contamination is bounded. The main difficulty here is that the Lasso encoder solves a problem that is not strongly convex – due to the overcompleteness of the dictionary – and thus showing that the encoded features satisfy a Lipschitz property w.r.t the model parameters, particularly in the adversarial setting, is not trivial. As a result, this section will be dedicated to showing that if the model perturbation and adversarial contamination is small enough and there exist a positive encoder margin, then some sparsity is retained in the features after the respective perturbations. With this result at hand, the proof of Lemma 4.2 follows directly from the proof of Theorem 4 in [Mehta and Gray, 2013], albeit with different constants which account for the perturbation $\mathbf{v}$. This *preservation of sparsity* result is formalized later in Lemma B.6, and the following immediate lemmata will build some intermediate results needed for it.

We first make a few remarks about the encoded features. We assume that $\mathbf{x} \in \mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$, and recall that $\varphi_{\mathbf{D}}(\mathbf{x}) = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1$. We are interested in the result of the encoder when contaminated with an energy-bounded perturbation, namely

$$\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v}) = \arg\min_{\mathbf{z}} \frac{1}{2}\|(\mathbf{x}_0 + \mathbf{v}) - \mathbf{D}\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1, \quad (31)$$

where $\mathbf{v} \in \Delta_\nu = \{\mathbf{v} : \|\mathbf{v}\|_2 \leq \nu < 1\}$. We will often denote $\mathbf{x} = \mathbf{x}_0 + \mathbf{v}$ for simplicity. Also, note that there exist natural bounds for the penalty parameter, $0 \leq \lambda \leq (1+\nu)$. The upper bound follows from the observation that as long as $\lambda > \|\mathbf{D}^T\mathbf{x}\|_\infty$, the solution of Eq.(31) is the zero vector. Since the columns of $\mathbf{D}$ are normalized, $\|\mathbf{D}^T(\mathbf{x}_0+\mathbf{v})\|_\infty \leq \|\mathbf{x}_0+\mathbf{v}\|_2 \leq 1+\nu$.

Recall that from optimality conditions of Lasso, the solution $\varphi_{\mathbf{D}}(\mathbf{x})$ satisfies

$$|\mathbf{D}_i^T(\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}))| = \lambda \quad \text{if} \quad [\varphi_{\mathbf{D}}(\mathbf{x})]_i \neq 0 \quad (32)$$

$$|\mathbf{D}_i^T(\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}))| \leq \lambda \quad \text{if} \quad [\varphi_{\mathbf{D}}(\mathbf{x})]_i = 0. \quad (33)$$

Lastly, recall that the encoder gap assumption ($\tau_s \geq \tau* > 0$) implies that there exist a set of inactive $(p-s)$ atoms $\mathcal{I}$ so that

$$|\mathbf{D}_i^T(\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}))| < \lambda - \tau_s$$

for all $i \in \mathcal{I}$.

Let us now formalize a few properties on the solution of the Lasso solution that will be used throughout.

Remark B.2. *For the setting above, we have that*

a) $\|(\mathbf{x}_0 + \mathbf{v}) - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2 \leq (1+\nu)$,

b) $\|\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2 \leq (1+\nu)^2/(2\lambda)$,

c) $\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2 \leq (1+\nu)$,

Proof. Remarks $a)$ and $b)$ can be shown by noting that, by definition of the encoder,

$$\frac{1}{2}\|(\mathbf{x}_0 + \mathbf{v}) - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2^2 + \lambda\|\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_1 \leq \frac{1}{2}\|(\mathbf{x}_0 + \mathbf{v})\|_2^2 \leq \frac{1}{2}(1+\nu)^2. \tag{34}$$

The above follows since the LHS is the minimum function value, attained precisely $\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})$, whereas $\frac{1}{2}\|(\mathbf{x}_0 + \mathbf{v})\|_2^2$ is the function value for the alternative choice of $\mathbf{z} = 0$. The right-most inequality follows from the triangle inequality on $\|\mathbf{x}_0 + \mathbf{v}\|_2$.

For remark $c)$, denote $\mathbf{x} = \mathbf{x}_0 + \mathbf{v}$, and note that the minimizer of the above optimization problem satisfies (as follows from optimality of the minimizer [Mehta and Gray, 2013, Lemma 13 of Supplementary])

$$\frac{1}{2}\|\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 + \lambda\|\varphi_{\mathbf{D}}(\mathbf{x})\|_1 = \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2. \tag{35}$$

We expand the LHS and obtain a lower bound by Cauchy-Schwarz (and dropping the ℓ_1 term)

$$\frac{1}{2}\|\mathbf{x}\| + \frac{1}{2}\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 - \mathbf{x}^T\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}) + \lambda\|\varphi_{\mathbf{D}}(\mathbf{x})\|_1 \geq \frac{1}{2}\|\mathbf{x}\| + \frac{1}{2}\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 - \|\mathbf{x}\|_2\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2 \tag{36}$$

$$\geq \frac{1}{2}\|\mathbf{x}\| + \frac{1}{2}\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 - (1+\nu)\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2 \tag{37}$$

Thus, together with (35), we have that $\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x} + \mathbf{v})\|_2 \leq (1+\nu)$. □

Lemma B.3. *If* $\|\mathbf{D} - \mathbf{D}'\| \leq \epsilon \leq 2\lambda/(1+\nu)^2$*, then*

$$\max_{\mathbf{v}\in\Delta_\nu} \left| \|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2^2 - \|\mathbf{D}'\varphi_{\mathbf{D}'}(\mathbf{x}_0 + \mathbf{v})\|_2^2 \right| \leq \frac{5\epsilon}{2\lambda}(1+\nu)^2. \tag{38}$$

The proof mimics that in [Mehta and Gray, 2013, Lemma 10-11], though accommodating for the adversarial perturbation. We include it here for completeness. Note that the above assumption on $\epsilon \leq 2\lambda/(1+\nu)^2$ is mild, and it will hold under the setting of later lemmata.

