Machine Learning (2019) 108:1975–2008
https://doi.org/10.1007/s10994-019-05805-2

Risk bound of transfer learning using parametric feature mapping and its application to sparse coding

Wataru Kumagai[1] · **Takafumi Kanamori**[2]

Received: 15 May 2018 / Accepted: 30 April 2019 / Published online: 20 May 2019

Abstract
In this study, we consider a transfer-learning problem using the parameter transfer approach, in which a suitable parameter of feature mapping is learned through one task and applied to another objective task. We introduce the notion of local stability and parameter transfer learnability of parametric feature mapping, and derive an excess risk bound for parameter transfer algorithms. As an application of parameter transfer learning, we discuss the performance of sparse coding in self-taught learning. Although self-taught learning algorithms with a large volume of unlabeled data often show excellent empirical performance, their theoretical analysis has not yet been studied. In this paper, we also provide a theoretical excess risk bound for self-taught learning. In addition, we show that the results of numerical experiments agree with our theoretical analysis.

Keywords Transfer learning · Sparse coding · Risk bound

1 Introduction

In traditional machine learning, it is assumed that data are identically drawn from a single distribution. However, this assumption does not always hold in real-world applications. Therefore, it is imperative to develop methods that are capable of incorporating samples drawn from different distributions. In this case, *transfer learning* provides a general way to accommodate these situations. In transfer learning, apart from the few samples that are available related to an objective task, abundant samples from another domain that are not necessarily drawn from an identical distribution can be used. The domain related to the objective task is called the target domain and the other domain is called the source domain.

Editor: Steve Hanneke.

✉ Wataru Kumagai
wataru.kumagai@riken.jp

[1] Center for Advanced Intelligence Project, RIKEN, 1-4-1, Nihonbashi, Chuo, Tokyo 103-0027, Japan

[2] Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Transfer learning aims to extract some useful knowledge from the source domain and apply this knowledge to achieve high task performance in the target domain.

Transfer learning is categorized in Pan and Yang (2010) into three areas: inductive transfer learning, transductive transfer learning, and unsupervised transfer learning. Inductive transfer learning corresponds to the setting in which labeled samples are available in the target domain. In addition, when labeled samples in the source domain are unavailable, the setting is called self-taught learning (Raina et al. 2007). In particular, self-taught learning can be applied to the case in which tasks are different in the source and target domains. Transductive transfer learning corresponds to the setting in which labeled samples are available only in the source domain. Then, tasks in both domains are typically assumed to be the same as in a covariate shift (Shimodaira 2000; Sugiyama et al. 2008) and sample selection bias (Zadrozny 2004; Huang et al. 2007). Domain adaptation (Daume and Marcu 2006; Blitzer et al. 2006) can be regarded as transfer learning in which tasks are the same in both domains; this is closely related to transductive transfer learning. Unsupervised transfer learning corresponds to the setting where labeled samples are unavailable in both domains. In this setting, the purpose is not to achieve high predictive performance but to perform an unsupervised task well in the target domain.

In accordance with the type of knowledge that is transferred, approaches for solving transfer-learning problems can be classified into types such as instance transfer, feature representation transfer, and parameter transfer (Pan and Yang 2010). In recent years, the parameter transfer approach has particularly attracted much attention in fine-tuning network weights of a deep neural network trained on source domains. In the setting of the parameter transfer approach, some kind of parametric models are supposed in both domains and the transferred knowledge is encoded into parameters. Biased regularization has been studied as a typical method in the parameter transfer approach, where the regularization term to an empirical loss has a non-zero center (e.g., $\|\mathbf{w} - \mathbf{w}_0\|^2$ instead of $\|\mathbf{w}\|^2$) and the center is learnt on the source domain (Ben-David and Urner 2013; Pentina and Lampert 2014; Tommasi et al. 2014). Recently, generalization of the biased regularization was proposed and theoretically analyzed (Kuzborskij and Orabona 2013, 2017). Owing to its flexibility, the parameter transfer approach can be applied to other algorithms such as sparse coding (Raina et al. 2007; Maurer et al. 2013), multiple kernel learning (Duan et al. 2012), and deep learning (Yosinski et al. 2014).

As the parameter transfer approach typically requires many samples to accurately learn a suitable parameter in the source domain, unsupervised methods are often utilized for the learning process. In this sense, self-taught learning is compatible with the parameter transfer approach. The sparse coding-based method was used in Raina et al. (2007), in which self-taught learning was first introduced. Moreover, in this work, the parameter transfer approach was used with regard to a dictionary learnt from images as the parameter to be transferred. However, although self-taught learning has been studied in various contexts (Dai et al. 2008; Lee et al. 2009; Wang et al. 2013; Zhu et al. 2013) and many algorithms based on the parameter transfer approach have empirically demonstrated impressive performance in self-taught learning, some fundamental problems remain. First, the theoretical aspects of the parameter transfer approach have not been sufficiently studied. For example, in the context of the parameter transfer approach, the generalization error bound applicable to self-taught learning has not been considered except in a few studies (Kuzborskij and Orabona 2013, 2017). Furthermore, existing studies only treat restricted hypothesis sets, limiting applicability in areas such as sparse coding, multiple kernel learning, and neural network. Second, although it is believed that a large amount of unlabeled data helps improve the performance of the objective task in self-taught learning, the exact sample size has not been sufficiently clarified.

Third, although sparsity-based methods are typically employed in self-taught learning, it is unknown how the sparsity works to guarantee the performance of self-taught learning.

In this study, we aimed to shed light on the above problems.[1] In this paper, we focus on inductive transfer learning and consider a general model of parametric feature mapping in the parameter transfer approach. We newly formulate the local stability of parametric feature mapping and the parameter transfer learnability for this mapping, and provide an excess risk bound for parameter transfer learning algorithms based on the notions. Furthermore, we consider the stability of sparse coding. Finally, we discuss parameter transfer learnability by dictionary learning under the sparse model. By applying the excess risk bound for parameter transfer learning algorithms, we derive an excess risk bound for the sparse coding algorithm in self-taught learning. Moreover, we show that the results of numerical experiments on handwritten digits datasets are in good agreement with the theoretical analysis of transfer learning with sparse coding. Note that our setting differs from the environment-based setting (Baxter 2000; Maurer 2009), where distribution over a set of distributions on labeled samples, known as an environment, is assumed. In our formulation, the existence of the environment is not assumed and presence of labeled data in the source domain is not required.

The remainder of the paper is organized as follows. In Sect. 2, we formulate the stability and parameter transfer learnability of the parametric feature mapping. Then, we present an excess risk bound for parameter transfer learning. In Sect. 3, we show the stability of sparse coding under perturbation of the dictionaries. By imposing sparsity assumptions on samples and considering dictionary learning, we derive the parameter transfer learnability for sparse coding. In particular, an excess risk bound is obtained for sparse coding in the setting of self-taught learning. Section 4 is devoted to numerical experiments of transfer learning with sparse coding. We conclude the paper with Sect. 5.

2 Excess risk bound for parameter transfer learning

2.1 Problem setting of parameter transfer learning

We formulate parameter transfer learning in this section. We first briefly introduce notations and terminology in transfer learning (Pan and Yang 2010). Let $\mathcal{X}$ and $\mathcal{Y}$ represent a sample space and label space, respectively. In addition, let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ be a hypothesis space and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ represent a loss function. Then, the expected risk and the empirical risk are defined as $\mathcal{R}(h) := \mathbb{E}_{(\mathbf{x},y)\sim P}[\ell(y, h(\mathbf{x}))]$ and $\widehat{\mathcal{R}}_n(h) := \frac{1}{n}\sum_{j=1}^{n} \ell(y_j, h(\mathbf{x}_j))$, respectively. In the transfer learning setting, it is assumed that, apart from samples from a domain of interest (i.e., *target domain*), samples from another domain (i.e., *source domain*) are also available. We distinguish between the target and source domains by adding a subscript $\mathcal{T}$ or $\mathcal{S}$ to each notation introduced above, (e.g., $P_{\mathcal{T}}$, $\mathcal{R}_{\mathcal{S}}$). The homogeneous setting (i.e., $\mathcal{X}_{\mathcal{S}} = \mathcal{X}_{\mathcal{T}}$) is not assumed in general, and thus, the heterogeneous setting (i.e., $\mathcal{X}_{\mathcal{S}} \neq \mathcal{X}_{\mathcal{T}}$) is used here. We note that self-taught learning, which is discussed in Sect. 3, corresponds to the case in which the label space $\mathcal{Y}_{\mathcal{S}}$ in the source task is the set of a single element.

We consider the parameter transfer approach in which the knowledge to be transferred is encoded in a parameter. The parameter transfer approach aims to learn a hypothesis with low expected risk for the target task by obtaining some knowledge about an effective parameter

[1] A short version of this article was published as a conference paper (Kumagai 2016). In this article, we present new theoretical results that extend our previous work, and include more detailed analysis of excess risk bound. Furthermore, results are provided from numerical experiments conducted to confirm the validity of the theoretical analysis in practical situations.

in the source domain and transferring it to the target domain. We suppose that there are parametric models on both the source and target domains and their parameter spaces are partly shared. Our strategy is to learn an effective parameter in the source domain and then transfer a part of the parameter to the target domain. Next, we describe the formulation. In the target domain, we assume that $\mathcal{Y}_\mathcal{T} \subset \mathbb{R}$ and there is a parametric feature mapping $\psi_{\boldsymbol{\theta}} : \mathcal{X}_\mathcal{T} \to \mathbb{R}^m$ on the target domain such that each hypothesis $h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}} : \mathcal{X}_\mathcal{T} \to \mathcal{Y}_\mathcal{T}$ is represented by

$$h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}}(\mathbf{x}) := \langle \mathbf{w}, \psi_{\boldsymbol{\theta}}(\mathbf{x}) \rangle, \tag{1}$$

with parameters $\boldsymbol{\theta} \in \Theta$ and $\mathbf{w} \in \mathcal{W}_\mathcal{T}$, where Θ is a subset of a normed space with norm $\|\cdot\|$ and $\mathcal{W}_\mathcal{T}$ is a subset of $\mathbb{R}^m$. Then, the hypothesis set in the target domain is parameterized as

$$\mathcal{H}_\mathcal{T} = \{h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}} | \boldsymbol{\theta} \in \Theta, \mathbf{w} \in \mathcal{W}_\mathcal{T}\}.$$

In the following discussion, we simply denote $\mathcal{R}_\mathcal{T}(h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}})$ and $\widehat{\mathcal{R}}_{\mathcal{T},n}(h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}})$ by $\mathcal{R}_\mathcal{T}(\boldsymbol{\theta}, \mathbf{w})$ and $\widehat{\mathcal{R}}_{\mathcal{T},n}(\boldsymbol{\theta}, \mathbf{w})$, respectively. In the source domain, we suppose that there exists some kind of parametric model such as a sample distribution $P_{\mathcal{S},\boldsymbol{\theta},\mathbf{w}}$ or a hypothesis $h_{\mathcal{S},\boldsymbol{\theta},\mathbf{w}}$ with parameters $\boldsymbol{\theta} \in \Theta$ and $\mathbf{w} \in \mathcal{W}_\mathcal{S}$, and a part Θ of the parameter space is shared with the target domain. Then, let $\boldsymbol{\theta}^*_\mathcal{S} \in \Theta$ and $\mathbf{w}^*_\mathcal{S} \in \mathcal{W}_\mathcal{S}$ be parameters that are supposed to be effective in the source domain (e.g., the true parameter of the sample distribution, the parameter of the optimal hypothesis with respect to the expected risk $\mathcal{R}_\mathcal{S}$). Here, the parameters $\boldsymbol{\theta}^*_\mathcal{S}$ and $\mathbf{w}^*_\mathcal{S}$ may be taken mathematically arbitrarily (i.e. there are no mathematical restrictions) and we do not use any specific property on $\boldsymbol{\theta}^*_\mathcal{S}$ and $\mathbf{w}^*_\mathcal{S}$. Then, the parameter transfer algorithm treated in this paper is described as follows. Let N- and n-samples be available in the source and target domains, respectively. First, a parameter transfer algorithm outputs the estimator $\widehat{\boldsymbol{\theta}}_N \in \Theta$ of $\boldsymbol{\theta}^*_\mathcal{S}$ by using N-samples. Next, for the parameter

$$\mathbf{w}^*_\mathcal{T} := \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}_\mathcal{T}} \mathcal{R}_\mathcal{T}\left(\boldsymbol{\theta}^*_\mathcal{S}, \mathbf{w}\right) \tag{2}$$

in the target domain, the algorithm outputs its estimator

$$\widehat{\mathbf{w}}_{N,n} := \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}_\mathcal{T}} \widehat{\mathcal{R}}_{\mathcal{T},n}(\widehat{\boldsymbol{\theta}}_N, \mathbf{w}) + \rho r(\mathbf{w}) \tag{3}$$

by using n-samples, where $r(\mathbf{w})$ is a 1-strongly convex function with respect to $\|\cdot\|_2$ and $\rho > 0$. If the source domain relates to the target domain in some sense, the effective parameter $\boldsymbol{\theta}^*_\mathcal{S}$ in the source domain is also expected to be useful for the target task. In the next section, we regard $\mathcal{R}_\mathcal{T}\left(\boldsymbol{\theta}^*_\mathcal{S}, \mathbf{w}^*_\mathcal{T}\right)$ as the baseline of predictive performance and derive an excess risk bound. The validity of the baseline is discussed in Sect. 2.2.

2.2 Excess risk bound based on stability and learnability

We introduce two new metrics, the local stability and parameter transfer learnability, as described below. These notions are essential to derive an excess risk bound in Theorem 1.

Definition 1 (*Local Stability*) A parametric feature mapping $\psi_{\boldsymbol{\theta}}$ is said to be locally stable if there exist $\epsilon_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}_{>0}$ for each $\boldsymbol{\theta} \in \Theta$ and $L_\psi > 0$ such that, for $\boldsymbol{\theta}' \in \Theta$,

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq \epsilon_{\boldsymbol{\theta}}(\mathbf{x}) \Rightarrow \|\psi_{\boldsymbol{\theta}}(\mathbf{x}) - \psi_{\boldsymbol{\theta}'}(\mathbf{x})\|_2 \leq L_\psi \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

Local stability implies that the feature is not significantly affected by the parameter shift. We term $\epsilon_{\boldsymbol{\theta}}(\mathbf{x})$ as the permissible radius of perturbation for $\boldsymbol{\theta}$ at $\mathbf{x}$. For samples $\mathbf{X}^n = \{\mathbf{x}_1, \dots \mathbf{x}_n\}$, we have $\epsilon_{\boldsymbol{\theta}}(\mathbf{X}^n) := \min_{j\in[n]} \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_j)$, where $[n] := \{1, \dots, n\}$ for a positive integer n.

Next, we formulate the parameter transfer learnability based on the local stability.

Definition 2 (*Parameter Transfer Learnability*) Suppose that N-samples are available in the source domain and a sample $\mathbf{x}$ is available in the target domain. Let the parametric feature mapping $\{\psi_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$ be locally stable. For $\bar{\delta}_N \in [0, 1)$, $\{\psi_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$ is said to be parameter transfer learnable with probability $1-\bar{\delta}_N$ if there exists an algorithm that depends only on N-samples in the source domain such that the output $\widehat{\boldsymbol{\theta}}_N$ of the algorithm satisfies

$$\Pr\left[\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\| \le \epsilon_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\right] \ge 1 - \bar{\delta}_N.$$

The parameter $\bar{\delta}_N$ is written as $\bar{\delta}$ for short as long as no conflict arises.

The parameter transfer learnability describes whether the effective parameter is properly transformed on the target domain with high probability. For n-samples $\mathbf{X}^n = \{\mathbf{x}_1, \dots \mathbf{x}_n\}$ in the target domain, the union bound ensures that the inequality $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\| \le \epsilon_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{X}^n)$ holds with probability greater than or equal to $1 - n\bar{\delta}_N$.

Given training samples $\{(\mathbf{x}_j, y_j) : j = 1, \dots, n\}$ in the target domain, let us consider the learning method

$$\min_{\mathbf{w}\in\mathcal{W}_{\mathcal{T}}} \frac{1}{n}\sum_{j=1}^{n} \ell(y_j, \langle \mathbf{w}, \psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}_j)\rangle) + \rho r(\mathbf{w}),$$

where $\widehat{\boldsymbol{\theta}}_N$ is the estimated parameter in the source domain using N training samples. The optimal parameter in $\mathcal{W}_{\mathcal{T}}$ is denoted as $\widehat{\mathbf{w}}_{N,n}$. Then, the following excess risk bound is obtained.

Theorem 1 (Excess Risk Bound) *We assume the following conditions.*

1. *The parametric feature mapping $\psi_\theta(\mathbf{x})$ is bounded and locally stable with the parameter L_ψ. Suppose that $\sup_{\boldsymbol{\theta}\in\Theta,\mathbf{x}\in\mathcal{X}} \|\psi_{\boldsymbol{\theta}}(\mathbf{x})\|_2 \le R_\psi$ holds for some positive constant R_ψ.*
2. *The estimator $\widehat{\boldsymbol{\theta}}_N$ on the source domain satisfies the transfer learnability with probability $1-\bar{\delta}$.*
3. *The non-negative loss $\ell(\cdot,\cdot)$ on the target domain is L_ℓ-Lipschitz and convex in the second argument. Moreover, we assume that $\sup_y \ell(y, 0)$ is bounded above by a positive constant L_0.*
4. *The non-negative regularization term $r(\mathbf{w})$ is 1-strongly convex and $r(\mathbf{0}) = 0$ holds.*

Then, the excess risk is bounded above by

$$\begin{aligned}\mathcal{R}_{\text{excess}} &:= \mathcal{R}_{\mathcal{T}}(\widehat{\boldsymbol{\theta}}_N, \widehat{\mathbf{w}}_{N,n}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}^*_{\mathcal{T}}) \\ &\le c\left\{\frac{\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|}{\sqrt{\rho}} + \frac{1}{\sqrt{n'\rho}} + \frac{\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|^{1/2}}{\rho^{3/4}} + \frac{1}{n\rho} + \rho\right\} \end{aligned} \tag{4}$$

with probability $1-\delta-(n+n')\bar{\delta}$, where n' is an arbitrary natural number and c is a positive constant expressed as a polynomial in L_ψ, R_ψ, L_ℓ, L_0, $r(\mathbf{w}^*_{\mathcal{T}})$, and $\log(1/\delta)$.

Proof In the proof, we define c_i $(i = 1, 2, 3, 4, 5)$ as a positive number depending on L_ψ, R_ψ, L_ℓ, L_0, and $\log(1/\delta)$.

