Bilevel Sparse Coding for Coupled Feature Spaces

Jianchao Yang[†], Zhaowen Wang[†], Zhe Lin[‡], Xianbiao Shu[†], Thomas Huang[†]
[†]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois
[‡]Adobe Systems Inc., San Jose, California
[†]{jyang29, wang308, xshu2, huang}@illinois.edu, [‡]zlin@adobe.com

Abstract

In this paper, we propose a bilevel sparse coding model for coupled feature spaces, where we aim to learn dictionaries for sparse modeling in both spaces while enforcing some desired relationships between the two signal spaces. We first present our new general sparse coding model that relates signals from the two spaces by their sparse representations and the corresponding dictionaries. The learning algorithm is formulated as a generic bilevel optimization problem, which is solved by a projected first-order stochastic gradient descent algorithm. This general sparse coding model can be applied to many specific applications involving coupled feature spaces in computer vision and signal processing. In this work, we tailor our general model to learning dictionaries for compressive sensing recovery and single image super-resolution to demonstrate its effectiveness. In both cases, the new sparse coding model remarkably outperforms previous approaches in terms of recovery accuracy.

1. Introduction

In the areas of signal processing and pattern recognition, it has always been paramount to look for meaningful data representations, *e.g*., in compression, we want the representation to account for the essential content of the signal with as few coefficients as possible. In the past decades, analytic representations from orthogonal bases have been prevalent in signal processing techniques due to their mathematical simplicity and computational efficiency, *e.g*., wavelets for compression (JPEG2000) and denoising [7]. Despite their simplicity, these dictionaries are limited in representing the natural signals efficiently, which are well known to be mixtures of diverse phenomena. Over-complete bases are thus explored [17], which offer the flexibility to represent much wider range of signals with more elementary basis atoms than the signal dimension [18].

Sparse and redundant data modeling seeks the representation of signals as a linear combination of a few number of atoms from a pre-defined dictionary. Recently, there has been fast growing interest in dictionary training, *i.e*., using machine learning techniques to learn over-complete dictionaries directly from data, so that the most relevant properties of the signals can be efficiently captured. Most learning algorithms employ the ℓ^0- or ℓ^1-norm as the sparsity penalty measure for representations, which lead to simple optimization formulations and allow the use of recent developed efficient sparse coding techniques. Example works include the Method of Optimal Directions (MOD) with ℓ^0 sparsity measure proposed by Engan *et al*. [10], the greedy K-SVD algorithm by Aharon *et al*. [1], an efficient formulation with ℓ^1 sparsity measure by Lee *et al*. [13], and an online dictionary learning algorithm by Mairal *et al*. [15]. Compared with the conventional mathematically defined dictionaries, the learned dictionaries are more adaptive to the signal distribution, which have attained state-of-the-art performances on many signal processing tasks, *e.g*., denoising [1], inpainting [16] and super-resolution [21].

Most existing dictionary learning methods are reconstruction based and only consider sparse modeling in a single signal space[13]. In many problems, we have two coupled signal spaces, *e.g*., high- and low-resolution patch spaces in patch-based image super-resolution, source and target image spaces in texture transfer, and original and compressed signal spaces in compressive sensing. The two coupled spaces are usually related by some mapping function, which could be complicated and even unknown. In such cases, it is often desirable to learn representations that can not only well represent each signal space individually, but also capture their relationships through the underlying sparse representations. The resultant *coupled dictionaries* are potentially useful for many tasks such as signal recovery, information fusion, and transfer learning. However, dictionary learning across different signal spaces has received little attention in the literature. Yang *et al*. [21] proposed a joint dictionary training method to learn dictionaries for coupled signal spaces, which essentially concatenates the two signal spaces and converts the problem to a conventional sparse coding formulation. The coupled

dictionaries obtained in this way are not indeed customized to each individual space; and they cannot capture possible complex relationships between different signal spaces arising in various scenarios. Some efforts are also devoted to supervised dictionary learning [3, 22, 2] for classification, which only model the mapping from a high dimensional feature space to the discrete categorical label space.

In this paper, we first propose a general bilevel sparse coding model for learning dictionaries across coupled signal spaces, which potentially could model various relationships between the two signal spaces. The algorithm is formulated as a bilevel optimization [6], which can be solved efficiently using our projected first-order stochastic gradient descent. It is worth noting that a similar optimization scheme has been used by Yang *et al*. [22] on supervised dictionary learning for image classification, which is later extended to general regression tasks by Mairal *et al*. [14]. In this work, we focus on sparse coding across different feature spaces, and give an explicit form of the gradient for stochastic learning with theoretical proof. Tailored to specific applications, we apply our general model to learn dictionaries for compressive sensing recovery and patch-wise single image super-resolution. In both cases, our bilevel sparse coding model remarkably outperforms the corresponding baselines in terms of recovery accuracy.

The remainder of the paper is organized as follows. Section 2 briefly reviews the conventional sparse coding in a single feature space. Section 3 presents our general coupled sparse coding model and the learning algorithm. In Section 4, we tailor our general model to two specific applications, *i.e*., compressive sensing and image super-resolution, and demonstrate significant improvements over the corresponding baselines. Finally, Section 5 concludes our paper with discussions.

