The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Optimal Margin Distribution Clustering*

Teng Zhang, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
Collaborative Innovation Center of
Novel Software Technology and Industrialization, Nanjing 210023, China
{zhangt, zhouzh}@lamda.nju.edu.cn

Abstract

Maximum margin clustering (MMC), which borrows the large margin heuristic from support vector machine (SVM), has achieved more accurate results than traditional clustering methods. The intuition is that, for a good clustering, when labels are assigned to different clusters, SVM can achieve a large *minimum margin* on this data. Recent studies, however, disclosed that maximizing the minimum margin does not necessarily lead to better performance, and instead, it is crucial to optimize the *margin distribution*. In this paper, we propose a novel approach ODMC (Optimal margin Distribution Machine for Clustering), which tries to cluster the data and achieve optimal margin distribution simultaneously. Specifically, we characterize the margin distribution by the first- and second-order statistics, i.e., the margin mean and variance, and extend a stochastic mirror descent method to solve the resultant minimax problem. Moreover, we prove theoretically that ODMC has the same convergence rate with state-of-the-art cutting plane based algorithms but involves much less computation cost per iteration, so our method is much more scalable than existing approaches. Extensive experiments on UCI data sets show that ODMC is significantly better than compared methods, which verifies the superiority of optimal margin distribution learning.

Introduction

Clustering is an important research area in machine learning, data mining and pattern recognition that aims at grouping data points which are similar. It arises in a wide range of domains including information retrieval, computer version, bioinformatics, etc., and various clustering algorithms have been proposed over past decades (Jain and Dubes 1988; Xu and Wunsch 2005; Jain 2010).

A recently proposed method for clustering, referred to as maximum margin clustering (MMC), is based on the large margin heuristic of support vector machine (SVM) (Cortes and Vapnik 1995; Vapnik 1995). The intuition is that, for a good clustering, when labels are assigned to different clusters, SVM can achieve a large minimum margin on this data. Since the resultant minimax problem involves labeling each instance from the set $\{+1, -1\}$, it's no longer a convex optimization problem but a mixed-integer programming which is much more difficult to handle. From then on, a lot of efforts have been devoted to solve this problem, which can be roughly classified into two groups. The first group applies various convex relaxation techniques. Xu et al. (2005) first relaxed it as a convex semi-definite programming (SDP), in which a positive semi-definite matrix with linear constraints is used to approximate the matrix of label outer product. Soon after, Valizadegan and Jin (2006) introduced a new formulation, whose number of variables is significantly reduced, although it's still a SDP problem. Finally, Li et al. (2009; 2013) proposed a tighter minimax relaxation than SDP formulation, which can be solved by iteratively generating the most violated labels and then combining them via multiple kernel learning. The second group directly optimizes the original problem via variants of non-convex optimization. Examples include alternative optimization (Zhang, Tsang, and Kwok 2007; 2009), in which clustering is preformed by sequentially finding labels and optimizing a support vector regression (SVR), and constrained convex-concave procedure (CCCP) (Zhao, Wang, and Zhang 2008; Wang, Zhao, and Zhang 2010), in which the non-convex objective function or constraints are expressed as the a difference between two convex functions, and the latter is further replaced by a linear approximation so that the whole is convex. Moreover, several researchers also tried to extend MMC to more general learning settings. For example, Zhou et al. (2013) assumed that the data has latent variables and developed the LMMC framework. Niu et al. (2013) showed an alternative principle to MMC, called maximum volume clustering (MVC), is more theoretically advantageous. The incremental version of MMC is also proposed (Vijaya Saradhi and Charly Abraham 2016).

Aforementioned MMC algorithms are all based on the large margin principle, i.e., trying to maximize the minimum margin of training instances. However, recent studies on margin theory (Gao and Zhou 2013) disclosed that maximizing the minimum margin does not necessarily lead to better performance, and instead, it is crucial to optimize the margin distribution. Inspired by this recognition, Zhang and Zhou (2014; 2016; 2017) proposed optimal margin distribution machine (ODM) which can achieve better gen-

*This research was supported by the 973 Program and NSFC (61333014) and the Program B for outstanding PhD candidate of Nanjing University.

eralization performance than large margin based methods. Later, Zhou and Zhou (2016) extends the idea to an approach which is able to exploit unlabeled data and handle unequal misclassification cost. The success of optimal margin distribution learning suggests that there may still exist large space to further enhance for MMC.

In this paper, we propose a novel approach ODMC (Optimal margin Distribution Machine for Clustering), which tries to cluster the data and achieve optimal margin distribution simultaneously. Specifically, we characterize the margin distribution by the first- and second-order statistics, i.e., the margin mean and variance, and then apply the minimax convex relaxation proposed in (Li et al. 2009), which is proven to be tighter than SDP relaxations, to get a convex reformulation. For the optimization of the resultant minimax problem, we propose a stochastic mirror descent method which can converge quickly in practice. Moreover, we prove theoretically that ODMC has the same convergence rate with state-of-the-art cutting plane based algorithms but involves much less computation cost per iteration, so our method is much more scalable than existing approaches. Extensive experiments on UCI data sets show that ODMC is significantly better than compared methods, which verifies the superiority of optimal margin distribution learning.

The rest of this paper is organized as follows. We first introduce some preliminaries and then present the ODMC method. Next we show the experimental studies. Finally we conclude this paper with future work.

Preliminaries

We start with a simpler scenario of supervised learning. Denote $\mathcal{X}$ as the instance space and $\mathcal{Y} = \{+1, -1\}$ as the label set. Let $\mathcal{D}$ be an unknown (underlying) distribution over $\mathcal{X} \times \mathcal{Y}$. A training set $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ is drawn identically and independently (i.i.d.) according to $\mathcal{D}$. Let $\phi : \mathcal{X} \mapsto \mathbb{H}$ be a feature mapping associated to some positive definite kernel κ. The hypothesis is defined based on the linear model $h(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x})$ and the predicted label of instance $\boldsymbol{x}$ is the sign of $h(\boldsymbol{x})$, then the decision function naturally leads to the definition of margin for a labeled instance, i.e., $\gamma(\boldsymbol{x}, y) = y\boldsymbol{w}^\top \phi(\boldsymbol{x})$ (Cristianini and Shawe-Taylor 2000). Thus h misclassifies $(\boldsymbol{x}, y)$ if and only if it produces a negative margin. Given a hypothesis set $\mathcal{H}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$ and the labeled training set $\mathcal{S}$, our goal is to learn a function $h \in \mathcal{H}$ such that the generalization error $R(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[1_{\mathrm{sign}(h(\boldsymbol{x}))\neq y}]$ is small, where $1_{(\cdot)}$ is the indicator function that returns 1 when the argument holds, and 0 otherwise.

Optimal margin distribution machine

It is well known that SVM employs the large margin principle to select h and tries to maximize the minimum margin of training data, i.e., the smallest distance from the instances to the decision boundary. As a result, the solution of SVM just consists of a small amount of data, that is support vectors (SV), and the rest (non-SVs) are totally ignored, which may be misleading in some situations (Zhou 2014).

