Transductive Learning with Multi-class Volume Approximation

Gang Niu NIUGANG@BAIDU.COM
Tokyo Institute of Technology, Tokyo, 152-8552, Japan
Baidu Inc., Beijing, 100085, China

Bo Dai BODAI@GATECH.EDU
Georgia Institute of Technology, Atlanta, GA 30332, USA

Marthinus Christoffel du Plessis CHRISTO@SG.CS.TITECH.AC.JP
Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Tokyo Institute of Technology, Tokyo, 152-8552, Japan

Abstract

Given a hypothesis space, the *large volume principle* by Vladimir Vapnik prioritizes equivalence classes according to their volume in the hypothesis space. The *volume approximation* has hitherto been successfully applied to binary learning problems. In this paper, we propose a novel generalization to multiple classes, allowing applications of the large volume principle on more learning problems such as *multi-class*, *multi-label* and *serendipitous* learning in a transductive manner. Although the resultant learning method involves a non-convex optimization problem, the globally optimal solution is almost surely unique and can be obtained using $O(n^3)$ time. Novel theoretical analyses are presented for the proposed method, and experimental results show it compares favorably with the one-vs-rest extension.

1. Introduction

The history of the *large volume principle* (LVP) goes back to the early age of the statistical learning theory when Vapnik (1982) introduced it for the case of hyperplanes. But it did not gain much attention until a creative approximation was proposed in El-Yaniv et al. (2008) to implement LVP for the case of soft response vectors. From then on, it has been applied to various binary learning problems successfully, such as binary transductive learning (El-Yaniv et al., 2008), binary clustering (Niu et al., 2013a), and outlier detection (Li & Ng, 2013).

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

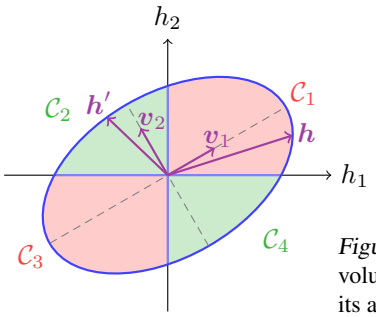


Figure 1. The large volume principle and its approximation.

LVP is a learning-theoretic principle which views learning as *hypothesis selecting* from a certain *hypothesis space* $\mathcal{H}$. Regardless of the hypothesis form, $\mathcal{H}$ can always be partitioned into a finite number of equivalence classes on some observed data set, where each *equivalence class* is a set of hypotheses that generate the same labeling of the observed data. LVP, as one of the learning-theoretic principles from the statistical learning theory, prioritizes those equivalence classes according to the volume they occupy in $\mathcal{H}$. See the illustration in Figure 1: The blue ellipse represents $\mathcal{H}$, and it is partitioned into $\mathcal{C}_1, \ldots, \mathcal{C}_4$ each occupying a quadrant of the Cartesian coordinate system $\mathbb{R}^2$ intersected with $\mathcal{H}$; LVP claims that $\mathcal{C}_1$ and $\mathcal{C}_3$ are more preferable than $\mathcal{C}_2$ and $\mathcal{C}_4$, since $\mathcal{C}_1$ and $\mathcal{C}_3$ have larger volume than $\mathcal{C}_2$ and $\mathcal{C}_4$.

In practice, the hypothesis space $\mathcal{H}$ cannot be as simple as $\mathcal{H}$ in Figure 1. It is often located in very high-dimensional spaces where exact or even quantifiable *volume estimation* is challenging. Therefore, El-Yaniv et al. (2008) proposed a *volume approximation* to bypass the volume estimation. Instead of focusing on the equivalence classes of $\mathcal{H}$, it directly focuses on the hypotheses in $\mathcal{H}$ since learning is regarded as hypothesis selecting in LVP. It defines $\mathcal{H}$ via an ellipsoid, measures the angles from hypotheses to the principal axes of $\mathcal{H}$, and then prefers hypotheses near the long

principal axes to those near the short ones. This manner is reasonable, since the long principal axes of $\mathcal{H}$ lie in large-volume regions. In Figure 1, $\boldsymbol{h}$ and $\boldsymbol{h}'$ are two hypotheses and $\boldsymbol{v}_1$/$\boldsymbol{v}_2$ is the long/short principal axis; LVP advocates that $\boldsymbol{h}$ is more preferable than $\boldsymbol{h}'$ as $\boldsymbol{h}$ is close to $\boldsymbol{v}_1$ and $\boldsymbol{h}'$ is close to $\boldsymbol{v}_2$. We can adopt this volume approximation to regularize our loss function, which has been demonstrated helpful for various binary learning problems.

Nevertheless, the volume approximation in El-Yaniv et al. (2008) only fits binary learning problem settings despite its potential advantage. In this paper we extend it naturally to a more general definition, which can be applied to several transductive problem settings, including but not limited to *multi-class learning* (Zhou et al., 2003), *multi-label learning* (Kong et al., 2013), and *serendipitous learning* (Zhang et al., 2011). We adopt the same strategy as El-Yaniv et al. (2008): For n data and c labels, a hypothesis space is defined in $\mathbb{R}^{n\times c}$ and linked to an ellipsoid in $\mathbb{R}^{nc}$, such that the equivalence classes and the volume approximation can be defined accordingly. We name the learning method that realizes the above approximation *multi-class approximate volume regularization* (MAVR). It involves a non-convex optimization problem, but the globally optimal solution is almost surely unique and accessible in $O(n^3)$ time following Forsythe & Golub (1965). Moreover, we theoretically provide novel stability and error analyses for MAVR, and experimentally show that MAVR compares favorably with the one-vs-rest extension of El-Yaniv et al. (2008).

The rest of this paper is organized as follows. In Section 2 the problem settings are discussed. In Section 3 the binary volume approximation is reviewed and the multi-class volume approximation is derived. Then the proposed method MAVR is developed and analyzed in Section 4. At last the experimental results are reported in Section 5.

2. Transductive Problem Settings

Recall the setting of transductive binary problems (Vapnik, 1998, p. 341). Suppose that $\mathcal{X}$ is the domain of input data, and most often but not necessarily, $\mathcal{X} \subset \mathbb{R}^d$ where d is a natural number. A fixed set $X_n = \{x_1, \ldots, x_n\}$ of n points from $\mathcal{X}$ is observed, and the labels $y_1, \ldots, y_n \in \{-1, +1\}$ of these points are also fixed but unknown. A subset $X_l \subset X_n$ of size l is picked uniformly at random, and then y_i is revealed if $x_i \in X_l$. We call $S_l = \{(x_i, y_i) \mid x_i \in X_l\}$ the labeled data and $X_u = X_n \setminus X_l$ the unlabeled data. Using S_l and X_u, the goal is to predict y_i of $x_i \in X_u$ (while any unobserved $x \in \mathcal{X} \setminus X_n$ is currently left out of account).