Proof. Denote $\mathbf{x} = \mathbf{x}_0 + \mathbf{v}$. Let us further denote the optimal value attained by the encoders with one and other model as

$$v_{\mathbf{D}}^* = \min_z \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1,$$

$$v_{\mathbf{D}'}^* = \min_z \frac{1}{2}\|\mathbf{x} - \mathbf{D}'\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1.$$

Then, since this cost is only increased if using a different representation, we have that:

$$
\begin{aligned}
v^*_{\mathbf{D}'} \leq & \frac{1}{2}\|\mathbf{x} - \mathbf{D}'\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 + \lambda\|\varphi_{\mathbf{D}}(\mathbf{x})\|_1 && (39)\\
= & \frac{1}{2}\|\mathbf{x} - \mathbf{D}'\varphi_{\mathbf{D}}(\mathbf{x}) + (\mathbf{D} - \mathbf{D})\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 + \lambda\|\varphi_{\mathbf{D}}(\mathbf{x})\|_1 && (40)\\
\leq & \frac{1}{2}\|\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 + |\langle \mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}), (\mathbf{D} - \mathbf{D}')\varphi_{\mathbf{D}}(\mathbf{x})\rangle| + \frac{1}{2}\|(\mathbf{D} - \mathbf{D}')\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 + \dots && (41)\\
& \cdots + \lambda\|\varphi_{\mathbf{D}}(\mathbf{x})\|_1 && (42)\\
\leq & v^*_{\mathbf{D}} + \|\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2\|\mathbf{D} - \mathbf{D}'\|_2\|\varphi_{\mathbf{D}}(\mathbf{x})\|_2 + \frac{1}{2}\left(\|\mathbf{D} - \mathbf{D}'\|_2\|\varphi_{\mathbf{D}}(\mathbf{x})\|_2\right)^2 \quad \text{by C.Swz.} && (43)\\
\leq & v^*_{\mathbf{D}} + (1+\nu)\epsilon\frac{(1+\nu)^2}{2\lambda} + \frac{1}{2}\left(\frac{\epsilon(1+\nu)^2}{2\lambda}\right)^2 \quad \text{by Remark B.2} && (44)
\end{aligned}
$$

We further simplify the expression above by noting that $\nu < 1$ and that $\frac{\epsilon(1+\nu)^2}{2\lambda} \leq 1$ by assumption, obtaining

$$
v^*_{\mathbf{D}'} \leq v^*_{\mathbf{D}} + \frac{5\epsilon}{4\lambda}(1+\nu)^2. \tag{45}
$$

Thus, from this (and a symmetric argument) follows that

$$
|v^*_{\mathbf{D}'} - v^*_{\mathbf{D}}| \leq \frac{5\epsilon}{4\lambda}(1+\nu)^2. \tag{46}
$$

Lastly, recall from Eq. (35) that $v^*_{\mathbf{D}} = \frac{1}{2}\|\mathbf{x}\|_2^2 - \|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2$. Thus, using this expression for $v^*_{\mathbf{D}}$ and $v^*_{\mathbf{D}'}$ above, we get

$$
\left|\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 - \|\mathbf{D}'\varphi_{\mathbf{D}'}(\mathbf{x})\|_2^2\right| \leq 2\left|v^*_{\mathbf{D}'} - v^*_{\mathbf{D}}\right| \leq \frac{5\epsilon}{2\lambda}(1+\nu)^2. \tag{47}
$$

□

We now show that if the dictionaries are close, then the reconstructions from one and other encoded representation are not too far either.

Lemma B.4. *If* $\|\mathbf{D} - \tilde{\mathbf{D}}\| \leq \epsilon \leq 2\lambda/(1+\nu)^2$, *then*

$$
\max_{\mathbf{v}\in\Delta_\nu} \|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v}) - \mathbf{D}\varphi_{\tilde{\mathbf{D}}}(\mathbf{x}_0 + \mathbf{v})\|_2^2 \leq \frac{25\epsilon}{\lambda}(1+\nu)^2. \tag{48}
$$

Proof. For simplicity, denote $\mathbf{x} = \mathbf{x}_0 + \mathbf{v}$, as well as $\boldsymbol{\alpha} = \varphi_{\mathbf{D}}(\mathbf{x})$ and $\tilde{\boldsymbol{\alpha}} = \varphi_{\tilde{\mathbf{D}}}(\mathbf{x})$, where $\|\mathbf{D} - \tilde{\mathbf{D}}\|_2 \leq \epsilon$. We first upper bound $\left|\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 - \|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2\right|$ by a sequence of algebraic manipulations:

$$\left|\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 - \|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2\right| \leq \left|\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 - \|\tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\|_2^2\right| + \left|\|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2 - \|\tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\|_2^2\right| \qquad \pm\|\tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\|_2^2, \text{ triang. inq.} \tag{49}$$

$$\text{Lemma B.3, } \pm\tilde{\mathbf{D}} \leq \frac{5\epsilon}{2\lambda}(1+\nu)^2 + \left|\langle \mathbf{D}\tilde{\boldsymbol{\alpha}}, (\mathbf{D} - \tilde{\mathbf{D}} + \tilde{\mathbf{D}})\tilde{\boldsymbol{\alpha}}\rangle - \langle \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\rangle\right| \tag{50}$$

$$= \frac{5\epsilon}{2\lambda}(1+\nu)^2 + \left|\langle \mathbf{D}\tilde{\boldsymbol{\alpha}}, (\mathbf{D} - \tilde{\mathbf{D}})\tilde{\boldsymbol{\alpha}}\rangle + \langle \mathbf{D}\tilde{\boldsymbol{\alpha}} - \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\rangle\right| \tag{51}$$