A more robust strategy is to consider the whole data, i.e., optimizing the margin distribution. To characterize the distribution, the two most straightforward statistics are the first- and second-order statistics, that is, the margin mean and variance. Moreover, a recent study (Gao and Zhou 2013) on margin theory proved that maximizing the margin mean and minimizing the margin variance simultaneously can yield a tighter generalization bound, so we arrive at the following formulation,

$$\begin{aligned}\min_{\boldsymbol{w},\bar{\gamma},\xi_i,\epsilon_i} \ & \frac{1}{2}\|\boldsymbol{w}\|_{\mathbb{H}}^2 - \eta\bar{\gamma} + \frac{\lambda}{m}\sum_{i=1}^m(\xi_i^2 + \epsilon_i^2),\\ \text{s.t.} \ & \gamma(\boldsymbol{x}_i, y_i) \geq \bar{\gamma} - \xi_i,\\ & \gamma(\boldsymbol{x}_i, y_i) \leq \bar{\gamma} + \epsilon_i, \ \forall i,\end{aligned}$$

where $\bar{\gamma}$ is the margin mean, η and λ are trading-off parameters, ξ_i and ϵ_i are the deviation of $\gamma(\boldsymbol{x}_i, y_i)$ to the margin mean. It's evident that $\sum_{i=1}^m(\xi_i^2 + \epsilon_i^2)/m$ is exactly the margin variance.

First, by scaling $\boldsymbol{w}$ which doesn't affect the final classification results, the margin mean can be fixed as 1, then the deviation of $\gamma(\boldsymbol{x}_i, y_i)$ to the margin mean is $|y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) - 1|$. Secondly, the hyperplane $y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) = 1$ divides the feature space into two parts and for each instance, no matter which part it lies in, it will suffer a loss which is quadratic with the deviation. So it is more reasonable to set different weights for the two kinds of deviations because the instances lie in $\{\boldsymbol{x} \mid y\boldsymbol{w}^\top\phi(\boldsymbol{x}) < 1\}$ are much easier to be misclassified than the other. Thirdly, according to representer theorem (Schölkopf and Smola 2001), the optimal solution is spanned only by SVs. To achieve a sparse solution, we introduce a θ-insensitive loss like SVR, i.e., the instances whose deviation is smaller than θ are tolerated and only those whose deviation is larger than θ will suffer a loss. Finally, we obtain the formulation of ODM,

$$\begin{aligned}\min_{\boldsymbol{w},\xi_i,\epsilon_i} \ & \frac{1}{2}\|\boldsymbol{w}\|_{\mathbb{H}}^2 + \frac{\lambda}{m}\sum_{i=1}^m\frac{\xi_i^2 + \nu\epsilon_i^2}{(1-\theta)^2},\\ \text{s.t.} \ & y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \geq 1 - \theta - \xi_i,\\ & y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \leq 1 + \theta + \epsilon_i, \ \forall i.\end{aligned} \tag{1}$$

where ν is a parameter for trading-off different kinds of deviations, θ is a parameter for controlling the sparsity of the solution, and $(1-\theta)^2$ in the denominator is to scale the second term to be a surrogate loss for 0-1 loss.

Optimal margin distribution clustering

In clustering setting, labels are no longer available, and so also need to be optimized. Let $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_m] \in \{\pm 1\}^m$ denotes a vector of the unknown labels. The basic idea of ODMC is to minimize the objective function in Eq. (1) w.r.t. both the labeling $\hat{\boldsymbol{y}}$ and decision function parameter $\boldsymbol{w}$, ξ_i, ϵ_i. Hence, Eq. (1) is extended to

$$\begin{aligned}\min_{\hat{\boldsymbol{y}}\in\mathcal{B}} \min_{\boldsymbol{w},\xi_i,\epsilon_i} \ & \frac{1}{2}\|\boldsymbol{w}\|_{\mathbb{H}}^2 + \frac{\lambda}{m}\sum_{i=1}^m\frac{\xi_i^2 + \nu\epsilon_i^2}{(1-\theta)^2},\\ \text{s.t.} \ & \hat{y}_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \geq 1 - \theta - \xi_i,\\ & \hat{y}_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \leq 1 + \theta + \epsilon_i, \ \forall i,\end{aligned} \tag{2}$$

where $\mathcal{B}$ is a set of candidate label assignments obtained from some domain knowledge. For example, when the positive and negative instances are known to be approximately balanced, we can set $\mathcal{B} = \{\hat{\boldsymbol{y}} \mid -\beta \leq \boldsymbol{e}^\top \hat{\boldsymbol{y}} \leq \beta\}$ where β is a small constant controlling the class imbalance. When a set of "must-link" constraints M which requires two instances should be associated with the same cluster, or a set of "cannot-link" constraints C which requires two instances should be associated with different clusters, are provided, we can set $\mathcal{B} = \{\hat{\boldsymbol{y}} \mid \hat{y}_i = \hat{y}_j, \hat{y}_j \neq \hat{y}_k, \forall(\boldsymbol{x}_i, \boldsymbol{x}_j) \in M, \forall(\boldsymbol{x}_j, \boldsymbol{x}_k) \in C\}$.

To avoid the curse of dimensionality, the inner minimization problem of Eq. (2) is usually cast in the dual form. Denote $\boldsymbol{X}$ as the data matrix whose i-th column is $\phi(\boldsymbol{x}_i)$, i.e., $\boldsymbol{X} = [\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_m)]$, and introduce the dual variables $\boldsymbol{\alpha} \succeq \mathbf{0}$, the Lagrangian of Eq. (2) leads to

$$\begin{aligned}\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}\max_{\boldsymbol{\alpha}\succeq\mathbf{0}}\ & -\frac{1}{2}\boldsymbol{\alpha}^\top \begin{bmatrix} \boldsymbol{K}\odot\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top & -\boldsymbol{K}\odot\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top \\ -\boldsymbol{K}\odot\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top & \boldsymbol{K}\odot\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top\end{bmatrix}\boldsymbol{\alpha} \\ & - \frac{m(1-\theta)^2}{4\lambda}\boldsymbol{\alpha}^\top\begin{bmatrix}\boldsymbol{I} & \mathbf{0}\\ \mathbf{0} & \frac{1}{\nu}\boldsymbol{I}\end{bmatrix}\boldsymbol{\alpha} - \begin{bmatrix}(\theta-1)\boldsymbol{e}\\(\theta+1)\boldsymbol{e}\end{bmatrix}^\top\boldsymbol{\alpha},\end{aligned} \tag{3}$$

where $\boldsymbol{K} = \boldsymbol{X}^\top\boldsymbol{X}$ is the kernel matrix, $\odot$ denotes the element-wise product, and $\boldsymbol{e}$ stands for the all-one vector. Note that the objective function is a negative definite quadratic form whose stationary point can't locate at the infinity, so we can replace the constraint $\{\boldsymbol{\alpha} \mid \boldsymbol{\alpha} \succeq \mathbf{0}\}$ by a bounded box $\mathcal{A} = \{\boldsymbol{\alpha} \mid \mathbf{0} \preceq \boldsymbol{\alpha} \preceq \tau\boldsymbol{e}\}$, where the auxiliary parameter τ is introduced for the sake of mathematical soundness. For a sufficiently large τ, the new problem is equal to the original problem.

