机器学习导论
综合能力测试

141242006, 袁帅, 141242006@smail.nju.edu.cn

2017 年 6 月 15 日

1 [40pts] Exponential Families

指数分布族 (Exponential Families) 是一类在机器学习和统计中非常常见的分布族, 具有良好的性质。在后文不引起歧义的情况下, 简称为指数族。

指数分布族是一组具有如下形式概率密度函数的分布族群:

$$f_X(x|\theta) = h(x)\exp\left(\eta(\theta)\cdot T(x) - A(\theta)\right) \tag{1.1}$$

其中, $\eta(\theta)$, $A(\theta)$ 以及函数 $T(\cdot)$, $h(\cdot)$ 都是已知的。

(1) **[10pts]** 试证明多项分布 (Multinomial distribution) 属于指数分布族。

(2) **[10pts]** 试证明多元高斯分布 (Multivariate Gaussian distribution) 属于指数分布族。

(3) **[20pts]** 考虑样本集 $\mathcal{D} = \{x_1, \cdots, x_n\}$ 是从某个已知的指数族分布中独立同分布地 (i.i.d.) 采样得到, 即对于 $\forall i \in [1, n]$, 我们有 $f(x_i|\boldsymbol{\theta}) = h(x_i)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}T(x_i) - A(\boldsymbol{\theta})\right)$.

对参数 $\boldsymbol{\theta}$, 假设其服从如下先验分布:

$$p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right) \tag{1.2}$$

其中, $\boldsymbol{\chi}$ 和 ν 是 $\boldsymbol{\theta}$ 生成模型的参数。请计算其后验, 并证明后验与先验具有相同的形式。(**Hint**: 上述又称为"共轭"(Conjugacy), 在贝叶斯建模中经常用到)

Solution.

(1) *Proof.* The Multinomial Distribution's probability mass function can be rewritten as

$$\begin{aligned}
P(\boldsymbol{x}|n, p_1, p_2, \cdots, p_k) &= \frac{n!}{x_1!x_2!\cdots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k} \\
&= \frac{n!}{x_1!x_2!\cdots x_k!}\exp(x_1\ln p_1 + x_2\ln p_2 + \cdots + x_k\ln p_k) \\
&= h(\boldsymbol{x})\exp(\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{\theta})\cdot \boldsymbol{T}(\boldsymbol{X}) - A(\boldsymbol{\theta})),
\end{aligned} \tag{1.3}$$

in which $\boldsymbol{\theta} = \{n, p_1, p_2, \ldots, p_k\}$, $h(\boldsymbol{x}) = \frac{n!}{x_1!x_2!\cdots x_k!}$, $\boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{bmatrix}\ln p_1 \\ \ln p_2 \\ \vdots \\ \ln p_k\end{bmatrix}$, $\boldsymbol{T}(\boldsymbol{x}) = \begin{bmatrix}x_1 \\ x_2 \\ \vdots \\ x_k\end{bmatrix}$, $A(\boldsymbol{\theta}) = 0$, and the multiplication $\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{\theta})\cdot\boldsymbol{T}(\boldsymbol{X})$ is the dot product of two vectors. □

(2) *Proof.* The Multivariate Gaussian Distribution's probability density function is

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \\
&= \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})-\frac{d}{2}\ln(2\pi)-\frac{1}{2}\ln|\boldsymbol{\Sigma}|] \\
&= \exp[-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{d}{2}\ln(2\pi)-\frac{1}{2}\ln|\boldsymbol{\Sigma}|] \\
&= \exp[-\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}(\boldsymbol{\Sigma}^{-1})_{ij}x_ix_j+\sum_{i=1}^{d}(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1})_ix_i-A(\boldsymbol{\theta})] \\
&= \exp[-\frac{1}{2}\mathrm{Vec}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1})\cdot\mathrm{Vec}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})+(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{\mathrm{T}}\cdot\boldsymbol{x}-A(\boldsymbol{\theta})] \\
&= h(\boldsymbol{x})\exp(\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{\theta})\cdot\boldsymbol{T}(\boldsymbol{X})-A(\boldsymbol{\theta})), \qquad (1.4)
\end{aligned}
$$