2. Sparse Coding in a Single Feature Space

The goal of sparse modeling is to represent an input signal $\boldsymbol{x} \in \mathbb{R}^d$ approximately as a linear combination of a few elementary signals called basis atoms, often chosen from an over-complete dictionary $D \in \mathbb{R}^{d\times K}$ $(d < K)$. Sparse coding is the method to automatically discover such a good set of basis atoms. Given training data $\{\boldsymbol{x}_i\}_{i=1}^N$, the problem of learning dictionary for sparse coding, in its most popular form, is solved by minimizing the energy function that combines square reconstruction errors and ℓ^1 sparsity penalties on the representations:

$$\begin{aligned} &\min_{D,\{\boldsymbol{\alpha}_i\}_{i=1}^N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - D\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1 \\ &s.t.\ \ \|D(:,k)\|_2 \le 1,\ \ \forall k \in \{1,2,...,K\}, \end{aligned} \tag{1}$$

where $D(:,k)$ is the k-th column of D and λ is a parameter controlling the sparsity penalty. The above optimization problem is convex with either D or $\{\boldsymbol{\alpha}_i\}_{i=1}^N$ fixed, but not with both. When D is fixed, inference for $\{\boldsymbol{\alpha}_i\}_{i=1}^N$ is known as Lasso problem in statistics literatures; when $\{\boldsymbol{\alpha}_i\}_{i=1}^N$ are fixed, solving D becomes a standard quadratically constrained quadratic programming problem. A practical solution to Eqn. (1) is to alternatively optimize over D and $\{\boldsymbol{\alpha}_i\}_{i=1}^N$, and at each step the cost function can be guaranteed to decrease [13].

3. Bilevel Sparse Coding for Coupled Feature Spaces

In this section, we propose our generic bilevel sparse coding model in two related signal spaces and an efficient algorithm for learning the dictionaries. To keep the model general, we do not specify the form of cost function until we discuss specific applications in Section 4.

3.1. The Learning Model

Suppose we have two coupled signal spaces $\mathcal{X} \in \mathbb{R}^{d_1}$ and $\mathcal{Y} \in \mathbb{R}^{d_2}$, where the signals are sparse in their high-dimensional spaces, *i.e*., the signals have sparse representations in terms of certain dictionaries. There exists a mapping function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ (not necessarily linear and probably unknown) that relates a signal in $\mathcal{X}$ to its corresponding signal in $\mathcal{Y}$.[1] We assume that the mapping function is at least nearly injective. Given the training samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$, where $\boldsymbol{y}_i = \mathcal{F}(\boldsymbol{x}_i)$, our coupled sparse coding model aims to train dictionaries for one or both of these signal spaces so that the relationships between them are captured for particular signal modeling problems. Concretely, we formulate the coupled sparse coding model as a generic bilevel optimization problem:

$$\begin{aligned} \min_{D_x,D_y}\ & \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{z}_i^x, D_x; \boldsymbol{z}_i^y, D_y) \\ \text{s.t.}\ \ & \boldsymbol{z}_i^x = \arg\min_{\alpha} \|\alpha\|_1,\ \text{s.t.}\ \ \|\boldsymbol{x}_i - D_x\alpha\|_2^2 \le \epsilon_x,\ \forall i, \\ & \boldsymbol{z}_i^y = \arg\min_{\alpha} \|\alpha\|_1,\ \text{s.t.}\ \ \|\boldsymbol{y}_i - D_y\alpha\|_2^2 \le \epsilon_y,\ \forall i, \\ & \|D_x(:,k)\|_2 \le 1,\ \forall k, \\ & \|D_y(:,k)\|_2 \le 1,\ \forall k, \end{aligned} \tag{2}$$

where D_x and D_y are the sparse dictionaries for spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, and $\mathcal{L}$ is some smooth cost function designed to capture the desired relationships between the two signal spaces. For example, we can define $\mathcal{L}_i = \mathcal{L}(\boldsymbol{z}_i^x, D_x; \boldsymbol{z}_i^y, D_y) = \|\boldsymbol{z}_i^x - \boldsymbol{z}_i^y\|_2^2$ in order to learn two dictionaries, such that the coupled signals $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ have the

[1] The definitions of the signal spaces $\mathcal{X}, \mathcal{Y}$ and the mapping function $\mathcal{F}$ depend on specific applications.

same sparse representation with respect to their own dictionaries, which could be useful in many sparse recovery problems. As another example, we can define $\mathcal{L}_i = \|Z\|_{\ell^1/\ell^2}$, where $Z = [\boldsymbol{z}_i^x, \boldsymbol{z}_i^y]$, and the ℓ^1/ℓ^2 norm is used to enforce group sparsity such that the two sparse codes share the same representation supports, but their representation coefficients could disagree. Such a model is more flexible, and could be useful in image space transformation applications, *e.g.*, intrinsic image estimation [12]. In Section 4, we will talk about two specific applications of our generic model in detail. Before that, we first describe the learning algorithm in the following.

3.2. The Learning Algorithm

The problem in Eqn. (2) is a bilevel optimization problem [6], where optimization problems (ℓ^1-norm minimizations in this case) appear in the constraints. In our problem, the upper-level problem $\mathcal{L}$ selects the dictionaries D_x and D_y, and the lower-level ℓ^1-norm minimizations return the sparse codes $\boldsymbol{z}_i^x$ and $\boldsymbol{z}_i^y$ to the upper-level $\mathcal{L}$ in order to evaluate the objective function value. Being generically non-convex and non-differentiable, bilevel optimization programs are intrinsically difficult [6]. In this subsection, we develop an efficient optimization procedure based on the first-order projected stochastic gradient descent, which turns out to be very effective in practice.