$$\text{by } \pm\mathbf{D} \quad = \frac{5\epsilon}{2\lambda}(1+\nu)^2 + \left|\langle \mathbf{D}\tilde{\boldsymbol{\alpha}}, (\mathbf{D} - \tilde{\mathbf{D}})\tilde{\boldsymbol{\alpha}}\rangle + \langle \mathbf{D}\tilde{\boldsymbol{\alpha}} - \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}, (\tilde{\mathbf{D}} - \mathbf{D} + \mathbf{D})\tilde{\boldsymbol{\alpha}}\rangle\right| \tag{52}$$

$$= \frac{5\epsilon}{2\lambda}(1+\nu)^2 + \left|\langle \mathbf{D}\tilde{\boldsymbol{\alpha}}, (\mathbf{D} - \tilde{\mathbf{D}})\tilde{\boldsymbol{\alpha}}\rangle + \langle (\mathbf{D} - \tilde{\mathbf{D}})\tilde{\boldsymbol{\alpha}}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle - \langle (\mathbf{D} - \tilde{\mathbf{D}})\tilde{\boldsymbol{\alpha}}, (\mathbf{D} - \tilde{\mathbf{D}})\tilde{\boldsymbol{\alpha}}\rangle\right| \tag{53}$$

$$\leq \frac{5\epsilon}{2\lambda}(1+\nu)^2 + 2\left|\langle \mathbf{D}\tilde{\boldsymbol{\alpha}}, (\mathbf{D} - \tilde{\mathbf{D}})\tilde{\boldsymbol{\alpha}}\rangle\right| \qquad \text{by dropping } -\|(\mathbf{D} - \tilde{\mathbf{D}})\tilde{\alpha}\|_2^2 \tag{54}$$

$$\leq \frac{5\epsilon}{2\lambda}(1+\nu)^2 + 2\|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2\|\mathbf{D} - \tilde{\mathbf{D}}\|_2\|\tilde{\boldsymbol{\alpha}}\|_2 \qquad \text{by C.S. and operator norm} \tag{55}$$

$$\leq \frac{5\epsilon}{2\lambda}(1+\nu)^2 + 2\frac{\epsilon}{2\lambda}(1+\nu)^2\|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2 \qquad \text{by Remark B.2} \tag{56}$$

The term $\|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2$ cannot be directly bounded via Remark B.2 because $\tilde{\boldsymbol{\alpha}}$ is the representation computed via $\tilde{\mathbf{D}}$ (not $\mathbf{D}$). Instead, by letting $\Delta = \mathbf{D} - \tilde{\mathbf{D}}$, we can simplify the above bound as $\|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2 \leq \|\tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\|_2 + \|\Delta\tilde{\boldsymbol{\alpha}}\|_2 \leq (1+\nu) + \epsilon(1+\nu)^2/(2\lambda) \leq 2 + 1 = 3$. Then, resuming above,

$$\left|\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 - \|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2\right| \leq \frac{5\epsilon}{2\lambda}(1+\nu)^2 + \frac{6\epsilon}{2\lambda}(1+\nu)^2 = \frac{11\epsilon}{2\lambda}(1+\nu)^2. \tag{57}$$

Now, by definition of $\boldsymbol{\alpha}$ (as the minimizer of the Lasso problem), we have that

$$v_{\mathbf{D}}^*(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \leq \frac{1}{2}\left\|\mathbf{x} - \mathbf{D}\left(\frac{\boldsymbol{\alpha} + \tilde{\boldsymbol{\alpha}}}{2}\right)\right\|_2^2 + \lambda\left\|\frac{\boldsymbol{\alpha} + \tilde{\boldsymbol{\alpha}}}{2}\right\|_1. \tag{58}$$

We now expand the RHS above through the same algebraic manipulations:

$$v_{\mathbf{D}}^*(\mathbf{x}) \leq \frac{1}{2}\left\|\mathbf{x} - \mathbf{D}\left(\frac{\boldsymbol{\alpha} + \tilde{\boldsymbol{\alpha}}}{2}\right)\right\|_2^2 + \lambda\left\|\frac{\boldsymbol{\alpha} + \tilde{\boldsymbol{\alpha}}}{2}\right\|_1 \tag{59}$$

$$= \frac{1}{2}\left(\|\mathbf{x}\|_2^2 - \langle \mathbf{x}, (\mathbf{D}\boldsymbol{\alpha} + \mathbf{D}\tilde{\boldsymbol{\alpha}})\rangle + \frac{1}{4}\|\mathbf{D}\boldsymbol{\alpha} + \mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2\right) + \lambda\left\|\frac{\boldsymbol{\alpha} + \tilde{\boldsymbol{\alpha}}}{2}\right\|_1 \tag{60}$$

$$= \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\boldsymbol{\alpha}\rangle - \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \frac{1}{8}\left(\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2 + 2\langle \mathbf{D}\boldsymbol{\alpha}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle\right) + \lambda\left\|\frac{\boldsymbol{\alpha} + \tilde{\boldsymbol{\alpha}}}{2}\right\|_1 \tag{61}$$

$$\leq \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\boldsymbol{\alpha}\rangle - \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \frac{1}{4}\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{1}{4}\langle \mathbf{D}\boldsymbol{\alpha}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \dots \tag{62}$$

$$\cdots + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_1 + \frac{\lambda}{2}\|\tilde{\boldsymbol{\alpha}}\|_1 + \frac{11}{16}\frac{\epsilon}{\lambda}(1+\nu)^2, \tag{63}$$

where the last step follows by adding and subtracting $\|\mathbf{D}\boldsymbol{\alpha}\|_2^2$ and employing the bound obtained above in (57). Now, from Eq. (35), we can write

$$\lambda\|\boldsymbol{\alpha}\|_1 = \langle \mathbf{x} - \mathbf{D}\boldsymbol{\alpha}, \mathbf{D}\boldsymbol{\alpha}\rangle = \langle \mathbf{x}, \mathbf{D}\boldsymbol{\alpha}\rangle - \|\mathbf{D}\boldsymbol{\alpha}\|_2^2. \tag{64}$$

