Theoretical Analysis of Learning Local Anchors for Classification

Junbiao Pang[†], Qingming Huang[‡,§], Baocai Yin[†], Lei Qin[‡], Dan Wang[†]
[†]*College of Computer Science and Technology, Beijing University of Technology, China*
[‡]*Key Lab. of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, China*
[§]*Graduate University of Chinese Academy of Sciences, Chinese Academy of Sciences, China*
[†]*{jbpang, ybc, wangdan}@bjut.edu.cn,* [‡]*{qmhuang, lqin}@jdl.ac.cn*

Abstract

In this paper, we present a theoretical analysis on learning anchors for local coordinate coding (LCC), which is a method to model functions for data lying on non-linear manifolds. In our analysis several local coding schemes, i.e., orthogonal coordinate coding (OC-C), local Gaussian coding (LGC), local Student coding (LSC), are theoretically compared, in terms of the upper-bound locality error on any high-dimension data; this provides some insight to understand the local coding for classification tasks. We further give some interesting implications of our results, such as tradeoff between locality and approximation ability in learning anchors.

1. Introduction

Local Coordinate Coding (LCC) [10] is a method that approximates any non-linear (α, β, p)-Lipschitz smooth function over the data manifold using linear functions. There are two components in this method: 1) a set of anchors (data points) which build local coordinates; and 2) local coding schemes for each data determined by these anchors. LCC has been successfully applied in many challenging problems, e.g., utilizing very high-dimension data in VOC competitions.

Although the success of LCC, its classification performance is highly depended on the number of anchors, as suggested by the theoretical bound [10] and the practical observations [9]. These anchors should be "local" enough to encode data on the manifold accurately, which sometimes means that the number of anchors in real applications would increase explosively. Moreover, theoretically analysis suggests that locality is more essential that sparsity in terms of non-linear function approximation ability. Therefore, coding schemes are critical to balance between accuracy and complexity.

A few approach has been proposed for learning anchors, motivated by non-linear approximation or not, but the solutions have focused on data quantization or compression. For instance, sparse coding using the Lagrange dual [4], online dictionary learning with stochastic approximation [5], the k-means clustering [3]. However, these methods do not provide the theoretical connection between the number of anchors and the approximation ability of (α, β, p)-Lipschitz function.

Several existing publications have been aware of this pitfall. For instance, orthogonal coordinate coding (OC-C) [11] provides the theoretical analysis for the upper-bounded approximation ability of (α, β, p)-Lipschitz function. However, we still do not know the approximation ability for more coding schemes, e.g., sparse coding [4].

In this paper, we propose two local coding schemes (LCSs), and further give theoretical analysis and comparison with OCC. Intuitively, if a sample $\mathbf{x}$ is closer to the anchor $\mathbf{v}$, the value of local coding function $\gamma_v(\mathbf{x})$ should be also larger [6]; thus, two different types of LCS are proposed to achieve this purpose:

1. *Local Gaussian coding (LGC)* presumes the relation between samples and anchors as,

$$\gamma_v^{lgc}(\mathbf{x};\sigma) = \frac{\exp\left(\frac{-\|\mathbf{v}-\mathbf{x}\|^2}{\sigma^2}\right)}{\sum_{\mathbf{v}\in\mathcal{C}} \exp\left(\frac{-\|\mathbf{v}-\mathbf{x}\|^2}{\sigma^2}\right)}, \tag{1}$$

where $\mathbf{v} \in \mathbb{R}^D$, is the anchor, and the bias σ is the hyper parameter to control the decay ability of $\gamma_m^{lgc}(\mathbf{x})$.

2. *Local Student coding (LSC)* uses Student t-distribution with one degree of freedom, which is the same as Cauchy distribution,

$$\gamma_v^{lsc}(\mathbf{x};\sigma) = \frac{(\sigma^2 + \|\mathbf{v}-\mathbf{x}\|^2)^{-1}}{\sum_{\mathbf{v}\in\mathcal{C}}(\sigma^2 + \|\mathbf{v}-\mathbf{x}\|^2)^{-1}}, \tag{2}$$

Table 1: Some notations in this paper.