Using the boundedness of the non-negative loss $\ell(\cdot,\cdot)$ and the strong convexity of $r(\mathbf{w})$ with some other conditions, we have

$$\begin{aligned}\frac{\rho}{2}\|\widehat{\mathbf{w}}_{N,n}\|^2 &\leq \frac{1}{n}\sum_{j=1}^{n}\ell(y_j, \langle\widehat{\mathbf{w}}_{N,n},\ \psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}_j)\rangle) + \rho r(\widehat{\mathbf{w}}_{N,n})\\ &\leq \frac{1}{n}\sum_{j=1}^{n}\ell(y_j, 0) + \rho r(\mathbf{0}) \leq L_0.\end{aligned}$$

Thus, $\|\widehat{\mathbf{w}}_{N,n}\|$ is bounded above by $\sqrt{2L_0/\rho}$. Let $\widehat{\mathbf{w}}^*_n$ be the optimal solution of

$$\min_{\mathbf{w}\in\mathcal{W}_{\mathcal{T}}}\frac{1}{n}\sum_{j=1}^{n}\ell(y_j,\ \langle\mathbf{w},\psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x}_j)\rangle) + \rho r(\mathbf{w}).$$

Likewise, we see that the norm of $\widehat{\mathbf{w}}^*_n$ has the same upper bound.

The excess risk is decomposed to the following three terms.

$$\begin{aligned}&\mathcal{R}_{\mathcal{T}}(\widehat{\boldsymbol{\theta}}_N,\ \widehat{\mathbf{w}}_{N,n}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}^*_{\mathcal{S}},\ \mathbf{w}^*_{\mathcal{T}})\\ &\quad= \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y, \langle\widehat{\mathbf{w}}_{N,n}, \psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x})\rangle)\right] - \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y, \langle\widehat{\mathbf{w}}_{N,n}, \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right]\\ &\qquad+ \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y, \langle\widehat{\mathbf{w}}_{N,n}, \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right] - \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y, \langle\widehat{\mathbf{w}}^*_n, \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right]\\ &\qquad+ \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y, \langle\widehat{\mathbf{w}}^*_n, \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right] - \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y, \langle\mathbf{w}^*_{\mathcal{T}}, \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right].\end{aligned}$$

Let us consider the upper bound of each term.

For the first term of the excess risk, the following inequality holds:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y, \langle\widehat{\mathbf{w}}_{N,n}, \psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x})\rangle)\right] - \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y, \langle\widehat{\mathbf{w}}_{N,n}, \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right]\\ &\quad\leq L_\ell\sqrt{\frac{2L_0}{\rho}}\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\|\psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}) - \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\|\right].\end{aligned} \tag{5}$$

Here, for an arbitrary natural number n', we introduce independent random variables $\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_{n'}$ (called ghost samples) such that the probability distribution of each $\bar{\mathbf{x}}_j$ is the marginal distribution of $P_{\mathcal{T}}$. Then, we have the following bound with probability greater than $1-\delta/2$ by Hoeffding's inequality:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\|\psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}) - \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\|\right]\\ &\quad\leq \frac{1}{n'}\sum_{i=1}^{n'}\|\psi_{\widehat{\boldsymbol{\theta}}_N}(\bar{\mathbf{x}}_i) - \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\bar{\mathbf{x}}_i)\| + R_\psi\sqrt{\frac{2\log(2/\delta)}{n'}}.\end{aligned} \tag{6}$$

Moreover, since it holds that $\|\psi_{\widehat{\boldsymbol{\theta}}_N}(\bar{\mathbf{x}}_i) - \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\bar{\mathbf{x}}_i)\| \leq L_\psi\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|$ with probability greater than $1-\bar{\delta}$ by local stability and parameter transfer learnability, we have the following bound with probability greater than $1-n'\bar{\delta}$ by the union bound:

$$\frac{1}{n'}\sum_{i=1}^{n'}\|\psi_{\widehat{\boldsymbol{\theta}}_N}(\bar{\mathbf{x}}_i) - \psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\bar{\mathbf{x}}_i)\| \leq L_\psi\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|. \tag{7}$$

From (5)–(7), with probability greater than $1-\delta/2-n'\bar{\delta}$, we obtain

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y,\langle\widehat{\mathbf{w}}_{N,n},\psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x})\rangle)\right]-\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y,\langle\widehat{\mathbf{w}}_{N,n},\psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right]\\
&\leq L_\ell L_\psi\sqrt{\frac{2L_0}{\rho}}\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}^*_{\mathcal{S}}\|+L_\ell R_\psi\sqrt{\frac{2L_0}{\rho}}\sqrt{\frac{2\log(2/\delta)}{n'}}\\
&=c_1\frac{\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}^*_{\mathcal{S}}\|}{\sqrt{\rho}}+c_2\frac{1}{\sqrt{n'\rho}}.
\end{aligned}$$

Next, we provide an upper bound of the second term of the decomposed excess risk. To do so, we first provide an upper bound of $\|\widehat{\mathbf{w}}^*_n-\widehat{\mathbf{w}}_{N,n}\|$ in the following. The 1-strong convexity of r leads to ρ-strong convexity of the empirical loss with the regularization term in the parameter $\mathbf{w}$. Hence, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\ell(y_j,\langle\widehat{\mathbf{w}}^*_n,\psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x}_j)\rangle)+\rho r(\widehat{\mathbf{w}}^*_n)+\frac{\rho}{2}\|\widehat{\mathbf{w}}^*_n-\widehat{\mathbf{w}}_{N,n}\|^2\\
&\quad\leq\frac{1}{n}\sum_{j=1}^{n}\ell(y_j,\langle\widehat{\mathbf{w}}_{N,n},\psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x}_j)\rangle)+\rho r(\widehat{\mathbf{w}}_{N,n})
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\ell(y_j,\langle\widehat{\mathbf{w}}_{N,n},\psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}_j)\rangle)+\rho r(\widehat{\mathbf{w}}_{N,n})+\frac{\rho}{2}\|\widehat{\mathbf{w}}^*_n-\widehat{\mathbf{w}}_{N,n}\|^2\\
&\quad\leq\frac{1}{n}\sum_{j=1}^{n}\ell(y_j,\langle\widehat{\mathbf{w}}^*_n,\psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}_j)\rangle)+\rho r(\widehat{\mathbf{w}}^*_n).
\end{aligned}$$

Summing up the above two inequalities, we have

$$\begin{aligned}
\rho\|\widehat{\mathbf{w}}^*_n-\widehat{\mathbf{w}}_{N,n}\|^2&\leq\frac{1}{n}\sum_{j=1}^{n}\left(\ell(y_j,\langle\widehat{\mathbf{w}}^*_n,\psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}_j)\rangle)-\ell(y_j,\langle\widehat{\mathbf{w}}^*_n,\psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x}_j)\rangle)\right)\\
&\quad+\frac{1}{n}\sum_{j=1}^{n}\left(\ell(y_j,\langle\widehat{\mathbf{w}}_{N,n},\psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x}_j)\rangle)-\ell(y_j,\langle\widehat{\mathbf{w}}_{N,n},\psi_{\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}_j)\rangle)\right)\\
&\leq 2L_\ell L_\psi\sqrt{\frac{2L_0}{\rho}}\left\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}^*_{\mathcal{S}}\right\|.
\end{aligned}$$

The last inequality holds with probability greater than or equal to $1-n\bar{\delta}$ owing to the parameter transfer learnability and local stability. Thus, $\|\widehat{\mathbf{w}}^*_n-\widehat{\mathbf{w}}_{N,n}\|\leq 2^{3/4}(L_\ell L_\psi L_0^{1/2})^{1/2}\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}^*_{\mathcal{S}}\|^{1/2}/\rho^{3/4}$ holds. Hence, the second term of the decomposed excess risk is bounded above by

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y,\langle\widehat{\mathbf{w}}_{N,n},\psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right]-\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y,\langle\widehat{\mathbf{w}}^*_n,\psi_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})\rangle)\right]\\
&\leq L_\ell R_\psi\|\widehat{\mathbf{w}}_{N,n}-\widehat{\mathbf{w}}^*_n\|\\
&\leq 2^{3/4}L_\ell^{3/2}L_\psi^{1/2}L_0^{1/4}R_\psi\frac{\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}^*_{\mathcal{S}}\|^{1/2}}{\rho^{3/4}}=c_3\frac{\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}^*_{\mathcal{S}}\|^{1/2}}{\rho^{3/4}}
\end{aligned}$$

with probability $1-n\bar{\delta}$.

For the third term of the excess risk, we obtain the following upper bound with probability $1-\delta/2$ by Theorem 1 of Sridharan et al. (2009):

$$\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y,\langle\widehat{\mathbf{w}}_n^*,\psi_{\boldsymbol{\theta}_{\mathcal{S}}^*}(\mathbf{x})\rangle)\right]-\mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{T}}}\left[\ell(y,\langle\mathbf{w}_{\mathcal{T}}^*,\psi_{\boldsymbol{\theta}_{\mathcal{S}}^*}(\mathbf{x})\rangle)\right]$$
$$\leq\frac{8L_\ell^2R_\psi^2(32+\log(2/\delta))}{n\rho}+\rho r(\mathbf{w}_{\mathcal{T}}^*)=\frac{c_4}{n\rho}+c_5\rho.$$

Combining the above results, we obtain

$$\mathcal{R}_{\text{excess}}\leq c\left\{\frac{\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_{\mathcal{S}}^*\|}{\sqrt{\rho}}+\frac{1}{\sqrt{n'\rho}}+\frac{\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_{\mathcal{S}}^*\|^{1/2}}{\rho^{3/4}}+\frac{1}{n\rho}+\rho\right\}$$

with probability of at least $1-\delta-(n+n')\bar{\delta}$. □

We mention the relation between our formulation and a fast rate result of the excess risk in "Appendix A". The optimal ρ is obtained by minimizing the upper bound of the excess risk.

Corollary 1 *Suppose that the conditions 1, 3, and 4 in Theorem 1 hold. In addition, we assume that there exist a real number $\beta\geq 1$ and a sequence τ_N such that*

$$\mathbb{E}\left[\epsilon_{\boldsymbol{\theta}_{\mathcal{S}}^*}(\mathbf{x})^{-\beta}\right]<\infty,\quad\text{and}\quad\mathbb{E}[\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_{\mathcal{S}}^*\|^\beta]\leq\tau_N^\beta\longrightarrow 0\ \ (N\to\infty).\tag{8}$$

When $n\tau_N^\beta$ is sufficiently small, the asymptotic upper bound of the excess risk is given as

$$\mathcal{R}_{\text{excess}}\leq c\max\{n^{-1/2},\ \tau_N^{2/7}\},$$

by setting $\rho=\Theta(\max\{n^{-1/2},\ \tau_N^{2/7}\})$.

Proof The assumptions (8) and Markov's inequality lead to $\Pr\left[\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_{\mathcal{S}}^*\|/\tau_N\geq a\right]\leq a^{-\beta}$ and

$$\Pr\left[\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_{\mathcal{S}}^*\|\geq\epsilon_{\boldsymbol{\theta}_{\mathcal{S}}^*}(\mathbf{x})\right]\leq C\tau_N^\beta,\tag{9}$$

where C is a positive constant. Here, the independence of the source and target samples is used. The second inequality denotes that parameter transfer learnability holds by setting $\bar{\delta}_N=C\tau_N^\beta$. From the first inequality, we have $\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_{\mathcal{S}}^*\|/\sqrt{\rho}=O_p(\tau_N/\sqrt{\rho})$ and $\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_{\mathcal{S}}^*\|^{1/2}/\rho^{3/4}=O_p(\tau_N^{1/2}/\rho^{3/4})$, where O_p denotes the probabilistic order. Let δ be a small positive constant, and define n' by $n'=\delta/\bar{\delta}_N=\delta/(C\tau_N^\beta)$. We have

$$\frac{\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_{\mathcal{S}}^*\|}{\sqrt{\rho}}+\frac{1}{\sqrt{n'\rho}}=O_p(\tau_N^{\min\{1,\beta/2\}}/\sqrt{\rho}).$$

Suppose that $\rho\to 0$ and $n\rho\to\infty$ hold as $n\to\infty$ and $\tau_N\to 0$ while keeping $n\bar{\delta}_N=Cn\tau_N^\beta$ sufficiently small. For large n and small τ_N, we have $\tau_N^{\min\{1,\beta/2\}}/\sqrt{\rho}\leq\tau_N^{1/2}/\rho^{3/4}$. Hence, we obtain

$$\mathcal{R}_{\text{excess}} \leq c \left\{ \frac{\tau_N^{1/2}}{\rho^{3/4}} + \frac{1}{n\rho} + \rho \right\}$$

with probability greater than $1 - 2\delta - n\bar{\delta}_N$. Substituting

$$\rho = \Theta\left(\max\left\{n^{-1/2},\ \tau_N^{2/7}\right\}\right)$$

which satisfies the above condition, we have $\mathcal{R}_{\text{excess}} \leq c \max\{n^{-1/2},\ \tau_N^{2/7}\}$ with high probability. □

The upper bound of the excess risk is expressed by the bias term τ_N induced from the source domain and the sample complexity bound on the target domain. If τ_N is large, additional training samples on the target domain will not help attain high prediction accuracy. On the contrary, when the bias term τ_N is sufficiently small, the excess risk is bounded above by $\mathcal{O}(n^{-1/2})$, which is the standard asymptotic order of the supervised learning using n i.i.d. samples.

Remark 1 Suppose that the bias τ_N on the source domain is of the order $N^{-1/2}$, which is the standard order in the parameter estimation.[2] When $n\tau_N^{\beta}$ is sufficiently small for some $\beta \geq 1$, we have $n = \mathcal{O}(N^{\beta/2})$. If $c' N^{2/7} \leq n \leq c'' N^{\beta/2}$ holds for some constants c', c'', the excess risk is of the order $\mathcal{O}(N^{-1/7})$. For $n = \mathcal{O}(N^{2/7})$, we have $\mathcal{R}_{\text{excess}} = \mathcal{O}(n^{-1/2})$. Given an acceptable level of the excess risk, the above result provides a rough estimate of the required sample size on both the source and target domains.

One can regard $\mathcal{R}_{\mathcal{T}}^* = \min_{\boldsymbol{\theta}, \mathbf{w}} \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}, \mathbf{w})$ as the baseline instead of $\mathcal{R}_{\mathcal{T}}\left(\boldsymbol{\theta}_{\mathcal{S}}^*, \mathbf{w}_{\mathcal{T}}^*\right)$. In this case, the risk bound is decomposed into

$$\mathcal{R}_{\mathcal{T}}(\widehat{\boldsymbol{\theta}}_N, \widehat{\mathbf{w}}_{N,n}) - \mathcal{R}_{\mathcal{T}}^* = \underbrace{\left(\mathcal{R}_{\mathcal{T}}(\widehat{\boldsymbol{\theta}}_N, \widehat{\mathbf{w}}_{N,n}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}_{\mathcal{S}}^*, \mathbf{w}_{\mathcal{T}}^*)\right)}_{\mathcal{R}_{\text{excess}}} + \underbrace{\left(\mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}_{\mathcal{S}}^*, \mathbf{w}_{\mathcal{T}}^*) - \mathcal{R}_{\mathcal{T}}^*\right)}_{\mathcal{R}_{\text{gap}}}.$$

The first term, $\mathcal{R}_{\text{excess}}$, denotes the excess risk to transfer learning with optimal parameter on the source domain and its upper bounded is presented in Theorem 1 and Corollary 1. The second term, $\mathcal{R}_{\text{gap}}$, is interpreted as the difference between the source and target domains.

In an ideal situation, transfer learning is regarded as a method to reduce the bias of the model; this is explained next. Suppose that $\mathcal{R}_{\text{gap}}$ is close to zero and N is sufficiently large. Then, self-taught learning with the optimal parameter $\boldsymbol{\theta}_{\mathcal{S}}^*$ is approximately realized. However, in the common learning setup using samples from only the target domain, the optimal feature representation $\psi_{\boldsymbol{\theta}_{\mathcal{S}}^*}$ will not be available. This is thought to be the main reason why transfer learning is advantageous over the standard learning methods.

On the contrary, if $\mathcal{R}_{\text{gap}}$ is much larger than $\mathcal{R}_{\text{excess}}$, *negative transfer* can occur easily, i.e., transfer learning actually decreases the prediction performance. This is because the parameter $\boldsymbol{\theta}$ that is superior to $\boldsymbol{\theta}_{\mathcal{S}}^*$ will be effortlessly found.

[2] For example, let us consider the case of fine-tuning of deep learning, which is a typical transfer learning method. Then, probabilistic models constructed by neural networks are pre-trained in the source domain and fine-tuned in the target domain. We note that deep neural networks with Lipschitz activation functions (e.g. ReLU, sigmoid and softmax) satisfy local stability and parameter transfer learnability for $\epsilon_\theta(\mathbf{x}) \equiv \infty$. If the models satisfy regularity conditions around the optimal parameter $\boldsymbol{\theta}^*$ in the source domain, the bias τ_N can attain the order $\mathcal{O}(N^{-1/2})$. However, the models based on neural networks are known to have many singular points. Then, the bias around a singular point is thought to have a different order and its evaluation is the future work which relates to parameter transfer learning.

Example 1 As an example of $\mathcal{R}_{\mathrm{gap}}$, let us consider the regression analysis using the basis function $\psi_{\boldsymbol{\theta}}$. We assume that the labels in source and target domains are given as $y = \mathbf{w}_{\mathcal{S}}^{\top}\psi_{\boldsymbol{\theta}_{\mathcal{S}}} + \xi$ and $y = \mathbf{w}_{\mathcal{T}}^{\top}\psi_{\boldsymbol{\theta}_{\mathcal{T}}} + \epsilon$ respectively, where ξ and ϵ are noise random variables with mean 0. In addition, let the loss function be $\ell(y, y') := |y - y'|$ and effective parameters in source domain be $\boldsymbol{\theta}_{\mathcal{S}}^{*} = \boldsymbol{\theta}_{\mathcal{S}}, \mathbf{w}_{\mathcal{S}}^{*} = \mathbf{w}_{\mathcal{S}}$. Then, it holds that

$$\begin{aligned}
\mathcal{R}_{\mathrm{gap}} &:= \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}_{\mathcal{S}}^{*}, \mathbf{w}_{\mathcal{T}}^{*}) - \mathcal{R}_{\mathcal{T}}^{*} \\
&= \mathbb{E}_{\mathcal{T}}[|\mathbf{w}_{\mathcal{T}}^{\top}\psi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}) + \epsilon - \mathbf{w}_{\mathcal{T}}^{*\top}\psi_{\boldsymbol{\theta}_{\mathcal{S}}}(\mathbf{x})|] - \mathbb{E}_{\mathcal{T}}[|\mathbf{w}_{\mathcal{T}}^{\top}\psi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}) + \epsilon - \mathbf{w}_{\mathcal{T}}^{\top}\psi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x})|] \\
&\leq \mathbb{E}_{\mathcal{T}}[|\mathbf{w}_{\mathcal{T}}^{\top}\psi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}) - \mathbf{w}_{\mathcal{T}}^{*\top}\psi_{\boldsymbol{\theta}_{\mathcal{S}}}(\mathbf{x})|] + \mathbb{E}[|\epsilon|] - \mathbb{E}[|\epsilon|] \\
&\leq \mathbb{E}_{\mathcal{T}}[|\mathbf{w}_{\mathcal{T}}^{\top}(\psi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}) - \psi_{\boldsymbol{\theta}_{\mathcal{S}}}(\mathbf{x}))|] + \mathbb{E}_{\mathcal{T}}[|(\mathbf{w}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{*})^{\top}\psi_{\boldsymbol{\theta}_{\mathcal{S}}}(\mathbf{x})|] \\
&\leq \|\mathbf{w}_{\mathcal{T}}\|\mathbb{E}_{\mathcal{T}}[\|\psi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}) - \psi_{\boldsymbol{\theta}_{\mathcal{S}}}(\mathbf{x})\|] + R_{\psi}\|\mathbf{w}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{*}\|.
\end{aligned}$$

Thus, it is found from this upper bound of $\mathcal{R}_{\mathrm{gap}}$ that, if the parameter $\boldsymbol{\theta}_{\mathcal{S}}$ of the optimal feature map in source domain is distant from that $\boldsymbol{\theta}_{\mathcal{T}}$ in target domain, the first term can be large, and accordingly, the second term can be also large since $\mathbf{w}_{\mathcal{T}}^{*}$ depends on $\boldsymbol{\theta}_{\mathcal{S}}$.