To overcome the difficulty of this mixed-integer programming, many relaxations have been proposed, among which the minimax convex relaxation proposed in (Li et al. 2009; 2013) is proven to be the tightest. So in this paper, we also employ this method to deal with the mixed-integer problem, i.e., interchanging the order of $\max_{\boldsymbol{\alpha}\in\mathcal{A}}$ and $\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}$, then we can obtain

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}\ G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}),$$

where $G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}})$ is the objective function of Eq. (3), and this can be further transformed into

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\delta}\ (-\delta) \quad \text{s.t. } G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k) \geq \delta,\ \forall \hat{\boldsymbol{y}}_k \in \mathcal{B}. \tag{4}$$

For the inner optimization in Eq. (4), introduce the dual variables $\boldsymbol{\mu}^\top = [\mu_1, \ldots, \mu_{|\mathcal{B}|}] \succeq \mathbf{0}$, the Lagrangian leads to

$$\max_{\boldsymbol{\mu}\succeq\mathbf{0}}\min_{\delta}\ \{-\delta - \sum_{k:\hat{\boldsymbol{y}}_k\in\mathcal{B}} \mu_k(G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k) - \delta)\},$$

By setting the partial derivative of δ to zero, we can obtain $\sum_{k:\hat{\boldsymbol{y}}_k\in\mathcal{B}} \mu_k = 1$ and the dual turns into

$$\max_{\boldsymbol{\mu}\in\mathcal{M}}\ \{-\sum_{k:\hat{\boldsymbol{y}}_k\in\mathcal{B}} \mu_k G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k)\}, \tag{5}$$

where $\mathcal{M} = \{\boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{B}|} \mid \boldsymbol{e}^\top\boldsymbol{\mu} = 1\}$ is the simplex in $\mathbb{R}^{|\mathcal{B}|}$. By substituting Eq. (5) into Eq. (4) and denoting $\varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) = \sum_{k:\hat{\boldsymbol{y}}_k\in\mathcal{B}} \mu_k G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k)$, Eq. (4) can be rewritten as

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\boldsymbol{\mu}\in\mathcal{M}}\ \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}).$$

Note that $\varphi(\boldsymbol{\mu}, \boldsymbol{\alpha})$ is a convex combination of negative definite quadratic forms, so it's convex in $\boldsymbol{\mu}$ and concave in $\boldsymbol{\alpha}$. According to Sion's minimax theorem (Sion 1958), there exists a saddle point $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) \in \mathcal{M} \times \mathcal{A}$ such that

$$\begin{aligned}\min_{\boldsymbol{\mu}\in\mathcal{M}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) &\leq \max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\hat{\boldsymbol{\mu}}, \boldsymbol{\alpha}) = \varphi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) \\ &= \min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}}) \leq \max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}),\end{aligned} \tag{6}$$

By combining with the following minimax inequality (Kim and Boyd 2008),

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) \leq \min_{\boldsymbol{\mu}\in\mathcal{M}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}),$$

we can realize that all the equalities hold in Eq. (6) and arrive at the final formulation of ODMC:

$$\min_{\boldsymbol{\mu}\in\mathcal{M}}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\ \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}). \tag{7}$$

Optimization

In this section, we commence with a simple introduction to minimax problem, followed by a stochastic mirror descent method to find the saddle point. Finally we give a convergence rate analysis.

Minimax problem

Since $\varphi(\cdot, \boldsymbol{\alpha})$ is convex and $\varphi(\boldsymbol{\mu}, \cdot)$ is concave, according to the convex inequality, for any pair $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}}) \in \mathcal{M} \times \mathcal{A}$ we have

$$\begin{aligned}\varphi(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}}) - \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) &\leq \partial_{\boldsymbol{\mu}}\varphi(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}})^\top(\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}),\ \forall \boldsymbol{\mu} \in \mathcal{M}, \\ \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \varphi(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}}) &\leq -\partial_{\boldsymbol{\alpha}}\varphi(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}})^\top(\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}),\ \forall \boldsymbol{\alpha} \in \mathcal{A}.\end{aligned}$$

By adding the above two inequalities together we have

$$\varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \leq g(\bar{\boldsymbol{w}})^\top(\bar{\boldsymbol{w}} - \boldsymbol{w}),\ \forall \boldsymbol{\mu}, \boldsymbol{\alpha}, \tag{8}$$

where $\boldsymbol{w} = (\boldsymbol{\mu}, \boldsymbol{\alpha}), \bar{\boldsymbol{w}} = (\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}}) \in \mathcal{M} \times \mathcal{A}$ and $g(\bar{\boldsymbol{w}}) = (\partial_{\boldsymbol{\mu}}\varphi(\bar{\boldsymbol{w}}), -\partial_{\boldsymbol{\alpha}}\varphi(\bar{\boldsymbol{w}}))$. Note that Eq. (8) holds for any $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$, in particular we have

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \leq g(\bar{\boldsymbol{w}})^\top(\bar{\boldsymbol{w}} - \boldsymbol{w}). \tag{9}$$

The left hand side is referred to as the "duality gap", which can be decomposed into two parts, i.e.,

$$\begin{aligned}&\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \\ &= \max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \varphi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) + \varphi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) - \min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}}) \\ &= \underbrace{\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\bar{\boldsymbol{\mu}}, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu}\in\mathcal{M}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha})}_{primal\ gap} \\ &\qquad + \underbrace{\max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \bar{\boldsymbol{\alpha}})}_{dual\ gap}.\end{aligned}$$

As can be seen, the primal gap and the dual gap are both nonnegative and the more closer to the saddle point, the smaller both gaps. So duality gap can be viewed as a measure to evaluate the closeness of current point $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\alpha}})$ to the saddle point $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})$.

Stochastic mirror descent

For ODMC, the feasible set of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are simplex and bounded box, respectively, so the most suitable mirror maps for the two domains are $\Phi_{\mathcal{M}}(\boldsymbol{\mu}) = \sum_k \mu_k \log \mu_k$ and $\Phi_{\mathcal{A}}(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_2^2/2$. Introduce the joint map $\Phi(\boldsymbol{w}) = a\Phi_{\mathcal{M}}(\boldsymbol{\mu}) + b\Phi_{\mathcal{A}}(\boldsymbol{\alpha})$, where a and b are parameters to be specified later. It can be shown that $\nabla\Phi_{\mathcal{M}}(\boldsymbol{\mu}) = \log\boldsymbol{\mu} + \boldsymbol{e}$, $\nabla\Phi_{\mathcal{A}}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}$ and $\nabla\Phi(\boldsymbol{w}) = (a\log\boldsymbol{\mu} + a\boldsymbol{e}, b\boldsymbol{\alpha})$.