3.2.1 The Formulation

A large class of approaches for solving the bilevel optimization problem is based on the descent method [6]. In problem (2), $\boldsymbol{z}_i^x$ and $\boldsymbol{z}_i^y$ are the outputs of the lower-level ℓ^1-norm minimization based on D_x and D_y. Assuming that we can define $\boldsymbol{z}_i^x$ and $\boldsymbol{z}_i^y$ as implicit functions $\boldsymbol{z}_i^x(D_x)$ and $\boldsymbol{z}_i^y(D_y)$ of D_x and D_y depending on the inputs $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$, problem (2) may be viewed solely in terms of the upper-level variables D_x and D_y. Given a feasible point for D_x and D_y, the descent method makes an attempt to find a feasible (descent) direction along which the upper-level objective decreases. The major issue about descent method is the availability of the gradient of the upper-level objective, $\nabla\mathcal{L}_{D_x}$ and $\nabla\mathcal{L}_{D_y}$, at a feasible point. Applying the chain rule, we have, whenever $\partial\boldsymbol{z}_i^x/\partial D_x$ and $\partial\boldsymbol{z}_i^y/\partial D_y$ are well defined:

$$\begin{aligned}(\nabla\mathcal{L}_i)_{D_x} &= \frac{\partial\mathcal{L}_i}{\partial D_x} + \frac{\partial\mathcal{L}_i}{\partial\boldsymbol{z}_i^x}\frac{\partial\boldsymbol{z}_i^x}{\partial D_x},\\ (\nabla\mathcal{L}_i)_{D_y} &= \frac{\partial\mathcal{L}_i}{\partial D_y} + \frac{\partial\mathcal{L}_i}{\partial\boldsymbol{z}_i^y}\frac{\partial\boldsymbol{z}_i^y}{\partial D_y},\end{aligned} \tag{3}$$

where the functions are evaluated at the current iteration. However, there is no analytical link between $\boldsymbol{z}_i^x$ and D_x or $\boldsymbol{z}_i^y$ and D_y for direct evaluation of $\partial\boldsymbol{z}_i^x/\partial D_x$ and $\partial\boldsymbol{z}_i^y/\partial D_y$. In the following section, we will see that the sparse codes $\boldsymbol{z}_i^x$ and $\boldsymbol{z}_i^y$ are almost differentiable with respect to their depending dictionaries D_x and D_y, and thus the gradients can be evaluated by implicit differentiation [3, 22].

3.2.2 Derivatives in the ℓ^1-norm minimization

Note that the ℓ_1-norm minimization problems in Eqn. (2) can be equivalently reformulated as an unconstrained optimization problem for a properly chosen λ

$$\boldsymbol{z} = \arg\min_{\alpha} \|\boldsymbol{x} - D\alpha\|_2^2 + \lambda\|\alpha\|_1, \tag{4}$$

known as the Lasso in statistics literature [9]. We denote Λ as the active set of the Lasso solution $\boldsymbol{z}$, *i.e.*, $\Lambda = \{k : \boldsymbol{z}(k) \neq 0\}$, for the following presentation. In order to compute the gradient of $\boldsymbol{z}$ with respect to D, we first introduce the following lemmas.

Lemma 1. *For a given response vector $\boldsymbol{x}$, there is a finite sequence of λ's, $\lambda_0 > \lambda_1 > \cdots > \lambda_K = 0$, such that if λ is in the interval of $(\lambda_m, \lambda_{m+1})$, the active set Λ and sign vector $\mathrm{Sgn}(\boldsymbol{z}_\Lambda)$ are constant with respect to λ.*

These characteristics of the Lasso solution has been shown by Efron *et al*. [9]. The active set changes at $\{\lambda_m\}$, hence they are called *transition points* [23]. Any $\lambda \in [0, \inf)\backslash\{\lambda_m\}$ is called a nontransition point.

Lemma 2. *$\forall\lambda$, $\boldsymbol{z}$ is a continuous function of D and $\boldsymbol{x}$.*

Instead of a formal proof, we simply state that function $f(\boldsymbol{x}, \alpha, D) = \|\boldsymbol{x} - D\alpha\|_2^2 + \lambda\|\alpha\|_1$ is continuous in $\boldsymbol{x}$, α and D, and thus $\boldsymbol{z}$ is a continuous function of $\boldsymbol{x}$ and D [23, 15].

Lemma 3. *Fix any $\lambda > 0$, and λ is not a transition point for $\boldsymbol{x}$, the active set Λ and the sign vector $\mathrm{Sgn}(\boldsymbol{z}_\Lambda)$ are locally constant with respect to both $\boldsymbol{x}$ and D (please refer to the appendix for a proof).*

For λ being a nontransition point, we have the equiangular conditions [9]:

$$\frac{\partial\|\boldsymbol{x} - D\boldsymbol{z}\|_2^2}{\partial\boldsymbol{z}(k)} + \lambda\,\mathrm{sign}(\boldsymbol{z}(k)) = 0, \text{ for } k \in \Lambda, \tag{5}$$

$$\left|\frac{\partial\|\boldsymbol{x} - D\boldsymbol{z}\|_2^2}{\partial\boldsymbol{z}(k)}\right| < \lambda, \text{ for } k \notin \Lambda. \tag{6}$$

Eqn. (5) is the stationary condition for $\boldsymbol{z}$ to be optimal, which links $\boldsymbol{z}$ and D analytically on the active set Λ. We rewrite this condition as

$$D_\Lambda^T D_\Lambda \boldsymbol{z}_\Lambda - D_\Lambda^T\boldsymbol{x} + \lambda\mathrm{Sgn}(\boldsymbol{z}_\Lambda) = 0, \tag{7}$$

where D_Λ consists of the columns of D in the active set Λ. Based on Lemma 3, the active set Λ and sign vector