A simple way to avoid the negative transfer is to assess the $\mathcal{R}_{\mathrm{gap}}$. A naive statistic,

$$\widehat{\mathcal{R}}_{\mathrm{gap}} = \widehat{\mathcal{R}}_{\mathcal{T},n}(\widehat{\boldsymbol{\theta}}_{N}, \widehat{\mathbf{w}}_{N,n}) - \min_{\boldsymbol{\theta},\mathbf{w}} \widehat{\mathcal{R}}_{\mathcal{T},n}(\boldsymbol{\theta}, \mathbf{w}),$$

is available to estimate $\mathcal{R}_{\mathrm{gap}}$. When $\widehat{\mathcal{R}}_{\mathrm{gap}}$ is significantly larger than the order of $\mathcal{O}(n^{-1/2})$, we will need more elaborate learning on the source domain or fine tuning (Goodfellow et al. 2016, Sec. 8.7.4) of the parameter $\boldsymbol{\theta}$ using samples on the target domain. The domain adaptation is also another promising method to avoid a large $\mathcal{R}_{\mathrm{gap}}$ when samples in the source and target domains are simultaneously available. We do not go into the details for this case here. In this paper, we assume that $\mathcal{R}_{\mathrm{gap}}$ is sufficiently small and we focus on the excess risk $\mathcal{R}_{\mathrm{excess}}$ via local stability and parameter transfer learnability.

3 Stability and learnability in sparse coding

In this section, we consider sparse coding in self-taught learning, where the source domain essentially consists of the sample space $\mathcal{X}_{\mathcal{S}}$ without the label space $\mathcal{Y}_{\mathcal{S}}$. We assume that the sample spaces in both domains are $\mathbb{R}^d$. Then, the sparse coding method considered here consists of a two-stage procedure, where a dictionary is learnt on the source domain, and then sparse coding with the learnt dictionary is used for a predictive task in the target domain.

First, we show that sparse coding satisfies the local stability in Sect. 3.1 and then explain how appropriate dictionary learning algorithms satisfy the parameter transfer learnability in Sect. 3.3. As a consequence of Theorem 1, we obtain the excess risk bound of self-taught learning algorithms based on sparse coding. We note that the results in this section are useful independent of transfer learning.

Next, we summarize the notations used in this section. Let $\|\cdot\|_p$ be the p-norm on $\mathbb{R}^d$. We define $\mathrm{supp}(\mathbf{a}) := \{i \in [m] | a_i \neq 0\}$ for $\mathbf{a} \in \mathbb{R}^m$. We denote the number of elements of a set S by $|S|$. When a vector $\mathbf{a}$ satisfies $\|\mathbf{a}\|_0 = |\mathrm{supp}(\mathbf{a})| \leq k$, $\mathbf{a}$ is said to be k-sparse. We set $\mathcal{D} := \{\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_m] \in \mathbb{R}^{d\times m} | \ \|\mathbf{d}_j\|_2 = 1 \ (i = 1, \ldots, m)\}$ and each $\mathbf{D} \in \mathcal{D}$ represents a dictionary of size m.

Definition 3 (*Induced matrix norm*)[3] For an arbitrary matrix $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_m] \in \mathbb{R}^{d\times m}$, the induced matrix norm is defined by $\|\mathbf{E}\|_{1,2} := \max_{i\in[m]} \|\mathbf{e}_i\|_2$.

We adopt $\|\cdot\|_{1,2}$ to measure the difference in dictionaries since it is typically used in the framework of dictionary learning. We note that $\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2} \leq 2$ holds for arbitrary dictionaries $\mathbf{D}, \tilde{\mathbf{D}} \in \mathcal{D}$.

3.1 Local stability of sparse representation

In this section, we show the local stability of sparse representation under a sparse model. A sparse representation with dictionary parameter $\mathbf{D}$ of a sample $\mathbf{x} \in \mathbb{R}^d$ is expressed as follows:

$$\varphi_{\mathbf{D}}(\mathbf{x}) := \underset{\mathbf{z}\in\mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1, \tag{10}$$

where $\lambda > 0$ is a regularization parameter that induces sparsity. This situation corresponds to the case where $\boldsymbol{\theta} = \mathbf{D}$ and $\psi_{\boldsymbol{\theta}} = \varphi_{\mathbf{D}}$ in the setting of Sect. 2.1.

We define some notions used in the discussion on stability of sparse representation. The following k-margin was introduced by Mehta and Gray (2013).

Definition 4 (*k-margin*) Given a dictionary $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_m] \in \mathcal{D}$ and a point $\mathbf{x} \in \mathbb{R}^d$, the k-margin of $\mathbf{D}$ on $\mathbf{x}$ is

$$\mathcal{M}_k(\mathbf{D}, \mathbf{x}) := \max_{\mathcal{I}\subset[m], |\mathcal{I}|=m-k} \min_{j\in\mathcal{I}} \left\{\lambda - |\langle \mathbf{d}_j, \mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\rangle|\right\}.$$

The following μ-incoherence is not equal to the k-incoherence defined in Mehta and Gray (2013), although these are related to each other as stated in Remark 2.

Definition 5 (*μ-incoherence*) A dictionary matrix $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_m] \in \mathcal{D}$ is said to be μ-incoherent if $|\langle \mathbf{d}_i, \mathbf{d}_j\rangle| \leq \mu/\sqrt{d}$ for all $i \neq j$.

Then, the following theorem is obtained.

Theorem 2 (Local Stability of Sparse Coding) *Let $\mathbf{D} \in \mathcal{D}$ be μ-incoherent for $\mu < \sqrt{d}/k$ and $\|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2} \leq \lambda$. When*

$$\|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2} \leq \epsilon_{k,\mathbf{D}}(\mathbf{x}) := \frac{\mathcal{M}_k(\mathbf{D}, \mathbf{x})^2\lambda}{64\max\{1, \|\mathbf{x}\|\}^4}, \tag{11}$$

the following stability bound holds.

$$\left\|\varphi_{\mathbf{D}}(\mathbf{x}) - \varphi_{\tilde{\mathbf{D}}}(\mathbf{x})\right\|_2 \leq \frac{2\sqrt{k}\,(1 + 2\|\mathbf{x}\|_2/\lambda)\,\|\mathbf{x}\|_2}{1 - \mu k/\sqrt{d}}\|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2}$$

From Theorem 2, $\epsilon_{k,\mathbf{D}}(\mathbf{x})$ becomes the permissible radius of perturbation in Definition 1.

Remark 2 We mention the relation between the μ-incoherence defined above and k-incoherence of a dictionary, which is the assumption of the sparse coding stability in Mehta and Gray (2013). For $k \in [m]$ and $\mathbf{D} \in \mathcal{D}$, the k-incoherence $s_k(\mathbf{D})$ is defined as

$$s_k(\mathbf{D}) := (\min\{\varsigma_k(\mathbf{D}_\Lambda) | \Lambda \subset [m], |\Lambda| = k\})^2,$$

[3] In general, the (p, q)-induced norm for $p, q \geq 1$ is defined by $\|\mathbf{E}\|_{p,q} := \sup_{\mathbf{v}\in\mathbb{R}^m, \|\mathbf{v}\|_p=1} \|\mathbf{E}\mathbf{v}\|_q$. Then, $\|\cdot\|_{1,2}$ in this general definition coincides with that in Definition 3 by Lemma 17 of Vainsencher et al. (2011).

where $\varsigma_k(\mathbf{D}_\Lambda)$ is the kth singular value of $\mathbf{D}_\Lambda = [\mathbf{d}_{i_1}, \ldots, \mathbf{d}_{i_k}]$ for $\Lambda = \{i_1, \ldots, i_k\}$. From Lemma 9 in "Appendix B", when a dictionary $\mathbf{D}$ is μ-incoherent, the k-incoherence of $\mathbf{D}$ satisfies

$$s_k(\mathbf{D}) \geq 1 - \frac{\mu k}{\sqrt{d}}.$$

Thus, a μ-incoherent dictionary has positive k-incoherence when $d > (\mu k)^2$. On the other hand, when $k \geq 2$, if a dictionary $\mathbf{D}$ has positive k-incoherence $s_k(\mathbf{D})$, there is $0 < \mu < \sqrt{d}$ such that the dictionary is μ-incoherent.[4] However, we note that positive k-incoherence $s_k(\mathbf{D})$ does not imply that $\mathbf{D}$ is μ-incoherent and $\mu < \sqrt{d}/k$ in general.[5]

Here, we refer to the relation with the sparse coding stability (Theorem 4) of Mehta and Gray (2013) in which the difference of dictionaries was measured by $\|\cdot\|_{2,2}$ instead of $\|\cdot\|_{1,2}$ and the permissible radius of perturbation was given by $\mathcal{M}_k(\mathbf{D}, \mathbf{x})^2\lambda$ except for a constant factor. Applying the simple inequality $\|\mathbf{E}\|_{2,2} \leq \sqrt{m}\|\mathbf{E}\|_{1,2}$ for $\mathbf{E} \in \mathbb{R}^{d\times m}$, we can obtain a variant of the sparse coding stability with norm $\|\cdot\|_{1,2}$. However, then the dictionary size m affects the permissible radius of perturbation and the stability bound of sparse coding stability. On the other hand, the factor of m does not appear in Theorem 2, and thus, the result is effective even for a large m. In addition, whereas $\|\mathbf{x}\| \leq 1$ is assumed in Mehta and Gray (2013), Theorem 2 does not assume that $\|\mathbf{x}\| \leq 1$ and clarifies the dependency for the norm $\|\mathbf{x}\|$. The Lipschitz constant L_ψ is obtained independent of $\mathbf{x}$ for a bounded sample space.

In existing studies related to sparse coding, the sparse representation $\varphi_{\mathbf{D}}(\mathbf{x})$ is modified as $\varphi_{\mathbf{D}}(\mathbf{x}) \otimes \mathbf{x}$ (Mairal et al. 2009) or $\varphi_{\mathbf{D}}(\mathbf{x}) \otimes (\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x}))$ (Raina et al. 2007), where $\otimes$ is the tensor product. Owing to the stability of sparse representation (Theorem 2), it can be shown that such modified representations also have local stability.

3.2 Sparse modeling and margin bound

In this section, we assume a sparse structure for samples $\mathbf{x} \in \mathbb{R}^d$ and specify a lower bound for the k-margin used in (11). The result obtained in this section plays an essential role in demonstrating the parameter transfer learnability in Sect. 3.3.

Assumption 1 (*Model*) There exists a dictionary matrix $\mathbf{D}^*$ such that every sample $\mathbf{x}$ is independently generated by a representation $\mathbf{a}$ and noise $\boldsymbol{\xi}$ as

$$\mathbf{x} = \mathbf{D}^*\mathbf{a} + \boldsymbol{\xi}.$$

Moreover, we impose the following three assumptions on the above model.

Assumption 2 (*Dictionary*) The dictionary matrix $\mathbf{D}^* = [\mathbf{d}_1, \ldots, \mathbf{d}_m] \in \mathcal{D}$ is μ-incoherent.

Assumption 3 (*Representation*) The representation $\mathbf{a}$ is a random variable that is k-sparse (i.e., $\|\mathbf{a}\|_0 \leq k$) and the non-zero entries are lower bounded by $C > 0$ (i.e., $a_i \neq 0$ satisfies $|a_i| \geq C$).

[4] Since $s_k(\mathbf{D})^2 = \min_{\mathbf{b}:k\text{-sparse}} \mathbf{b}^\top\mathbf{D}^\top\mathbf{D}\mathbf{b}$, we have $s_2(\mathbf{D}) \geq s_k(\mathbf{D})$ if $k \geq 2$. Moreover, $s_2(\mathbf{D})^2 = 1 - \max_{i\neq j} |\langle \mathbf{d}_i, \mathbf{d}_j\rangle|$ from the direct calculation. Thus, if $s_k(\mathbf{D})$ is positive, $\max_{i\neq j} |\langle \mathbf{d}_i, \mathbf{d}_j\rangle| < 1$. In other words, it holds that $|\langle \mathbf{d}_i, \mathbf{d}_j\rangle| \leq \mu/\sqrt{d}$ for all $i \neq j$ when $\mu := \sqrt{d}\max_{i\neq j} |\langle \mathbf{d}_i, \mathbf{d}_j\rangle| < \sqrt{d}$.

[5] For example, when $d = k = 2$ and $\mathbf{D} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 4/3 \\ 1 & \sqrt{2}/3 \end{bmatrix}$, it holds that $s_k(\mathbf{D}) > 0$. However, there is no μ such that $\mathbf{D}$ is μ-incoherent and $\mu < \sqrt{d}/k$.

Assumption 4 (*Noise*) The noise $\boldsymbol{\xi}$ is independent across coordinates and Gaussian with zero mean and a maximum variance $\sigma^2 k/d$ on each component, where $\sigma > 0$ is a constant.

Remark 3 Note that Assumption 4 is valid under Assumptions 1–3 and the condition $\mu \leq \sqrt{d}/k$ if we assume a situation where dictionary learning is possible. To learn the true dictionary $\mathbf{D}^*$ and true signal $\mathbf{a}$ from a sample $\mathbf{x}$, it is necessary that the noise $\boldsymbol{\xi}$ must be much smaller than the signal $\mathbf{D}^*\mathbf{a}$ with high probability. This condition is represented by $\|\boldsymbol{\xi}\| \leq \|\mathbf{D}^*\mathbf{a}\|$. Here, it holds that

$$\|\boldsymbol{\xi}\|^2 \leq \|\mathbf{D}\mathbf{a}\|^2 \leq |\mathbf{a}|^\top \left(I + \frac{\mu}{\sqrt{d}}\mathbf{1}\right)|\mathbf{a}| \leq a_{max}^2 k\left(1 + \mu\frac{k}{\sqrt{d}}\right) \leq 2a_{max}^2 k, \tag{12}$$

where $|\mathbf{a}|$ is the vector whose components are absolute values of those of $\mathbf{a}$ and $a_{max} := \max_{1\leq i\leq m}|a_i|$. Then, each component ξ_i of $\boldsymbol{\xi}$ approximately satisfies, with high probability,

$$|\xi_i|^2 \simeq \frac{\|\boldsymbol{\xi}\|^2}{d} = \tilde{\mathcal{O}}(k/d). \tag{13}$$

Thus, since each component is Gaussian, its variance should be $\tilde{\mathcal{O}}(k/d)$.

In transfer learning, samples on the source and target domains are not necessarily identically distributed. Indeed, independent but non-identical distributions are allowed under Assumptions 3 and 4. This is essential because samples in the source and target domains cannot be assumed to be identically distributed in transfer learning.

Theorem 3 (Margin Bound) *Let* $0 < t < 1$. *We set*

$$\begin{aligned}\delta_{t,\lambda} := {} & \frac{2\sigma\sqrt{k}m}{(1-t)\sqrt{d}\lambda}\exp\left(-\frac{(1-t)^2 d\lambda^2}{8\sigma^2 k}\right) + \frac{2\sigma\sqrt{k}m}{\sqrt{d}\lambda}\exp\left(-\frac{d\lambda^2}{8\sigma^2 k}\right) \\ & + \frac{4\sigma k^{3/2}}{C\sqrt{d(1-\mu k/\sqrt{d})}}\exp\left(-\frac{C^2 d(1-\mu k/\sqrt{d})}{8\sigma^2 k}\right) \\ & + \frac{8\sigma\sqrt{k}(d-k)}{\sqrt{d}\lambda}\exp\left(-\frac{d\lambda^2}{32\sigma^2 k}\right).\end{aligned} \tag{14}$$

We assume that $d \geq \left\{\left(1 + \frac{6}{(1-t)}\right)\mu k\right\}^2$ *and* $\lambda = d^{-\tau}$ *for arbitrary* $1/4 \leq \tau \leq 1/2$. *Under Assumptions 1–4, the following inequality holds with a probability of at least* $1 - \delta_{t,\lambda}$.

$$\mathcal{M}_k(\mathbf{D}^*, \mathbf{x}) \geq t\lambda \tag{15}$$

We provide the proof of Theorem 3 in "Appendix C".

Note that the failure probability of the margin bound in (14) decreases as the dimension increases since the variance of the noise gets smaller because of Assumption 4.

Next, we analyze the regularization parameter λ. An appropriate reflection of the sparsity of samples requires the regularization parameter λ to be set suitably. This is according to Theorem 4 of Zhao and Yu (2006)[6] when samples follow the sparse model as in Assumptions 1–4 and $\lambda \cong d^{-\tau}$ for $1/4 \leq \tau \leq 1/2$. The representation $\varphi_{\mathbf{D}}(\mathbf{x})$ reconstructs the true

[6] Theorem 4 of Zhao and Yu (2006), which was stated for Gaussian noise. However, it can be easily generalized to sub-Gaussian noise. Our setting corresponds to the case in which $c_1 = 1/2, c_2 = 1, c_3 = (\log\kappa + \log\log d)/\log d$ for some $\kappa > 1$ (i.e., $e^{d^{c_3}} \cong d^\kappa$) and $c_4 = c$ in Theorem 4 of Zhao and Yu (2006). Note that our regularization parameter λ corresponds to λ_d/d in Zhao and Yu (2006).

sparse representation $\mathbf{a}$ of sample $\mathbf{x}$ with a small error. In particular, when $\tau = 1/4$ (i.e., $\lambda \cong d^{-1/4}$) in Theorem 3, the failure probability $\delta_{t,\lambda} \cong e^{-\sqrt{d}}$ on the margin is guaranteed to become sub-exponentially small with respect to dimension d and is negligible for the high-dimensional case. On the other hand, the typical choice $\tau = 1/2$ (i.e., $\lambda \cong d^{-1/2}$) does not provide a useful result because $\delta_{t,\lambda}$ is not small at all.