At the t-th iteration, we first map $\boldsymbol{w}_t = (\boldsymbol{\mu}_t, \boldsymbol{\alpha}_t)$ into the dual space $\nabla\Phi(\boldsymbol{w}_t) = (a\log\boldsymbol{\mu}_t + a\boldsymbol{e}, b\boldsymbol{\alpha}_t)$, followed by one step of stochastic gradient descent in the dual space,

$$\begin{aligned}&\nabla\Phi(\bar{\boldsymbol{w}}_{t+1}) = \nabla\Phi(\boldsymbol{w}_t) - \eta\widetilde{g}(\boldsymbol{w}_t)\\ &= (a\log\boldsymbol{\mu}_t + a\boldsymbol{e} - \eta\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t), b\boldsymbol{\alpha}_t + \eta\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t))\end{aligned}$$

where $\partial_{\boldsymbol{\mu}}\widetilde{\varphi}$, $\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}$ and $\widetilde{g}$ are the noisy unbiased estimation of $\partial_{\boldsymbol{\mu}}\varphi$, $\partial_{\boldsymbol{\alpha}}\varphi$ and g, respectively, and η is the step size. Next, we map $\nabla\Phi(\bar{\boldsymbol{w}}_{t+1})$ back to the primal space, i.e., to find $\bar{\boldsymbol{w}}_{t+1} = (\boldsymbol{u}_{t+1}, \boldsymbol{v}_{t+1})$ such that

$$\begin{aligned}a\log\boldsymbol{u}_{t+1} + a\boldsymbol{e} &= a\log\boldsymbol{\mu}_t + a\boldsymbol{e} - \eta\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t),\\ b\boldsymbol{v}_{t+1} &= b\boldsymbol{\alpha}_t + \eta\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t)),\end{aligned}$$

which implies that $\boldsymbol{u}_{t+1} = \boldsymbol{\mu}_t\exp(-\eta\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t)/a)$ and $\boldsymbol{v}_{t+1} = \boldsymbol{\alpha}_t + \eta\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t)/b$. Finally, we project $(\boldsymbol{u}_{t+1}, \boldsymbol{v}_{t+1})$ back to $\mathcal{M}\times\mathcal{A}$ based on Kullback-Leibler divergence and Euclidean distance, respectively, i.e., we solve the following two optimization problems:

$$\boldsymbol{\mu} = \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathcal{M}} \boldsymbol{\mu}^\top \log\frac{\boldsymbol{\mu}}{\boldsymbol{u}_{t+1}},\ \boldsymbol{\alpha} = \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathcal{A}} \|\boldsymbol{\alpha} - \boldsymbol{v}_{t+1}\|_2^2,$$

Fortunately, both problems have a closed-form solution. The latter is to project $\boldsymbol{v}_{t+1}$ onto the bounded box, so we have $\boldsymbol{\alpha}_{t+1} = \max\{\min\{\boldsymbol{v}_{t+1}, \tau\boldsymbol{e}\}, \boldsymbol{0}\}$. For the former, the Lagrangian function leads to $\boldsymbol{\mu}^\top\log(\boldsymbol{\mu}/\boldsymbol{u}_{t+1}) + \zeta(\boldsymbol{e}^\top\boldsymbol{\mu} - 1)$, where ζ is the dual variable. By setting the partial derivative of $\boldsymbol{\mu}$ to zero, i.e., $\log(\boldsymbol{\mu}/\boldsymbol{u}_{t+1}) + \boldsymbol{e} + \zeta\boldsymbol{e} = \boldsymbol{0}$, we have $\boldsymbol{\mu}_{t+1} = \boldsymbol{u}_{t+1}\exp(-1-\zeta)$. Since $\boldsymbol{\mu}_{t+1}$ belongs to a simplex, hence $1 = \boldsymbol{e}^\top\boldsymbol{\mu}_{t+1} = \boldsymbol{e}^\top\boldsymbol{u}_{t+1}\exp(-1-\zeta) = \|\boldsymbol{u}_{t+1}\|_1\exp(-1-\zeta)$, which implies that $\exp(-1-\zeta) = 1/\|\boldsymbol{u}_{t+1}\|_1$, thus we have $\boldsymbol{\mu}_{t+1} = \boldsymbol{u}_{t+1}/\|\boldsymbol{u}_{t+1}\|_1$. Figure 1 illustrates one iteration of this procedure.

The remaining question is how to find the stochastic gradient $\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t)$ and $\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t)$. Note that $\varphi(\boldsymbol{\mu},\boldsymbol{\alpha}) = \sum_{k:\hat{\boldsymbol{y}}_k\in\mathcal{B}}\mu_k G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k)$, so we have

$$\begin{aligned}&\partial_{\boldsymbol{\mu}}\varphi(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t) = [G(\boldsymbol{\alpha}_t,\hat{\boldsymbol{y}}_1),\ldots,G(\boldsymbol{\alpha}_t,\hat{\boldsymbol{y}}_{|\mathcal{B}|})],\\ &\partial_{\boldsymbol{\alpha}}\varphi(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t) = [\partial_{\boldsymbol{\alpha}}G(\boldsymbol{\alpha}_t,\hat{\boldsymbol{y}}_1),\ldots,\partial_{\boldsymbol{\alpha}}G(\boldsymbol{\alpha}_t,\hat{\boldsymbol{y}}_{|\mathcal{B}|})]\boldsymbol{\mu}_t.\end{aligned}$$

By uniformly choosing an index i_t from $\{1,2,\ldots,|\mathcal{B}|\}$, we can obtain $\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,i_t) = [0,\ldots,|\mathcal{B}|G(\boldsymbol{\alpha}_t,\hat{\boldsymbol{y}}_{i_t})\ldots,0]$. On the other hand, by randomly sampling an index j_t according to the distribution $\boldsymbol{\mu}_t$ on $\{1,2,\ldots,|\mathcal{B}|\}$, we can obtain $\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,j_t) = \partial_{\boldsymbol{\alpha}}G(\boldsymbol{\alpha}_t,\hat{\boldsymbol{y}}_{j_t})$. It can be shown that

$$\begin{aligned}\mathbb{E}[\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,i_t)\mid\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t] &= \partial_{\boldsymbol{\mu}}\varphi(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t),\\ \mathbb{E}[\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,j_t)\mid\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t] &= \partial_{\boldsymbol{\alpha}}\varphi(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t),\end{aligned}$$

and $\widetilde{g}(\boldsymbol{w}_t) = (\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,i_t), -\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,j_t))$ is an unbiased estimation of $g(\boldsymbol{w}_t)$.

Algorithm 1 summarizes the pseudo-code of ODMC.