Notation	Definition
$\mathbf{v} \in \mathbb{R}^d$	A d-dimension anchor
$\mathcal{C} \subset \mathbb{R}^d$	A set of all anchors
$\gamma_v(\mathbf{x}) \in \mathbb{R}$	The local coding of a data $\mathbf{x}$ with the anchor $\mathbf{v}$
$\gamma(\mathbf{x}) \in \mathbb{R}^d$	The approximation vector of a data $\mathbf{x}$
$\gamma_x \in \mathbb{R}^{\lvert\mathcal{C}\rvert}$	The coding vector of a data $\mathbf{x}$ by all anchors
γ	A map of a data $\mathbf{x}$ to γ_x
$(\gamma, \mathcal{C})$	A coordinate coding scheme

where the bias σ is the hyper parameter to control the decay ability of $\gamma_v^{lsc}(\mathbf{x})$. Student t-distribution has the nice property that $(\sigma^2 + \|\mathbf{v} - \mathbf{x}\|^2)^{-1}$ approaches an inverse square law for large pairwise distances $\|\mathbf{v} - \mathbf{x}\|^2$.

2. Theoretical Analysis of the Upper Bound

For clarification, we summarize some notations in Table 1 which are used in this paper. We first revisit some definitions and conclusions in LCC. Note that in the following sections, $\| \cdot \|$ denotes the ℓ_2 norm.

Definition 1 (Lipschitz smoothness [10]). *A function $f(x) \in \mathbb{R}^d$ is (α, β, p)-Lipschitz smooth with respect to the $\| \cdot \|$ norm, if $|f(\mathbf{x}') - f(\mathbf{x})| \leq \alpha \|\mathbf{x} - \mathbf{x}'\|$ and $|f(\mathbf{x}') - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{x}' - \mathbf{x})| \leq \beta \|\mathbf{x} - \mathbf{x}'\|^{1+p}$, where we assume $\alpha, \beta > 0$ and $p \in (0, 1]$.*

Definition 2 (Coordinate coding [10]). *Let $(\gamma, \mathcal{C})$ be an arbitrary coordinate coding on $\mathbb{R}^d$. Let f be an (α, β, p)-Lipschitz smooth function. We have for all $\mathbf{x} \in \mathbb{R}^d$: $\gamma(\mathbf{x}) = \sum_{\mathbf{v} \in \mathcal{C}} \gamma_v(\mathbf{x})\mathbf{v}$.*

Lemma 1 (Linearization [10]). *Let $(\gamma, \mathcal{C})$ be an arbitrary coordinate coding on $\mathbb{R}^d$. Let f be an (α, β, p)-Lipschitz smooth function. We have for all $\mathbf{x} \in \mathbb{R}^d$:*

$$\left| f(\mathbf{x}) - \sum_{\mathbf{v} \in \mathcal{C}} \gamma_v(\mathbf{x}) f(\mathbf{v}) \right| \leq \alpha \|\mathbf{x} - \gamma(\mathbf{x})\| + \beta \sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \|\mathbf{v} - \gamma(\mathbf{x})\|^{1+p} \tag{3}$$

2.1 Analysis of the Locality Bound

In this subsection, lGC and LSC schemes would firstly be analyzed and further be compared with OCC; moreover, the σ-related bound of LGC and LSC would be presented and discussed.