3.3 Parameter transfer learnability for dictionary learning

When a true dictionary $\mathbf{D}^*$ exists as in Assumption 1, we show that the output $\widehat{\mathbf{D}}_N$ of a suitable dictionary learning algorithm from N-unlabeled samples satisfies the parameter transfer learnability for the sparse coding $\varphi_{\mathbf{D}}$. Then, Theorem 1 guarantees the excess risk bound in self-taught learning since the discussion in this section does not assume the label space in the source domain. This situation corresponds to the case where $\boldsymbol{\theta}^*_{\mathcal{S}} = \mathbf{D}^*, \widehat{\boldsymbol{\theta}}_N = \widehat{\mathbf{D}}_N$ and $\|\cdot\| = \|\cdot\|_{1,2}$ in Sect. 2.1.

We show that an appropriate dictionary learning algorithm satisfies parameter transfer learnability for the sparse coding $\varphi_{\mathbf{D}}$ by focusing on the permissible radius of perturbation in (11) under some assumptions. When Assumptions 1–4 hold and $\lambda = d^{-\tau}$ for $1/4 \le \tau \le 1/2$, the margin bound (15) for $\mathbf{x} \in \mathcal{X}$ holds with probability $1 - \delta_{t,\lambda}$, and we have

$$\epsilon_{k,\mathbf{D}^*}(\mathbf{x}) \;\ge\; \frac{t^2\lambda^3}{64\max\{1, \|\mathbf{x}\|\}^4} \;=\; \Theta(d^{-3\tau}).$$

Thus, if a dictionary learning algorithm outputs the estimator $\widehat{\mathbf{D}}_N$ such that

$$\|\widehat{\mathbf{D}}_N - \mathbf{D}^*\|_{1,2} \;=\; \mathcal{O}(d^{-3\tau}) \tag{16}$$

with probability $1 - \delta_N$, the estimator $\widehat{\mathbf{D}}_N$ of $\mathbf{D}^*$ satisfies the parameter transfer learnability for the sparse coding $\varphi_{\mathbf{D}}$ with probability $\bar{\delta}_N := \delta_N + \delta_{t,\lambda}$. Then, by local stability of sparse representation and parameter transfer learnability of such a dictionary learning, Theorem 1 guarantees that sparse coding in self-taught learning satisfies the excess risk bound in (4). For n-samples $\mathbf{X}^n = \{\mathbf{x}_1, \ldots \mathbf{x}_n\}$ in the target domain, detailed analysis reveals that the inequality $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\| \le \epsilon_{\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{X}^n)$ holds with probability $1-(\delta_N + n\delta_{t,\lambda})$, which is sharper than $1 - n\bar{\delta}_N$.

Theorem 1 applies to any dictionary learning algorithm as long as (16) is satisfied. For example, from Theorem 12 in Arora et al. (2015), when some conditions[7] including Assumptions 1–4 are assumed, there is an iterative algorithm [Algorithm 5 in Arora et al. (2015)] whose output $\mathbf{D}^s$ at iteration s satisfies

$$\|\mathbf{D}^s - \mathbf{D}^*\|_{1,2}^2 \;\le\; \gamma^s \|\mathbf{D}^0 - \mathbf{D}^*\|_{1,2}^2 + \mathcal{O}(d^{-2}) \tag{17}$$

for some $1/2 < \gamma < 1$. When $s \ge C \log d$ for a large constant C and dimension d is large enough, it holds that

$$\|\mathbf{D}^s - \mathbf{D}^*\|_{1,2} \;=\; \mathcal{O}(d^{-1}).$$

We note that the algorithm requires infinite number of samples at each iteration. However, modifying Appendix G of Arora et al. (2015), it is expected that there is a large constant C' and an alternative stochastic algorithm whose output $\widehat{\mathbf{D}}^s \cong \mathbf{D}^s$ at each iteration $s \ge C' \log d$ satisfies (16) for $1/4 < \tau < 1/3$.

Note that, although we imposed the hard-sparsity assumption (Assumption 3) as in Arora et al. (2015), we focused on the LASSO-based encoder $\varphi_{\mathbf{D}}$ instead of the hard-sparsity encoder

[7] See the page 4 in Arora et al. (2015).

treated in Arora et al. (2015). Under the hard-sparsity assumption, we could derive the lower bound of the permissible radius of perturbation $\epsilon_{k,\mathbf{D}}$ about the LASSO-based encoder and use the result about the estimation error in dictionary learning in Arora et al. (2015).

4 Numerical experiments

We report numerical experiments using US postal service (USPS) and MNIST handwritten digits datasets and compare the results with our theoretical conclusions. Especially, we intensively investigate the relationship among N, n, $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|$, ρ, and the prediction performance of the transfer learning.

The USPS dataset is composed of $d = 256$ dimensional 7291 training images and 2007 test images, and each element of data vectors ranges from -1 to 1. The MNIST data set has 60,000 training images and 10,000 test images of dimension $d = 784$, and each element of the data vectors range from 0 to 255. In numerical experiments, the MNIST data is scaled such that each element takes a value in the interval $[0, 1]$. In both datasets, each image with the ℓ_∞-norm 1 has a label in $\{0, 1, \ldots, 9\}$.

4.1 Prediction accuracy of learning with sparse representation

Let us describe the setup of numerical experiments in which the self-taught learning was applied to the USPS dataset. N images out of USPS training images were randomly chosen as training samples on the source domain. In the experiments, N was set to 3000. The data matrix of the source domain was represented by $\mathbf{X}_{\mathcal{S}} = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_N) \in \mathbb{R}^{d\times N}$.

The dictionary $\widehat{\mathbf{D}} \in \mathbb{R}^{d\times m}$ was obtained by solving the problem

$$\min_{\mathbf{D},\mathbf{Z}} \frac{1}{2}\|\mathbf{X}_{\mathcal{S}} - \mathbf{D}\mathbf{Z}\|_2^2 + \lambda\|\mathbf{Z}\|_1, \quad \text{s.t. } \mathbf{D} \in \mathcal{D},\ \mathbf{Z} \in \mathbb{R}^{m\times N},$$

where $\|\mathbf{A}\|_p$ was $(\sum_{i,j} |a_{ij}|^p)^{1/p}$ for the matrix $\mathbf{A} = (a_{ij})$. The feature map was defined by the sparse representation of $\mathbf{x} \in \mathbb{R}^d$ in (10). The regularization parameter λ for the sparse representation was set to $\lambda = 1$. The dimension m of the dictionary varied from 16 to 512. The problem on the target domain was a 10-class digits classification of the image data. In experiments, n images were randomly chosen out of the remaining 4291 USPS training images. Here, n was varied from 500 to 3000. The prediction accuracy of the classifier was evaluated using all test images. The sparse representation of the training images, $\{(\varphi_{\widehat{\mathbf{D}}}(\mathbf{x}_i),\ y_i) : i = 1, \ldots, n\}$, was used to train the linear SVM (Huang and Aviyente 2006; Yang et al. 2009). The Lasso-type sparse representation (10) was employed, since it is quite popular in the dictionary learning Zhang et al. (2015). In addition, a computationally efficient implementation of the dictionary learning called `spams` is available as an R package (Mairal et al. 2009). The classifier was provided by `kernlab` package of `R` language (Karatzoglou et al. 2004; R Core Team 2016), and a one-vs-one strategy was used to deal with multi-class classification problems.

In addition, we evaluated the influence of the *dictionary shift* by analyzing how the estimation error on the source domain affected learning results on the target domain. A shifted dictionary of $\widehat{\mathbf{D}}$ was obtained by adding a random matrix $\mathbf{M}$ to $\widehat{\mathbf{D}}$. In experiments, each element of $\mathbf{M}$ was assumed to be an i.i.d. copy of Gaussian noise with mean 0 and standard deviation σ. Each column of the perturbed matrix $\widehat{\mathbf{D}}+\mathbf{M}$ was normalized to obtain the shifted dictionary $\widetilde{\mathbf{D}} \in \mathcal{D}$. The feature map $\varphi_{\widetilde{\mathbf{D}}}(\mathbf{x})$ with the shifted dictionary was used to obtain the

sparse representation. Numerical experiments were conducted to reveal the relation among the test error, regularization parameter ρ, and noise level σ.

For the MNIST dataset, N was set to 30,000 and n was varied from 500 to 10,000. The dimension of the dictionary, m, was varied from 16 to 1568. All test images were used to evaluate the test error on the target domain. Furthermore, the effect of the dictionary shift was evaluated.

The results are shown in Figs. 1 and 2. The test error on the target domain is plotted in the left column of each figure. Furthermore, the test error of the linear SVM using only samples on the target domain are reported. The regularization parameter ρ in the linear SVM was set to an optimal value that achieved the smallest test error. In the right column, the optimal ρ of learning with sparse representation is shown as a function of the sample size n.

When large dictionaries were used, we found that the test error of the SVM using sparse representation was smaller than that got by implementing standard SVM that used only target samples without the sparse representation. Hence, samples on the source domain were effectively used to improve the prediction accuracy.

Furthermore, we investigated the usefulness of samples on the source domain by comparing with a variant of supervised dictionary learning (SDL) proposed by Mairal et al. (2009). In the common SDL, the dictionary and classifier are simultaneously optimized based on only samples from the target domain. In the experiments, we employed a simple variant of the SDL to reduce the computational cost. In the simplified SDL, the dictionary $\mathbf{D}$ and feature $\mathbf{Z}$ were obtained using the data matrix of the target domain instead of the source domain. Then, the sparse feature representation, $\mathbf{Z}$, was fed into the linear SVM as training input with the output label.

Table 1 shows the result. The column "source and target" shows the test error of transfer learning using information on the source and target domains. On the other hand, the column "target (SDL)" shows the results of simplified SDL using only samples on the target domain. For the MNIST dataset, the SDL with $m = 1568$ was dropped, because the computational cost of training $\mathbf{D}$ in each iteration was too high. Overall, learning using both the source and target domains performed better than the simplified SDL, especially for small n. Therefore, transfer learning using $\mathbf{D}$ trained on the source domain is expected to be practically useful.

4.2 Setting of regularization parameters

Next, we investigate the relationship between the dictionary shift on the source domain and the regularization parameter ρ on the target domain. For $m = 16$ of USPS data in Fig. 1, a large optimal regularization parameter ρ was required to deal with the large noise level σ. Theoretical analysis in Sect. 2.2 showed that a large regularization parameter was needed to suppress the large perturbation of the feature map. Hence, numerical results for small m agreed with our theoretical results. On the contrary, when $m = 512$, the small optimal regularization parameter efficiently worked to suppress the large noise level σ. The same tendency was observed in the result on MNIST dataset in Fig. 2. In summary, when the dictionary is small, a larger regularization parameter is required to deal with larger noise level. For a large dictionary, the opposite is true.

Next, we explain the relation between the noise level σ and regularization parameter ρ. Theorem 2 shows the relation between the sensitivity of the sparse representation and the dictionary shift. The upper bound of the sensitivity depends mainly on the degree of incoherence and amount of dictionary shift. Let $\mu_{\mathbf{D}}$ be the degree of incoherence for dictionary $\mathbf{D}$; see Definition 5. For $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_m]$, $\mu_{\mathbf{D}}$ is computed as $\sqrt{d} \max\{|\langle \mathbf{d}_i, \mathbf{d}_j \rangle| : i \neq j\}$.

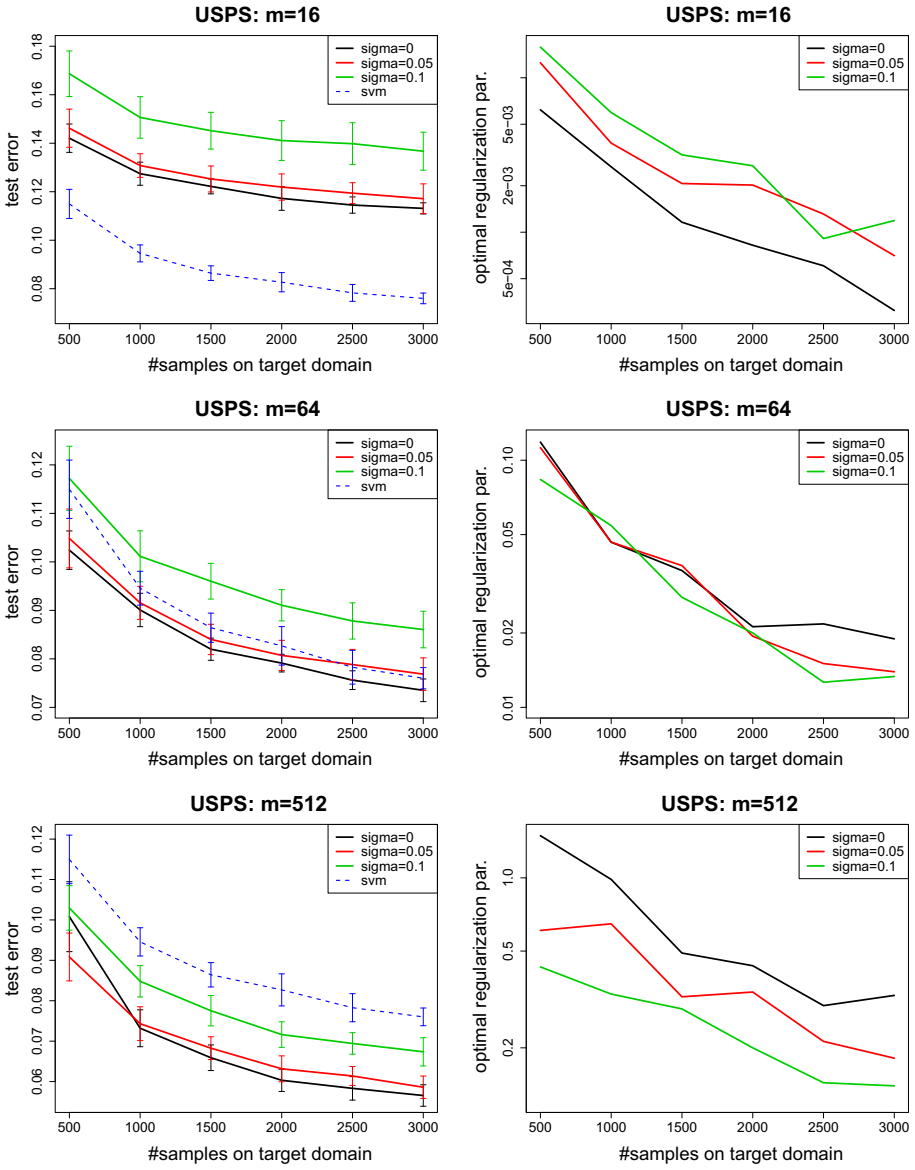

Fig. 1 Plot of test errors (left column) and optimal regularization parameters (right column) on USPS dataset. The dimension of dictionary $\mathbf{D} \in \mathbb{R}^{d \times m}$ with $d = 256$ was set to $m = 16$, 64, and 512 and the noise level, σ, was varied from 0 to 0.1. Curves "svm" in the left column show test error of SVM using only target samples

Then, the difference $\|\varphi_{\widehat{\mathbf{D}}}(\mathbf{x}) - \varphi_{\widetilde{\mathbf{D}}}(\mathbf{x})\|_2$ is bounded above by $\|\widehat{\mathbf{D}} - \widetilde{\mathbf{D}}\|_{1,2}/(1 - \mu_{\widetilde{\mathbf{D}}}/\sqrt{d})$ up to a positive factor depending on $\mathbf{x}$ when the 1-margin is used. We numerically confirmed that the upper bound using $\mu_{\widetilde{\mathbf{D}}}$ is tighter than the upper bound using $\mu_{\widehat{\mathbf{D}}}$. This is because $\widehat{\mathbf{D}}$ includes some similar column vectors and the random noise relaxes such similarity. As shown in the proof of Theorem 1, the shift of the feature map $\|\varphi_{\widehat{\mathbf{D}}}(\mathbf{x}) - \varphi_{\widetilde{\mathbf{D}}}(\mathbf{x})\|_2$ directly affects the upper bound of the generalization ability. Figure 3 depicts the average value of

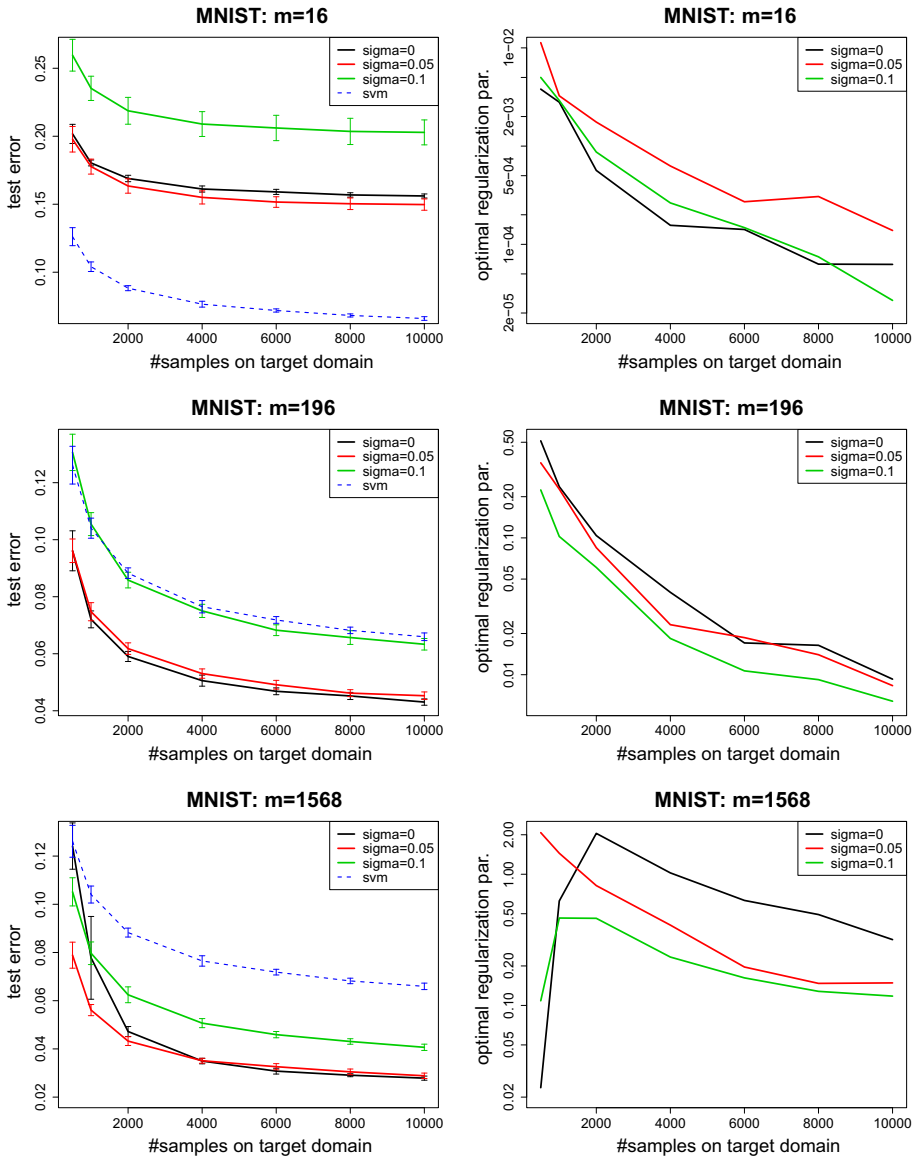

Fig. 2 Plot of test errors (left column) and optimal regularization parameters (right column) on MNIST dataset. The dimension of dictionary $\mathbf{D} \in \mathbb{R}^{d\times m}$ with $d = 784$ was set to $m = 16$, 196, and 1568, and the noise level, σ, was varied from 0 to 0.1. Curves "svm" in the left column show test error of SVM using only target samples

$\|\widehat{\mathbf{D}} - \widetilde{\mathbf{D}}\|_{1,2}/(1 - \mu_{\widetilde{\mathbf{D}}}/\sqrt{d})$ over 20 different random matrices as a function of σ. Generally, larger σ leads to larger $\|\widehat{\mathbf{D}} - \widetilde{\mathbf{D}}\|_{1,2}$ and smaller $\mu_{\widetilde{\mathbf{D}}}$. When the size of the dictionary, m, is small, the effect of $\|\widehat{\mathbf{D}} - \widetilde{\mathbf{D}}\|_{1,2}$ dominates that of $\mu_{\widetilde{\mathbf{D}}}$. On the contrary, for a large dictionary, the decrease of $\mu_{\widetilde{\mathbf{D}}}$ dominates the increase of $\|\widehat{\mathbf{D}} - \widetilde{\mathbf{D}}\|_{1,2}$. As a result, the upper bound of $\|\varphi_{\widehat{\mathbf{D}}}(\mathbf{x}) - \varphi_{\widetilde{\mathbf{D}}}(\mathbf{x})\|_2$ becomes small for large noise level.