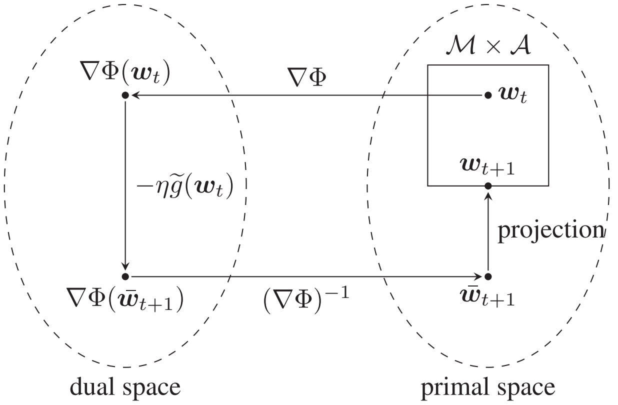
$\nabla\Phi(\boldsymbol{w}_t)$
$\nabla\Phi$
$\mathcal{M}\times\mathcal{A}$
$\boldsymbol{w}_t$
$\boldsymbol{w}_{t+1}$
$-\eta\widetilde{g}(\boldsymbol{w}_t)$
projection
$\nabla\Phi(\bar{\boldsymbol{w}}_{t+1})$
$(\nabla\Phi)^{-1}$
$\bar{\boldsymbol{w}}_{t+1}$
dual space
primal space

Figure 1: Illustration of one iteration of stochastic mirror descent.

Algorithm 1 Stochastic mirror descent for ODMC

```
Input: data set $\boldsymbol{X}$, maximum iteration number $T$, ODM parameters $\lambda$, $\nu$, $\theta$, upper bound $\tau$, stopping criteria $\iota$.
Initialize $\boldsymbol{\mu}_0 \leftarrow [1/|\mathcal{B}|,\ldots,1/|\mathcal{B}|]$, $\boldsymbol{\alpha}_0 \leftarrow \boldsymbol{0}$, $t \leftarrow 0$.
while $t < T$ do
    Uniformly select $i_t$ from $\{1,2,\ldots,|\mathcal{B}|\}$.
    $\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,i_t) \leftarrow [0,\ldots,|\mathcal{B}|G(\boldsymbol{\alpha}_t,\hat{\boldsymbol{y}}_{i_t})\ldots,0]$.
    Select $j_t$ from $\{1,2,\ldots,|\mathcal{B}|\}$ according to $\boldsymbol{\mu}_t$.
    $\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,j_t) \leftarrow \partial_{\boldsymbol{\alpha}}G(\boldsymbol{\alpha}_t,\hat{\boldsymbol{y}}_{j_t})$.
    $\boldsymbol{u}_{t+1} \leftarrow \boldsymbol{\mu}_t\exp(-\eta\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,i_t)/a)$.
    $\boldsymbol{v}_{t+1} \leftarrow \boldsymbol{\alpha}_t + \eta\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{\mu}_t,\boldsymbol{\alpha}_t,j_t)/b$.
    $\boldsymbol{\mu}_{t+1} \leftarrow \boldsymbol{u}_{t+1}/\|\boldsymbol{u}_{t+1}\|_1$.
    $\boldsymbol{\alpha}_{t+1} \leftarrow \max\{\min\{\boldsymbol{v}_{t+1},\tau\boldsymbol{e}\},\boldsymbol{0}\}$.
    $t \leftarrow t+1$.
    if duality gap is smaller than $\iota$ then
        Break.
    end if
end while
Output: $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$.
```

Recovering the cluster assignment

Once the saddle point $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})$ is found, we can obtain the cluster assignment according to $\operatorname{sign}(\sum_{k:\hat{\boldsymbol{y}}_k\in\mathcal{B}}\mu_k^\star\hat{\boldsymbol{y}}_k)$.

Convergence rate

For the cutting-plane based algorithms, it has been proven that the time complexity is $O(1/\epsilon^2)$ (Zhao, Wang, and Zhang 2008), i.e., to get a solution with ϵ accuracy, the algorithm needs run $O(1/\epsilon^2)$ iterations. In this section, we show that ODMC has the same convergence rate, but note that in each iteration, cutting-plane based algorithms need to find the most violated label and then train a SVM model, whereas, ODMC just preforms a random sampling and a step of stochastic gradient descent, followed by a projection with closed-form solutions, so our method is much more scalable.

Theorem 1. *Assume $|G(\boldsymbol{\alpha},\hat{\boldsymbol{y}}_k)|$ and $\|\partial_{\boldsymbol{\alpha}}G(\boldsymbol{\alpha},\hat{\boldsymbol{y}}_k)\|_2$ are upper bounded by L_1 and L_2 respectively for any $\hat{\boldsymbol{y}}_k\in\mathcal{B}$ and $\boldsymbol{\alpha}\in\mathcal{A}$. Let $a = |\mathcal{B}|L_1/\sqrt{\log|\mathcal{B}|}$, $b = \sqrt{2}L_2/\sqrt{m}\tau$ and*

$\eta = \sqrt{2/T}$, the expectation of duality gap at the average point $(\sum_{t=1}^{T} \boldsymbol{\mu}_t/T, \sum_{t=1}^{T} \boldsymbol{\alpha}_t/T)$ satisfies

$$\mathbb{E}\left[\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\mu}_t, \boldsymbol{\alpha}\right) - \min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi\left(\boldsymbol{\mu}, \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}_t\right)\right]$$
$$\leq (|\mathcal{B}|L_1\sqrt{2\log|\mathcal{B}|} + L_2\tau\sqrt{m})/\sqrt{T}.$$

Proof. Similar to the analysis of vanilla gradient descent, we can prove the duality gap at the average point is upper bounded by the sum of inner product between the gradient $\widetilde{g}(\boldsymbol{w}_t)$ and the gap $\boldsymbol{w}_t - \boldsymbol{w}$,

$$\begin{aligned}
&\mathbb{E}\left[\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\mu}_t, \boldsymbol{\alpha}\right) - \min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi\left(\boldsymbol{\mu}, \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}_t\right)\right] \\
&\leq \mathbb{E}\left[\max_{\boldsymbol{\alpha}\in\mathcal{A}} \frac{1}{T}\sum_{t=1}^{T}\varphi(\boldsymbol{\mu}_t, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu}\in\mathcal{M}} \frac{1}{T}\sum_{t=1}^{T}\varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}_t)\right] \\
&\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\max_{\boldsymbol{\alpha}\in\mathcal{A}} \varphi(\boldsymbol{\mu}_t, \boldsymbol{\alpha}) - \min_{\boldsymbol{\mu}\in\mathcal{M}} \varphi(\boldsymbol{\mu}, \boldsymbol{\alpha}_t)\right)\right] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[g(\boldsymbol{w}_t)^\top(\boldsymbol{w}_t - \boldsymbol{w})] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}[\widetilde{g}(\boldsymbol{w}_t)^\top(\boldsymbol{w}_t - \boldsymbol{w}) \mid \boldsymbol{w}_t]] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\widetilde{g}(\boldsymbol{w}_t)^\top(\boldsymbol{w}_t - \boldsymbol{w})],
\end{aligned}$$

where the first inequality holds since $\varphi(\boldsymbol{\mu}, \boldsymbol{\alpha})$ is convex in $\boldsymbol{\mu}$ and concave in $\boldsymbol{\alpha}$, the second inequality is owing to the sub-additivity of $\max$, the third inequality is Eq. (9), and the final equality is according to the law of total expectation.