Transductive learning (TL) (e.g., Blum & Chawla, 2001; Szummer & Jaakkola, 2001; Joachims, 2003; Zhou et al., 2003; El-Yaniv et al., 2008) slightly differs from *semi-supervised learning* (SSL) (e.g., Bennett & Demiriz, 1998; Zhu et al., 2003; Grandvalet & Bengio, 2004; Belkin et al., 2006; Li et al., 2009; Li & Zhou, 2011; Niu et al., 2013b): TL focuses on predicting X_u while SSL aims at predicting $\mathcal{X} \setminus X_l$, and TL is distribution free but SSL is not.[1] More specifically, TL generally makes no assumption about the underlying distributions, and the true labels are deterministic; SSL usually assumes that S_l is sampled from $p(x, y)$ and X_u is sampled from $p(x)$, and then the true labels are stochastic. Moreover, if there is any distributional change, SSL should specify the form of the change, but TL might deal with it directly. To sum up, SSL is inductive learning in nature, and the advantage of TL over inductive learning is conceptually critical for us.

As an extension of El-Yaniv et al. (2008), the volume approximation to be proposed can be applied to many transductive problem settings, where the differences are the encoding of labels and the decoding of hypotheses. The first setting is *multi-class learning*: Instead of $y_i \in \{-1, +1\}$, we have $y_i \in \mathcal{Y}$ where $\mathcal{Y} = \{1, \ldots, c\}$ and c is a natural number. Each of the c labels here have some labeled data in spite of any distributional change. The second setting is *multi-label learning*: $y_i \subseteq \mathcal{Y}$ with $\mathcal{Y} = \{1, \ldots, c\}$ where y_i is a label set, or $y_i \in \mathcal{Y}$ with $\mathcal{Y} = \{-1, 0, 1\}^c$ where y_i is a label vector (cf. Kong et al., 2013). The third setting is *serendipitous learning* which is a multi-class setting with missing classes in S_l, that is, some of the c labels have no labeled data (cf. Zhang et al., 2011). It is non-trivial to see the distributional change is *covariate shift* (Yamada et al., 2010) or *class-prior change* (du Plessis & Sugiyama, 2012) from semi-supervised point of view, whereas it is unnecessary to specify the form of the change in our settings.

In principle, all transductive methods can solve multi-class problems with the one-vs-rest extension. But this may not be a good idea for methods defined in terms of non-convex optimizations like El-Yaniv et al. (2008). Furthermore, the encoding of labels for multi-label and serendipitous problems is an issue when using the one-vs-rest extension. The volume approximation to be proposed can handle all these settings in a unified manner, but in this paper we focus on multi-class and serendipitous learning since they do not require sophisticated post-processing as Kong et al. (2013).

3. Volume Approximations

In this section we review the binary volume approximation and propose our multi-class volume approximation.

3.1. Binary volume approximation

The binary volume approximation involves a few key concepts (El-Yaniv et al., 2008): The soft response vector, the hypothesis space and the equivalence class, and the power

[1] Some methods lie between them, e.g., Wang et al. (2013).

and volume of equivalence classes.

Given a set of n data $X_n = \{x_1, \ldots, x_n\}$ where $x_i \in \mathcal{X}$, a *soft response vector* is an n-dimensional vector

$$\boldsymbol{h} := (h_1, \ldots, h_n)^\top \in \mathbb{R}^n, \tag{1}$$

so that h_i stands for a soft or confidence-rated label of x_i. For binary problems, $\boldsymbol{h}$ suggests that x_i is from the positive class if $h_i > 0$, x_i is from the negative class if $h_i < 0$, and the above two cases are equally possible if $h_i = 0$.

A *hypothesis space* is a collection of hypotheses. The volume approximation requires a symmetric positive-definite matrix $Q \in \mathbb{R}^{n\times n}$ which contains the pairwise information about X_n. Consider the hypothesis space

$$\mathcal{H}_Q := \{\boldsymbol{h} \mid \boldsymbol{h}^\top Q \boldsymbol{h} \leq 1\}, \tag{2}$$

where the hypotheses are soft response vectors. The set of sign vectors $\{\mathrm{sign}(\boldsymbol{h}) \mid \boldsymbol{h} \in \mathcal{H}_Q\}$ contains all of $N = 2^n$ possible dichotomies of X_n, and $\mathcal{H}_Q$ can be partitioned into a finite number of *equivalence classes* $\mathcal{C}_1, \ldots, \mathcal{C}_N$, such that for fixed k, all hypotheses in $\mathcal{C}_k$ will generate the same labeling of X_n.

Then, in statistical learning theory, the *power* of an equivalence class $\mathcal{C}_k$ is defined as the probability mass of all hypotheses in it (Vapnik, 1998, p. 708), i.e.,

$$\mathcal{P}(\mathcal{C}_k) := \int_{\mathcal{C}_k} p(\boldsymbol{h}) \mathrm{d}\boldsymbol{h}, \quad k = 1, \ldots, N,$$

where $p(\boldsymbol{h})$ is the underlying probability density of $\boldsymbol{h}$ over $\mathcal{H}_Q$. The hypotheses in $\mathcal{C}_k$ which has a large power should be preferred according to Vapnik (1998).

When no specific domain knowledge is available (i.e., $p(\boldsymbol{h})$ is unknown), it would be natural to assume the continuous uniform distribution $p(\boldsymbol{h}) = 1/\sum_{k=1}^N \mathcal{V}(\mathcal{C}_k)$, where

$$\mathcal{V}(\mathcal{C}_k) := \int_{\mathcal{C}_k} \mathrm{d}\boldsymbol{h}, \quad k = 1, \ldots, N,$$

is the *volume* of $\mathcal{C}_k$. That is, the volume of an equivalence class is defined as the geometric volume of all hypotheses in it. As a result, $\mathcal{P}(\mathcal{C}_k)$ is proportional to $\mathcal{V}(\mathcal{C}_k)$, and the larger the value $\mathcal{V}(\mathcal{C}_k)$ is, the more confident we are of the hypotheses chosen from $\mathcal{C}_k$.

However, it is very hard to accurately compute the geometric volume of even a single convex body in $\mathbb{R}^n$, let alone all 2^n convex bodies, so El-Yaniv et al. (2008) introduced an efficient approximation. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of Q, and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be the associated orthonormal eigenvectors. Actually, the hypothesis space $\mathcal{H}_Q$ in Eq. (2) is geometrically an origin-centered ellipsoid in $\mathbb{R}^n$ with $\boldsymbol{v}_i$ and $1/\sqrt{\lambda_i}$ as the direction and length of its i-th principal axis. Note that a small angle from a hypothesis $\boldsymbol{h}$ in $\mathcal{C}_k$ to some $\boldsymbol{v}_i$ with a small/large index i (i.e., a long/short principal axis) implies that $\mathcal{V}(\mathcal{C}_k)$ is large/small (cf. Figure 1). Based on this crucial observation, we define

$$V(\boldsymbol{h}) := \sum_{i=1}^n \lambda_i \left(\frac{\boldsymbol{h}^\top \boldsymbol{v}_i}{\|\boldsymbol{h}\|_2}\right)^2 = \frac{\boldsymbol{h}^\top Q \boldsymbol{h}}{\|\boldsymbol{h}\|_2^2}, \tag{3}$$

where $\boldsymbol{h}^\top \boldsymbol{v}_i / \|\boldsymbol{h}\|_2$ means the cosine of the angle between $\boldsymbol{h}$ and $\boldsymbol{v}_i$. We subsequently expect $V(\boldsymbol{h})$ to be small when $\boldsymbol{h}$ lies in a large-volume equivalence class, and conversely to be large when $\boldsymbol{h}$ lies in a small-volume equivalence class.