Table 1 Test errors and standard deviation of transfer learning in percent. The dictionary $\mathbf{D}$ and feature $\mathbf{Z}$ are trained using samples on "source and target" domain or only "target" domain. In the latter method, samples on the target domain are used to learn the dictionary

Data	Source and target			Target (SDL)		
m	16	64	512	16	64	512
USPS data						
$n = 500$	14.2 ± 0.58	10.2 ± 0.40	10.1 ± 0.86	15.7 ± 0.60	12.1 ± 0.68	10.8 ± 0.72
$n = 1000$	12.7 ± 0.48	9.0 ± 0.34	7.3 ± 0.46	13.7 ± 0.68	9.9 ± 0.61	7.7 ± 0.50
$n = 3000$	11.3 ± 0.23	7.4 ± 0.23	5.7 ± 0.27	11.9 ± 0.58	7.7 ± 0.33	5.9 ± 0.29

Data	Source and target		Target (SDL)	
m	16	196	16	196
MNIST data				
$n = 1000$	18.0 ± 0.22	7.2 ± 0.30	18.7 ± 0.75	9.0 ± 0.44
$n = 2000$	16.9 ± 0.24	5.9 ± 0.18	17.2 ± 0.75	7.0 ± 0.27
$n = 10{,}000$	15.6 ± 0.15	4.3 ± 0.11	15.6 ± 0.38	4.6 ± 0.17

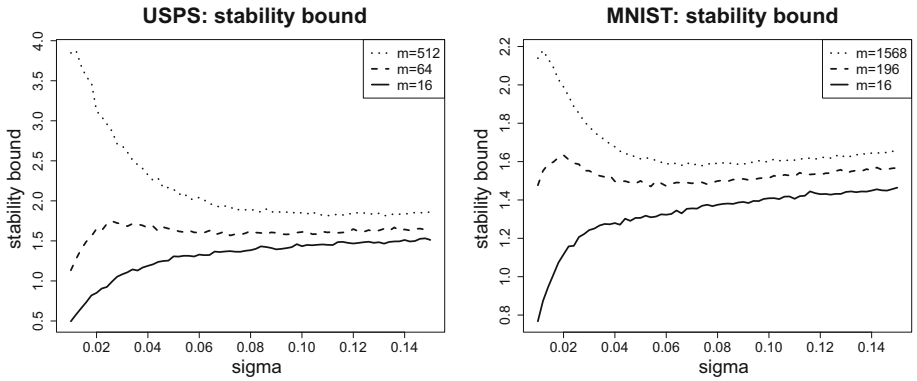

Fig. 3 Plot of average of upper bound, $\|\widehat{\mathbf{D}} - \widetilde{\mathbf{D}}\|_{1,2}/(1 - \mu_{\widetilde{\mathbf{D}}}/\sqrt{d})$, for USPS and MNIST datasets. Horizontal axis denotes the noise level σ. The size of the dictionary was varied from 16 to 512 for USPS and from 16 to 1568 for MNIST

Based on the above consideration, the numerical results in Figs. 1 and 2 can be interpreted as follows. Let $\mathbf{D}^*$ be the *true* dictionary. The difference $\|\varphi_{\mathbf{D}^*}(\mathbf{x}) - \varphi_{\widetilde{\mathbf{D}}}(\mathbf{x})\|_2$ will affect the test error and optimal regularization parameter. Let us consider the upper bound of the difference, $\|\varphi_{\mathbf{D}^*}(\mathbf{x}) - \varphi_{\widehat{\mathbf{D}}}(\mathbf{x})\|_2 + \|\varphi_{\widehat{\mathbf{D}}}(\mathbf{x}) - \varphi_{\widetilde{\mathbf{D}}}(\mathbf{x})\|_2$. In numerical experiments, the first term $\|\varphi_{\mathbf{D}^*}(\mathbf{x}) - \varphi_{\widehat{\mathbf{D}}}(\mathbf{x})\|_2$ is fixed and the second term $\|\varphi_{\widehat{\mathbf{D}}}(\mathbf{x}) - \varphi_{\widetilde{\mathbf{D}}}(\mathbf{x})\|_2$ affects the learning results. As shown above, the effect of the dictionary shift presented in Figs. 1 and 2 agree with our theoretical findings.

5 Conclusion

We derived an excess risk bound (Theorem 1) for a parameter transfer learning problem based on local stability and parameter transfer learnability, which were newly introduced in this

paper. By applying the proposed model to a sparse coding-based algorithm under a sparse model (Assumptions 1–4), we obtained the first theoretical guarantee of excess risk bound in self-taught learning. Numerical experiments in Sect. 4 showed that transfer learning with appropriate regularization parameter worked efficiently to achieve high prediction accuracy on the target domain. Moreover, we confirmed that the theoretical analysis of local stability for sparse coding was useful for understanding the relationship between size of the dictionary and regularization parameter and prediction accuracy.

The framework of parameter transfer learning included not only sparse coding, but also other promising algorithms such as multiple kernel learning and deep neural networks. Our results are expected to be effective in analyzing the theoretical performance of these algorithms. Finally, we noted that our excess risk bound could be applied to applications other than self-taught learning because Theorem 1 included a case in which labeled samples were available in the source domain.

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers 16K00044, 15H03636, 15H01678, 17K12653.

A Fast rate in hypothesis transfer learning

We have been studied the two-step learning algorithm, where first an effective parameter $\boldsymbol{\theta}^*_{\mathcal{S}}$ is learnt in the source domain, and then, the optimal parameter $\mathbf{w}^*_{\mathcal{T}}$ based on an estimator of $\boldsymbol{\theta}^*_{\mathcal{S}}$ is learnt in the target domain. Here, we consider hypothesis transfer learning in Kuzborskij and Orabona (2017) as a special case of the setting introduced in Sect. 2.1, and show that the expected risk of the two-step algorithm has a fast rate when the number of samples in the target domain is not large.

Let the sample space $\mathcal{X} \subset \mathbb{R}^d$ and the label space $\mathcal{Y} = [-1, 1]$ be common in the source and target domains. In addition, let $r : \mathbb{R}^d \to \mathbb{R}$ and $r' : \mathbb{R}^K \to \mathbb{R}$ be non-negative 1-strongly convex functions and satisfy $r(\mathbf{0}) = 0$ and $r'(\mathbf{0}) = 0$, respectively. We set $\mathcal{W} := \left\{\mathbf{w} \in \mathbb{R}^d | r(\mathbf{w}) \leq D\right\}$ and $\Theta := \left\{\boldsymbol{\theta} \in \mathbb{R}^K | r'(\boldsymbol{\theta}) \leq B\right\}$. Kuzborskij and Orabona (2017), it is supposed that there exist finite hypotheses $\{h_k : \mathcal{X} \to \mathcal{Y}\}_{k=1}^K$ such that the source hypothesis is represented by

$$h_{\mathcal{S},\boldsymbol{\theta}}(\mathbf{x}) := \sum_{k=1}^{K} \theta_k h_k(\mathbf{x}) \tag{18}$$

and the target hypothesis is represented by

$$h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}}(\mathbf{x}) := \langle \mathbf{w}, \mathbf{x} \rangle + h_{\mathcal{S},\boldsymbol{\theta}}(\mathbf{x}), \tag{19}$$

for $\boldsymbol{\theta} \in \Theta$ and $\mathbf{w} \in \mathcal{W}$.

Here, we set the feature mapping

$$\psi_{\boldsymbol{\theta}}(\mathbf{x}) := \begin{pmatrix} \mathbf{x} \\ h_{\mathcal{S},\boldsymbol{\theta}}(\mathbf{x}) \end{pmatrix}$$

and the set

$$\mathcal{W}_{\mathcal{T}} := \left\{ \begin{pmatrix} \mathbf{w} \\ 1 \end{pmatrix} \middle| \mathbf{w} \in \mathcal{W} \right\}.$$

Then, identifying $\mathbf{w} \in \mathcal{W}$ with $(\mathbf{w}^\top, 1)^\top \in \mathcal{W}_{\mathcal{T}}$, the hypothesis in (1) coincides with that in (19). In this sense, the formulation in Kuzborskij and Orabona (2017) is covered by ours.

In the above setting, let us consider the excess risk:

$$\begin{aligned} \mathcal{R}_{\mathrm{excess}} &:= \mathcal{R}_{\mathcal{T}}(\widehat{\boldsymbol{\theta}}_N, \widehat{\mathbf{w}}_{N,n}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}^*_{\mathcal{T}}) \\ &= (\mathcal{R}_{\mathcal{T}}(\widehat{\boldsymbol{\theta}}_N, \widehat{\mathbf{w}}_{N,n}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}^*_{\mathcal{S}}, \widehat{\mathbf{w}}_{N,n})) \\ &\quad + (\mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}^*_{\mathcal{S}}, \widehat{\mathbf{w}}_{N,n}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}^*_{\mathcal{T}})), \end{aligned} \tag{20}$$

where

$$\boldsymbol{\theta}^*_{\mathcal{S}} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_{\mathcal{S}}(h_{\mathcal{S},\boldsymbol{\theta}}) \tag{21}$$

and $\mathbf{w}^*_{\mathcal{T}}$ was defined in (2). The first term in (20) is bounded above as follows:

$$\begin{aligned} \mathcal{R}_{\mathcal{T}}(\widehat{\boldsymbol{\theta}}_N, \widehat{\mathbf{w}}_{N,n}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta}^*_{\mathcal{S}}, \widehat{\mathbf{w}}_{N,n}) &\le L_\ell D \mathbb{E}[|h_{\mathcal{S},\widehat{\boldsymbol{\theta}}_N}(\mathbf{x}) - h_{\mathcal{S},\boldsymbol{\theta}^*_{\mathcal{S}}}(\mathbf{x})|] \\ &= L_\ell D \mathbb{E}[|\langle \widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}, \boldsymbol{h}(\mathbf{x}) \rangle|] \\ &\le L_\ell D \|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|_1 \mathbb{E}[\|\boldsymbol{h}(\mathbf{x})\|_\infty] \\ &\le L_\ell D \|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|_1, \end{aligned}$$

where $\boldsymbol{h}(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_K(\mathbf{x})]^\top$, where we used $\|\boldsymbol{h}(\mathbf{x})\|_\infty \le 1$ because of $h_k(\mathbf{x}) \in \mathcal{Y} = [-1, 1]$. The second term in (20) can be evaluated by the following theorem.

Theorem 4 [Theorem 2 of Kuzborskij and Orabona (2017)] *Let* $\mathcal{R}_{src} := \mathcal{R}_{\mathcal{T}}\left(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{0}\right)$. *When the loss function* ℓ *is* κ*-smooth, it holds that*

$$\begin{aligned} &\mathcal{R}_{\mathcal{T}}\left(\boldsymbol{\theta}^*_{\mathcal{S}}, \widehat{\mathbf{w}}_{N,n}\right) - \min_{r(\mathbf{w}) \le B} \widehat{\mathcal{R}}_{\mathcal{T},n}\left(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}\right) \\ &\quad = O\left(C_n \max\left\{ \mathcal{R}_{src}, \frac{1}{n} \right\}^{1/8} \frac{\mathcal{R}_{src}^{1/8}}{n^{1/4}} + \frac{1}{n} \right), \end{aligned} \tag{22}$$

where

$$C_n := \sqrt{\kappa B}\left(\mathcal{R}_{src}^{1/4} + B^{1/4}\right) + \kappa^{1/4}\sqrt{B}\left(\mathcal{R}_{src}^{1/8} + B^{1/8}\right) + \frac{\mathcal{R}_{src}^{3/8}}{n^{1/8}}$$

which is positive and has a constant order for a small $\mathcal{R}_{src}$.

We assume that $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|_1 = O(1/n)$ holds.[8] Then, from the above discussion, as long as the number of samples in the target domain satisfies $n = O(1/\mathcal{R}_{src}^{1/5})$, the excess risk has the following order:

$$\mathcal{R}_{\mathrm{excess}} = O(1/n).$$

In other words, when the estimation error in the source domain is small enough and the number of samples n in the target domain is not so large, the excess risk decreases in the order of $O(1/n)$, which is faster than the conventional asymptotic order $O(1/\sqrt{n})$ in a non-transfer setting.

[8] When the number of samples N is large enough, this condition typically holds for an appropriate estimator $\widehat{\boldsymbol{\theta}}_N$.

B Proof of sparse coding stability

The proof of Theorem 2 is almost the same as that of Theorem 1 in Mehta and Gray (2012). However, since a part of the proof cannot be applied to our setting, we provide the full proof of Theorem 2 in this section.

Lemma 1 *Let* $\mathbf{a} \in \mathbb{R}^m$ *and* $\mathbf{E} \in \mathbb{R}^{d \times m}$. *Then,* $\|\mathbf{E}\mathbf{a}\|_2 \leq \|\mathbf{E}\|_{1,2}\|\mathbf{a}\|_1$.

Proof

$$\|\mathbf{E}\mathbf{a}\|_2 = \|\sum_{i=1}^{m} a_i \mathbf{e}_i\|_2 \leq \sum_{i=1}^{m} |a_i| \|\mathbf{e}_i\|_2 \leq \|\mathbf{E}\|_{1,2} \sum_{i=1}^{m} |a_i| = \|\mathbf{E}\|_{1,2}\|\mathbf{a}\|_1.$$

□

Lemma 2 *The sparse representation* $\varphi_{\mathbf{D}}(\mathbf{x})$ *satisfies* $\|\varphi_{\mathbf{D}}(\mathbf{x})\|_1 \leq \frac{\|\mathbf{x}\|_2^2}{2\lambda}$.

Proof

$$\begin{aligned}
\lambda \|\varphi_{\mathbf{D}}(\mathbf{x})\|_1 &\leq \frac{1}{2}\|\mathbf{x} - \mathbf{D}\varphi_{\mathbf{D}}(\mathbf{x})\|_2^2 + \lambda \|\varphi_{\mathbf{D}}(\mathbf{x})\|_1 \\
&= \min_{z \in \mathbb{R}^m} \frac{1}{2}\|\mathbf{x} - \mathbf{D}z\|_2^2 + \lambda \|z\|_1 \\
&\leq \frac{1}{2}\|\mathbf{x}\|_2^2
\end{aligned}$$

□

We prepare the following notation:

$$v_{\mathbf{D}}(\mathbf{z}) := \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1.$$

Let $\mathbf{a}^*$ and $\tilde{\mathbf{a}}^*$ denote the solutions to the LASSO problems for the dictionary $\mathbf{D}$ and $\tilde{\mathbf{D}}$, respectively.

$$\mathbf{a}^* := \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^m} v_{\mathbf{D}}(\mathbf{z}) \quad \tilde{\mathbf{a}}^* := \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^m} v_{\tilde{\mathbf{D}}}(\mathbf{z})$$

Then, the following equation holds owing to the subgradient of $v_{\mathbf{D}}(\mathbf{z})$ with respect to $\mathbf{z}$ [e.g. (2.8) of Osborne et al. (2000)].