Next we bound each term in the summation separately. With the update rule of stochastic mirror descent, we have

$$\begin{aligned}
&\widetilde{g}(\boldsymbol{w}_t)^\top(\boldsymbol{w}_t - \boldsymbol{w}) \\
&= \frac{1}{\eta}(\nabla\Phi(\boldsymbol{w}_t) - \nabla\Phi(\bar{\boldsymbol{w}}_{t+1}))^\top(\boldsymbol{w}_t - \boldsymbol{w}) \\
&= \frac{1}{\eta}(\Delta_\Phi(\boldsymbol{w}, \boldsymbol{w}_t) + \Delta_\Phi(\boldsymbol{w}_t, \bar{\boldsymbol{w}}_{t+1}) - \Delta_\Phi(\boldsymbol{w}, \bar{\boldsymbol{w}}_{t+1})) \\
&\leq \frac{1}{\eta}(\Delta_\Phi(\boldsymbol{w}, \boldsymbol{w}_t) + \Delta_\Phi(\boldsymbol{w}_t, \bar{\boldsymbol{w}}_{t+1}) - \Delta_\Phi(\boldsymbol{w}, \boldsymbol{w}_{t+1}) \\
&\qquad - \Delta_\Phi(\boldsymbol{w}_{t+1}, \bar{\boldsymbol{w}}_{t+1})),
\end{aligned}$$

where $\Delta_\Phi(\boldsymbol{w}, \boldsymbol{w}_t) = \Phi(\boldsymbol{w}) - \Phi(\boldsymbol{w}_t) - \nabla\Phi(\boldsymbol{w}_t)^\top(\boldsymbol{w} - \boldsymbol{w}_t)$ is the Bregman divergence. Since $\boldsymbol{w}_{t+1}$ is the projection of $\bar{\boldsymbol{w}}_{t+1}$ onto the convex set $\mathcal{M}\times\mathcal{A}$, the final inequality holds true for any $\boldsymbol{w} \in \mathcal{M}\times\mathcal{A}$ according to the generalized triangle inequality. Note that the first and third term will lead to a telescopic sum when summing over $t = 0$ to $t = T$, it remains to bound the other two terms,

$$\begin{aligned}
&\Delta_\Phi(\boldsymbol{w}_t, \bar{\boldsymbol{w}}_{t+1}) - \Delta_\Phi(\boldsymbol{w}_{t+1}, \bar{\boldsymbol{w}}_{t+1}) \\
&= \Phi(\boldsymbol{w}_t) - \Phi(\boldsymbol{w}_{t+1}) - \nabla\Phi(\bar{\boldsymbol{w}}_{t+1})^\top(\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) \\
&\leq \eta\widetilde{g}(\boldsymbol{w}_t)^\top(\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) - \frac{1}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{.}^2 \\
&\leq \eta\|\widetilde{g}(\boldsymbol{w}_t)\|_*\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{.} - \frac{1}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{.}^2 \\
&= -\frac{1}{2}(\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{.} - \eta\|\widetilde{g}(\boldsymbol{w}_t)\|_*)^2 + \frac{1}{2}(\eta\|\widetilde{g}(\boldsymbol{w}_t)\|_*)^2 \\
&\leq \frac{\eta^2\|\widetilde{g}(\boldsymbol{w}_t)\|_*^2}{2}
\end{aligned}$$

where $\|\boldsymbol{w}\|_{.}^2 = a\|\boldsymbol{\mu}\|_1^2 + b\|\boldsymbol{\alpha}\|_2^2$ and $\|\cdot\|_*^2 = \|\boldsymbol{\mu}\|_\infty^2/a + \|\boldsymbol{\alpha}\|_2^2/b$ is the dual norm. The first inequality holds since $\Phi(\cdot)$ is 1-strongly convex function w.r.t. the norm $\|\cdot\|_{.}$, and the second inequality is according to Hölder's inequality.

Note that $|G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k)| \leq L_1$ and $\|\partial_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}, \hat{\boldsymbol{y}}_k)\|_2 \leq L_2$, hence we have

$$\begin{aligned}
\|\widetilde{g}(\boldsymbol{w}_t)\|_*^2 &= \frac{1}{a}\|\partial_{\boldsymbol{\mu}}\widetilde{\varphi}(\boldsymbol{w}_t, i_t)\|_\infty^2 + \frac{1}{b}\|\partial_{\boldsymbol{\alpha}}\widetilde{\varphi}(\boldsymbol{w}_t, j_t)\|_2^2 \\
&\leq \frac{1}{a}|\mathcal{B}|^2 L_1^2 + \frac{1}{b}L_2^2 = |\mathcal{B}|L_1\sqrt{\log|\mathcal{B}|} + L_2\tau\sqrt{m/2}.
\end{aligned}$$

Further note that

$$\begin{aligned}
&\Delta_\Phi(\boldsymbol{w}, \boldsymbol{w}_1) = \Phi(\boldsymbol{w}) - \Phi(\boldsymbol{w}_1) - \nabla\Phi(\boldsymbol{w}_1)^\top(\boldsymbol{w} - \boldsymbol{w}_1) \\
&= a\boldsymbol{\mu}\log(\boldsymbol{\mu}/\boldsymbol{\mu}_1) + b\|\boldsymbol{\mu} - \boldsymbol{\mu}_1\|_2^2/2 \leq a\log|\mathcal{B}| + bm\tau^2/2 \\
&= |\mathcal{B}|L_1\sqrt{\log|\mathcal{B}|} + L_2\tau\sqrt{m/2}.
\end{aligned}$$

Combine together and we have,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\widetilde{g}(\boldsymbol{w}_t)^\top(\boldsymbol{w}_t - \boldsymbol{w})] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\frac{\eta\|\widetilde{g}(\boldsymbol{w}_t)\|_*^2}{2} + \frac{\Delta_\Phi(\boldsymbol{w}, \boldsymbol{w}_1) - \Delta_\Phi(\boldsymbol{w}, \boldsymbol{w}_{T+1})}{\eta T} \\
&= (|\mathcal{B}|L_1\sqrt{\log|\mathcal{B}|} + L_2\tau\sqrt{m/2})\left(\frac{\eta}{2} + \frac{1}{\eta T}\right).
\end{aligned}$$

By substituting $\eta = \sqrt{2/T}$ we can conclude the proof. □

This theorem shows that the duality gap decays as the rate $O(1/\sqrt{T})$. By setting $O(1/\sqrt{T}) = \epsilon$ we can obtain the time complexity $T = O(1/\epsilon^2)$, which is the same with cutting-plane based methods.

Empirical Study

In this section, we empirically evaluate the proposed method on 24 UCI data sets. Table 1 summarizes the statistics of these data sets. As can be seen, the number of instance is ranged from 62 to 3175, and the dimensionality is ranged from 2 to 4702, covering a broad range of properties.

Evaluation criteria

We set the number of clusters equal to the true number of classes for all the methods. To evaluate their performance, we compare the generated clusters with the true classes by computing the following two performance measures.

Clustering Accuracy (Acc) (Xu et al. 2005). First remove the labels for all instances, and then predict the clusters via performing clustering methods, finally measure the classification accuracy according to true label.