Theorem 1 (Localization Error of LGC or LSC). *Let $(\gamma, \mathcal{C})$ either be a LGC or LSC on $\mathbb{R}^d$ data manifolds, where the number of anchors is M, i.e., $|\mathcal{C}| = M$. Let f be an (α, β, p)-Lipschitz smooth function. Without losing generalization, assuming $\forall \mathbf{x} \in \mathbb{R}^d$, $||\mathbf{x}|| \leq 1$ and $\forall \mathbf{v} \in \mathcal{C}$, $1 \leq ||\mathbf{v}|| \leq h (h \geq 1)$, then the localization error in Lemma 1 is bounded by:*

$$\sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \parallel \mathbf{v} - \gamma(\mathbf{x}) \parallel^{1+p} \leq \left[h^2 + h^2 M^2\right]^{\frac{1+p}{2}} \tag{4}$$

Proof. Let $\gamma_v(\mathbf{x})$ either be (1) or (2), then

$$\begin{aligned}
&\sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \parallel \mathbf{v} - \gamma(\mathbf{x}) \parallel^{1+p} \\
&= \sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \left[\|\mathbf{v}\|^2 - 2\gamma(\mathbf{x})\mathbf{v} + \left(\sum_{\mathbf{v} \in \mathcal{C}} \gamma_v(\mathbf{x})\mathbf{v} \right)^2 \right]^{\frac{1+p}{2}} \\
&\leq \sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \left[\|\mathbf{v}\|^2 - 2 \sum_{\mathbf{v} \in \mathcal{C}} \gamma_v(\mathbf{x}) \|\mathbf{v}\|^2 + \sum_{\mathbf{v} \in \mathcal{C}} \sum_{\mathbf{v} \in \mathcal{C}} \gamma_v(\mathbf{x})^2 \|\mathbf{v}\|^2 \right]^{\frac{1+p}{2}} \\
&= \sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \Bigg[\parallel \mathbf{v} \parallel^2 - 2 \sum_{\mathbf{v} \in \mathcal{C}} \gamma_v(\mathbf{x}) \|\mathbf{v}\|^2 \\
&\qquad + \left(\max_{\mathbf{v} \in \mathcal{C}} \parallel \mathbf{v} \parallel^2 \right) \sum_{\mathbf{v} \in \mathcal{C}} \sum_{\mathbf{v} \in \mathcal{C}} \gamma_v(\mathbf{x})^2 \Bigg]^{\frac{1+p}{2}}
\end{aligned} \tag{5}$$

Because $\forall \mathbf{x} \in \mathbb{R}^d$, $\parallel \mathbf{v} \parallel \leq 1$ and $\forall \mathbf{v} \in \mathcal{C}$, $1 \leq \parallel \mathbf{v} \parallel \leq h$, $\sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| = 1$, so $|\gamma_v(\mathbf{x})| \leq 1$, $\sum_{\mathbf{v} \in \mathcal{C}} \gamma_v(\mathbf{x})^2 \leq M$. Therefore, (5) can be

$$\begin{aligned}
&\sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \|\mathbf{v} - \gamma(\mathbf{x})\|^{1+p} \\
&\leq \sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \left[h^2 - 2h^2 \sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| + h^2 M^2 \right]^{\frac{1+p}{2}} \\
&\leq \sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \left[h^2 + h^2 M^2 \right]^{\frac{1+p}{2}} \\
&= \left[h^2 + h^2 M^2 \right]^{\frac{1+p}{2}} \cdot \sum_{\mathbf{v} \in \mathcal{C}} |\gamma_v(\mathbf{x})| \\
&= \left[h^2 + h^2 M^2 \right]^{\frac{1+p}{2}}
\end{aligned}$$

□

Discussion among OCC, LGC and LSC: It is obvious that $\left[h^2 + h^2 M^2\right]^{\frac{1+p}{2}}$ is lower than $[(M + 1)h]^{1+p}$, the upper-bound of OCC [11]. Thus, LGC or LSC theoretically obtain more lower upper-bound approximation error than OCC. Although an anchor plane in OCC could contain infinite anchors, most of anchors do not necessarily live on the same plane; in other words, anchors do not densely live on a plane. Therefore, the number of planes in OCC would naturally be larger than the number of anchors in lGC or LSC.