The expression for $\|\tilde{\boldsymbol{\alpha}}\|_1$ is expanded similarly but then upper bounded via Lemma B.3 by adding and subtracting $\|\mathbf{D}\boldsymbol{\alpha}\|_2^2$:

$$\lambda\|\tilde{\boldsymbol{\alpha}}\|_1 = \langle \mathbf{x} - \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\rangle \leq \langle \mathbf{x}, \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\rangle - \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{5\epsilon}{2\lambda}(1+\nu)^2 \tag{65}$$

$$= \langle \mathbf{x}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \langle \mathbf{x}, (\tilde{\mathbf{D}} - \mathbf{D})\tilde{\boldsymbol{\alpha}}\rangle - \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{5\epsilon}{2\lambda}(1+\nu)^2 \qquad \text{by } \pm\mathbf{D} \tag{66}$$

$$\text{by C.S.} \leq \langle \mathbf{x}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \|\mathbf{x}\|_2\|\tilde{\mathbf{D}} - \mathbf{D}\|_2\|\tilde{\boldsymbol{\alpha}}\|_2 - \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{5\epsilon}{2\lambda}(1+\nu)^2 \tag{67}$$

$$\leq \langle \mathbf{x}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + (1+\nu)\epsilon\frac{(1+\nu)^2}{2\lambda} - \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{5\epsilon}{2\lambda}(1+\nu)^2 \tag{68}$$

$$\leq \langle \mathbf{x}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle - \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{7\epsilon}{2\lambda}(1+\nu)^2. \tag{69}$$

Thus, we can now upper bound the expression for $v^*_{\mathbf{D}}$ in Eq. (35) by combining Eq. (62), (64) and (69) as follows. From Eq. (35) and the bound in Eq. (62) we get:

$$v^*_{\mathbf{D}} = \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 \tag{70}$$

$$\leq \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\boldsymbol{\alpha}\rangle - \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \frac{1}{4}\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{1}{4}\langle \mathbf{D}\boldsymbol{\alpha}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \ldots... \tag{71}$$

$$\cdots + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_1 + \frac{\lambda}{2}\|\tilde{\boldsymbol{\alpha}}\|_1 + \frac{11}{16}\frac{\epsilon}{\lambda}(1+\nu)^2. \tag{72}$$

Replacing now the expression for $\lambda\|\boldsymbol{\alpha}\|_1$ from (64) and the upper bound for $\lambda\|\tilde{\boldsymbol{\alpha}}\|_1$ from (69):

$$v^*_{\mathbf{D}} = \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{3}{4}\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{1}{4}\langle \mathbf{D}\boldsymbol{\alpha}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \frac{39}{16}\frac{\epsilon}{\lambda}(1+\nu)^2, \tag{73}$$

which leads to

$$-\frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq -\frac{3}{4}\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{1}{4}\langle \mathbf{D}\boldsymbol{\alpha}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \frac{39}{16}\frac{\epsilon}{\lambda}(1+\nu)^2, \tag{74}$$

and so

$$\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \langle \mathbf{D}\boldsymbol{\alpha}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle + \frac{39}{4}\frac{\epsilon}{\lambda}(1+\nu)^2. \tag{75}$$

Finally, with this expression we can now bound the distance

$$\|\mathbf{D}\boldsymbol{\alpha} - \mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2 = \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2 - 2\langle \mathbf{D}\boldsymbol{\alpha}, \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle \tag{76}$$

$$\leq \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2 - 2\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{39}{2}\frac{\epsilon}{\lambda}(1+\nu)^2 \qquad \text{by the expression above} \tag{77}$$

$$= \|\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_2^2 - \|\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \frac{39}{2}\frac{\epsilon}{\lambda}(1+\nu)^2 \tag{78}$$

$$\leq 25\frac{\epsilon}{\lambda}(1+\nu)^2, \tag{79}$$

where the last inequality follows from Eq. (57).

□

We will now show that if the dictionaries are close enough and if the solution of one of them was at most s-sparse and has a positive encoder gap, then the solution with the perturbed model retains the sparsity. This is inspired by the result in [Mehta and Gray, 2013]. However, because of the adversarial perturbation, extra work is required to provide a condition on the *unperturbed* gap, i.e. that which will withstand adversarial energy-bounded perturbation. The following Lemma will be necessary to show the result:

Lemma B.5. *If* $\|\mathbf{v}\|_2 \leq \nu$, *then*

$$\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}) - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})\|_2^2 \leq \nu^2. \tag{80}$$

Let us postpone the proof of this result for later. We are now ready to state and prove the preservation of sparsity result:

Lemma B.6. *(Preservation of sparsity under model deviation and adversarial perturbations) Consider $\varphi_{\mathbf{D}}(\mathbf{x}_0+\mathbf{v})$, for $\|\mathbf{v}\|_2 \leq \nu$, and an alternative dictionary $\tilde{\mathbf{D}}$ so that $\|\mathbf{D}-\tilde{\mathbf{D}}\|_2 \leq \epsilon \leq 2\lambda/(1+\nu)^2$. If there exist a set of inactive $(p-s)$ atoms $\mathcal{I}$ so that*

$$|\mathbf{D}_i^T(\mathbf{x}_0 - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0))| < \lambda - \tau_s \tag{81}$$

for all $i \in \mathcal{I}$, and

$$\tau_s > 2\nu + \sqrt{\epsilon}\left(\sqrt{\frac{25}{\lambda}}(1+\nu) + 2\left(\frac{(1+\nu)}{\lambda}+1\right)\right), \tag{82}$$

then $[\varphi_{\tilde{\mathbf{D}}}(\mathbf{x}_0+\mathbf{v})]_i = 0\ \forall i \in \mathcal{I}$, where (reminder)

$$\varphi_{\tilde{\mathbf{D}}}(\mathbf{x}_0+\mathbf{v}) = \arg\min_{\mathbf{z}} \frac{1}{2}\|(\mathbf{x}_0+\mathbf{v}) - \tilde{\mathbf{D}}\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1. \tag{83}$$