3.2. Multi-class volume approximation

The multi-class volume approximation could deal with the aforementioned transductive problem settings in a unified manner. In order to extend the definition Eq. (3), we need only to extend the hypothesis and the hypothesis space.

To begin with, we allocate a soft response vector in Eq. (1) for each of the c labels:

$$\boldsymbol{h}_1 = (h_{1,1}, \ldots, h_{n,1})^\top, \ldots, \boldsymbol{h}_c = (h_{1,c}, \ldots, h_{n,c})^\top.$$

The value $h_{i,j}$ is a soft or confidence-rated label of x_i concerning the j-th label and it suggests that

- x_i should possess the j-th label, if $h_{i,j} > 0$;
- x_i should not possess the j-th label, if $h_{i,j} < 0$;
- the above two cases are equally possible, if $h_{i,j} = 0$.

For multi-class and serendipitous problems, y_i is predicted by $\hat{y}_i = \arg\max_j h_{i,j}$. For multi-label problems, we need a threshold T_h that is either preset or learned since usually positive and negative labels are imbalanced, and y_i can be predicted by $\hat{y}_i = \{j \mid h_{i,j} \geq T_h\}$; or we can employ the prediction methods proposed in Kong et al. (2013). Then, a *soft response matrix* as our transductive hypothesis is an n-by-c matrix defined by

$$H = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_c) \in \mathbb{R}^{n\times c}, \tag{4}$$

and a *stacked soft response vector* as an equivalent hypothesis is an nc-dimensional vector defined by

$$\boldsymbol{h} = \mathrm{vec}(H) = (\boldsymbol{h}_1^\top, \ldots, \boldsymbol{h}_c^\top)^\top \in \mathbb{R}^{nc},$$

where $\mathrm{vec}(H)$ is the vectorization of H formed by stacking its columns into a single vector.

As the binary definition of the hypothesis space, a symmetric positive-definite matrix $Q \in \mathbb{R}^{n\times n}$ which contains the pairwise information about X_n is provided, and we assume further that a symmetric positive-definite matrix $P \in \mathbb{R}^{c\times c}$ which contains the pairwise information about $\mathcal{Y}$ is available. Consider the hypothesis space

$$\mathcal{H}_{P,Q} := \{H \mid \mathrm{tr}(H^\top Q H P) \leq 1\}, \tag{5}$$

where the hypotheses are soft response matrices. Let $P \otimes Q \in \mathbb{R}^{nc \times nc}$ be the *Kronecker product* of P and Q. Due to the symmetry and the positive definiteness of P and Q, the Kronecker product $P \otimes Q$ is also symmetric and positive definite, and $\mathcal{H}_{P,Q}$ in (5) could be defined equivalently as

$$\mathcal{H}_{P,Q} := \{H \mid \operatorname{vec}(H)^\top (P \otimes Q) \operatorname{vec}(H) \leq 1\}. \quad (6)$$

The equivalence between Eqs. (5) and (6) comes from the fact that $\operatorname{tr}(H^\top Q H P) = \operatorname{vec}(H)^\top (P \otimes Q) \operatorname{vec}(H)$ following the well-known identity (see, e.g., Theorem 13.26 of Laub, 2005)

$$(P^\top \otimes Q) \operatorname{vec}(H) = \operatorname{vec}(QHP).$$

As a consequence, there is a bijection between $\mathcal{H}_{P,Q}$ and

$$\mathcal{E}_{P,Q} := \{\boldsymbol{h} \mid \boldsymbol{h}^\top (P \otimes Q) \boldsymbol{h} \leq 1\}$$

which is geometrically an origin-centered ellipsoid in $\mathbb{R}^{nc}$. The set of sign vectors $\{\operatorname{sign}(\boldsymbol{h}) \mid \boldsymbol{h} \in \mathcal{E}_{P,Q}\}$ spreads over all the $N = 2^{nc}$ quadrants of $\mathbb{R}^{nc}$, and thus the set of sign matrices $\{\operatorname{sign}(H) \mid H \in \mathcal{H}_{P,Q}\}$ contains all of N possible dichotomies of $X_n \times \{1, \ldots, c\}$. In other words, $\mathcal{H}_{P,Q}$ can be partitioned into N equivalence classes $\mathcal{C}_1, \ldots, \mathcal{C}_N$, such that for fixed k, all soft response matrices in $\mathcal{C}_k$ will generate the same labeling of $X_n \times \{1, \ldots, c\}$.

The definition of the power is same as before, and so is the definition of the volume:

$$\mathcal{V}(\mathcal{C}_k) := \int_{\mathcal{C}_k} \mathrm{d}H, \quad k = 1, \ldots, N.$$

Because of the bijection between $\mathcal{H}_{P,Q}$ and $\mathcal{E}_{P,Q}$, $\mathcal{V}(\mathcal{C}_k)$ is likewise the geometric volume of all stacked soft response vectors in the intersection of the k-th quadrant of $\mathbb{R}^{nc}$ and $\mathcal{E}_{P,Q}$. By a similar argument to the definition of $V(\boldsymbol{h})$, we define

$$V(H) := \frac{\boldsymbol{h}^\top (P \otimes Q) \boldsymbol{h}}{\|\boldsymbol{h}\|_2^2} = \frac{\operatorname{tr}(H^\top Q H P)}{\|H\|_{\mathrm{Fro}}^2}, \quad (7)$$

where $\boldsymbol{h} = \operatorname{vec}(H)$ and $\|H\|_{\mathrm{Fro}}$ means the Frobenius norm of H. We subsequently expect $V(H)$ to be small when H lies in a large-volume equivalence class, and conversely to be large when H lies in a small-volume equivalence class.

Note that $V(H)$ and $V(\boldsymbol{h})$ are consistent for binary learning problems. When $c = 2$, we may constrain $\boldsymbol{h}_1 + \boldsymbol{h}_2 = \mathbf{0}_n$ where $\mathbf{0}_n$ means the all-zero vector in $\mathbb{R}^n$. Let $P = I_2$ where I_2 means the identity matrix of size 2, then

$$V(H) = \frac{\boldsymbol{h}_1^\top Q \boldsymbol{h}_1 + \boldsymbol{h}_2^\top Q \boldsymbol{h}_2}{\|\boldsymbol{h}_1\|_2^2 + \|\boldsymbol{h}_2\|_2^2} = \frac{\boldsymbol{h}_1^\top Q \boldsymbol{h}_1}{\|\boldsymbol{h}_1\|_2^2} = V(\boldsymbol{h}_1),$$

which coincides with $V(\boldsymbol{h})$ defined in Eq. (3). Similarly to $V(\boldsymbol{h})$, for two soft response matrices H and H' from the same equivalence class, $V(H)$ and $V(H')$ may not necessarily be the same value. In addition, the domain of $V(H)$ could be extended to $\mathbb{R}^{n \times c}$ though the definition of $V(H)$ is originally null for H outside $\mathcal{H}_{P,Q}$.