$\mathrm{Sgn}(\boldsymbol{z}_\Lambda)$ are constant in a local neighborhood of D, and therefore, Eqn. (7) and Eqn. (6) hold for a sufficient small perturbation of D. Denoting Ω as the nonactive set, we can now evaluate the full gradient of $\boldsymbol{z}$ with respect to D in three parts:

1. As $\boldsymbol{z}$ is a continuous function of D, Λ and $\mathrm{Sgn}(\boldsymbol{z}_\Lambda)$ are locally constant with respect to D, we can apply implicit differentiation to Eqn. (7) to get the partial derivative $\boldsymbol{z}_\Lambda$ with respect to D_Λ, [2]

$$\frac{\partial \boldsymbol{z}_\Lambda}{\partial D_\Lambda} = \left(D_\Lambda^T D_\Lambda\right)^{-1}\left(\frac{\partial D_\Lambda^T \boldsymbol{x}}{\partial D_\Lambda} - \frac{\partial D_\Lambda^T D_\Lambda}{\partial D_\Lambda}\boldsymbol{z}_\Lambda\right). \tag{8}$$

2. As $\boldsymbol{z}_\Lambda$ is only related with D_Λ, a perturbation on D_Ω would not change its value, and therefore, we have $\partial \boldsymbol{z}_\Lambda / \partial D_\Omega = 0$.

3. As Λ and $\mathrm{Sgn}(\boldsymbol{z}_\Lambda)$ are constant for a small perturbation of D, $\boldsymbol{z}_\Omega$ stays as zero, so we have $\partial \boldsymbol{z}_\Omega / \partial D = 0$.

In summary, based on the assumption that λ is not a transition point, $\partial \boldsymbol{z}/\partial D$ is very sparse and the nonzero part is given only by $\partial \boldsymbol{z}_\Lambda / \partial D_\Lambda$, making it very efficient to evaluate in practice.

For λ being a transition point, the above derivatives are not exact any more. However, we have the following Lemma proved in [23],

Lemma 4. *$\forall \lambda > 0$, $\exists$ a null set $\mathcal{N}_\lambda$ which is a finite collection of hyperplanes in $\mathbb{R}^d$. Then $\forall \boldsymbol{x} \in \mathbb{R}^d \backslash \mathcal{N}_\lambda$, λ is not any of the transition points of $\boldsymbol{x}$.*

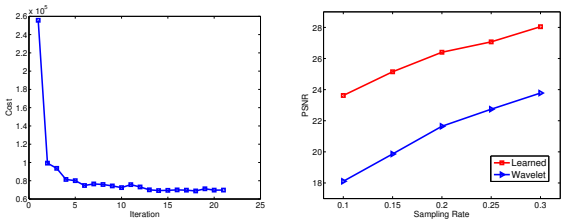
Based on this Lemma, for a reasonable assumption on the distribution of the input vectors $\boldsymbol{x}$, the chance that λ is a transition point for $\boldsymbol{x}$ is low and thus is neglectable. On the other hand, from a practical point of view, even if we could not evaluate the exact full gradient, as long as we find a feasible direction from the partial derivative, we can still decrease the objective function value using the descent method.

3.2.3 Stochastic Gradient Descent

Now we could evaluate $(\nabla \mathcal{L}_i)_{D_x}$ and $(\nabla \mathcal{L}_i)_{D_y}$ in Eqn. (3) based on $\partial \boldsymbol{z}_\Lambda / \partial D_\Lambda$ for stochastic gradient descent. The dictionary updating rule is simply

$$\begin{aligned} D_x^{n+1} &= D_x^n - r\frac{(\nabla \mathcal{L}_i)_{D_x}}{\|(\nabla \mathcal{L}_i)_{D_x}\|_2}, \\ D_y^{n+1} &= D_y^n - r\frac{(\nabla \mathcal{L}_i)_{D_y}}{\|(\nabla \mathcal{L}_i)_{D_y}\|_2}. \end{aligned} \tag{9}$$

[2] $D_\Lambda^T D_\Lambda$ is well conditioned for inverse if $\boldsymbol{z}$ is unique. In practice, we find that this is not a problem.



(a) (b)

Figure 1. (a) Objective function value vs. iteration number for 10% sampling rate. The optimization converges very fast; (b) Recovery accuracy comparisons on the test image patches in terms of PSNR for Harr Wavelet basis and our learned dictionary on different sampling rates. In all cases, our learning scheme improves the baseline by around 5 dB.

with

$$r = \frac{r_0}{\sqrt{n/N + 1}}, \tag{10}$$

where n is the cumulative counts of the data samples fed into the learning algorithm, N is the total number of iterations, and r_0 is the initial learning rate.