Proof. Let $\mathbf{x} = \mathbf{x}_0+\mathbf{v}$, as well as $\boldsymbol{\alpha} = \varphi_{\mathbf{D}}(\mathbf{x})$, $\tilde{\boldsymbol{\alpha}} = \varphi_{\tilde{\mathbf{D}}}(\mathbf{x})$, and let $\mathcal{I}$ be the set of $(p-s)$ inactive atoms with positive gap τ_s.

In order for the inactive set of atoms $\mathcal{I}$ to remain inactive, we need to show that $\forall i \in \mathcal{I}$,

$$\left|\langle \tilde{\mathbf{D}}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\rangle\right| < \lambda.$$

Consider the following upper bound to the LHS above:

$$\begin{aligned}
\left|\langle \tilde{\mathbf{D}}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\rangle\right| &\leq \left|\langle \mathbf{D}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\rangle\right| + \|\tilde{\mathbf{D}}_i - \mathbf{D}_i\|_2\|\mathbf{x} - \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\|_2 && \text{by } \pm\mathbf{D}_i \text{ and C.S.} && (84)\\
&\leq \left|\langle \mathbf{D}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\boldsymbol{\alpha}}\rangle\right| + \epsilon(1+\nu) && && (85)\\
&\leq |\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle| + \left|\langle \mathbf{D}_i, (\tilde{\mathbf{D}}-\mathbf{D})\tilde{\boldsymbol{\alpha}}\rangle\right| + \epsilon(1+\nu) && \text{by } \pm\mathbf{D} \text{ and triang ineq.} && (86)\\
&\leq |\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle| + \|\mathbf{D}_i\|_2\|\tilde{\mathbf{D}}-\mathbf{D}\|_2\|\tilde{\boldsymbol{\alpha}}\|_2 + \epsilon(1+\nu) && && (87)\\
&\leq |\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle| + \frac{\epsilon}{2\lambda}(1+\nu)^2 + \epsilon(1+\nu). && \text{by Remark B.2 and unit-norm columns} && (88)
\end{aligned}$$

Thus, it is sufficient to show that

$$|\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle| < \lambda - \epsilon(1+\nu)\left(\frac{(1+\nu)}{2\lambda}+1\right). \tag{89}$$

Let us now replace $\mathbf{x}$, $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$ by their definitions and upper bound the left hand side above by using Lemma B.4 and Lemma B.5

$$\begin{aligned}
|\langle \mathbf{D}_i, \mathbf{x} - \mathbf{D}\tilde{\boldsymbol{\alpha}}\rangle| &= |\langle \mathbf{D}_i, (\mathbf{x}_0+\mathbf{v}) - \mathbf{D}\varphi_{\tilde{\mathbf{D}}}(\mathbf{x}_0+\mathbf{v})\rangle| && && (90)\\
&\leq \left|\langle \mathbf{D}_i, (\mathbf{x}_0+\mathbf{v}) - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0+\mathbf{v})\right| + |\langle \mathbf{D}_i, \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0+\mathbf{v}) - \mathbf{D}\varphi_{\tilde{\mathbf{D}}}(\mathbf{x}_0+\mathbf{v})| && \text{by } \pm\boldsymbol{\alpha} && (91)\\
&\leq \overbrace{|\langle \mathbf{D}_i, \mathbf{x}_0 - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0+\mathbf{v})| + \|\mathbf{D}_i\|_2\|\mathbf{v}\|_2}^{} + \|\mathbf{D}_i\|_2\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0+\mathbf{v}) - \mathbf{D}\varphi_{\tilde{\mathbf{D}}}(\mathbf{x}_0+\mathbf{v})\|_2 && && (92)\\
&\leq |\langle \mathbf{D}_i, \mathbf{x}_0 - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0+\mathbf{v})| + \nu + \sqrt{\frac{25\epsilon}{\lambda}}(1+\nu) && \text{by Lemma B.4} && (93)\\
&\leq |\langle \mathbf{D}_i, \mathbf{x}_0 - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0)| + \|\mathbf{D}_i\|_2\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0) - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0+\mathbf{v})\|_2 + \nu + \sqrt{\frac{25\epsilon}{\lambda}}(1+\nu) && \text{by } \pm\varphi_{\mathbf{D}}(\mathbf{x}_0) && (94)\\
&\leq |\langle \mathbf{D}_i, \mathbf{x}_0 - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0)| + 2\nu + \sqrt{\frac{25\epsilon}{\lambda}}(1+\nu) && \text{by Lemma B.5} && (95)\\
&< \lambda - \tau_s + 2\nu + \sqrt{\frac{25\epsilon}{\lambda}}(1+\nu) && && (96)
\end{aligned}$$

where the last step follows from the assumption of the encoder gap in Eq. (81). Thus, merging with (89), we require

$$-\tau_s + 2\nu + \sqrt{\frac{25\epsilon}{\lambda}}(1+\nu) < -\epsilon(1+\nu)\left(\frac{(1+\nu)}{2\lambda}+1\right),$$

implying that as long as

$$\tau_s > 2\nu + \sqrt{\frac{25\epsilon}{\lambda}}(1+\nu) + \epsilon(1+\nu)\left(\frac{(1+\nu)}{2\lambda}+1\right) \tag{97}$$

the inactive set $\mathcal{I}$ remains inactive. For the sake of simplicity, we will make the above condition more stringent. Using the fact that $\nu < 1$ and that $\epsilon \leq 1$. Thus,

$$\tau_s > 2\nu + \sqrt{\epsilon}\left(\sqrt{\frac{25}{\lambda}}(1+\nu) + 2\left(\frac{(1+\nu)}{2\lambda}+1\right)\right). \tag{98}$$

□

The lemma above is central, as it guarantees that a sparsity of up to s non-zeros is retained under model deviations and adversarial perturbations. Lemma B.3 now follows directly from the proof of Theorem 4 in [Mehta and Gray, 2013], albeit with the constants provided by Remark B.2 which account for the perturbation $\mathbf{v}$.