Theorem 2 (σ-related upper bound for LGC). *Let $(\gamma, \mathcal{C})$ be a LGC (1) on $\mathbb{R}^d$, where the number of anchors is equal to M, i.e., $|\mathcal{C}| = M$. Let f be an (α, β, p)-Lipschitz smooth function. Without losing generalization, assuming $\forall \mathbf{x} \in \mathbb{R}^d$, $||\mathbf{x}|| \leq 1$ and*

$\forall \mathbf{v} \in \mathcal{C}$, $1 \le ||\mathbf{v}|| \le h(h \ge 1)$, and $d_l \le \|\mathbf{x}-\mathbf{v}\| \le d_u$, then the localization error in Lemma 1 is bounded by:

$$\sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \parallel \mathbf{v}-\gamma(\mathbf{x}) \parallel^{1+p} \le [h^2 + 2M^2h^2(\frac{d_u^2}{\sigma^2}-1) + M^2h^2]^{\frac{1+p}{2}} \tag{6}$$

Proof. Let $\gamma_v(\mathbf{x})$ be (1) and $s_{\mathbf{x}} = \sum_{\mathbf{v}\in\mathcal{C}} \exp\left(\frac{-\|\mathbf{v}-\mathbf{x}\|^2}{\sigma^2}\right)$, then

$$\begin{aligned}&\sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \parallel \mathbf{v}-\gamma(\mathbf{x}) \parallel^{1+p} \\ &= \sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \left[\parallel \mathbf{v} \parallel^2 - 2\gamma(\mathbf{x})\mathbf{v} + \left(\sum_{\mathbf{v}\in\mathcal{C}} \gamma(\mathbf{x})\mathbf{v}\right)^2\right]^{\frac{1+p}{2}}\end{aligned} \tag{7}$$

Reuse the derivation in (5) and the well-known inequality, i.e., $1-x \le e^{-x}$, (7) can be reformulated as

$$\begin{aligned}&\sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \parallel \mathbf{v}-\gamma(\mathbf{x}) \parallel^{1+p} \\ &\le \sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \left[h^2 - 2h^2 s_{\mathbf{x}} \sum_{\mathbf{v}\in\mathcal{C}} \gamma_v(\mathbf{x}) + M^2h^2\right]^{\frac{1+p}{2}} \\ &= \sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \left[h^2 + 2h^2 s_{\mathbf{x}} \sum_{\mathbf{v}\in\mathcal{C}} \left(\frac{\|\mathbf{x}-\mathbf{v}\|^2}{\sigma^2} - 1\right) \right. \\ &\qquad \left. + M^2h^2\right]^{\frac{1+p}{2}}\end{aligned}$$

Because $s_{\mathbf{x}} \le M$, and $d_l \le \|\mathbf{x}-\mathbf{v}\| \le d_u$, above inequality can be written as

$$\begin{aligned}&\le \sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \left[h^2 + 2M^2h^2 \left(\frac{d_u^2}{\sigma^2} - 1\right) + M^2h^2\right]^{\frac{1+p}{2}} \\ &= \left[h^2 + 2M^2h^2 \left(\frac{d_u^2}{\sigma^2} - 1\right) + M^2h^2\right]^{\frac{1+p}{2}}\end{aligned}$$

□

Next, we would give the σ-related upper bound for LSC (2), and compare its bound with the one of LGC in Lemma 2.

Theorem 3 (σ-related upper bound for LSC). *Let $(\gamma, \mathcal{C})$ be a LSC (2) on $\mathbb{R}^d$, where the number of anchors is equal to M, i.e., $|\mathcal{C}| = M$. Let f be an (α, β, p)-Lipschitz smooth function. Without losing generalization, assuming $\forall \mathbf{x} \in \mathbb{R}^d$, $||\mathbf{x}|| \le 1$ and $\forall \mathbf{v} \in \mathcal{C}$, $1 \le ||\mathbf{v}|| \le h(h \ge 1)$ and $d_l \le \|\mathbf{x}-\mathbf{v}\| \le d_u$, then the localization error in Lemma 1 is bounded by:*

$$\sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \parallel \mathbf{v}-\gamma(\mathbf{x}) \parallel^{1+p} \le \left[h^2 - \frac{2M^2h^2}{(\sigma+d_l)^2} + M^2h^2\right]^{\frac{1+p}{2}} \tag{8}$$