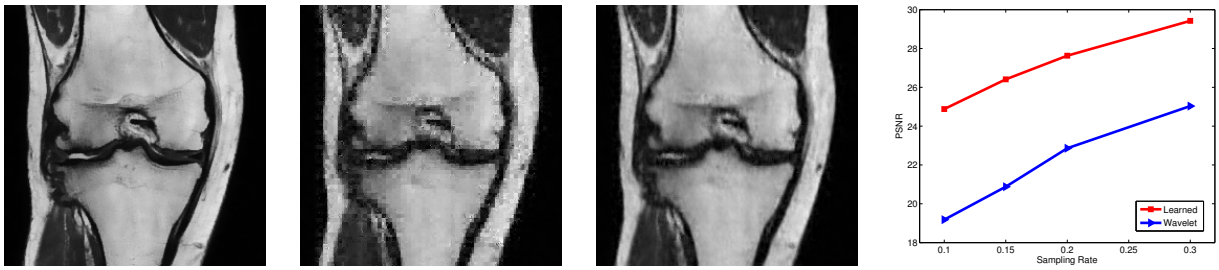
Since we have the norm constraints on each dictionary atom, we project the updated dictionary back onto the unitary ball after each update. Furthermore, to ensure that the learned dictionaries can *sparsely* represent the data samples well, we add the reconstruction constraint $\|\boldsymbol{x}_i - D_x \boldsymbol{z}_i^x\|_2^2$ and $\|\boldsymbol{y}_i - D_y \boldsymbol{z}_i^y\|_2^2$ to the cost function $\mathcal{L}$ as an additional regularization. Since the optimization problem in Eqn. (2) is highly nonlinear and highly nonconvex, we can only expect this projected first-order stochastic gradient procedure to find a local minimum. In practice, we find that our algorithm is quite efficient and effective with proper initialization.

4. Applications

In many of signal processing and computer vision applications, we will deal with sparse high-dimensional coupled signal spaces, where our bilevel sparse coding model could potentially be helpful. In this Section, we tailor our generic model discussed above to two specific applications– compressive sensing and single image super-resolution.

4.1. Compressive Sensing

Compressive sensing is about acquiring a sparse signal in the most efficient way possible with the help of an incoherent projecting basis [5]. Unlike traditional sampling methods, compressive sensing provides a new framework for sampling signals in a multiplexed manner [4], and states that sparse signals, can be exactly recovered from a number of linear projections of dimension considerably lower than the number of samples required by the Shannon-Nyquist theorem. Compressive sensing relies on two fundamental principles: sparsity of the signal and incoherent sampling.



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

Figure 2. Recovery comparison on the "bone" image with 20% measurements. (a): ground truth image; (b): recovery with wavelet basis (22.8 dB); (c): recovery with learned dictionary (27.6 dB); and (d) recovery PSNR comparison under different sampling rates.

Given $\boldsymbol{x} \in \mathbb{R}^{d_1}$ a d_1-dimensional signal, compressive sensing requires that the signal has a sparse representation in terms of some dictionary D_x. Let Φ be an $m \times d_1$ sampling matrix ($m \ll d_1$), such as $\boldsymbol{y} = \Phi \boldsymbol{x}$ is an m-dimensional vector of linear measurements of the underlying signal $\boldsymbol{x}$. Compressive sensing requires that the sensing matrix Φ and the sparse representation matrix D_x to be as incoherent as possible. The recovery of $\boldsymbol{x}$ from its linear measurement $\boldsymbol{y}$ can be done by ℓ^1-norm minimization under conditions related to the sparsity of the signal and the incoherence of the sensing matrix,

$$\boldsymbol{z} = \arg\min_{\alpha} \|\alpha\|_1, \;\; \text{s.t.} \;\; \boldsymbol{y} = \Phi D_x \alpha, \;\; \boldsymbol{x} = D_x \alpha. \quad (11)$$

Therefore, the choices of the sensing matrix Φ and the sparse representation matrix D_x are both critical for the success of compressive sensing recovery, especially when only few measurements are available. A method for simultaneously learning the sensing matrix and the sparse representation matrix is proposed in [8] by reducing the mutual coherence of the dictionary. In practice, the sensing matrix is usually constrained by the hardware implementation, and therefore, we fix our sensing matrix in this work, and try to optimize the sparse representation matrix D_x. Fitting into the generic bilevel sparse coding model, our optimization over the dictionary D_x can be formulated in the following:

$$\begin{aligned} \min_{D_x} \;\; & \sum_{i=1}^{N} \|\boldsymbol{x}_i - D_x \boldsymbol{z}_i\|_2^2 \\ \text{s.t.} \;\; & \boldsymbol{y}_i = \Phi \boldsymbol{x}_i, \; \forall i, \\ & \boldsymbol{z}_i = \arg\min_{\alpha} \|\alpha\|_1, \; \text{s.t.} \;\; \|\boldsymbol{y}_i - D_y \alpha\|_2^2 \le \epsilon, \; \forall i \\ & D_y = \Phi D_x, \\ & \|D_x(:,k)\|_2 \le 1, \; \forall k. \end{aligned} \quad (12)$$

It is easy to see that the above model is a special case of the generic model in Eqn. (2) by defining $\mathcal{L} = \|\boldsymbol{x}_i - D_x \boldsymbol{z}_i\|_2^2$ and optimizing only over D_x (D_y is defined by ΦD_x). In this model, instead of improving the based on the compressive sensing theory as in [8], we directly optimize the dictionary D_x to achieve low sparse recovery errors.

For experimental evaluation, we randomly sample $10,000$ image patches of size 16×16 for training and 5000 image patches of the same size for testing from two sets of MRI bone medical images. We use the Haar wavelet basis as our baseline, which also serves as the initial dictionary for our optimization. For the sampling matrix, we use Bernoulli random matrix with sampling rate at 10%, 15%, 20%, 25%, 30% for the measurements. Figure 1 (a) draws how the objective function value drops with the optimization iterations for sampling rate 10%, which shows that the algorithm converges very fast, typically in 10 iterations (faster than standard sparse coding). Even with only one iteration, the algorithm already provides a reasonable solution. Figure 1 (b) demonstrates the average recovery accuracy comparisons on the 5000 testing image patches between wavelet and our learned dictionary across different sampling rates. In all cases, our learned dictionary outperforms the wavelet basis by a remarkable margin of around 5 dB. This is especially striking for small sampling rates, *e.g.*, our algorithm dramatically reduces the RMSE from 31.8 to 16.8 for sampling rate 10%. In Figure 2, we perform patch-wise sparse recovery on the whole "bone" test image for sampling rate 20%. The result from wavelet basis (b) shows obvious blocky artifacts, while our result (c) is much more accurate and informative. Figure 2 (d) shows the recovery accuracy comparisons under different sampling rates. Again, our learned dictionary outperforms the wavelet baseline by around 5 dB in all cases.