Rand Index (RI) (Rand 1971). Let $\mathcal{C}$ be the set of clustering results, and denotes $\mathcal{L}$ as the set of true classes. Rand index represents the frequency of occurrence of agreements over all the instance pairs, i.e., the probability that $\mathcal{C}$ and $\mathcal{L}$ will agree on a randomly chosen instance pair.

Compared methods

ODMC is compared with k-means (KM) method, normalized cut (NC) method (Shi and Malik 2000), GMMC (Valizadegan and Jin 2006), IterSVR (Zhang, Tsang, and Kwok 2007), CPMMC (Zhao, Wang, and Zhang 2008) and LG-MMC (Li et al. 2009). MMC (Xu et al. 2005) is not chosen as a baseline since it can't return results in a reasonable time for most data sets.

For GMMC, IterSVR, CPMMC, LG-MMC, ODMC, the parameters C or λ is selected from $\{1, 10, 100, 1000\}$. For ODMC, ν and θ are selected from $[0.2, 0.4, 0.6, 0.8]$. For all data sets, both the linear and Gaussian kernels are used. In particular, the width σ of Gaussian kernel is picked from $\{0.25\sqrt{\gamma}, 0.5\sqrt{\gamma}, \sqrt{\gamma}, 2\sqrt{\gamma}, 4\sqrt{\gamma}\}$, where γ is the average distance between instances. The parameter of normalized cut is chosen from the same range of σ. The balance constraint is set in the same manner as in (Zhang, Tsang, and Kwok 2007), i.e., $0.03m$ for balanced data set and $0.3m$ for imbalanced data set. All the experiments are repeated 10 times and the average performance is reported with the best parameter setting.

Table 1: Characteristics of experimental data sets.

ID	Data set	#Instance	#Feature
1	*echocardiogram*	62	8
2	*dbworld*	64	4,702
3	*hepatitis*	80	19
4	*colic*	188	13
5	*house*	232	16
6	*heart-h*	261	10
7	*heart*	270	9
8	*heart-statlog*	270	13
9	*breast*	277	9
10	*cylinder-bands*	277	39
11	*heart-c*	296	13
12	*haberman*	306	14
13	*ionosphere*	351	33
14	*vehicle*	435	16
15	*credit-a*	653	15
16	*diabetes*	768	8
17	*fourclass*	862	2
18	*tic-tac-toe*	958	9
19	*credit-g*	1,000	20
20	*german*	1,000	59
21	*optdigits*	1,143	42
22	*svmguide3*	1,284	22
23	*sick*	2,643	28
24	*splice*	3,175	60

Performance

Table 2 summarizes the results on 24 UCI data sets. GMMC did not return results on some data sets due to the high computation cost. As can be seen, for both measures, ODMC achieves the best performance on 17 data sets and shows significant improvement over existing MMC approaches on most data sets.

Time cost

We compare the average single iteration time cost of our method with IterSVR, CPMMC and LG-MMC on some representative data sets. All the experiments are performed with MATLAB 2017b on a machine with 8×2.60 GHz CPUs and 32GB main memory. As shown in Figure 2, our method achieves the lowest time cost for most data sets and is only slightly worse than compared methods on data set credit-g. Note that the sub-problem of IterSVR and LG-MMC are solved by LIBSVM (Chang and Lin 2011), which is a fast implementation of both SVR and SVM, this shows that our method is also computationally efficient.

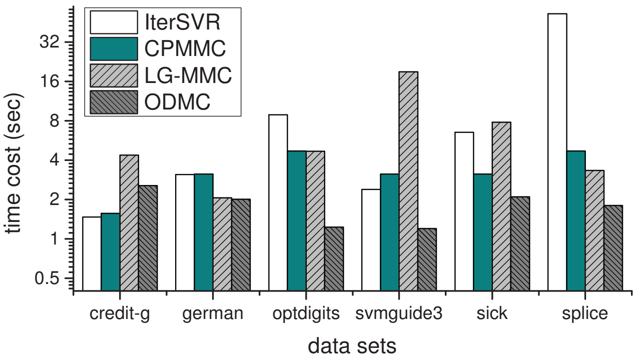

Figure 2: Average single iteration time cost of IterSVR, CPMMC, LG-MMC, ODMC.

Conclusions

Maximum margin clustering (MMC), which employs the large margin heuristic from support vector machine, have achieved more accurate results than traditional clustering methods. Recent studies disclosed that instead of minimum margin, it is more crucial to optimize the margin distribution for SVM-style learning algorithms. Inspired by this recognition, we propose a novel approach ODMC for clustering by optimizing the margin distribution. To solve the resultant minimax problem, we extend a stochastic mirror descent method which is much more scalable than existing cutting-plane based approaches. Experimental results in various data sets show that our method achieves promising clustering performance, which further verifies the superiority of optimal margin distribution learning. In the future, we will apply importance sampling (Schmidt et al. 2015) to further accelerate our method and extend it to other learning settings, i.e., semi-supervised learning.

Table 2: Clustering Accuracy (Acc) and Rand Index (RI) comparisons on 24 UCI data sets. The best performance on each data set is bolded. ●/○ indicates ODMC is significantly better/worse than compared methods (paired t-tests at 95% significance level). The win/tie/loss counts for ODMC are summarized in the last two rows. GMMC did not return results on some data sets in two days.