4. Multi-class Approximate Volume Regularization

The proposed volume approximation motivates a family of new transductive methods taking it as a regularization. We develop and analyze an instantiation in this section whose optimization problem is non-convex but can be solved exactly and efficiently.

4.1. Model

First of all, we define the label indicator matrix $Y \in \mathbb{R}^{n \times c}$ for convenience whose entries can be from either $\{0, 1\}$ or $\{-1, 0, 1\}$ depending on the problem settings and whether negative labels ever appear. Specifically, we can set $Y_{i,j} = 1$ if x_i is labeled to have the j-th label and $Y_{i,j} = 0$ otherwise, or alternatively we can set $Y_{i,j} = 1$ if x_i is labeled to have the j-th label, $Y_{i,j} = -1$ if x_i is labeled to not have the j-th label, and $Y_{i,j} = 0$ otherwise.

Let $\Delta(Y, H)$ be our loss function measuring the difference between Y and H. The multi-class volume approximation motivates the following family of transductive methods:

$$\min_{H \in \mathcal{H}_{P,Q}} \Delta(Y, H) + \gamma \cdot \frac{\operatorname{tr}(H^\top Q H P)}{\|H\|_{\mathrm{Fro}}^2},$$

where $\gamma > 0$ is a regularization parameter. The denominator $\|H\|_{\mathrm{Fro}}^2$ is annoying so we would like to get rid of it as in El-Yaniv et al. (2008) or Niu et al. (2013a). We fix $\tau > 0$ as a scale parameter, constrain H to be of norm τ, replace the feasible region $\mathcal{H}_{P,Q}$ with $\mathbb{R}^{n \times c}$, and it becomes

$$\begin{aligned} \min_{H \in \mathbb{R}^{n \times c}} \ & \Delta(Y, H) + \gamma \operatorname{tr}(H^\top Q H P) \\ \text{s.t.} \ & \|H\|_{\mathrm{Fro}} = \tau. \end{aligned} \quad (8)$$

Although the optimization is done in $\mathbb{R}^{n \times c}$, the regularization is relative to $\mathcal{H}_{P,Q}$, since $\operatorname{tr}(H^\top Q H P)$ is a weighted sum of squared cosines between $\operatorname{vec}(H)$ and the principal axes of $\mathcal{E}_{P,Q}$ under the constraint $\|H\|_{\mathrm{Fro}} = \tau$.

Subsequently, we denote by $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ and $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ the c-dimensional vectors that satisfy $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$ and $H = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)^\top$. Consider the following loss functions to be $\Delta(Y, H)$ in optimization (8):

1. Squared losses over all data $\sum_{X_n} \|\boldsymbol{y}_i - \boldsymbol{r}_i\|_2^2$;
2. Squared losses over labeled data $\sum_{X_l} \|\boldsymbol{y}_i - \boldsymbol{r}_i\|_2^2$;
3. Linear losses over all data $\sum_{X_n} -\boldsymbol{y}_i^\top \boldsymbol{r}_i$;
4. Linear losses over labeled data $\sum_{X_l} -\boldsymbol{y}_i^\top \boldsymbol{r}_i$;

They or their binary counterparts have been used in Zhou et al. (2003), El-Yaniv et al. (2008) and Niu et al. (2013a). Actually, the third and fourth ones are identical since $\boldsymbol{y}_i$ is zero for $x_i \in X_u$, and the first one is equivalent to them

in (8) since $\sum_{X_n} \|\boldsymbol{y}_i\|_2^2$ and $\sum_{X_l} \|\boldsymbol{y}_i\|_2^2$ are constants and $\sum_{X_n} \|\boldsymbol{r}_i\|_2^2 = \tau^2$ is also a constant. The second one is undesirable due to an issue of the time complexity. Thus, we instantiate $\Delta(Y, H) := \sum_{X_n} \|\boldsymbol{y}_i - \boldsymbol{r}_i\|_2^2 = \|Y - H\|_{\text{Fro}}^2$, and optimization (8) becomes

$$\begin{aligned} \min_{H \in \mathbb{R}^{n \times c}} \ & \|Y - H\|_{\text{Fro}}^2 + \gamma \operatorname{tr}(H^\top Q H P) \\ \text{s.t.} \ & \|H\|_{\text{Fro}} = \tau. \end{aligned} \tag{9}$$

We refer to constrained optimization problem (9) as *multi-class approximate volume regularization* (MAVR). An unconstrained version of MAVR is then

$$\min_{H \in \mathbb{R}^{n \times c}} \|Y - H\|_{\text{Fro}}^2 + \gamma \operatorname{tr}(H^\top Q H P). \tag{10}$$

4.2. Algorithm

Optimization (9) is non-convex, but we can rewrite it using the stacked soft response vector $\boldsymbol{h} = \operatorname{vec}(H)$ as

$$\min_{\boldsymbol{h} \in \mathbb{R}^{nc}} \|\boldsymbol{y} - \boldsymbol{h}\|_2^2 + \gamma \boldsymbol{h}^\top (P \otimes Q) \boldsymbol{h} \quad \text{s.t.} \ \|\boldsymbol{h}\|_2 = \tau, \tag{11}$$

where $\boldsymbol{y} = \operatorname{vec}(Y)$ is the vectorization of Y. In this representation, the objective is a second-degree polynomial and the constraint is an origin-centered sphere, and fortunately we could solve it exactly and efficiently following Forsythe & Golub (1965). To this end, a fundamental property of the Kronecker product is necessary (see, e.g., Theorems 13.10 and 13.12 of Laub, 2005):

Theorem 1. *Let $\lambda_{Q,1} \leq \cdots \leq \lambda_{Q,n}$ be the eigenvalues and $\boldsymbol{v}_{Q,1}, \ldots, \boldsymbol{v}_{Q,n}$ be the associated orthonormal eigenvectors of Q, $\lambda_{P,1} \leq \cdots \leq \lambda_{P,c}$ and $\boldsymbol{v}_{P,1}, \ldots, \boldsymbol{v}_{P,c}$ be those of P, and the eigen-decompositions of Q and P be $Q = V_Q \Lambda_Q V_Q^\top$ and $P = V_P \Lambda_P V_P^\top$. Then, the eigenvalues of $P \otimes Q$ are $\lambda_{P,j}\lambda_{Q,i}$ associated with orthonormal eigenvectors $\boldsymbol{v}_{P,j} \otimes \boldsymbol{v}_{Q,i}$ for $j = 1, \ldots, c$, $i = 1, \ldots, n$, and the eigen-decomposition of $P \otimes Q$ is $P \otimes Q = V_{PQ} \Lambda_{PQ} V_{PQ}^\top$, where $\Lambda_{PQ} = \Lambda_P \otimes \Lambda_Q$ and $V_{PQ} = V_P \otimes V_Q$.*