in which $\mathrm{Vec}(\cdot)$ is the vectorization operator, $\boldsymbol{\theta}=\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$, $h(\boldsymbol{x})=1$, $\boldsymbol{\eta}(\boldsymbol{\theta})=\begin{bmatrix}-\frac{1}{2}\mathrm{Vec}(\boldsymbol{\Sigma}^{-1})\\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\end{bmatrix}$, $\boldsymbol{T}(\boldsymbol{x})=\begin{bmatrix}\mathrm{Vec}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\\ \boldsymbol{x}\end{bmatrix}$, $A(\boldsymbol{\theta})=\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln|\boldsymbol{\Sigma}|$, and the multiplication $\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{\theta})\cdot\boldsymbol{T}(\boldsymbol{X})$ is the dot product of two vectors. □

(3) According to Wikipedia[1], The posterior distribution could be given as

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D};\boldsymbol{\chi},\nu) &= \frac{p(\boldsymbol{\theta}|\boldsymbol{\chi},\nu)p(\mathcal{D}|\boldsymbol{\theta};\boldsymbol{\chi},\nu)}{p(\mathcal{D}|\boldsymbol{\chi},\nu)} \\
&= \frac{f(\boldsymbol{\chi},\nu)\exp(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi}-\nu A(\boldsymbol{\theta}))\prod_{i=1}^{n}h(\boldsymbol{x}_i)\exp(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{T}(\boldsymbol{x}_i)-A(\boldsymbol{\theta}))}{\int_{\boldsymbol{\alpha}}f(\boldsymbol{\chi},\nu)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\chi}-\nu A(\boldsymbol{\alpha}))\prod_{i=1}^{n}h(\boldsymbol{x}_i)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{T}(\boldsymbol{x}_i)-A(\boldsymbol{\alpha}))} \\
&= \frac{\exp[\boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{\chi}+\sum_{i=1}^{n}\boldsymbol{T}(\boldsymbol{x}_i))-(\nu+n)A(\boldsymbol{\theta})]}{\int_{\boldsymbol{\alpha}}\exp[\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{\chi}+\sum_{i=1}^{n}\boldsymbol{T}(\boldsymbol{x}_i))-(\nu+n)A(\boldsymbol{\alpha})]} \\
&= \hat{f}(\hat{\boldsymbol{\chi}},\hat{\nu})\exp(\boldsymbol{\theta}^{\mathrm{T}}\hat{\boldsymbol{\chi}}-\hat{\nu}\hat{A}(\boldsymbol{\theta})), \qquad (1.5)
\end{aligned}
$$

where $\hat{\boldsymbol{\chi}}=\boldsymbol{\chi}+\sum_{i=1}^{n}\boldsymbol{T}(\boldsymbol{x}_i)$, $\hat{\nu}=\nu+n$, $\hat{A}(\boldsymbol{\theta})=A(\boldsymbol{\theta})$, $\hat{f}(\hat{\boldsymbol{\chi}},\hat{\nu})=\frac{1}{\int_{\boldsymbol{\alpha}}\exp[\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\chi}}-\hat{\nu}A(\boldsymbol{\alpha})]}$. Therefore, the posterior distribution is of the same form as the prior.

2 [40pts] Decision Boundary

考虑二分类问题, 特征空间 $X\in\mathcal{X}=\mathbb{R}^d$, 标记 $Y\in\mathcal{Y}=\{0,1\}$. 我们对模型做如下生成式假设:

- attribute conditional independence assumption: 对已知类别, 假设所有属性相互独立, 即每个属性特征独立地对分类结果发生影响;
- Bernoulli prior on label: 假设标记满足 Bernoulli 分布先验, 并记 $\Pr(Y=1)=\pi$.

(1) **[20pts]** 假设 $P(X_i|Y)$ 服从指数族分布, 即

$$\Pr(X_i=x_i|Y=y)=h_i(x_i)\exp(\theta_{iy}\cdot T_i(x_i)-A_i(\theta_{iy}))$$

请计算后验概率分布 $\Pr(Y|X)$ 以及分类边界 $\{x \in \mathcal{X} : P(Y=1|X=x) = P(Y=0|X=x)\}$. (**Hint**: 你可以使用 sigmoid 函数 $\mathcal{S}(x) = 1/(1+e^{-x})$ 进行化简最终的结果).