Data set	Measure	KM	NC	GMMC	IterSVR	CPMMC	LG-MMC	ODMC
echocardiogram	Acc	0.696●	0.677●	0.694●	0.537●	0.710●	0.774●	**0.792**
	RI	0.572●	0.556●	0.568●	0.498●	0.581●	0.645●	**0.661**
dbworld	Acc	0.599●	0.625●	0.578●	0.840	0.547●	**0.859**	0.857
	RI	0.534●	0.524●	0.504●	0.743●	0.497●	**0.754**	0.748
hepatitis	Acc	0.693●	0.525●	0.575●	0.638●	0.838	0.838	**0.841**
	RI	0.576●	0.495●	0.505●	0.534●	0.724	0.724	**0.731**
colic	Acc	0.749●	0.622●	0.537●	0.740●	0.622●	0.840●	**0.862**
	RI	0.622●	0.527●	0.500●	0.613●	0.527●	0.730●	**0.752**
house	Acc	0.893	0.534●	0.737●	0.901	0.534●	**0.905**	0.902
	RI	0.808●	0.500●	0.611●	0.821	0.500●	**0.828**	0.825
heart-h	Acc	0.742●	0.536●	0.617●	0.780	0.625●	0.789	**0.795**
	RI	0.636●	0.501●	0.525●	0.662	0.529●	0.666	**0.669**
heart	Acc	0.670●	0.567●	0.563●	0.684●	0.556●	0.744●	**0.772**
	RI	0.572●	0.507●	0.506●	0.589●	0.504●	0.618●	**0.637**
heart-statlog	Acc	0.765●	0.504●	0.741●	0.761●	0.556●	0.796●	**0.811**
	RI	0.649●	0.498●	0.614●	0.645●	0.504●	0.674	**0.681**
breast	Acc	0.629●	0.592●	0.538●	0.575●	0.708●	**0.726**	**0.726**
	RI	0.550●	0.515●	0.501●	0.518●	0.585●	**0.600**	**0.600**
cylinder-bands	Acc	0.634●	0.534●	0.574●	0.626●	0.643●	0.657●	**0.675**
	RI	0.534●	0.501●	0.509●	0.530●	0.539●	0.548●	**0.571**
heart-c	Acc	0.662●	0.551●	0.588●	**0.780**	0.541●	0.777	0.775
	RI	0.559●	0.503●	0.514●	**0.657**	0.502●	0.652	0.652
haberman	Acc	0.604●	0.735	0.582●	0.520●	0.735	**0.739**	0.736
	RI	0.530●	0.609	0.512●	0.500●	0.609	**0.613**	0.611
ionosphere	Acc	0.708●	0.541●	0.721●	0.691●	0.641●	0.738●	**0.754**
	RI	0.586●	0.502●	0.596●	0.572●	0.538●	0.612●	**0.636**
vehicle	Acc	0.659●	0.510●	0.554●	0.693●	0.501●	0.715●	**0.742**
	RI	0.556●	0.499●	0.505●	0.581●	0.499●	0.591●	**0.624**
credit-a	Acc	0.730●	0.576●	0.522●	0.771	0.547●	**0.772**	0.770
	RI	0.636●	0.511●	0.500●	**0.669**	0.504●	0.647●	0.662
diabetes	Acc	0.667●	0.637●	0.663●	0.629●	0.651●	0.733●	**0.745**
	RI	0.555●	0.537●	0.552●	0.533●	0.545●	0.608●	**0.625**
fourclass	Acc	0.659●	0.624●	0.515●	0.640●	0.644●	0.763●	**0.788**
	RI	0.552●	0.530●	0.500●	0.546●	0.541●	0.638●	**0.666**
tic-tac-toe	Acc	0.547●	0.551●	0.511●	0.548●	0.653	0.653	**0.658**
	RI	0.504●	0.505●	0.500●	0.505●	0.547	0.547	**0.551**
credit-g	Acc	0.539●	0.683●	N/A	0.534●	0.700	0.704	**0.710**
	RI	0.507●	0.567●	N/A	0.509●	0.580	0.583	**0.585**
german	Acc	0.569●	**0.700**	N/A	0.546●	**0.700**	**0.700**	**0.700**
	RI	0.510●	**0.580**	N/A	0.505●	**0.580**	**0.580**	**0.580**
optdigits	Acc	0.962●	0.501●	N/A	**0.995**○	0.500●	0.978	0.980
	RI	0.952	0.500●	N/A	**0.989**○	0.500●	0.957	0.958
svmguide3	Acc	0.593●	0.657●	N/A	0.572●	0.738	0.739	**0.741**
	RI	0.518●	0.549●	N/A	0.511●	0.613	0.614	**0.618**
sick	Acc	0.731●	0.518●	N/A	0.515●	**0.920**	**0.920**	**0.920**
	RI	0.628●	0.500●	N/A	0.501●	**0.852**	**0.852**	**0.852**
splice	Acc	**0.665**	0.519●	N/A	0.630●	0.519●	0.641	0.655
	RI	**0.554**	0.501●	N/A	0.534●	0.501●	0.539	0.550
ODMC: w/t/l	Acc	22/2/0	22/2/0	18/0/0	18/5/1	17/7/0	9/15/0	
	RI	22/2/0	22/2/0	18/0/0	19/4/1	17/7/0	9/15/0	

References

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20(3):273–297.

Cristianini, N., and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learn-*

ing Methods. Cambridge, UK: Cambridge University Press.

Gao, W., and Zhou, Z.-H. 2013. On the doubt about margin explanation of boosting. *Artificial Intelligence* 203:1–18.

Jain, A. K., and Dubes, R. C. 1988. *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall.

Jain, A. K. 2010. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* 31(8):651–666.

Kim, S.-J., and Boyd, S. 2008. A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal on Optimization* 19(3):1344–1367.

Li, Y.-F.; Tsang, I. W.; Kwok, J. T.; and Zhou, Z.-H. 2009. Tighter and convex maximum margin clustering. In *In Proceeding of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5, 344–351.

Li, Y.-F.; Tsang, I. W.; Kwok, J. T.; and Zhou, Z.-H. 2013. Convex and scalable weakly labeled svms. *Journal of Machine Learning Research* 14(1):2151–2188.

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

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336):846–850.

Schmidt, M.; Babanezhad, R.; Ahmed, M.; Defazio, A.; Clifton, A.; and Sarkar, A. 2015. Non-uniform stochastic average gradient method for training conditional random fields. In Lebanon, G., and Vishwanathan, S. V. N., eds., *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, 819–828. San Diego, CA: PMLR.

Schölkopf, B., and Smola, A. J. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905.

Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics* 8(1):171–176.

Valizadegan, H., and Jin, R. 2006. Generalized maximum margin clustering and unsupervised kernel learning. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 1417–1424. Cambridge, MA, USA: MIT Press.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vijaya Saradhi, V., and Charly Abraham, P. 2016. Incremental maximum margin clustering. *Pattern Analysis and Applications* 19(4):1057–1067.

Wang, F.; Zhao, B.; and Zhang, C. 2010. Linear time maximum margin clustering. *IEEE Transactions on Neural Networks* 21(2):319–332.

Xu, R., and Wunsch, II, D. 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks* 16(3):645–678.

Xu, L.; Neufeld, J.; Larson, B.; and Schuurmans, D. 2005. Maximum margin clustering. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 17*. MIT Press. 1537–1544.

Zhang, T., and Zhou, Z.-H. 2014. Large margin distribution machine. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 313–322.

Zhang, T., and Zhou, Z.-H. 2016. Optimal margin distribution machine. *CoRR, abs/1604.03348*.

Zhang, T., and Zhou, Z.-H. 2017. Multi-class optimal distribution machine. In *Proceedings of the 34th International Conference on Machine Learning*, 4063–4071.

Zhang, K.; Tsang, I. W.; and Kwok, J. T. 2007. Maximum margin clustering made practical. In *Proceedings of the 24th International Conference on Machine Learning*, 1119–1126.

Zhang, K.; Tsang, I. W.; and Kwok, J. T. 2009. Maximum margin clustering made practical. *Trans. Neur. Netw.* 20(4):583–596.

Zhao, B.; Wang, F.; and Zhang, C. 2008. Efficient maximum margin clustering via cutting plane algorithm. In *Siam International Conference on Data Mining*, 751–762.

Zhou, Y.-H., and Zhou, Z.-H. 2016. Large margin distribution learning with cost interval and unlabeled data. *IEEE Transactions on Knowledge and Data Engineering* 28(7):1749–1763.

Zhou, G.-T.; Lan, T.; Vahdat, A.; and Mori, G. 2013. Latent maximum margin clustering. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 28–36. USA: Curran Associates Inc.

Zhou, Z.-H. 2014. Large margin distribution learning. In *Proceedings of the 6th IAPR International Workshop on Artificial Neural Networks in Pattern Recognition*, 1–11.