After we ignore the constants $\|\boldsymbol{y}\|_2^2$ and $\|\boldsymbol{h}\|_2^2$ in the objective of optimization (11), the Lagrange function is

$$\Phi(\boldsymbol{h}, \rho) = -2\boldsymbol{h}^\top \boldsymbol{y} + \gamma \boldsymbol{h}^\top (P \otimes Q) \boldsymbol{h} - \rho(\boldsymbol{h}^\top \boldsymbol{h} - \tau^2),$$

where $\rho \in \mathbb{R}$ is the Lagrangian multiplier for $\|\boldsymbol{h}\|_2^2 = \tau^2$. The stationary conditions are

$$\partial \Phi / \partial \boldsymbol{h} = -\boldsymbol{y} + \gamma (P \otimes Q) \boldsymbol{h} - \rho \boldsymbol{h} = \boldsymbol{0}_{nc}, \tag{12}$$

$$\partial \Phi / \partial \rho = \boldsymbol{h}^\top \boldsymbol{h} - \tau^2 = 0. \tag{13}$$

Hence, for any locally optimal solution $(\boldsymbol{h}, \rho)$ where ρ/γ is not an eigenvalue of $P \otimes Q$, we have

$$\begin{aligned} \boldsymbol{h} &= (\gamma P \otimes Q - \rho I_{nc})^{-1} \boldsymbol{y} && (14) \\ &= V_{PQ} (\gamma \Lambda_{PQ} - \rho I_{nc})^{-1} V_{PQ}^\top \boldsymbol{y} \\ &= (V_P \otimes V_Q)(\gamma \Lambda_{PQ} - \rho I_{nc})^{-1} \operatorname{vec}(V_Q^\top Y V_P) && (15) \end{aligned}$$

Algorithm 1 MAVR

Input: P, Q, Y, γ and τ
Output: H and ρ

1: Eigen-decompose P and Q;
2: Construct the function $g(\rho)$;
3: Find the smallest root of $g(\rho)$;
4: Recover $\boldsymbol{h}$ using ρ and reshape $\boldsymbol{h}$ to H.

based on Eq. (12) and Theorem 1. Next, we search for the feasible ρ for (12) and (13) which will lead to the globally optimal $\boldsymbol{h}$. Let $\boldsymbol{z} = \operatorname{vec}(V_Q^\top Y V_P)$, then plugging (15) into (13) gives us

$$\boldsymbol{z}^\top (\gamma \Lambda_{PQ} - \rho I_{nc})^{-2} \boldsymbol{z} - \tau^2 = 0. \tag{16}$$

Let us sort the eigenvalues $\lambda_{P,1}\lambda_{Q,1}, \ldots, \lambda_{P,c}\lambda_{Q,n}$ into a non-descending sequence $\{\lambda_{PQ,1}, \ldots, \lambda_{PQ,nc}\}$, rearrange $\{z_1, \ldots, z_{nc}\}$ accordingly, and find the smallest k_0 which satisfies $z_{k_0} \neq 0$. As a result, Eq. (16) implies that

$$g(\rho) = \sum_{k=k_0}^{nc} \frac{z_k^2}{(\gamma \lambda_{PQ,k} - \rho)^2} - \tau^2 = 0 \tag{17}$$

for any stationary ρ. By Theorem 4.1 of Forsythe & Golub (1965), the smallest root of $g(\rho)$ determines a unique $\boldsymbol{h}$ so that $(\boldsymbol{h}, \rho)$ is the globally optimal solution to $\Phi(\boldsymbol{h}, \rho)$, i.e., $\boldsymbol{h}$ minimizes the objective of (11) globally. Here, the only exception where we cannot determine $\boldsymbol{h}$ by Eq. (14) for a specific value of ρ is when ρ/γ is an eigenvalue of $P \otimes Q$. This, however, happens with probability zero. The theorem below points out the location of the optimal ρ (the proof is in the appendix):

Theorem 2. *The function $g(\rho)$ defined in Eq. (17) has exactly one root in the interval $[\rho_0, \gamma\lambda_{PQ,k_0})$ and no root in the interval $(-\infty, \rho_0)$, where $\rho_0 = \gamma\lambda_{PQ,k_0} - \|\boldsymbol{y}\|_2/\tau$.*

The algorithm of MAVR is summarized in Algorithm 1. It is easy to see that fixing $\rho = -1$ in Algorithm 1 instead of finding the smallest root of $g(\rho)$ suffices to solve optimization (10). Moreover, for a special case $P = I_c$ where I_c is the identity matrix of size c, any stationary H is simply

$$H = (\gamma Q - \rho I_n)^{-1} Y = V_Q (\gamma \Lambda_Q - \rho I_n)^{-1} V_Q^\top Y.$$

Let $\boldsymbol{z} = V_Q^\top Y \boldsymbol{1}_c$ where $\boldsymbol{1}_c$ means the all-one vector in $\mathbb{R}^c$, and k_0 is the smallest number that satisfies $z_{k_0} \neq 0$. Then the smallest root of $g(\rho) = \sum_{k=k_0}^{n} z_k^2/(\gamma\lambda_{Q,k} - \rho)^2 - \tau^2$ gives us the feasible ρ leading to the globally optimal H.

The asymptotic time complexity of Algorithm 1 is $O(n^3)$. More specifically, eigen-decomposing Q in the first step of Algorithm 1 costs $O(n^3)$, and this is the dominating computation time. Eigen-decomposing P just needs $O(c^3)$ and is negligible under the assumption that $n \gg c$ without loss

of generality. In the second step, it requires $O(nc\log(nc))$ for sorting the eigenvalues of $P \otimes Q$ and $O(n^2c)$ for computing $\boldsymbol{z}$. Finding the smallest root of $g(\rho)$ based on a binary search algorithm uses $O(\log(\|\boldsymbol{y}\|_2))$ in the third step. In the fourth step, recovering $\boldsymbol{h}$ is essentially same as computing $\boldsymbol{z}$ and costs $O(n^2c)$.

We would like to comment a little more on the asymptotic time complexity of MAVR. Given fixed P and Q but different Y, γ and τ, the computational complexity is $O(n^2c)$ if we reuse the eigen-decompositions of P and Q and the sorted eigenvalues of $P \otimes Q$. This property is particularly advantageous for trying different hyperparameters. It is also quite useful for choosing different $X_l \subset X_n$ to be labeled following transductive problem settings. Finally, the asymptotic time complexity $O(n^3)$ for solving MAVR exactly can hardly be improved based on existing techniques. Even if ρ is fixed in optimization (10), the stationary condition Eq. (12) is a *discrete Sylvester equation* which consumes $O(n^3)$ for solving it (Sima, 1996).

4.3. Theoretical analyses

We provide two theoretical results. Under certain assumptions, the stability analysis upper bounds the difference of two optimal H and H' trained with two different label indicator matrices Y and Y', and the error analysis bounds the difference of H from the ground truth.