4.2. Single Image Super-resolution

Image super-resolution is the class of techniques that construct a high-resolution image from one or several low-resolution observations [19]. Among all those techniques, patch-based single image super-resolution is one of the promising approaches for many practical applications. Many previous example-based super-resolution works [11] apply a non-parametric approach to the super-resolution problem with a large training patch set. Motivated by the recent compressive sensing theories [4], Yang *et al*. [21] formulate the problem as patch-wise sparse recovery. In order to train a compact model, they propose joint sparse coding

to learn two dictionaries D_x and D_y, for high- and low-resolution image patches respectively, such that the sparse representation $\boldsymbol{z}_y$ of a low-resolution image patch $\boldsymbol{y}$ is the same as the sparse representation $\boldsymbol{z}_x$ of the corresponding high-resolution image patch $\boldsymbol{x}$. For any given testing low-resolution patch $\boldsymbol{y}_i$, the algorithm first finds its sparse representation $\boldsymbol{z}_i$ in terms of D_y using ℓ^1-norm minimization, and then recover the underlying high-resolution image patch $\boldsymbol{x}_i$ as $\hat{\boldsymbol{x}}_i = D_x \boldsymbol{z}_i$. In the following, we introduce the joint sparse coding algorithm and state its problem, which motivates our bilevel dictionary training algorithm followed.

4.2.1 Joint Sparse Coding for Super-resolution

Unlike the standard sparse coding, joint sparse coding considers the problem of learning the coupled dictionaries D_x and D_y for given coupled feature spaces, $\mathcal{X}$ and $\mathcal{Y}$, tied by a certain mapping function, such that the sparse representation of $\boldsymbol{y}_i \in \mathcal{Y}$ in terms of D_y should be as close as possible to that of $\boldsymbol{x}_i \in \mathcal{X}$ in terms of D_x, where $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ is a coupled signal pair. Accordingly, if $\boldsymbol{y}_i$ is our observation signal, we can recover its underlying latent signal $\boldsymbol{x}_i$ via their common sparse representation. Yang *et al*. [21] addressed this problem by generalizing the basic sparse coding scheme as follows:

$$\min_{D_x, D_y, \{\boldsymbol{\alpha}_i\}_{i=1}^N} \sum_{i=1}^{N} \frac{1}{2} \left(\|\boldsymbol{x}_i - D_x \boldsymbol{\alpha}_i\|_2^2 + \|\boldsymbol{y}_i - D_y \boldsymbol{\alpha}_i\|_2^2 \right) + \lambda \|\boldsymbol{\alpha}_i\|_1, \tag{13}$$

which basically requires that the resulting common sparse representation should reconstruct both $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ well. However, such joint sparse coding can only be claimed to be optimal in the concatenated feature space of $\mathcal{X}$ and $\mathcal{Y}$, but not in each feature space separately. To see this, we can group the first two reconstruction errors together by denoting

$$\bar{\boldsymbol{x}}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{y}_i \end{bmatrix}, \quad \bar{D} = \begin{bmatrix} D_x \\ D_y \end{bmatrix}. \tag{14}$$

Then Eqn. (13) reduces to a standard sparse coding problem in the concatenated feature space of $\mathcal{X}$ and $\mathcal{Y}$. In testing, suppose we only observe $\boldsymbol{y}_i$, the desired sparse representation $\boldsymbol{z}_i$ is obtained from $\hat{\boldsymbol{z}}_i = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{y}_i - D_y \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1$ in [21]. Compared with the training stage, the term $\|\boldsymbol{x}_i - D_x \boldsymbol{\alpha}\|_2^2$ is missing, which is unknown ($\boldsymbol{x}_i$ is the signal to recover). Therefore, the recovery accuracy for $\boldsymbol{x}$ from $\hat{\boldsymbol{z}}_i$ is not guaranteed.

4.2.2 Bilevel Sparse Coding for Super-resolution

Let the signals of low-resolution image patches constitute the observation space $\mathcal{Y}$ and the high-resolution image patches constitute the latent space $\mathcal{X}$. We want to model the mapping between the two spaces by our coupled sparse coding, and then use the learned dictionaries to recover high-resolution patch $\boldsymbol{x}$ for any given low-resolution patch $\boldsymbol{y}$. Following the routine of [20], the low-resolution $\boldsymbol{y}$ is represented by the gradient features of its interpolated version by Bicubic. Therefore, different from compressive sensing, the mapping between high- and low-resolution image patches are no longer linear, but complicated and obscure. As a result, the high- and low-resolution dictionaries D_x and D_y are no longer linearly related, and have to be defined explicitly.

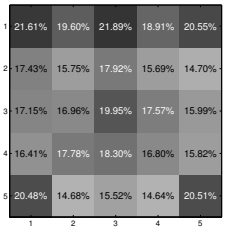


Figure 3. Average percentages of pixel-wise MSE reduced by our coupled training method compared with joint dictionary training method on the 5×5 patch.