Lemma 3

$$\lambda\|\mathbf{a}^*\|_1 = \langle \mathbf{x} - \mathbf{D}\mathbf{a}^*, \mathbf{D}\mathbf{a}^* \rangle$$

Let $v_{\mathbf{D}}$ and $v_{\tilde{\mathbf{D}}}$ be the optimal values of the LASSO problems for dictionary $\mathbf{D}$ and $\tilde{\mathbf{D}}$.

$$\begin{aligned}
v_{\mathbf{D}} &:= \min_{\mathbf{z} \in \mathbb{R}^m} v_{\mathbf{D}}(\mathbf{z}) = \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{a}^*\|_2^2 + \lambda\|\mathbf{a}^*\|_1, \\
v_{\tilde{\mathbf{D}}} &:= \min_{\mathbf{z} \in \mathbb{R}^m} v_{\tilde{\mathbf{D}}}(\mathbf{z}) = \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^*\|_2^2 + \lambda\|\tilde{\mathbf{a}}^*\|_1
\end{aligned}$$

Lemma 4 (Optimal Value Stability) *If* $\|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2} \leq \lambda$, *then*

$$|v_{\mathbf{D}} - v_{\tilde{\mathbf{D}}}| \leq \frac{1}{2}\left(1 + \frac{\|\mathbf{x}\|_2}{4}\right)\|\mathbf{x}\|_2^3 \frac{\|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2}}{\lambda}.$$

Proof

$$
\begin{aligned}
v_{\tilde{\mathbf{D}}} &\leq \frac{1}{2}\|\mathbf{x}-\tilde{\mathbf{D}}\mathbf{a}^*\|_2^2+\lambda\|\mathbf{a}^*\|_1 \\
&= \frac{1}{2}\|\mathbf{x}-\mathbf{D}\mathbf{a}^*+(\mathbf{D}-\tilde{\mathbf{D}})\mathbf{a}^*\|_2^2+\lambda\|\mathbf{a}^*\|_1 \\
&\leq \frac{1}{2}(\|\mathbf{x}-\mathbf{D}\mathbf{a}^*\|_2^2+2\|\mathbf{x}-\mathbf{D}\mathbf{a}^*\|_2\|(\mathbf{D}-\tilde{\mathbf{D}})\mathbf{a}^*\|_2+\|(\mathbf{D}-\tilde{\mathbf{D}})\mathbf{a}^*\|_2^2)+\lambda\|\mathbf{a}^*\|_1 \\
&\leq \frac{1}{2}\|\mathbf{x}-\mathbf{D}\mathbf{a}^*\|_2^2+\lambda\|\mathbf{a}^*\|_1+\|\mathbf{x}\|_2\left(\frac{\|\mathbf{x}\|_2^2\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{2\lambda}\right)+\frac{1}{2}\left(\frac{\|\mathbf{x}\|_2^2\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{2\lambda}\right)^2 \\
&\leq v_{\mathbf{D}}+\left(1+\frac{\|\mathbf{x}\|_2}{4}\right)\frac{\|\mathbf{x}\|_2^3}{2\lambda}\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2},
\end{aligned}
$$

where we use

$$
\|\mathbf{x}-\mathbf{D}\mathbf{a}^*\|_2=\sqrt{\|\mathbf{x}-\mathbf{D}\mathbf{a}^*\|^2}\leq\sqrt{\|\mathbf{x}-\mathbf{D}\mathbf{a}^*\|^2+\lambda\|\mathbf{a}^*\|_1}\leq\sqrt{\|\mathbf{x}\|_2^2}=\|\mathbf{x}\|_2.
$$

□

The following Lemma 5 is obtained by the proof of Lemma 11 in Mehta and Gray (2012).

Lemma 5 (Stability of Norm of Reconstructor) *If* $\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}\leq\lambda$, *then*

$$
\left|\|\mathbf{D}\mathbf{a}^*\|_2^2-\|\tilde{\mathbf{D}}\tilde{\mathbf{a}}^*\|_2^2\right| \;=\; 2|v_{\mathbf{D}}-v_{\tilde{\mathbf{D}}}| \;\leq\; \left(1+\frac{\|\mathbf{x}\|_2}{4}\right)\|\mathbf{x}\|_2^3\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{\lambda}.
$$

Lemma 6 *If* $\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}\leq\lambda$, *then*

$$
\left|\|\mathbf{D}\mathbf{a}^*\|_2^2-\|\mathbf{D}\tilde{\mathbf{a}}^*\|_2^2\right|\leq(\|\mathbf{x}\|_2+3)\,\|\mathbf{x}\|_2^3\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{\lambda}.
$$

Proof First, note that

$$
\|(\tilde{\mathbf{D}}-\mathbf{D})\tilde{\mathbf{a}}^*\|_2\leq\|(\tilde{\mathbf{D}}-\mathbf{D})\|_{1,2}\|\tilde{\mathbf{a}}^*\|_1\leq\|\mathbf{x}\|_2^2\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{2\lambda}
$$

and

$$
\begin{aligned}
\|\mathbf{D}\tilde{\mathbf{a}}^*\|_2 &\leq \|(\mathbf{D}-\tilde{\mathbf{D}})\tilde{\mathbf{a}}^*\|_2+\|\tilde{\mathbf{D}}\tilde{\mathbf{a}}^*-\mathbf{x}\|_2+\|\mathbf{x}\|_2 \\
&\leq \|\mathbf{x}\|_2^2\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{2\lambda}+2\|\mathbf{x}\|_2 \\
&\leq \left(\frac{1}{2}\|\mathbf{x}\|_2+2\right)\|\mathbf{x}\|_2,
\end{aligned}
$$

where we use Lemma 2. Then, we have

$$
\begin{aligned}
&\left|\|\mathbf{D}\tilde{\mathbf{a}}^*\|_2^2-\|\tilde{\mathbf{D}}\tilde{\mathbf{a}}^*\|_2^2\right| \\
&\quad\leq 2\left|\langle\mathbf{D}\tilde{\mathbf{a}}^*,(\tilde{\mathbf{D}}-\mathbf{D})\tilde{\mathbf{a}}^*\rangle\right|+\|(\tilde{\mathbf{D}}-\mathbf{D})\tilde{\mathbf{a}}^*\|_2^2 \\
&\quad\leq 2\|\mathbf{D}\tilde{\mathbf{a}}^*\|_2\|(\tilde{\mathbf{D}}-\mathbf{D})\tilde{\mathbf{a}}^*\|_2+\|(\tilde{\mathbf{D}}-\mathbf{D})\tilde{\mathbf{a}}^*\|_2^2
\end{aligned}
$$

$$\leq 2\left(\frac{1}{2}\|\mathbf{x}\|_2+2\right)\|\mathbf{x}\|_2\left(\frac{\|\mathbf{x}\|_2^2\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{2\lambda}\right)+\left(\frac{\|\mathbf{x}\|_2^2\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{2\lambda}\right)^2$$
$$\leq \left(\frac{3}{4}\|\mathbf{x}\|_2+2\right)\|\mathbf{x}\|_2^3\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{\lambda}.$$

Combining this fact with Lemma 5, we have

$$\begin{aligned}
&\left|\|\mathbf{D}\mathbf{a}^*\|_2^2-\|\mathbf{D}\tilde{\mathbf{a}}^*\|_2^2\right| \\
&\quad\leq \left|\|\mathbf{D}\mathbf{a}^*\|_2^2-\|\tilde{\mathbf{D}}\tilde{\mathbf{a}}^*\|_2^2\right|+\left|\|\tilde{\mathbf{D}}\tilde{\mathbf{a}}^*\|_2^2-\|\mathbf{D}\tilde{\mathbf{a}}^*\|_2^2\right| \\
&\quad\leq \left(1+\frac{\|\mathbf{x}\|_2}{4}\right)\|\mathbf{x}\|_2^3\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{\lambda}+\left(\frac{3}{4}\|\mathbf{x}\|_2+2\right)\|\mathbf{x}\|_2^3\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{\lambda} \\
&\quad= \left(\|\mathbf{x}\|_2+3\right)\|\mathbf{x}\|_2^3\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{\lambda}.
\end{aligned}$$

□

Lemma 7 (Reconstructor Stability) *If* $\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}\leq\lambda$, *then*

$$\|\mathbf{D}\mathbf{a}^*-\mathbf{D}\tilde{\mathbf{a}}^*\|_2^2\leq 2\left(3\|\mathbf{x}\|_2^2+9\|\mathbf{x}\|_2+2\right)\|\mathbf{x}\|_2^2\frac{\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}}{\lambda}.$$

Proof We set $\bar{\mathbf{a}}^*:=\frac{1}{2}(\mathbf{a}^*+\tilde{\mathbf{a}}^*)$. From the optimality of $\mathbf{a}^*$, it follows that $v_{\mathbf{D}}(\mathbf{a}^*)\leq v_{\mathbf{D}}(\bar{\mathbf{a}}^*)$, i.e.,

$$\frac{1}{2}\|\mathbf{x}-\mathbf{D}\mathbf{a}^*\|_2^2+\lambda\|\mathbf{a}^*\|_1\leq\frac{1}{2}\|\mathbf{x}-\mathbf{D}\bar{\mathbf{a}}^*\|_2^2+\lambda\|\bar{\mathbf{a}}^*\|_1. \tag{23}$$

We denote $\epsilon:=\|\mathbf{D}-\tilde{\mathbf{D}}\|_{1,2}$, $c_{\mathbf{x}}:=\left(1+\frac{\|\mathbf{x}\|_2}{4}\right)\|\mathbf{x}\|_2^3$ and $c'_{\mathbf{x}}:=(\|\mathbf{x}\|_2+3)\|\mathbf{x}\|_2^3$.

By the convexity of the l_1-norm, the RHS of (23) obeys:

$$\begin{aligned}
&\frac{1}{2}\left\|\mathbf{x}-\mathbf{D}\left(\frac{\mathbf{a}^*+\tilde{\mathbf{a}}^*}{2}\right)\right\|_2^2+\lambda\left\|\frac{\mathbf{a}^*+\tilde{\mathbf{a}}^*}{2}\right\|_1 \\
&\quad\leq \frac{1}{2}\left\|\mathbf{x}-\frac{1}{2}(\mathbf{D}\mathbf{a}^*+\mathbf{D}\tilde{\mathbf{a}}^*)\right\|_2^2+\frac{\lambda}{2}\left\|\mathbf{a}^*\right\|_1+\frac{\lambda}{2}\left\|\tilde{\mathbf{a}}^*\right\|_1 \\
&\quad= \frac{1}{2}\left(\|\mathbf{x}\|_2^2-2\left\langle\mathbf{x},\frac{1}{2}(\mathbf{D}\mathbf{a}^*+\mathbf{D}\tilde{\mathbf{a}}^*)\right\rangle+\frac{1}{4}\|\mathbf{D}\mathbf{a}^*+\mathbf{D}\tilde{\mathbf{a}}^*\|_2^2\right)+\frac{\lambda}{2}\|\mathbf{a}^*\|_1+\frac{\lambda}{2}\|\tilde{\mathbf{a}}^*\|_1 \\
&\quad= \frac{1}{2}\|\mathbf{x}\|_2^2-\frac{1}{2}\left\langle\mathbf{x},\mathbf{D}\mathbf{a}^*\right\rangle-\frac{1}{2}\left\langle\mathbf{x},\mathbf{D}\tilde{\mathbf{a}}^*\right\rangle+\frac{1}{8}(\|\mathbf{D}\mathbf{a}^*\|_2^2+\|\mathbf{D}\tilde{\mathbf{a}}^*\|_2^2+2\langle\mathbf{D}\mathbf{a}^*,\mathbf{D}\tilde{\mathbf{a}}^*\rangle) \\
&\qquad+\frac{\lambda}{2}\|\mathbf{a}^*\|_1+\frac{\lambda}{2}\|\tilde{\mathbf{a}}^*\|_1 \\
&\quad\leq \frac{1}{2}\|\mathbf{x}\|_2^2-\frac{1}{2}\left\langle\mathbf{x},\mathbf{D}\mathbf{a}^*\right\rangle-\frac{1}{2}\left\langle\mathbf{x},\mathbf{D}\tilde{\mathbf{a}}^*\right\rangle+\frac{1}{4}\|\mathbf{D}\mathbf{a}^*\|_2^2+\frac{1}{4}\langle\mathbf{D}\mathbf{a}^*,\mathbf{D}\tilde{\mathbf{a}}^*\rangle \\
&\qquad+\frac{\lambda}{2}\|\mathbf{a}^*\|_1+\frac{\lambda}{2}\|\tilde{\mathbf{a}}^*\|_1+\frac{c'_{\mathbf{x}}}{8}\frac{\epsilon}{\lambda} \\
&\quad= \frac{1}{2}\|\mathbf{x}\|_2^2-\frac{1}{2}\left\langle\mathbf{x},\mathbf{D}\mathbf{a}^*\right\rangle-\frac{1}{2}\left\langle\mathbf{x},\mathbf{D}\tilde{\mathbf{a}}^*\right\rangle+\frac{1}{4}\|\mathbf{D}\mathbf{a}^*\|_2^2+\frac{1}{4}\langle\mathbf{D}\mathbf{a}^*,\mathbf{D}\tilde{\mathbf{a}}^*\rangle \\
&\qquad+\frac{1}{2}\left\langle\mathbf{x}-\mathbf{D}\mathbf{a}^*,\mathbf{D}\mathbf{a}^*\right\rangle+\frac{1}{2}\langle\mathbf{x}-\tilde{\mathbf{D}}\tilde{\mathbf{a}}^*,\tilde{\mathbf{D}}\tilde{\mathbf{a}}^*\rangle+\frac{c'_{\mathbf{x}}}{8}\frac{\epsilon}{\lambda}
\end{aligned}$$

$$\leq \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\mathbf{a}^*\rangle - \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\tilde{\mathbf{a}}^*\rangle + \frac{1}{4}\|\mathbf{D}\mathbf{a}^*\|_2^2 + \frac{1}{4}\langle \mathbf{D}\mathbf{a}^*, \mathbf{D}\tilde{\mathbf{a}}^*\rangle$$
$$+ \frac{1}{2}\langle \mathbf{x}, \mathbf{D}\mathbf{a}^*\rangle - \frac{1}{2}\|\mathbf{D}\mathbf{a}^*\|_2^2 + \frac{1}{2}\langle \mathbf{x}, \tilde{\mathbf{D}}\tilde{\mathbf{a}}^*\rangle - \frac{1}{2}\|\mathbf{D}\mathbf{a}^*\|_2^2 + \left(\frac{c'_{\mathbf{x}}}{8} + \frac{c_{\mathbf{x}}}{4}\right)\frac{\epsilon}{\lambda}$$
$$= \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{3}{4}\|\mathbf{D}\mathbf{a}^*\|_2^2 + \frac{1}{4}\langle \mathbf{D}\mathbf{a}^*, \mathbf{D}\tilde{\mathbf{a}}^*\rangle + \frac{1}{2}\langle \mathbf{x}, (\tilde{\mathbf{D}} - \mathbf{D})\tilde{\mathbf{a}}^*\rangle + \left(\frac{c'_{\mathbf{x}} + 2c_{\mathbf{x}}}{8}\right)\frac{\epsilon}{\lambda}, \quad (24)$$

where we use Lemma 3 in (24).

Now, taking the (expanded) LHS of (23) and the newly derived upper bound of the RHS of (23) yields the inequality:

$$\frac{1}{2}\|\mathbf{x}\|_2^2 - \langle \mathbf{x}, \mathbf{D}\mathbf{a}^*\rangle + \frac{1}{2}\|\mathbf{D}\mathbf{a}^*\|_2^2 + \lambda\|\mathbf{a}^*\|_1$$
$$\leq \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{3}{4}\|\mathbf{D}\mathbf{a}^*\|_2^2 + \frac{1}{4}\langle \mathbf{D}\mathbf{a}^*, \mathbf{D}\tilde{\mathbf{a}}^*\rangle + \frac{1}{2}\langle \mathbf{x}, (\tilde{\mathbf{D}} - \mathbf{D})\tilde{\mathbf{a}}^*\rangle + \left(\frac{c'_{\mathbf{x}} + 2c_{\mathbf{x}}}{8}\right)\frac{\epsilon}{\lambda}.$$

Replacing $\lambda\|\mathbf{a}^*\|_1$ with $\langle \mathbf{x} - \mathbf{D}\mathbf{a}^*, \mathbf{D}\mathbf{a}^*\rangle$ by Lemma 3 yields:

$$-\langle \mathbf{x}, \mathbf{D}\mathbf{a}^*\rangle + \frac{1}{2}\|\mathbf{D}\mathbf{a}^*\|_2^2 + \langle \mathbf{x} - \mathbf{D}\mathbf{a}^*, \mathbf{D}\mathbf{a}^*\rangle$$
$$\leq -\frac{3}{4}\|\mathbf{D}\mathbf{a}^*\|_2^2 + \frac{1}{4}\langle \mathbf{D}\mathbf{a}^*, \mathbf{D}\tilde{\mathbf{a}}^*\rangle + \frac{1}{2}\langle \mathbf{x}, (\tilde{\mathbf{D}} - \mathbf{D})\tilde{\mathbf{a}}^*\rangle + \left(\frac{c'_{\mathbf{x}} + 2c_{\mathbf{x}}}{8}\right)\frac{\epsilon}{\lambda}.$$

Hence,

$$\|\mathbf{D}\mathbf{a}^*\|_2^2 \leq \langle \mathbf{D}\mathbf{a}^*, \mathbf{D}\tilde{\mathbf{a}}^*\rangle + 2\langle \mathbf{x}, (\tilde{\mathbf{D}} - \mathbf{D})\tilde{\mathbf{a}}^*\rangle + \left(\frac{c'_{\mathbf{x}} + 2c_{\mathbf{x}}}{2}\right)\frac{\epsilon}{\lambda}$$
$$\leq \langle \mathbf{D}\mathbf{a}^*, \mathbf{D}\tilde{\mathbf{a}}^*\rangle + 2\frac{\|\mathbf{x}\|_2^3\epsilon}{2\lambda} + \left(\frac{c'_{\mathbf{x}} + 2c_{\mathbf{x}}}{2}\right)\frac{\epsilon}{\lambda}$$
$$= \langle \mathbf{D}\mathbf{a}^*, \mathbf{D}\tilde{\mathbf{a}}^*\rangle + \left(\frac{c'_{\mathbf{x}} + 2c_{\mathbf{x}} + 2\|\mathbf{x}\|_2^3}{2}\right)\frac{\epsilon}{\lambda}.$$

Then, we obtain

$$\|\mathbf{D}\mathbf{a}^* - \mathbf{D}\tilde{\mathbf{a}}^*\|_2^2$$
$$= \|\mathbf{D}\mathbf{a}^*\|_2^2 + \|\mathbf{D}\tilde{\mathbf{a}}^*\|_2^2 - 2\langle \mathbf{D}\mathbf{a}^*, \mathbf{D}\tilde{\mathbf{a}}^*\rangle$$
$$\leq \|\mathbf{D}\mathbf{a}^*\|_2^2 + \left(\|\mathbf{D}\mathbf{a}^*\|_2^2 + c'_{\mathbf{x}}\frac{\epsilon}{\lambda}\right) + \left(-2\|\mathbf{D}\mathbf{a}^*\|_2^2 + (c'_{\mathbf{x}} + 2c_{\mathbf{x}} + 2\|\mathbf{x}\|_2^3)\frac{\epsilon}{\lambda}\right)$$
$$\leq 2(c'_{\mathbf{x}} + c_{\mathbf{x}} + \|\mathbf{x}\|_2^3)\frac{\epsilon}{\lambda}.$$

□

Lemma 8 (Preservation of Sparsity) *If*

$$\mathcal{M}_k(\mathbf{D}, \mathbf{x}) > \left(1 + \frac{\|\mathbf{x}\|_2}{\lambda}\right)\|\mathbf{x}\|_2\|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2}$$
$$+ \sqrt{2\left(3\|\mathbf{x}\|_2^2 + 9\|\mathbf{x}\|_2 + 2\right)\|\mathbf{x}\|_2^2\frac{\|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2}}{\lambda}}, \quad (25)$$

then

$$\|\varphi_{\mathbf{D}}(\mathbf{x}) - \varphi_{\tilde{\mathbf{D}}}(\mathbf{x})\|_0 \leq k. \quad (26)$$

Proof In this proof, we denote $\varphi_{\mathbf{D}}(\mathbf{x})$ and $\varphi_{\tilde{\mathbf{D}}}(\mathbf{x})$ by $\mathbf{a}^* = [a_1^*, \ldots, a_m^*]^\top$ and $\tilde{\mathbf{a}}^* = [\tilde{a}_1^*, \ldots, \tilde{a}_m^*]^\top$, respectively. When $\tilde{\mathbf{D}} = \mathbf{D}$, Lemma 8 obviously holds. In the following, we assume $\tilde{\mathbf{D}} \neq \mathbf{D}$. Since $\mathcal{M}_k(\mathbf{D}, \mathbf{x}) > 0$ from (25), there is a $\mathcal{I} \subset [m]$ with $|\mathcal{I}| = m - k$, such that, for all $j \in \mathcal{I}$,

$$0 < \mathcal{M}_k(\mathbf{D}, \mathbf{x}) \leq \lambda - |\langle \mathbf{d}_j, \mathbf{x} - \mathbf{D}\mathbf{a}^* \rangle|. \tag{27}$$

To obtain (26), it is enough to show that $a_i^* = 0$ and $\tilde{a}_i^* = 0$ for all $i \in \mathcal{I}$.