We owe the proof of Lemma B.5. This Lemma will also be instrumental in the proof of Theorem C.1. We now re-state it, and proceed to prove it.

Lemma B.7. *5.4 (Norm Stability under adversarial perturbations)*
If $\|\mathbf{v}\|_2 \leq \nu$, *then*

$$\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}) - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v})\|_2^2 \leq \nu^2 \tag{99}$$

Proof. We will re-formulate the Lasso problem as an equivalent quadratic program, and then utilize optimality properties of its solution. Let us define the vector $\bar{\mathbf{z}} \in \mathbb{R}^{3p}$ such that $\bar{\mathbf{z}} = [\mathbf{z}, \mathbf{z}^+, \mathbf{z}^-]^T$, with $\mathbf{z}^+$ and $\mathbf{z}^-$ containing all positive and negative elements in $\mathbf{z}$, respectively. Define then the following quadratic cost

$$Q(\bar{\mathbf{z}}, \mathbf{x}) = \frac{1}{2}\bar{\mathbf{z}}^T \begin{bmatrix} \mathbf{D}^T\mathbf{D} & \mathbf{0}_{p\times 2p} \\ \mathbf{0}_{2p\times p} & \mathbf{0}_{2p\times 2p} \end{bmatrix} \bar{\mathbf{z}} - \bar{\mathbf{z}}^T \begin{bmatrix} \mathbf{D}^T \\ \mathbf{0}_{2p\times d} \end{bmatrix} \mathbf{x} + \lambda[\mathbf{0}_p^T, \mathbf{1}_{2p}^T]\bar{\mathbf{z}}. \tag{100}$$

With this definition, the Lasso problem can be re-formulated as the following quadratic program:

$$\min_{\bar{\mathbf{z}}\in\mathbb{R}^{3p}} Q(\bar{\mathbf{z}}, \mathbf{x}) \quad \text{subject to} \quad \bar{\mathbf{z}} \in \mathcal{K} = \{\bar{\mathbf{z}} : \mathbf{z} = \mathbf{z}^+ - \mathbf{z}^-;\ \mathbf{z}^+, \mathbf{z}^- \geq 0\}. \tag{101}$$

Let us denote $Q(\bar{\mathbf{z}}) = Q(\bar{\mathbf{z}}, \mathbf{x}_0)$ and $\tilde{Q}(\bar{\mathbf{z}}) = Q(\bar{\mathbf{z}}, \mathbf{x}_0 + \mathbf{v})$ for short, and $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ as the solution to the quadratic program with $Q(\bar{\mathbf{z}})$ and $\tilde{Q}(\bar{\mathbf{z}})$, respectively. Moreover, denote $\boldsymbol{\alpha} = \varphi_{\mathbf{D}}(\mathbf{x}_0)$ and $\tilde{\boldsymbol{\alpha}} = \varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})$. With this notation, note that

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^+ \\ \boldsymbol{\alpha}^- \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\beta}} = \begin{bmatrix} \tilde{\boldsymbol{\alpha}} \\ \tilde{\boldsymbol{\alpha}}^+ \\ \tilde{\boldsymbol{\alpha}}^- \end{bmatrix}.$$

Note that the above problem in (101) is the minimization of a convex differentiable function over a convex set and therefore, for every $\bar{\mathbf{z}} \in \mathcal{K}$,

$$(\bar{\mathbf{z}} - \boldsymbol{\beta})^T \nabla_{\mathbf{z}} Q(\boldsymbol{\beta}) \geq 0 \tag{102a}$$

$$(\bar{\mathbf{z}} - \tilde{\boldsymbol{\beta}})^T \nabla_{\mathbf{z}} \tilde{Q}(\tilde{\boldsymbol{\beta}}) \geq 0. \tag{102b}$$

This gradient can be written as

$$\nabla Q(\bar{\mathbf{z}}) = \begin{bmatrix} \mathbf{D}^T\mathbf{D} & \mathbf{0}_{p\times 2p} \\ \mathbf{0}_{2p\times p} & \mathbf{0}_{2p\times 2p} \end{bmatrix} \bar{\mathbf{z}} - \begin{bmatrix} \mathbf{D}^T \\ \mathbf{0}_{2p\times d} \end{bmatrix} \mathbf{x} + \lambda \begin{bmatrix} \mathbf{0}_p \\ \mathbf{1}_{2p} \end{bmatrix}. \tag{103}$$

Now, choosing $\tilde{\boldsymbol{\beta}}$ as $\bar{\mathbf{z}}$ in (102a), $\boldsymbol{\beta}$ as $\bar{\mathbf{z}}$ in (102b) and subtracting one from the other, we get

$$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \left(\nabla Q(\boldsymbol{\beta}) - \nabla \tilde{Q}(\tilde{\boldsymbol{\beta}}) \right) \geq 0 \tag{104}$$

which after employing the definitions for $\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}, \nabla Q$ and $\nabla \tilde{Q}$ results in

$$(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^T \left(\mathbf{D}^T \mathbf{D}(\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}) + \mathbf{D}^T \mathbf{v} \right) \geq 0. \tag{105}$$