Suppose the dictionary D_x is given[3] to sparsely represent high-resolution signals in $\mathcal{X}$. Our goal is to learn a "coupled" dictionary D_y over $\mathcal{Y}$, such that the sparse representation $\boldsymbol{z}$ of any $\boldsymbol{y} \in \mathcal{Y}$ in terms of D_y can be used to recover its corresponding $\boldsymbol{x} \in \mathcal{X}$ with dictionary D_x as $\hat{\boldsymbol{x}} = D_x \boldsymbol{z}$. Formally, the optimization for D_y can be formulated in the following:

$$\begin{aligned} \min_{D_y} \ & \sum_{i=1}^{N} \|D_x \boldsymbol{z}_i^y - \boldsymbol{x}_i\|_2^2 \\ \text{s.t.} \ & \boldsymbol{z}_i^y = \arg\min_{\alpha} \|\alpha\|_1, \ \text{s.t.} \ \|\boldsymbol{y}_i - D_y \alpha\|_2^2 \le \epsilon, \ \forall i \\ & \|D_y(:,k)\|_2 \le 1, \quad \forall k, \end{aligned} \tag{15}$$

where $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ are training examples randomly sampled from the coupled signal spaces $\{\mathcal{X}, \mathcal{Y}\}$. Again, the above model is a special case of the bilevel sparse coding model in Eqn. (2) with $f = \|D_x \boldsymbol{z}_i^y - \boldsymbol{x}_i\|_2^2$ and optimization only over D_y.

To train the dictionary D_y, we sample $100,000$ patches from high- and low-resolution image pairs to obtain the training data. The patch size is chosen as 5×5 to achieve sufficient sparsity while maintaining an affordable dictionary dimension. We use the dictionaries trained from joint sparse coding in [21] as the initialization for D_x and D_y. The learning algorithm quickly converges in less than 5 iterations. We compare the results of our bilevel sparse coding

[3] Either by standard sparse coding or mathematical derivation.

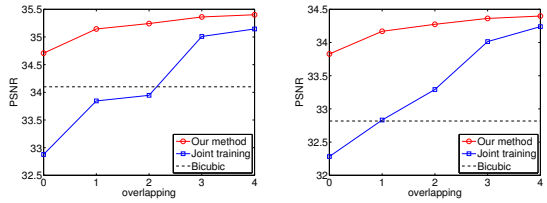


Figure 5. Recovery PSNRs using different dictionary training methods with different patch overlappings on the "Flower" (left) and "Lena" (right) test images from [21].

model with those of the joint sparse coding model, as [21] provides the state-of-the-art single super-resolution results.

The recovery accuracy of different sparse coding models is first evaluated on an independent validation set containing $100,000$ image patch pairs. Figure 3 shows the pixel-wise mean square error (MSE) reduction by using our bilevel dictionary compared with the joint dictionary training method. It can be seen that our approach significantly reduces the recovery errors in all pixel locations.

For super-resolution on the entire image, the low-resolution patches are sampled from the input image on a regular grid with overlapping for sparse recovery. The recovered high-resolution image patches are then aggregated by averaging the the overlapping pixels to obtain a final high-resolution image. Typically, higher accuracy could be achieved by increasing the amount of overlapping between adjacent patches, but at the expense of more computation cost. In Figure 4, we show the super-resolution results by a magnification factor of two on five natural images. The input patches are sampled with 0/1/2/3/4-pixel overlapping for test images from left to right. The top row shows the results of joint dictionary training, and the bottom row shows those of our bilevel sparse coding. In all cases, the joint dictionary training method produces visible visual artifacts, especially with smaller overlapping regions; on the contrary, none of those artifacts are observed in our bilevel sparse coding method. This indicates that the bilevel sparse coding method produces more accurate predictions, which is also evidenced by the reported PSNRs. Figure 5 shows the recovery PSNRs on two more test images "Flower" (left) and "Lena" (right) from [21] with different amount of pixel overlapping. For reference, the PSNRs of the respective "bicubic" interpolation are also plotted. In all cases, our method outperforms the other two substantially. More importantly, recovery using our coupled dictionary with 0-pixel patch overlapping can achieve approximately the same accuracy as the one given by joint dictionary with 3-pixel overlapping, implying computation time reduction by more than 6 times.

5. Conclusion

In this paper, we propose a general bilevel sparse coding model for learning dictionaries in coupled signal spaces. Our learning algorithm employs a theoretically proven stochastic descent procedure, which turns out to be both efficient and effective in practice. Tailored to specific applications, we apply our model to compressive sensing and single image super-resolution. In both cases, our bilevel sparse coding strategy achieves remarkable improvements over corresponding baselines, which demonstrates the effectiveness of our learning algorithm. As many problems in computer vision and signal processing can be addressed based on sparse representations, our bilevel sparse coding model can be generalized to many other applications for learning more meaningful dictionaries, such as image classification, image deblurring, and texture transfer.

Acknowledgments. This work is supported by U.S. ARL and ARO under grant number W911NF-09-1-0383. It is also supported in part by Adobe Systems, Inc. We would like to thank Xinqi Chu for useful discussions on compressive sensing.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Image Processing*, 54(11):4311–4322, 2006.

[2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[3] D. Bradley and J. A. D. Bagnell. Differentiable sparse coding. In *Proceedings of Neural Information Processing Systems 22*, December 2008.

[4] E. J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, 2006.

[5] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 5(2):489–509, 2006.

[6] B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of Operations Research*, pages 235–256, 2007.

[7] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41:613–627, 1995.

[8] J. M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Transactions on Image Processing*, 18:1395–1408, 2009.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annual of Statistics*, 32:407–499, 2004.

[10] K. Engan, S. O. Aase, and J. H. Husoy. Method of optimal directions for frame design. In *IEEE International Conference on Accoustics and Speech Signal Processing*, volume 5, pages 2443–2446, 1999.