Proof. Let $\gamma_v(\mathbf{x})$ be (2) and $s_{\mathbf{x}} = \sum_{\mathbf{v}\in\mathcal{C}} (\sigma^2 + \|\mathbf{v}-\mathbf{x}\|^2)^{-1}$. Following the procedure in Theorem 2, and using the inequality, i.e., $\frac{1}{(a+b)^2} \le \frac{1}{a^2+b^2}$, we can obtain the upper bound for LSC as: [1]

$$\begin{aligned}&\sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \|\mathbf{v}-\gamma(\mathbf{x})\|^{1+p} \\ &\le \sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \left[h^2 - \frac{2M^2h^2}{(\sigma+d_l)^2} + M^2h^2\right]^{\frac{1+p}{2}} \\ &= \left[h^2 - \frac{2M^2h^2}{(\sigma+d_l)^2} + M^2h^2\right]^{\frac{1+p}{2}}\end{aligned}$$

□

Discussion on LGC and LSC: As stated in Theorem 2, the upper bound of LGC is controlled by the σ-related term, i.e., $\left(\frac{d_u^2}{\sigma^2} - 1\right)$. If the choice of σ makes $\frac{d_u^2}{\sigma^2} < 1$, we would obtain more lower upper bound than other values of σ. Besides, $d_u^2 < \sigma^2$ theoretically shows that there exists an optimal hyper-parameter σ for lGC, if the number of anchors M is fixed.

For LSC (2), the term $-\frac{2M^2h^2}{(\sigma+d_l)^2}$ is always negative. If σ reduces, the bound of LSC also decreases. However, reducing σ would take danger to make the coding $\gamma_v(\mathbf{x})$ be zero. That is, although the locality term, $\sum_{\mathbf{v}\in\mathcal{C}} |\gamma_v(\mathbf{x})| \|\mathbf{v}-\gamma(\mathbf{x})\|^{1+p}$, in (3) has low error, the term, $\|x - \gamma(\mathbf{x})\|$, in (3) has large reconstruction error.

It should be noted that Theorem 2 and 3 do not discover the relation between the number of anchors M and the hyper parameter σ. Besides, all above conclusions are drawn on the assumption that the number of anchors M is fixed.

2.2 Learning Anchors in LGC and LSC

Since Theorem 2 or 3 proves that the locality error is bounded by a constant, we only consider the data reconstruction term to minimize the upper bound in Theorem 1 as:

$$\begin{aligned}&\min_{\mathbf{V}} \sum_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \gamma(\mathbf{x})\mathbf{V}\|^2 \\ &s.t.\ |\gamma(\mathbf{x})| = 1, \gamma(\mathbf{x}) \succeq 0, \forall m\end{aligned} \tag{9}$$

where $|\cdot|$ is the ℓ_1 norm, $\gamma(\mathbf{x}) \succeq 0$ means that all elements of $\gamma(\mathbf{x})$ stratify $\gamma_v(\mathbf{x}) \ge 0$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]^T$. It is quite similar to sparse coding, except the reconstruction term $\gamma(\mathbf{x})$ is determined by (1) or (2).

Optimizing (9) is beyond the scope of this paper, especially for large-scale and high-dimension visual data.

[1] Due to the limited length of paper, we omit the detailed derivation and directly present the conclusion.

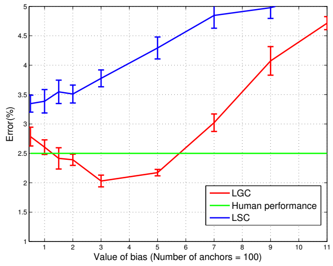

Figure 1: Comparison between LGC and LSC with different number of anchors.

An approximate solution for $\mathbf{V}$ is the clustering centers of k-means as [3] does.