The lemma is proven by finally expanding the above and employing Cauchy-Schwarz:

$$\|\mathbf{D}\tilde{\boldsymbol{\alpha}} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq (\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^T \mathbf{D}^T \mathbf{v} \leq \|\mathbf{v}\|_2 \|\mathbf{D}\tilde{\boldsymbol{\alpha}} - \mathbf{D}\boldsymbol{\alpha}\|_2 \leq \nu \|\mathbf{D}\tilde{\boldsymbol{\alpha}} - \mathbf{D}\boldsymbol{\alpha}\|_2. \tag{106}$$

□

C Robustness Certificate

Theorem C.1 (Robustness certificate for binary predictive sparse coding)**.** *Consider the predictor $f_{\mathbf{D},\mathbf{w}}(\mathbf{x})$, computed via $\varphi_{\mathbf{D}}(\mathbf{x})$ with an encoder gap of $\tau_s(\mathbf{x})$ and η_s-RIP dictionary $\mathbf{D}$. Then,*

$$\mathit{sign}(f_{\mathbf{D},\mathbf{w}}(\mathbf{x})) = \mathit{sign}(f_{\mathbf{D},\mathbf{w}}(\mathbf{x} + \mathbf{v})), \quad \forall \mathbf{v} : \|\mathbf{v}\|_2 \leq \nu \tag{107}$$

so long as $\nu < \min\{\, \tau_s(\mathbf{x})/2 \,,\, \rho_{\mathbf{x}} \sqrt{1 - \eta_s} \,\}$.

We now proceed to prove Theorem C.1. We first must show that if there exist a positive encoder gap for a particular inactive set, this set will remain inactive under adversarial perturbations. This follows as a particular case of Lemma B.6 with $\epsilon = 0$, i.e. when there is no difference between the dictionaries: $\|\mathbf{D} - \tilde{\mathbf{D}}\|_2 = 0$. We re-state it here for completeness in this simplified form.

Corollary C.2. *Consider $\varphi_{\mathbf{D}}(\mathbf{x}_0)$ and $\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})$, for $\|\mathbf{v}\|_2 \leq \nu$. If there exist a set of inactive $(p - s)$ atoms $\mathcal{I}$ in $\varphi_{\mathbf{D}}(\mathbf{x}_0)$ so that*

$$|\mathbf{D}_i^T (\mathbf{x}_0 - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0))| < \lambda - \tau_s \tag{108}$$

for all $i \in \mathcal{I}$, and

$$\tau_s > 2\nu, \tag{109}$$

then $[\varphi_{\tilde{\mathbf{D}}}(\mathbf{x}_0 + \mathbf{v})]_i = 0 \; \forall i \in \mathcal{I}$.

With this result, we now present a Lemma guaranteeing that the original and adversarially perturbed representation are not too far.

Lemma 5.2 (Stability of representations under adversarial perturbations)**.** *Let $\varphi_{\mathbf{D}}(\mathbf{x}_0)$ and $\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})$, for $\|\mathbf{v}\|_2 \leq \nu$. If $\varphi_{\mathbf{D}}(\mathbf{x}_0)$ has an encoder gap $\tau_s > 2\nu$, and the dictionary is RIP with constant η_s, then*

$$\|\varphi_{\mathbf{D}}(\mathbf{x}_0) - \varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2 \leq \frac{\nu}{\sqrt{1 - \eta_s}}. \tag{110}$$

Proof. The proof of this result is now simple given our previous developments. On one hand, we have from Lemma B.7 that

$$\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0) - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2^2 \leq \nu^2. \tag{111}$$

On the other hand, since $\varphi_{\mathbf{D}}(\mathbf{x}_0)$ has an encoder gap of $\tau_s > 2\nu$, there exist an inactive set of $(p - s)$ atoms that is retained in $\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})$ by Corollary C.2. Thus, $\|\varphi_{\mathbf{D}}(\mathbf{x}_0) - \varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_0 \leq s$. As a result, since $\mathbf{D}$ is η_s-RIP, we can write

$$\|\mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0) - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2^2 \geq (1 - \eta_s) \|\varphi_{\mathbf{D}}(\mathbf{x}_0) - \varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2^2. \tag{112}$$

Combining the lower and upper bounds proves the claim. □

We are now ready to prove the result in Theorem C.1.

Proof. The proof is simple and inspired by the analysis in [Romano et al., 2019]. Recall that the hypothesis is implemented by $f_{\mathbf{D},\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \varphi_{\mathbf{D}}(\mathbf{x}) \rangle$. Since $\varphi_{\mathbf{D}}(\mathbf{x})$ has an encoder gap $\tau_s \geq 2\nu$, then it follows from the above Lemma 5.2 that

$$\|\varphi_{\mathbf{D}}(\mathbf{x}_0) - \varphi_{\mathbf{D}}(\mathbf{x}_0 + \mathbf{v})\|_2 \leq \frac{\nu}{\sqrt{1-\eta_s}}. \tag{113}$$

Without loss of generality, consider the case when $f_{\mathbf{D},\mathbf{w}}(\mathbf{x}) > 0$. Let us lower bound $f_{\mathbf{D},\mathbf{w}}(\mathbf{x}+\mathbf{v})$ as follows:

$$\begin{aligned}
\langle \mathbf{w}, \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) \rangle &= \langle \mathbf{w}, \varphi_{\mathbf{D}}(\mathbf{x}) \rangle + \langle \mathbf{w}, \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) - \varphi_{\mathbf{D}}(\mathbf{x}) \rangle && (114)\\
&\geq \rho_{\mathbf{x}} \|\mathbf{w}\|_2 - |\langle \mathbf{w}, \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) - \varphi_{\mathbf{D}}(\mathbf{x}) \rangle| && (115)\\
&\geq \|\mathbf{w}\|_2 \left(\rho_{\mathbf{x}} - \frac{\nu}{\sqrt{1-\eta_s}} \right). && (116)\\
& && (117)
\end{aligned}$$