Theorem 2 guarantees that $\rho < \gamma\lambda_{PQ,k_0}$. In fact, with high probability over the choice of Y, it holds that $k_0 = 1$ (we did not meet $k_0 > 1$ in our experiments). For this reason, we make the following assumption:
Fix P and Q, and allow Y to change according to the partition of X_n into different X_l and X_u. There is $C_{\gamma,\tau} > 0$, which just depends on γ and τ, such that for all optimal ρ trained with different Y, $\rho \le \gamma\lambda_{PQ,1} - C_{\gamma,\tau}$.

Note that for unconstrained MAVR, there must be $C_{\gamma,\tau} > 1$ since $\gamma\lambda_{PQ,1} > 0$ and $\rho = -1$. We can prove the theorem below based on the assumption above and the lower bound of ρ in Theorem 2.

Theorem 3 (Stability of MAVR)**.** *Assume the existence of $C_{\gamma,\tau}$. Let (H,ρ) and (H',ρ') be two globally optimal solutions trained with two different label indicator matrices Y and Y' respectively. Then,*

$$\begin{aligned}\|H - H'\|_{\mathrm{Fro}} \le \|Y - Y'\|_{\mathrm{Fro}}/C_{\gamma,\tau} \\ + |\rho - \rho'|\min\{\|Y\|_{\mathrm{Fro}}, \|Y'\|_{\mathrm{Fro}}\}/C^2_{\gamma,\tau}.\end{aligned} \tag{18}$$

Consequently, for MAVR in optimization (9) we have

$$\begin{aligned}\|H - H'\|_{\mathrm{Fro}} \le \|Y - Y'\|_{\mathrm{Fro}}/C_{\gamma,\tau} \\ + \|Y\|_{\mathrm{Fro}}\|Y'\|_{\mathrm{Fro}}/\tau C^2_{\gamma,\tau},\end{aligned}$$

and for unconstrained MAVR in optimization (10) we have

$$\|H - H'\|_{\mathrm{Fro}} \le \|Y - Y'\|_{\mathrm{Fro}}/C_{\gamma,\tau}.$$

In order to present an error analysis, we assume there is a ground-truth soft response matrix H^* with two properties. Firstly, the value of $V(H^*)$ should be bounded, namely,

$$V(H^*) = \mathrm{tr}(H^{*\top}QH^*P)/\|H^*\|^2_{\mathrm{Fro}} \le C_h,$$

where $C_h > 0$ is a small number. This ensures that H^* lies in a large-volume region. Otherwise MAVR implementing the large volume principle can by no means learn some H close to H^*. Secondly, Y should contain certain information about H^*. MAVR makes use of P, Q and Y only and the meanings of P and Q are fixed already, so MAVR may access the information about H^* only through Y. To make Y and H^* correlated, we assume that $Y = H^* + E$ where $E \in \mathbb{R}^{n\times c}$ is a noise matrix of the same size as Y and H^*. All entries of E are independent with zero mean, and the variance of them is σ_l or σ_u depending on its correspondence to a labeled or an unlabeled position in Y. We could expect that $\sigma_l \ll \sigma_u$, such that the entries of Y in labeled positions are close to the corresponding entries of H^*, but the entries of Y in unlabeled positions are completely corrupted and uninformative for recovering H^*. Notice that we need this generating mechanism of Y even if C_h/γ is the smallest eigenvalue of $P \otimes Q$, since $P \otimes Q$ may have multiple smallest eigenvalues and $\pm H$ have totally different meanings. Based on these assumptions, we can prove the theorem below.

Theorem 4 (Accuracy of MAVR)**.** *Assume the existence of $C_{\gamma,\tau}$, C_h, and the generating process of Y from H^* and E. Let $\widetilde{l}$ and $\widetilde{u}$ be the numbers of the labeled and unlabeled positions in Y and assume that $\mathbb{E}_E\|Y\|^2_{\mathrm{Fro}} \le \widetilde{l}$ where the expectation is with respect to the noise matrix E. For each possible Y, let H be the globally optimal solution trained with it. Then,*

$$\begin{aligned}\mathbb{E}_E\|H - H^*\|_{\mathrm{Fro}} \le (\sqrt{C_h}\gamma\lambda_{PQ,1}/C_{\gamma,\tau})\|H^*\|_{\mathrm{Fro}} \\ + (\max\{\sqrt{\widetilde{l}}/\tau - \gamma\lambda_{PQ,1} - 1, \gamma\lambda_{PQ,1} - C_{\gamma,\tau} + 1\}/C_{\gamma,\tau}) \\ \cdot \|H^*\|_{\mathrm{Fro}} + \sqrt{\widetilde{l}\sigma_l^2 + \widetilde{u}\sigma_u^2}/C_{\gamma,\tau}\end{aligned} \tag{19}$$

for MAVR in optimization (9), and

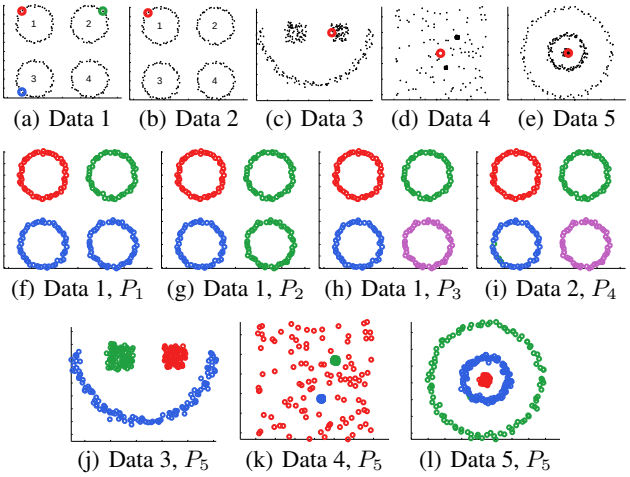
$$\mathbb{E}_E\|H - H^*\|^2_{\mathrm{Fro}} \le (C_h/4)\|H^*\|^2_{\mathrm{Fro}} + \widetilde{l}\sigma_l^2 + \widetilde{u}\sigma_u^2 \tag{20}$$

for unconstrained MAVR in optimization (10).

The proofs of Theorems 3 and 4 are in the appendix. Considering the instability bounds in Theorem 3 and the error bounds in Theorem 4, unconstrained MAVR is superior to constrained MAVR in both cases. That being said, bounds are just bounds. We will demonstrate the potential of constrained MAVR in the next section by experiments.

5. Experiments

In this section, we numerically evaluate MAVR. The baseline methods include the one-vs-rest extension of the *bina-*



(a) Data 1 (b) Data 2 (c) Data 3 (d) Data 4 (e) Data 5

(f) Data 1, P_1 (g) Data 1, P_2 (h) Data 1, P_3 (i) Data 2, P_4

(j) Data 3, P_5 (k) Data 4, P_5 (l) Data 5, P_5

Figure 2. Serendipitous learning by MAVR.

ry approximate volume regularization (BAVR) as well as a multi-class transductive method named *learning with local and global consistency* (LGC) (Zhou et al., 2003).