[11] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22:56–65, 2002.

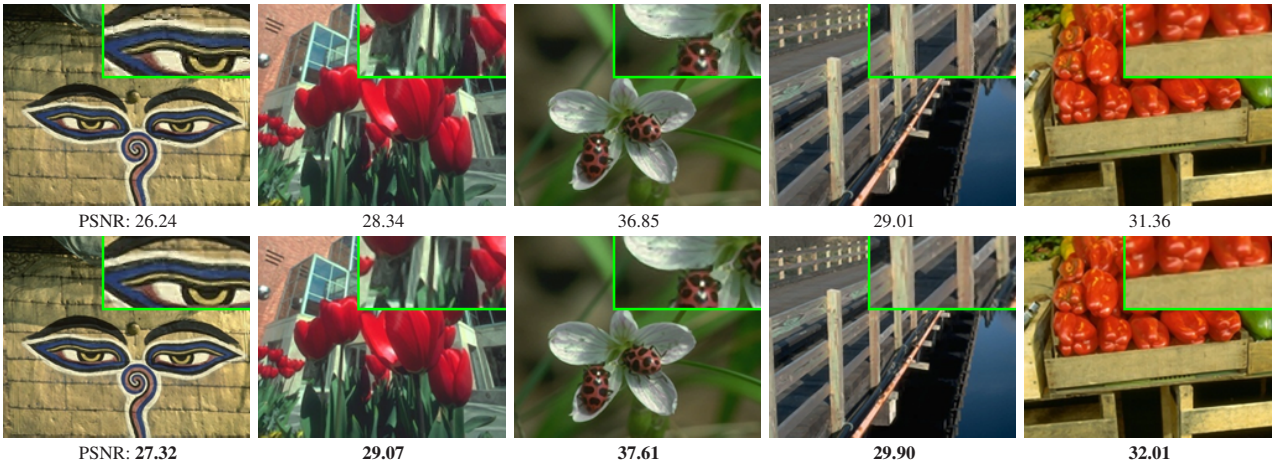


Figure 4. Super-resolution results up-scaled by magnification factor of 2, using joint dictionary training (top row) and bilevel sparse coding (bottom row), with 0/1/2/3/4-pixel overlapping between adjacent patches from left to right test images, respectively. In all cases, our bilevel sparse coding method generates artifact-free results while the joint dictionary training produces many visible artifacts. For better visual comparison, see the supplementary material for the original results.

[12] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.

[13] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *In NIPS*, pages 801–808. NIPS, 2007.

[14] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. to appear.

[15] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal fo Machine Learning Research*, 11:19–60, 2010.

[16] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17, 2008.

[17] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[18] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. In *Proceedings of IEEE*, volume 98, 2010.

[19] J. Yang and T. Huang. Image super-resolution: historical overview and future challenges. In P. Milanfar, editor, *Super-resolution imaging*, chapter 1. CRC Press, 2010.

[20] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[21] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):1–8, 2010.

[22] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[23] H. Zou, T. Hastie, and R. Tibshirani. On the "degree of freedom" of the lasso. *Annual of Statistics*, 35:2173–2192, 2007.

A. Proof of Lemma 3

Proof. Fixing $D \in \mathbb{R}^{d\times K}$, it is easy to show that Λ and $\mathrm{Sgn}(\boldsymbol{z}_\Lambda)$ are locally constant with respect to $\boldsymbol{x}$, given that λ is not a transition point for $\boldsymbol{x}$. The proof based on Lemma 2 and the equiangular conditions [9] is given in [23]. Therefore, for the given signal $\boldsymbol{x} \in \mathbb{R}^d$, there exists a d-dimensional Ball($\boldsymbol{x}$, ϵ) with center $\boldsymbol{x}$ and radius ϵ, such that Λ and $\mathrm{Sgn}(\boldsymbol{z}_\Lambda)$ are constant.

Now, fix the signal vector $\boldsymbol{x}$. Denote by Ball(D, ε) the dK-dimensional ball with center D and radius ε. Consider a perturbation E on the dictionary D such that $D_e = D+E \in \mathrm{Ball}(D, \varepsilon)$. Its Lasso formulation for $\boldsymbol{x}$ is

$$\min_\alpha \|\boldsymbol{x} - D_e\alpha\|_2^2 + \lambda\|\alpha\|_1, \tag{16}$$

which we reformulate as

$$\min_\alpha \|(\boldsymbol{x} - E\alpha) - D\alpha\|_2^2 + \lambda\|\alpha\|_1. \tag{17}$$

Denote $\boldsymbol{x}_e = \boldsymbol{x} - E\alpha$, which is $\boldsymbol{x}$ plus a perturbation $-E\alpha$. Since $\|\alpha\|_2^2 \leq B$ for some upper bound B and $\|E\|_2^2 \leq \varepsilon^2$, based on the Cauchy inequality, the perturbation vector $\|E\alpha\|_2^2 \leq d\varepsilon^2 B$. Therefore, there exists a sufficient small ε, such that for any perturbation vector $\|E\|_2 \leq \varepsilon$, we have $\|E\alpha\|_2 \leq \epsilon$, *i.e.*, $\boldsymbol{x}_e \in \mathrm{Ball}(\boldsymbol{x}, \epsilon)$ holds. Based on the local constancy property with respect to $\boldsymbol{x}$ in the above first step, we conclude that Λ and $\mathrm{Sgn}(\boldsymbol{z}_\Lambda)$ are also locally constant with respect to D. □