First, we show $a_i^* = 0$ for all $i \in \mathcal{I}$. From the optimality conditions for the LASSO (Fuchs 2004), we have

$$\begin{aligned} \langle \mathbf{d}_j, \mathbf{x} - \mathbf{D}\mathbf{a}^* \rangle &= \mathrm{sign}(a_j^*)\lambda \quad \text{if } a_j^* \neq 0, \\ |\langle \mathbf{d}_j, \mathbf{x} - \mathbf{D}\mathbf{a}^* \rangle| &\leq \lambda \quad \text{otherwise.} \end{aligned}$$

Note that the above optimality conditions imply that, if $a_j^* \neq 0$, then

$$|\langle \mathbf{d}_j, \mathbf{x} - \mathbf{D}\mathbf{a}^* \rangle| = \lambda. \tag{28}$$

Combining (28) with (27), it holds that $a_i^* = 0$ for all $i \in \mathcal{I}$.

Next, we show $\tilde{a}_i^* = 0$ for all $i \in \mathcal{I}$. To do so, it is sufficient to show that

$$|\langle \tilde{\mathbf{d}}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^* \rangle| < \lambda \tag{29}$$

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

$$\begin{aligned} |\langle \tilde{\mathbf{d}}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^* \rangle| &= |\langle \mathbf{d}_i + \tilde{\mathbf{d}}_i - \mathbf{d}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^* \rangle| \\ &\leq |\langle \mathbf{d}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^* \rangle| + \|\tilde{\mathbf{d}}_i - \mathbf{d}_i\|_2 \|\mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^*\|_2 \\ &\leq |\langle \mathbf{d}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^* \rangle| + \|\tilde{\mathbf{D}} - \mathbf{D}\|_{1,2} \|\mathbf{x}\|_2 \end{aligned}$$

and

$$\begin{aligned} |\langle \mathbf{d}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^* \rangle| &= |\langle \mathbf{d}_i, \mathbf{x} - (\mathbf{D} + \tilde{\mathbf{D}} - \mathbf{D})\tilde{\mathbf{a}}^* \rangle| \\ &\leq |\langle \mathbf{d}_i, \mathbf{x} - \mathbf{D}\tilde{\mathbf{a}}^* \rangle| + |\langle \mathbf{d}_i, (\tilde{\mathbf{D}} - \mathbf{D})\tilde{\mathbf{a}}^* \rangle| \\ &\leq |\langle \mathbf{d}_i, \mathbf{x} - \mathbf{D}\tilde{\mathbf{a}}^* \rangle| + \|\tilde{\mathbf{D}} - \mathbf{D}\|_{1,2} \|\tilde{\mathbf{a}}^*\|_1. \end{aligned}$$

Hence,

$$|\langle \tilde{\mathbf{d}}_i, \mathbf{x} - \tilde{\mathbf{D}}\tilde{\mathbf{a}}^* \rangle| \leq |\langle \mathbf{d}_i, \mathbf{x} - \mathbf{D}\tilde{\mathbf{a}}^* \rangle| + \left(1 + \frac{\|\mathbf{x}\|_2}{\lambda}\right) \|\mathbf{x}\|_2 \|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2}.$$

Now,

$$\begin{aligned} |\langle \mathbf{d}_i, \mathbf{x} - \mathbf{D}\tilde{\mathbf{a}}^* \rangle| &= |\langle \mathbf{d}_i, \mathbf{x} - \mathbf{D}\mathbf{a}^* + \mathbf{D}\mathbf{a}^* - \mathbf{D}\tilde{\mathbf{a}}^* \rangle| \\ &\leq |\langle \mathbf{d}_i, \mathbf{x} - \mathbf{D}\mathbf{a}^* \rangle| + |\langle \mathbf{d}_i, \mathbf{D}\mathbf{a}^* - \mathbf{D}\tilde{\mathbf{a}}^* \rangle| \\ &\leq \lambda - \mathcal{M}_k(\mathbf{D}, \mathbf{x}) + \|\mathbf{D}\mathbf{a}^* - \mathbf{D}\tilde{\mathbf{a}}^*\|_2 \\ &\leq \lambda - \mathcal{M}_k(\mathbf{D}, \mathbf{x}) + \sqrt{2\left(3\|\mathbf{x}\|_2^2 + 9\|\mathbf{x}\|_2 + 2\right)\|\mathbf{x}\|_2^2 \frac{\|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2}}{\lambda}}, \end{aligned} \tag{30}$$

where (30) is because of Lemma 7. Then, (29) is obtained from (25). □

Lemma 9 *When a dictionary* $\mathbf{D}$ *is* μ*-incoherent, the following bound holds for an arbitrary* k*-sparse vector* $\mathbf{b}$*.*

$$\mathbf{b}^\top \mathbf{D}^\top \mathbf{D}\mathbf{b} \geq \left(1 - \frac{\mu k}{\sqrt{d}}\right) \|\mathbf{b}\|_2^2.$$

Proof We set $\mathbf{G} := \mathbf{D}^\top \mathbf{D} - \mathbf{I}$, where $\mathbf{I}$ is the $m \times m$ identity matrix. Since $\mathbf{D}$ is μ-incoherent, the absolute value of each component of $\mathbf{G}$ is less than or equal to $\mu/\sqrt{d}$, and thus, $\mathbf{b}^\top \mathbf{G}\mathbf{b} \geq -\mu/\sqrt{d}\|\mathbf{b}\|_1^2$. Then, we obtain

$$\mathbf{b}^\top \mathbf{D}^\top \mathbf{D}\mathbf{b} \;=\; \mathbf{b}^\top (\mathbf{I} + \mathbf{G})\mathbf{b} \;\geq\; \|\mathbf{b}\|_2^2 - \frac{\mu}{\sqrt{d}}\|\mathbf{b}\|_1^2 \;\geq\; \left(1 - \frac{\mu k}{\sqrt{d}}\right)\|\mathbf{b}\|_2^2, \tag{31}$$

where we use $\|\mathbf{b}\|_1 \leq \sqrt{k}\|\mathbf{b}\|_2$ for the k-sparse vector $\mathbf{b}$ in the last inequality. □

Proof of Theorem 2 Following the notations of Mehta and Gray (2012), we denote $\varphi_{\mathbf{D}}(\mathbf{x})$ and $\varphi_{\tilde{\mathbf{D}}}(\mathbf{x})$ by z_* and t_*, respectively. From (23) of Mehta and Gray (2012), we have

$$\begin{aligned}
&(z_* - t_*)^\top \mathbf{D}^\top \mathbf{D}(z_* - t_*) \\
&\quad \leq (z_* - t_*)^\top \left((\tilde{\mathbf{D}}^\top \tilde{\mathbf{D}} - \mathbf{D}^\top \mathbf{D})t_* + 2(\mathbf{D} - \tilde{\mathbf{D}})^\top \mathbf{x}\right) \\
&\quad = (z_* - t_*)^\top (\tilde{\mathbf{D}}^\top \tilde{\mathbf{D}} - \mathbf{D}^\top \mathbf{D})t_* + 2(z_* - t_*)^\top (\mathbf{D} - \tilde{\mathbf{D}})^\top \mathbf{x}.
\end{aligned} \tag{32}$$

Note that the assumption (25) of Lemma 8 follows from (11), and thus, $\|z_* - t_*\|_0 \leq k$ holds from Lemma 8. Then, $\|z_* - t_*\|_1 \leq \sqrt{k}\|z_* - t_*\|_2$.

When $\mathbf{E} := \mathbf{D} - \tilde{\mathbf{D}}$, the first term in (32) is evaluated as follows.

$$\begin{aligned}
&(z_* - t_*)^\top (\mathbf{D}^\top \mathbf{D} - \tilde{\mathbf{D}}^\top \tilde{\mathbf{D}})t_* \\
&\quad \leq |(z_* - t_*)^\top (\mathbf{E}^\top \tilde{\mathbf{D}} + \tilde{\mathbf{D}}^\top \mathbf{E} + \mathbf{E}^\top \mathbf{E})t_*| \\
&\quad \leq |(z_* - t_*)^\top \mathbf{E}^\top \tilde{\mathbf{D}}t_*| + |(z_* - t_*)^\top \tilde{\mathbf{D}}^\top \mathbf{E}t_*| + |(z_* - t_*)^\top \mathbf{E}^\top \mathbf{E}t_*| \\
&\quad \leq \|\mathbf{E}(z_* - t_*)\|_2 \|\tilde{\mathbf{D}}t_*\|_2 + \|\tilde{\mathbf{D}}(z_* - t_*)\|_2 \|\mathbf{E}t_*\|_2 + \|\mathbf{E}(z_* - t_*)\|_2 \|\mathbf{E}t_*\|_2 \\
&\quad \leq (\|\mathbf{E}\|_{1,2}\|\tilde{\mathbf{D}}\|_{1,2}\|t_*\|_1 + \|\tilde{\mathbf{D}}\|_{1,2}\|\mathbf{E}\|_{1,2}\|t_*\|_1 + \|\mathbf{E}\|_{1,2}\|\mathbf{E}\|_{1,2}\|t_*\|_1)\|z_* - t_*\|_1 \\
&\quad \leq \left(\frac{\|\mathbf{x}\|_2^2 \|\mathbf{E}\|_{1,2}}{\lambda} + \frac{\|\mathbf{x}\|_2^2 \|\mathbf{E}\|_{1,2}}{\lambda} + \frac{\|\mathbf{x}\|_2^2 \|\mathbf{E}\|_{1,2}^2}{\lambda}\right)\sqrt{k}\|z_* - t_*\|_2 \\
&\quad \leq \left(\frac{4\|\mathbf{x}\|_2^2}{\lambda}\right)\|\mathbf{E}\|_{1,2}\sqrt{k}\|z_* - t_*\|_2,
\end{aligned}$$

where we use $\|\mathbf{E}\|_{1,2} \leq 2$ in the last inequality. The second term in (32) is evaluated as follows:

$$\begin{aligned}
2(z_* - t_*)^\top \mathbf{E}^\top \mathbf{x} &\leq 2\|\mathbf{E}(z_* - t_*)\|_2 \|\mathbf{x}\|_2 \\
&\leq 2\|\mathbf{E}\|_{1,2}\|z_* - t_*\|_1 \|\mathbf{x}\|_2 \\
&\leq 2\sqrt{k}\|\mathbf{E}\|_{1,2}\|z_* - t_*\|_2 \|\mathbf{x}\|_2.
\end{aligned}$$

Thus, we have

$$(z_* - t_*)^\top \mathbf{D}^\top \mathbf{D}(z_* - t_*) \leq 2\sqrt{k}\left(1 + \frac{2\|\mathbf{x}\|_2}{\lambda}\right)\|\mathbf{x}\|_2 \|\mathbf{E}\|_{1,2}\|z_* - t_*\|_2. \tag{33}$$

On the other hand, we have the following lower bound of (32) from the μ-incoherence of $\mathbf{D}$ and Lemma 9:

$$(z_* - t_*)^\top \mathbf{D}^\top \mathbf{D}(z_* - t_*) \geq \left(1 - \frac{\mu k}{\sqrt{d}}\right) \|z_* - t_*\|_2^2. \tag{34}$$

From (33) and (34), we obtain

$$\|z_* - t_*\|_2 \leq \frac{2\sqrt{k}\,(1 + 2\|\mathbf{x}\|_2/\lambda)\,\|\mathbf{x}\|_2}{1 - \mu k/\sqrt{d}} \|\mathbf{D} - \tilde{\mathbf{D}}\|_{1,2}.$$

□

C Proof of margin bound

In this proof, we set

$$\begin{aligned}
\delta_1 &:= \frac{2\sigma\sqrt{k}m}{(1-t)\sqrt{d}\lambda} \exp\left(-\frac{(1-t)^2 d\lambda^2}{8\sigma^2 k}\right),\\
\delta_2 &:= \frac{2\sigma\sqrt{k}m}{\sqrt{d}\lambda} \exp\left(-\frac{d\lambda^2}{8\sigma^2 k}\right),\\
\delta_3' &:= \frac{4\sigma k^{3/2}}{C\sqrt{d(1-\mu k/\sqrt{d})}} \exp\left(-\frac{C^2 d(1-\mu k/\sqrt{d})}{8\sigma^2 k}\right)\\
\delta_3'' &:= \frac{8\sigma\sqrt{k}(d-k)}{d\lambda} \exp\left(-\frac{d^2\lambda^2}{32\sigma^2 k}\right),\\
\delta_3 &:= \delta_3' + \delta_3''.
\end{aligned}$$

Then, $\delta_{t,\lambda} = \delta_1 + \delta_2 + \delta_3$.

The column vectors for a μ-incoherent dictionary are in general position. Thus, a solution of LASSO for a μ-incoherent dictionary is unique as per Lemma 3 in Tibshirani (2013).

The following notions are introduced in Zhao and Yu (2006). Let $\mathbf{a}$ be a k-sparse vector. Without loss of generality, we can assume that $\mathbf{a} = [a_1, \ldots, a_k, 0, \ldots, 0]^\top$. Then, $\mathbf{a}(1) = [a_1, \ldots, a_k]^\top$, $\mathbf{D}(1) = [\mathbf{d}_1, \ldots, \mathbf{d}_k]$ and $\mathbf{D}(2) = [\mathbf{d}_{k+1}, \ldots, \mathbf{d}_m]$. We define $\mathbf{C}_{ij} := \frac{1}{d}\mathbf{D}(i)^\top \mathbf{D}(j)$ for $i, j \in \{1, 2\}$. When a dictionary $\mathbf{D}$ is μ-incoherent and $(\mu k)^2/d < 1$, $\mathbf{C}_{11}$ is positive definite owing to Lemma 9 and hence invertible.

Definition 6 (*Strong Irrepresentation Condition*) There exists a positive vector $\boldsymbol{\eta}$ such that

$$|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathrm{sign}(\mathbf{a}(1))| \leq \mathbf{1} - \boldsymbol{\eta},$$

where $\mathrm{sign}(\mathbf{a}(1))$ maps positive entry of $\mathbf{a}(1)$ to 1, negative entry to -1 and 0 to 0, $\mathbf{1}$ is the $(d-k) \times 1$ vector of 1's, and the inequality holds element-wise.

Then, the following lemma is derived by modifying the proof of Corollary 2 of Zhao and Yu (2006).

Lemma 10 (Strong Irrepresentation Condition) *When a dictionary* $\mathbf{D}$ *is* μ*-incoherent and* $\sqrt{d} > \mu(2k-1)$ *holds, the strong irrepresentation condition holds with* $\boldsymbol{\eta} = (1 - \mu(2k-1)/\sqrt{d})\mathbf{1}$.

The following lemma is given in the proof of Theorem 3 and Theorem 4 of Zhao and Yu (2006).

Lemma 11 *When Assuptions 1–4 hold and* $\mathbf{D}$ *is* μ*-incoherent and* $\sqrt{d} > 2\mu(2k-1)$*, there exist Gaussian random variables* $\{z_i\}_{i=1}^k$ *and* $\{\zeta_i\}_{i=1}^{d-k}$[9] *such that their variances are bounded as* $\mathbf{E}[z_i^2] \leq \sigma^2 k/d(1-\mu k/\sqrt{d})$ *and* $\mathbf{E}[\zeta_i^2] \leq \sigma^2 k/d^2$ *and*

$$\begin{aligned}
&\Pr\left[\operatorname{sign}(\varphi_{\mathbf{D}}(\mathbf{x})) = \operatorname{sign}(\mathbf{a})\right] \\
&\geq 1 - \sum_{i=1}^{k} \Pr\left[|z_i| \geq \sqrt{d}\left(|a_i| - \frac{\sqrt{k}\lambda}{2(1-\mu k/\sqrt{d})d}\right)\right] \\
&\quad - \sum_{i=1}^{d-k} \Pr\left[|\zeta_i| \geq \frac{(1-\mu(2k-1)/\sqrt{d})\lambda}{2\sqrt{d}}\right]. \qquad (35)
\end{aligned}$$

Proof The following is derived in Proposition 1 of Zhao and Yu (2006).

Proposition 1 *Assume Strong Irrepresentable Condition holds with a constant* $\eta > 0$ *then*

$$\Pr(\operatorname{sign}(\varphi_{\mathbf{D}}(\mathbf{x})) = \operatorname{sign}(a)) \geq \Pr(A_d \cap B_d),$$

where

$$\begin{aligned}
W(1) &:= \frac{1}{\sqrt{d}}\mathbf{D}(1)^\top \boldsymbol{\xi} \text{ and } W(2) := \frac{1}{\sqrt{d}}\mathbf{D}(2)^\top \boldsymbol{\xi}, \\
A_d &:= \left\{ \|(\mathbf{C}_{11})^{-1} W^d(1)\|_2 < \sqrt{d}\left(\|\mathbf{a}(1)\|_2 - \frac{\lambda}{2d}\|(C_{11}^d)^{-1}\operatorname{sign}(\mathbf{a}(1))\|_2\right)\right\}, \\
B_d &:= \left\{ \|\mathbf{C}_{21}(\mathbf{C}_{11})^{-1} W^d(1) - W^d(2)\|_2 < \frac{\lambda}{2\sqrt{d}}\left(1 - \frac{\mu(2k-1)}{\sqrt{d}}\right)\right\}.
\end{aligned}$$

Here, we have

$$\Pr(A_d \cap B_d) \geq 1 - \Pr(A_d^c) - \Pr(B_d^c).$$

From Lemma 10, we obtain $\|(\mathbf{C}_{11}^d)^{-1}\operatorname{sign}(\mathbf{a}(1))\|_2 \leq \sqrt{k}/(1-\mu k/\sqrt{d})$. Thus, it holds that

$$\Pr(A_d^c) \leq \sum_{i=1}^{k} \Pr\left[|z_i| \geq \sqrt{d}\left(|a_i| - \frac{\lambda}{2d}\frac{\sqrt{k}}{(1-\mu k/\sqrt{d})}\right)\right], \qquad (36)$$

$$\Pr(B_d^c) \leq \sum_{i=1}^{d-k} \Pr\left[|\zeta_i| \geq \frac{\lambda}{2\sqrt{d}}\left(1 - \frac{\mu(2k-1)}{\sqrt{d}}\right)\right], \qquad (37)$$

where $\mathbf{z} = (z_1, ..., z_m)^\top := (C_{11}^d)^{-1} W^d(1)$ and $\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_m)^\top := C_{21}^d (C_{11}^d)^{-1} W^d(1) - W^d(2)$. Thus, it is enough to show that $\mathbf{E}[z_i^2] \leq \sigma^2 k/d(1-\mu k/\sqrt{d})$ and $\mathbf{E}[\zeta_i^2] \leq \sigma^2 k/d^2$.