Therefore, as long as $\rho_x > \frac{\nu}{\sqrt{1-\eta_s}}$ (and $\mathbf{w} \neq \mathbf{0}$), $sign(f_{\mathbf{D},\mathbf{w}}(\mathbf{x})) = sign(f_{\mathbf{D},\mathbf{w}}(\mathbf{x}+\mathbf{v}))$. □

Theorem 5.1 (Robustness Certificate for multiclass supervised sparse coding)**.** *Let ρ_x be the multiclass classifier margin of $f_{\mathbf{D},\mathbf{w}}(\mathbf{x})$, with $\rho_{\mathbf{x}} > 0$, composed of an encoder with gap of $\tau_s(\mathbf{x})$ and η_s-RIP dictionary $\mathbf{D}$. Furthermore, denote by $c_{\mathbf{W}} = \max_{i \neq j} \|\mathbf{W}_i - \mathbf{W}_j\|_2$ Then,*

$$\arg\max_{j \in [K]} \left[\mathbf{W}^T f_{\mathbf{D},\mathbf{w}}(\mathbf{x})\right] = \arg\max_{j \in [K]} \left[\mathbf{W}^T f_{\mathbf{D},\mathbf{w}}(\mathbf{x}+\mathbf{v})\right], \quad \forall\, \mathbf{v} : \|\mathbf{v}\|_2 \leq \nu \tag{118}$$

so long as $\nu \leq \min\{\tau_s(\mathbf{x})/2, \rho_{\mathbf{x}}\sqrt{1-\eta_s}/c_{\mathbf{W}}\}$.

Proof. Consider a sample with a positive multiclass margin:

$$\rho_{\mathbf{x}} = \mathbf{W}_y^T \varphi_{\mathbf{D}}(\mathbf{x}) - \max_{j \neq y} \mathbf{W}_j^T \varphi_{\mathbf{D}}(\mathbf{x}) > 0,$$

and let us lower-bound the margin on the perturbed input $f_{\mathbf{D},\mathbf{w}}(\mathbf{x}+\mathbf{v})$ as follows:

$$\begin{aligned}
\rho_{\mathbf{x}+\mathbf{v}} =& \mathbf{W}_y^T \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) - \max_{j \neq y} \mathbf{W}_j^T \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) && (119)\\
=& \min_{j \neq y} \langle \mathbf{W}_y - \mathbf{W}_j, \varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) \rangle && (120)\\
\geq& \min_{j \neq y} \langle \mathbf{W}_y - \mathbf{W}_j, \varphi_{\mathbf{D}}(\mathbf{x}) \rangle - \|\mathbf{W}_y - \mathbf{W}_j\|_2 \|\varphi_{\mathbf{D}}(\mathbf{x}+\mathbf{v}) - \varphi_{\mathbf{D}}(\mathbf{x})\|_2 && (121)\\
\geq& \rho_{\mathbf{x}} - c_{\mathbf{W}}\, \nu/\sqrt{1-\eta_s}, && (122)
\end{aligned}$$

where the second-to-last inequality follows from hypothesis and Lemma 5.2. Therefore, as long as $\rho_x > \frac{c_{\mathbf{W}}\nu}{\sqrt{1-\eta_s}}$, $\rho_{\mathbf{x}+\mathbf{v}} > 0$.

□

D Numerical Experiments Details

The models on images (MNIST and CIFAR10) were trained by minimizing the following regularized empirical risk

$$\min_{\mathbf{W},\mathbf{D}} \frac{1}{m} \sum_{i=1}^{m} \ell\left(y_i, \langle \mathbf{W}, \varphi_{\mathbf{D}}(\mathbf{x}_i) \rangle\right) + \alpha \|\mathbf{I} - \mathbf{D}^T\mathbf{D}\|_F^2 + \beta \|\mathbf{W}\|_F^2, \tag{123}$$

over the training set with m samples. We use the default training/testing split provided in the datasets. The difficulty in this optimization problem resides in computing the derivative of this loss w.r.t. the dictionary $\mathbf{D}$ via the solution of the encoder $\varphi_{\mathbf{D}}(\mathbf{x})$. Our approach relies on using an approximate but differentiable solution for $\varphi_{\mathbf{D}}(\mathbf{x})$: we compute the features (by solving the corresponding Lasso problem) via Fast Iterative Soft Thresholding [Beck and Teboulle, 2009]. This algorithm enjoys a fast convergence rate of $\mathcal{O}(1/T^2)$, and we use $T = 25$ iterations within the optimization problem above.

We found it useful to pre-train the model, $\mathbf{D}$, in an unsupervised manner first. This is done by simply minimizing a regression problem of the form

$$\min_{\mathbf{D}} \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}_i)\|_2^2.$$

Additionally, when performing the supervised learning stage, if progressively increase the value of λ through the iterations until the pre-specified target value (which where set to 0.2 and 0.3 in Figure 2c and Figure 2c, respectively). We employ Adam [Kingma and Ba, 2014] with a mini-batch size of 128, and train for 35 epochs. The dictionary is normalized after each weight-update. All other hyper-parameters are detailed in the accompanying code.

At deployment time, however, it is important that the solution computed by $\varphi_{\mathbf{D}}(\mathbf{x})$ is exact, because the encoder gap τ_s is defined in terms of these optimality conditions. Therefore, we use FISTA to find the estimated support of the solution, and then compute the exact solution analytically given this support.

All experiments were coded in Python and employing pytorch for GPU acceleration. All other employed packages are detailed in the accompanying code.