5.1. Serendipitous learning

We show how to handle serendipitous problems by MAVR directly without performing clustering (Hartigan & Wong, 1979; Ng et al., 2001; Sugiyama et al., 2014) or estimating the class-prior change (du Plessis & Sugiyama, 2012). The experimental results are displayed in Figure 2. There are 5 artificial data sets in total where the latter 3 data sets come from Zelnik-Manor & Perona (2004). The matrix Q was specified as the *normalized graph Laplacian* (see, e.g., von Luxburg, 2007)[2]

$$Q = L_{\mathrm{nor}} = I_n - D^{-1/2} W D^{-1/2},$$

where $W \in \mathbb{R}^{n \times n}$ is a similarity matrix and $D \in \mathbb{R}^{n \times n}$ is the degree matrix of W. The matrix P was specified by

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \; P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \; P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 \end{pmatrix},$$

$$P_4 = \begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 2 & 0 & 1/2 \\ 1/2 & 0 & 2 & 1/2 \\ 1/2 & 1/2 & 1/2 & 3 \end{pmatrix}, \; P_5 = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}.$$

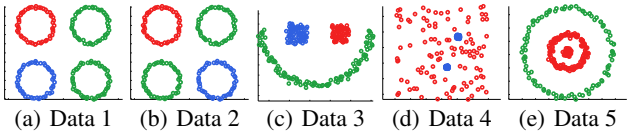
For data sets 1 and 2 we used the Gaussian similarity

$$W_{i,j} = \exp(-\|x_i - x_j\|_2^2/(2\sigma^2)), \; W_{i,i} = 0$$

with the kernel width $\sigma = 0.25$, and for data sets 3 to 5 we applied the local-scaling similarity (Zelnik-Manor & Perona, 2004)

$$W_{i,j} = \exp(-\|x_i - x_j\|_2^2/(2\sigma_i\sigma_j)), \; W_{i,i} = 0$$

[2]Though the graph Laplacian matrices have zero eigenvalues, they would not cause algorithmic problems when used as Q.



(a) Data 1 (b) Data 2 (c) Data 3 (d) Data 4 (e) Data 5

Figure 3. Serendipitous learning by LGC.

with the nearest-neighbor number $k = 7$, where each $\sigma_i = \|x_i - x_i^{(k)}\|_2$ is the scale parameter of x_i and $x_i^{(k)}$ is the k-th nearest neighbor of x_i in X_n. For the hyperparameters, we set $\gamma = 99$ and $\tau = \sqrt{l}$. Furthermore, a class-balance regularization was imposed for data sets 2 to 5, which tries to minimize

$$\gamma' \operatorname{tr}(H^\top (\mathbf{1}_n \mathbf{1}_n^\top) H (I_c - \mathbf{1}_c \mathbf{1}_c^\top / c)).$$

The detailed derivation is omitted due to the limited space, but the idea is to encourage balanced total responses among c classes. For this regularization, the regularization parameter was set to $\gamma' = 1$.

We can see that in Figure 2, MAVR successfully classified the data belonging to the known classes and simultaneously clustered the data belonging to the unknown classes. Moreover, we can control the influence of the known classes on the unknown classes by specifying different P, as shown in subfigures (f), (g) and (h) of Figure 2. On the other hand, BAVR cannot benefit from the class-balance regularization and LGC with the class-balance regularization for data sets 2 to 5 in Figure 3 was not as perfect as MAVR.

5.2. Multi-class learning

As commented in the end of our theoretical analyses, we would demonstrate the potential of constrained MAVR by experiments. Actually, LGC could be subsumed in MAVR as a special case of unconstrained MAVR: Although LGC is motivated by the label propagation point of view, it can be rewritten as the following optimization problem

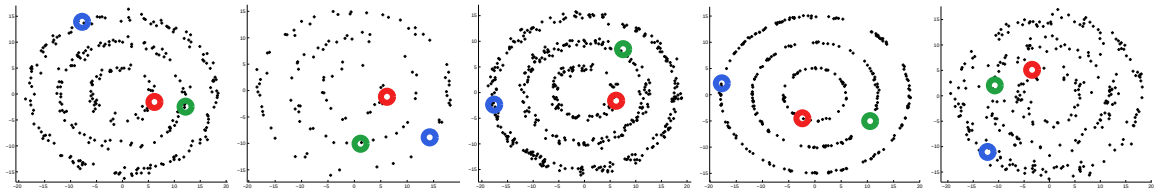
$$\min_{H \in \mathbb{R}^{n \times c}} \|Y - H\|_{\mathrm{Fro}}^2 + \gamma \operatorname{tr}(H^\top L_{\mathrm{nor}} H).$$

Therefore, unconstrained MAVR will be reduced exactly to LGC if $P = I_c$ and $Q = L_{\mathrm{nor}}$. Now we specify $P = I_c$ and $Q = L_{\mathrm{nor}}$ and illustrate the nuance of constrained MAVR, LGC, and BAVR using an artificial data set.

The artificial data set *3circles* is generated as follows. We have three classes with the class ratio $1/6 : 1/3 : 1/2$. Let y_i be the ground-truth label of x_i, then x_i is generated by

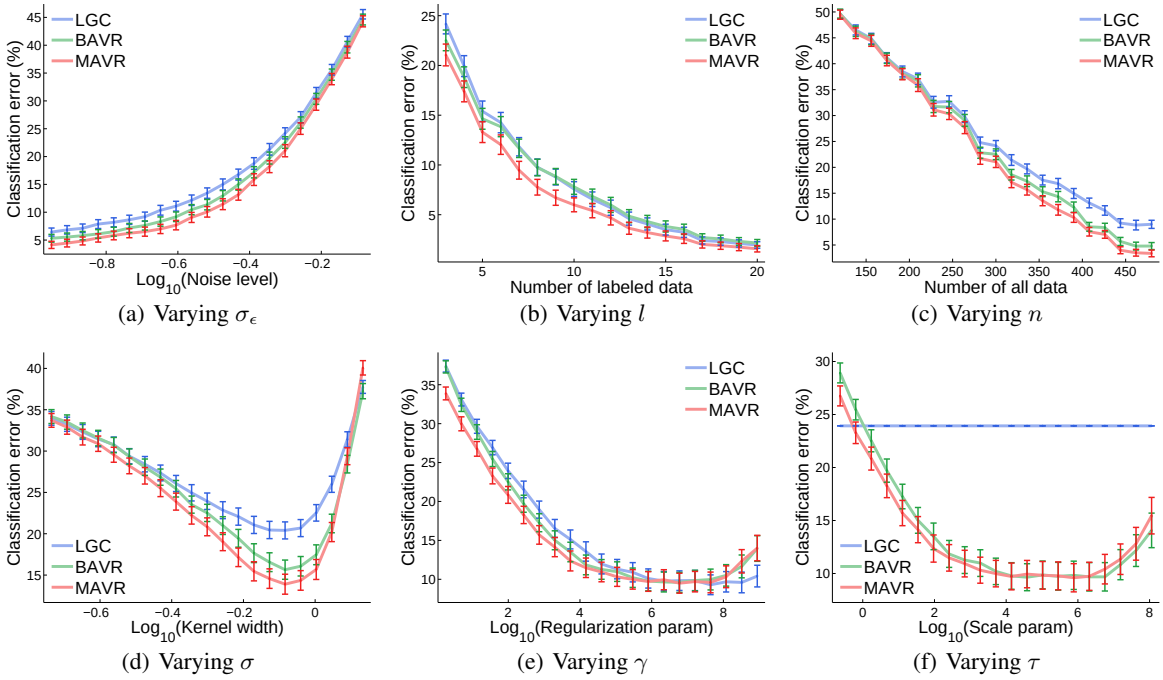
$$x_i = (6y_i \cos(a_i) + \epsilon_{i,1}, 5y_i \sin(a_i) + \epsilon_{i,2})^\top \in \mathbb{R}^2,$$