If we write $\mathbf{z} = \mathbf{H}_A^\top \boldsymbol{\xi}$ where $\mathbf{H}_A^\top = (\mathbf{h}_1^A, ..., \mathbf{h}_k^A)^\top = (\mathbf{C}_{11})^{-1}\sqrt{d}^{-1}\mathbf{D}(1)$, then

$$\mathbf{H}_A^\top \mathbf{H}_A = \frac{1}{d}(\mathbf{C}_{11})^{-1} \leq \frac{1}{1-\mu k/\sqrt{d}} I.$$

[9] The notations in Zhao and Yu (2006) and this paper relate as follows: $z_i^n := z_i, \zeta_i^n := \zeta_i, \beta_i^n := a_i, b_i^n \leq \frac{\sqrt{k}}{1-\mu k/\sqrt{d}}, \eta_i^n := 1-\mu(2k-1)/\sqrt{d}, \epsilon^n := \boldsymbol{\xi}, p := m, q := k, \lambda_n := \lambda, M_1 := 1/d, M_2 := 1-\mu k/\sqrt{d}$, respectively.

Therefore $z_i = h_1^{A\top}\boldsymbol{\xi}$ with

$$\mathbf{E}[z_i^2] = \frac{1}{1-\mu k/\sqrt{d}}\mathbf{E}[\|\boldsymbol{\xi}\|_2^2] \leq \frac{\sigma^2 k}{d(1-\mu k/\sqrt{d})}.$$

Similarly if we write $\boldsymbol{\zeta} = \mathbf{H}_B^\top \boldsymbol{\xi}$ where $\mathbf{H}_B^\top = (\mathbf{h}_1^B, ..., \mathbf{h}_k^B)^\top = \mathbf{C}_{21}(\mathbf{C}_{11})^{-1}\sqrt{d}^{-1}\mathbf{D}(1)^\top - \sqrt{d}^{-1}\mathbf{D}(2)^\top$, then

$$H_B^\top H_B = \frac{1}{d}\mathbf{D}(2)^\top(I - \mathbf{D}(1)^\top(\mathbf{D}(1)^\top\mathbf{D}(1))^{-1}\mathbf{D}(1)^\top)\mathbf{D}(2).$$

Then

$$\|\mathbf{h}_i^B\|_2^2 \leq \frac{1}{d}\mathbf{d}_i^\top(I - \mathbf{D}(1)^\top(\mathbf{D}(1)^\top\mathbf{D}(1))^{-1}\mathbf{D}(1)^\top)\mathbf{d}_i \leq \frac{1}{d}\mathbf{d}_i^\top\mathbf{d}_i = \frac{1}{d},$$

where we used the fact that $I - \mathbf{D}(1)^\top(\mathbf{D}(1)^\top\mathbf{D}(1))^{-1}\mathbf{D}(1)^\top$ has eigenvalues between 0 and 1. Therefore $\zeta_i = h_i^{B\top}\boldsymbol{\xi}$ with

$$\mathbf{E}[\zeta_i^2] \leq \mathbf{E}[\|\boldsymbol{\xi}\|_2^2\|\mathbf{h}_i^B\|_2^2] \leq \frac{\sigma^2 k}{d^2}.$$

□

Lemma 12 *Under Assuptions 1–4, when* $\mathbf{D}$ *is* μ*-incoherent and* $\sqrt{d} > 2\mu(2k-1)$*, the following holds:*

$$\Pr\left[|\mathrm{supp}(\mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x}))| \leq k\right] \geq 1 - \delta_3.$$

Proof The following inequality obviously holds:

$$\Pr\left[|\mathrm{supp}(\mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x}))| \leq k\right] \geq \Pr\left[\mathrm{sign}(\mathbf{a}) = \mathrm{sign}(\varphi_{\mathbf{D}}(\mathbf{x}))\right].$$

From Lemma 11, there exist Gaussian random variables $\{z_i\}_{i=1}^k$ and $\{\zeta_i\}_{i=1}^{d-k}$ such that their variances are bounded as $\mathbf{E}[z_i^2] \leq \sigma^2 k/d(1-\mu k/\sqrt{d})$ and $\mathbf{E}[\zeta_i^2] \leq \sigma^2 k/d^2$ and (35).

When $\lambda \leq (1-\mu k/\sqrt{d})Cd/\sqrt{k}$, the inequality $|a_i| - \frac{\sqrt{k}\lambda}{2(1-\mu k/\sqrt{d})d} \geq C/2$ holds since $|a_i| \geq C$. Then, since $1 - \mu(2k-1)/\sqrt{d} \geq 1/2$ holds, we obtain

$$\Pr\left[|z_i| \geq \sqrt{d}\left(|a_i| - \frac{\sqrt{k}\lambda}{2(1-\mu k/\sqrt{d})d}\right)\right] \leq \Pr\left[|z_i| \geq \frac{C\sqrt{d}}{2}\right] \leq \delta_3',$$

$$\Pr\left[|\zeta_i| \geq \frac{(1-\mu(2k-1)/\sqrt{d})\lambda}{2\sqrt{d}}\right] \leq \Pr\left[|\zeta_i| \geq \frac{\lambda}{4\sqrt{d}}\right] \leq \delta_3'',$$

where we use the fact that z_i and ζ_i are sub-Gaussian. Thus, the proof is completed. □

Lemma 13 *Let* $\mathbf{D}$ *be a dictionary. When* $\boldsymbol{\xi}$ *satisfies Assumption 4, the following holds:*

$$\Pr[\lambda \geq 2\|\mathbf{D}^\top\boldsymbol{\xi}\|_\infty] \geq 1 - \delta_2.$$

Proof Let ξ be a 1-dimensional Gaussian with variance $\sigma^2 k/d$. Then, it holds that, for $t > 0$,

$$\Pr[|\xi| > \lambda] \leq \frac{\sigma\sqrt{k}}{\sqrt{d}\lambda}\exp\left(-\frac{d\lambda^2}{2\sigma^2 k}\right). \tag{38}$$

Note that $\langle \mathbf{d}_j, \boldsymbol{\xi} \rangle$ is Gaussian with variance $\sigma^2 k/d$ because $\|\mathbf{d}_j\|_2 = 1$ for every $j \in [m]$ and components of $\boldsymbol{\xi}$ are independent and Gaussian with variance $\sigma^2 k/d$. Thus,

$$\Pr[\lambda < 2\|\mathbf{D}^\top \boldsymbol{\xi}\|_\infty] = \Pr\left[\cup_{j=1}^m \{\lambda < 2|\langle \mathbf{d}_j, \boldsymbol{\xi}\rangle|\}\right] \leq \sum_{j=1}^m \Pr[\lambda < 2|\langle \mathbf{d}_j, \boldsymbol{\xi}\rangle|] \leq \delta_2,$$

where we used (38) in the last inequality. □

Lemma 14 *Under Assuptions 1–4,*

$$\Pr\left[\|\mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x})\|_2 \leq \frac{3\sqrt{k}}{(1-\mu k/\sqrt{d})}\lambda\right] \geq 1 - \delta_2 - \delta_3.$$

Proof By Assumption 1, $\mathbf{x} = \mathbf{D}\mathbf{a} + \boldsymbol{\xi}$. We denote $\varphi_{\mathbf{D}}(\mathbf{x})$ by $\mathbf{a}^*$ and $\mathbf{a} - \mathbf{a}^*$ by Δ. We have the following inequality by the definition of $\mathbf{a}^*$:

$$\frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{a}^*\|_2^2 + \lambda\|\mathbf{a}^*\|_1 \leq \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1.$$

Substituting $\mathbf{x} = \mathbf{D}\mathbf{a} + \boldsymbol{\xi}$, we have

$$\begin{aligned}\frac{1}{2}\|\mathbf{D}\Delta\|_2^2 &\leq -\langle \mathbf{D}^\top \boldsymbol{\xi}, \Delta\rangle + \lambda(\|\mathbf{a}\|_1 - \|\mathbf{a}^*\|_1) \\ &\leq \|\mathbf{D}^\top \boldsymbol{\xi}\|_\infty \|\Delta\|_1 + \lambda(\|\mathbf{a}\|_1 - \|\mathbf{a}^*\|_1). \end{aligned} \tag{39}$$

Let Δ_k be the vector whose ith component equals that of Δ, if i is in the support of $\mathbf{a}$, and equals 0, otherwise. In addition, let $\Delta_k^\perp = \Delta - \Delta_k$. Using $\Delta = \Delta_k + \Delta_k^\perp$, we have

$$\|\mathbf{a}^*\| = \|\mathbf{a} - \Delta_k^\perp - \Delta_k\|_1 \geq \|\mathbf{a}\|_1 + \|\Delta_k^\perp\|_1 - \|\Delta_k\|_1.$$

Substituting the above inequality into (39), we have

$$\frac{1}{2}\|\mathbf{D}\Delta\|_2^2 \leq \|\mathbf{D}^\top \boldsymbol{\xi}\|_\infty \|\Delta\|_1 + \lambda(\|\Delta_k\|_1 - \|\Delta_k^\perp\|_1).$$

The inequality $\lambda \geq 2\|\mathbf{D}^\top \boldsymbol{\xi}\|_\infty$ holds with probability $1 - \delta_2$ due to Lemma 13, and then, the following inequality holds:

$$\begin{aligned}\frac{1}{2}\|\mathbf{D}\Delta\|_2^2 &\leq \frac{1}{2}\lambda(\|\Delta_k\|_1 + \|\Delta_k^\perp\|_1) + \lambda(\|\Delta_k\|_1 - \|\Delta_k^\perp\|_1) \\ &= \frac{3}{2}\lambda\|\Delta_k\|_1 - \frac{1}{2}\lambda\|\Delta_k^\perp\|_1 \\ &\leq \frac{3}{2}\lambda\|\Delta_k\|_1 \\ &\leq \frac{3}{2}\lambda\sqrt{k}\|\Delta_k\|_2.\end{aligned}$$

Thus, we have

$$\|\mathbf{D}\Delta\|_2^2 \leq 3\lambda\sqrt{k}\|\Delta_k\|_2 \leq 3\lambda\sqrt{k}\|\Delta\|_2.$$

Here, $\|\mathrm{supp}(\Delta)\|_0 \leq k$ with probability $1 - \delta_3$ due to Lemma 12 and the following inequality holds by the μ-incoherence of the dictionary $\mathbf{D}$:

$$(1 - \mu k/\sqrt{d})\|\Delta\|_2^2 \leq \|\mathbf{D}\Delta\|_2^2.$$

Thus,

$$\|\Delta\|_2 \leq \frac{3\lambda\sqrt{k}}{(1-\mu k/\sqrt{d})}.$$

□

Proof of Theorem 3 From Assumption 1, an arbitrary sample $\mathbf{x}$ is represented as $\mathbf{x} = \mathbf{D}^*\mathbf{a} + \boldsymbol{\xi}$. Then,

$$\begin{aligned}\langle \mathbf{d}_j, \mathbf{x} - \mathbf{D}^*\varphi_{\mathbf{D}}(\mathbf{x})\rangle &= \langle \mathbf{d}_j, \boldsymbol{\xi} + \mathbf{D}^*(\mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x}))\rangle \\ &= \langle \mathbf{d}_j, \boldsymbol{\xi}\rangle + \langle \mathbf{D}^{*\top}\mathbf{d}_j, \mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x})\rangle.\end{aligned}$$

We evaluate the probability that the first and second terms shown above are bounded by $\frac{1-t}{2}\lambda$.

We evaluate the probability for the first term. Since $\|\mathbf{d}_j\| = 1$ by definition, and $\boldsymbol{\xi}$ is drawn from a Gaussian distribution with variance $\sigma^2 k/d$, we have

$$\Pr\left[\max_{1\leq j\leq m} \langle \mathbf{d}_j, \boldsymbol{\xi}\rangle \leq \frac{1-t}{2}\lambda\right] \geq 1 - \delta_1.$$

With probability $1 - \delta_2 - \delta_3$, the second term is evaluated as follows:

$$\begin{aligned}|\langle \mathbf{D}^{*\top}\mathbf{d}_j, \mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x})\rangle| &= |\langle [\langle \mathbf{d}_1, \mathbf{d}_j\rangle, \ldots, \langle \mathbf{d}_m, \mathbf{d}_j\rangle]^\top, \mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x})\rangle| \\ &= |\langle \mathbf{1}_{\mathrm{supp}(\mathbf{a}-\varphi_{\mathbf{D}}(\mathbf{x}))} \circ [\langle \mathbf{d}_1, \mathbf{d}_j\rangle, \ldots, \langle \mathbf{d}_m, \mathbf{d}_j\rangle]^\top, \mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x})\rangle| \\ &\leq \|\mathbf{1}_{\mathrm{supp}(\mathbf{a}-\varphi_{\mathbf{D}}(\mathbf{x}))} \circ [\langle \mathbf{d}_1, \mathbf{d}_j\rangle, \ldots, \langle \mathbf{d}_m, \mathbf{d}_j\rangle]^\top\|_2 \|\mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x})\|_2 \\ &\leq \frac{\mu}{\sqrt{d}}\sqrt{|\mathrm{supp}(\mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x}))|}\|\mathbf{a} - \varphi_{\mathbf{D}}(\mathbf{x})\|_2 \\ &\leq \frac{3\mu k}{(1-\mu k/\sqrt{d})\sqrt{d}}\lambda \qquad (40) \\ &\leq \frac{1-t}{2}\lambda, \qquad (41)\end{aligned}$$

where we used Lemmas 12 and 14 in (40) and $d \geq \left\{\left(1 + \frac{6}{(1-t)}\right)\mu k\right\}^2$ in (41). Thus, with probability $1 - (\delta_1 + \delta_2 + \delta_3) = 1 - \delta_{t,\lambda}$,

$$\mathcal{M}_k(\mathbf{D}^*, \mathbf{x}) \geq \min_{1\leq j\leq m} \{\lambda - |\langle \mathbf{d}_j, \mathbf{x} - \mathbf{D}^*\varphi_{\mathbf{D}}(\mathbf{x})\rangle|\} \geq t\lambda.$$

Thus, the proof of Theorem 3 is completed. □

References

Arora, S., Ge, R., Ma, T., & Moitra, A. (2015). Simple, efficient, and neural algorithms for sparse coding. In *Conference on learning theory* (pp. 113–149).

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, *12*(149–198), 3.

Ben-David, S., & Urner, R. (2013). Domain adaptation as learning with auxiliary information. In *New directions in transfer and multi-task-workshop@ NIPS*.

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 120–128). Association for Computational Linguistics.

Dai, W., Yang, Q., Xue, G. R.,& Yu, Y. (2008). Self-taught clustering. In *Proceedings of the 25th international conference on Machine learning* (pp. 200–207). ACM.

Daume, H, I. I. I., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, *26*, 101–126.
Duan, L., Tsang, I. W., & Xu, D. (2012). Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(3), 465–479.
Fuchs, J. J. (2004). On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, *50*(6), 1341–1344.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.
Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., & Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems* (pp. 601–608).
Huang, K., & Aviyente, S. (2006). Sparse representation for signal classification. In *NIPS* (Vol. 19, pp. 609–616).
Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab: An S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9), 1–20.
Kumagai, W. (2016). Learning bound for parameter transfer learning. In *Advances in neural information processing systems* (pp. 2721–2729).
Kuzborskij, I., & Orabona, F. (2013). Stability and hypothesis transfer learning. In *Proceedings of the 30th international conference on machine learning (ICML-13)* (pp. 942–950).
Kuzborskij, I., & Orabona, F. (2017). Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, *106*(2), 171–195.
Lee, H., Raina, R., Teichman, A., & Ng, A. Y. (2009). Exponential family sparse coding with application to self-taught learning. In *IJCAI* (Vol. 9, pp. 1113–1119). Citeseer.
Mairal, J., Bach, F., Ponce, & J., Sapiro, G. (2009). Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning* (pp. 689–696). ACM.
Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., & Bach, F. R. (2009). Supervised dictionary learning. In *Advances in neural information processing systems* (pp. 1033–1040).
Maurer, A. (2009). Transfer bounds for linear feature learning. *Machine Learning*, *75*(3), 327–350.
Maurer, A., Pontil, M., & Romera-Paredes, B. (2013). Sparse coding for multitask and transfer learning. In *Proceedings of the 30th international conference on machine learning (ICML-13)* (pp. 343–351).
Mehta, N., & Gray, A. G. (2013). Sparsity-based generalization bounds for predictive sparse coding. In *Proceedings of the 30th international conference on machine learning (ICML-13)* (pp. 36–44).
Mehta, N. A., & Gray, A. G. (2012). On the sample complexity of predictive sparse coding. arXiv:1202.4050.
Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics*, *9*(2), 319–337.
Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.
Pentina, A., & Lampert, C. (2014). A PAC-Bayesian bound for lifelong learning. In *Proceedings of the 31st international conference on machine learning (ICML-14)* (pp. 991–999).
R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning* (pp. 759–766). ACM.
Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*(2), 227–244.
Sridharan, K., Shalev-Shwartz, S., & Srebro, N. (2009). Fast rates for regularized objectives. In *Advances in neural information processing systems* (pp. 1545–1552).
Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., & Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems* (pp. 1433–1440).
Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, *7*, 1456–1490.
Tommasi, T., Orabona, F., & Caputo, B. (2014). Learning categories from few examples with multi model knowledge transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(5), 928–941.
Vainsencher, D., Mannor, S., & Bruckstein, A. M. (2011). The sample complexity of dictionary learning. *The Journal of Machine Learning Research*, *12*, 3259–3281.
Wang, H., Nie, F., & Huang, H. (2013). Robust and discriminative self-taught learning. In *Proceedings of the 30th international conference on machine learning* (pp. 298–306).
Yang, J., Yu, K., Gong, Y., & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009* (pp. 1794–1801). IEEE.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning* (p. 114). ACM.

Zhang, Z., Xu, Y., Yang, J., Li, X., & Zhang, D. (2015). A survey of sparse representation: Algorithms and applications. *IEEE Access*, *3*, 490–530.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, *7*, 2541–2563.

Zhu, X., Huang, Z., Yang, Y., Shen, H. T., Xu, C., & Luo, J. (2013). Self-taught dimensionality reduction on the high-dimensional small-sized data. *Pattern Recognition*, *46*(1), 215–229.