3 Verification of Theoretical Results

We use USPS and MNIST data sets in the experiments. USPS consists of 7,291 training and 2,007 test gray-scale 16 × 16 images. Each label corresponds to "0"-"9" digits. MNIST contains 60,000 training and 10,000 28×28 gray-scale test images, which are reshaped directly into 784-dimensional vectors. Although these two data sets are relatively easy for classification, our primary purpose is to verify these theoretical conclusions in Subsection 2.1. As expected in Theorem 2, the error rate of LGC would first decrease until a critical point, and then increase [2] (see Figure 1). This observation well matches the conclusions in Theorem 2: there exists an optimal σ in LGC for classification. While for LSC, the error rate would raise as the value of σ increases. Obviously, Theorem 3 explains this phenomena well. That is, the σ should be as smaller as possible.

Table 2: Classification error rate $(\%)$ on MNIST and USPS.

Algorithms	MNIST	USPS
PEG-LLSVM+G-OCC(#vectors) [11]	1.81(40)	4.38(50)
PEG-LLSVM+C-OCC(#vectors) [11]	1.74(90)	4.09(80)
PEG-LLSVM+LGC(#anchors)	1.63(30)	**2.21(40)**
PEG-LLSVM+LSC(#anchors)	1.86(20)	2.64(20)
Lin.SVM+LCC(512) [10]	2.64	-
Lin.SVM+improved LCC(512) [9]	1.95	-
LA-SVM(2 passes) [2]	**1.36**	-
SVM_{struct} [8]	1.40	4.38
LL-SVM(10 passes, 100anchors) [3]	1.85	5.78

Table 2 summarizes our comparison results between our methods and some other SVMs-based approaches. The parameter of the RBF kernel used in the SVMs is the same as [1]. On USPS, we can see that LGC is better than LCC, improved LCC, OCC and LL-SVM; on MNIST, LGC is better than both LCC and LL-SVM, but slightly worse than LA-SVM. All of these results demonstrate that LGC or LSC is quite suitable to mode the non-linear anchors in LL-SVM for classification. On the other hand, LSC or LGC uses much less number of anchors compared to the one in LCC or OCC, while obtains better test accurate rate than other ones.

4 Conclusion

In this paper, we theoretically analyze the local coding, i.e., LSC and LGC, to encode high-dimension data. We prove that LSC and LGC can guarantee a lower locality error for any (α, β, p)-Lipschitz smooth function than previous methods. In future, we would like to learn the localized sparse coding (9) using stochastic gradient descent for large-scale and high-dimension visual data.

5 Acknowledgement

This work was supported by the National Natural Science Foundation of China No.60825203, No.61033004, No.61133003, No.61003165, in part by National Basic Research Program of China (973 Program): 2009CB320906, and Beijing Municipal Natural Science Foundation of China No.4122007. This work was also supported by the Startup Research Funding for Ph.D. Scholars, BJUT.

References

[1] A. Bordes, L. Bottou, P. Gallinari, and J. Weston. Solving multiclass support vector machines with larank. In *ICML*, 2007.

[2] A. Bordes, S. Ertekin, and L. Bottou. Fast kernel classifiers with online and active learning. *JMLR*, 2005.

[3] L. Ladický and P. Torr. Locally linear support vector machines. In *ICML*, 2011.

[4] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *NIPS*, 2007.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009.

[6] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000.

[7] S. Shalev-Shwartz, Y. Singer, and N. srebro. Pegasos: Primal estimated subgradient solver for svm. In *ICML*, 2007.

[8] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent out variables. *JMLR*, 2005.

[9] K. Yu and T. Zhang. Improved local coordinate coding using local tangents. In *ICML*, 2010.

[10] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *NIPS*, 2009.

[11] Z. Zhang, L. Ladický, P. H. Torr, and A. Saffari. Learning anchor planes for classification. In *NIPS*, 2011.

[2] We use PEGASOS [7] to optimize the LL-SVM in this paper.