where a_i is an angle drawn i.i.d. from the uniform distribution $\mathcal{U}(0, 2\pi)$, and $\epsilon_{i,1}$ and $\epsilon_{i,2}$ are noises drawn i.i.d. from the normal distribution $\mathcal{N}(0, \sigma_\epsilon^2)$. In our experiments, we varied one factor while fixed all other factors. The default



(a) Default (b) Small n (c) Large n (d) Small σ_ϵ (e) Large σ_ϵ

Figure 4. Visualization of the artificial data set 3circles.



(a) Varying σ_ϵ (b) Varying l (c) Varying n

(d) Varying σ (e) Varying γ (f) Varying τ

Figure 5. Means with standard errors of LGC, BAVR and MAVR on 3circles.

values of factors were $\sigma_\epsilon = 0.5$, $\sigma = 0.5$, $l = 3$, $n = 300$, $\gamma = 99$, and $\tau = \sqrt{l}$, and the ranges of these factors were

- $\sigma_\epsilon \in 0.5 \cdot \exp\{-1.5, -1.4, -1.3 \ldots, 0.5\}$;
- $l \in \{3, 4, 5, \ldots, 20\}$;
- $n \in \{120, 138, 156, \ldots, 480\}$;
- $\sigma \in 0.5 \cdot \exp\{-1, -0.9, -0.8, \ldots, 1\}$;
- $\gamma \in 99 \cdot \exp\{-4, -3, -2, \ldots, 16\}$;
- $\tau \in \sqrt{l} \cdot \exp\{-2, -1, 0, \ldots, 18\}$.

Note that there was a distributional change, since we sampled labeled data as balanced as possible across three classes. Figure 4 exhibits several realizations of 3circles given different values of factors.

Figure 5 shows the experimental results, where the means with the standard errors of the classification error rates are plotted. For each task that corresponds to a full specification of those factors, three methods were repeatedly run on 100 random samplings. We can see from Figure 5 that the performance of LGC or BAVR was usually not as good as MAVR. The drawback of LGC is that we would always have $\rho = -1$ since it is unconstrained. Though ρ is adaptive in BAVR, we would have c different ρ values since it is based on the one-vs-rest extension.

6. Conclusions

We proposed a multi-class volume approximation that can be applied to several transductive problem settings such as multi-class, multi-label and serendipitous learning. The resultant learning method is non-convex, but can however be solved exactly and efficiently. The method was theoretically justified by stability and error analyses and empirically demonstrated a promising approach via experiments.

Acknowledgments

GN was supported by the FIRST Program and the 973 Program No. 2014CB340505. MCdP was supported by KAKENHI 23120004, and MS was supported by KAKENHI 25700022 and AOARD.

References

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Bennett, K. and Demiriz, A. Semi-supervised support vector machines. In *NIPS*, 1998.

Blum, A. and Chawla, S. Learning from labeled and unlabeled data using graph mincuts. In *ICML*, 2001.

du Plessis, M. C. and Sugiyama, M. Semi-supervised learning of class balance under class-prior change by distribution matching. In *ICML*, 2012.

El-Yaniv, R., Pechyony, D., and Vapnik, V. Large margin vs. large volume in transductive learning. *Machine Learning*, 72(3):173–188, 2008.

Forsythe, G. and Golub, G. On the stationary values of a second-degree polynomial on the unit sphere. *Journal of the Society for Industrial and Applied Mathematics*, 13(4):1050–1068, 1965.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. In *NIPS*, 2004.

Hartigan, J. A. and Wong, M. A. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

Joachims, T. Transductive learning via spectral graph partitioning. In *ICML*, 2003.

Kong, X., Ng, M., and Zhou, Z.-H. Transductive multilabel learning via label set propagation. *IEEE Transaction on Knowledge and Data Engineering*, 25(3):704–719, 2013.

Laub, A. J. *Matrix Analysis for Scientists and Engineers*. Society for Industrial and Applied Mathematics, 2005.

Li, S. and Ng, W. Maximum volume outlier detection and its applications in credit risk analysis. *International Journal on Artificial Intelligence Tools*, 22(5), 2013.

Li, Y.-F. and Zhou, Z.-H. Towards making unlabeled data never hurt. In *ICML*, 2011.

Li, Y.-F., Kwok, J., and Zhou, Z.-H. Semi-supervised learning using label mean. In *ICML*, 2009.

Ng, A., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

Niu, G., Dai, B., Shang, L., and Sugiyama, M. Maximum volume clustering: A new discriminative clustering approach. *Journal of Machine Learning Research*, 14:2641–2687, 2013a.

Niu, G., Jitkrittum, W., Dai, B., Hachiya, H., and Sugiyama, M. Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning. In *ICML*, 2013b.

Sima, V. *Algorithms for Linear-Quadratic Optimization*. Marcel Dekker, 1996.

Sugiyama, M., Niu, G., Yamada, M., Kimura, M., and Hachiya, H. Information-maximization clustering based on squared-loss mutual information. *Neural Computation*, 26(1):84–131, 2014.

Szummer, M. and Jaakkola, T. Partially labeled classification with Markov random walks. In *NIPS*, 2001.

Vapnik, V. N. *Estimation of Dependences Based on Empirical Data*. Springer Verlag, 1982.

Vapnik, V. N. *Statistical Learning Theory*. John Wiley & Sons, 1998.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Wang, J., Jebara, T., and Chang, S.-F. Semi-supervised learning using greedy max-cut. *Journal of Machine Learning Research*, 14:771–800, 2013.

Yamada, M., Sugiyama, M., and Matsui, T. Semi-supervised speaker identification under covariate shift. *Signal Processing*, 90(8):2353–2361, 2010.

Zelnik-Manor, L. and Perona, P. Self-tuning spectral clustering. In *NIPS*, 2004.

Zhang, D., Liu, Y., and Si, L. Serendipitous learning: Learning beyond the predefined label space. In *KDD*, 2011.

Zhou, D., Bousquet, O., Navin Lal, T., Weston, J., and Schölkopf, B. Learning with local and global consistency. In *NIPS*, 2003.

Zhu, X., Ghahramani, Z., and Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.