(2) **[20pts]** 假设 $P(X_i|Y=y)$ 服从高斯分布, 且记均值为 μ_{iy} 以及方差为 σ_i^2 (注意, 这里的方差与标记 Y 是独立的), 请证明分类边界与特征 X 是成线性的。

Solution.

(1) The posterior distribution is given by

$$
\begin{aligned}
P(Y=0|\boldsymbol{X}=\boldsymbol{x}) &= \frac{P(Y=0)P(\boldsymbol{X}=\boldsymbol{x}|Y=0)}{P(\boldsymbol{X}=\boldsymbol{x})} \\
&= \frac{(1-\pi)\prod_{i=1}^d \exp(\boldsymbol{\theta}_{i0}^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - A_i(\boldsymbol{\theta}_{i0}))}{(1-\pi)\prod_{i=1}^d \exp(\boldsymbol{\theta}_{i0}^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - A_i(\boldsymbol{\theta}_{i0})) + \pi\prod_{i=1}^d \exp(\boldsymbol{\theta}_{i1}^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - A_i(\boldsymbol{\theta}_{i1}))} \\
&= \frac{1}{1+\frac{\pi}{1-\pi}\exp(\sum_{i=1}^d (\boldsymbol{\theta}_{i1}-\boldsymbol{\theta}_{i0})^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i1}) + \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i0}))} \\
&= \mathcal{S}(\sum_{i=1}^d (\boldsymbol{\theta}_{i0}-\boldsymbol{\theta}_{i1})^{\mathrm{T}} \boldsymbol{T}_i(x_i) - \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i0}) + \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i1}) + \ln\frac{1-\pi}{\pi}), \qquad (2.1)
\end{aligned}
$$

$$
\begin{aligned}
P(Y=1|\boldsymbol{X}=\boldsymbol{x}) &= \frac{P(Y=1)P(\boldsymbol{X}=\boldsymbol{x}|Y=1)}{P(\boldsymbol{X}=\boldsymbol{x})} \\
&= \frac{\pi\prod_{i=1}^d \exp(\boldsymbol{\theta}_{i1}^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - A_i(\boldsymbol{\theta}_{i1}))}{(1-\pi)\prod_{i=1}^d \exp(\boldsymbol{\theta}_{i0}^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - A_i(\boldsymbol{\theta}_{i0})) + \pi\prod_{i=1}^d \exp(\boldsymbol{\theta}_{i1}^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - A_i(\boldsymbol{\theta}_{i1}))} \\
&= \frac{1}{1+\frac{1-\pi}{\pi}\exp(\sum_{i=1}^d (\boldsymbol{\theta}_{i0}-\boldsymbol{\theta}_{i1})^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) + \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i1}) - \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i0}))} \\
&= \mathcal{S}(\sum_{i=1}^d (\boldsymbol{\theta}_{i1}-\boldsymbol{\theta}_{i0})^{\mathrm{T}} \boldsymbol{T}_i(x_i) + \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i0}) - \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i1}) - \ln\frac{1-\pi}{\pi}). \qquad (2.2)
\end{aligned}
$$

The decision boundary is determined by $P(Y=0|\boldsymbol{X}=\boldsymbol{x}) = P(Y=1|\boldsymbol{X}=\boldsymbol{x})$, which yields

$$P(Y=0)P(\boldsymbol{X}=\boldsymbol{x}|Y=0) = P(Y=1)P(\boldsymbol{X}=\boldsymbol{x}|Y=1), \qquad (2.3)$$

i.e.

$$(1-\pi)\prod_{i=1}^d \exp(\boldsymbol{\theta}_{i0}^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - A_i(\boldsymbol{\theta}_{i0})) = \pi\prod_{i=1}^d \exp(\boldsymbol{\theta}_{i1}^{\mathrm{T}} \cdot \boldsymbol{T}_i(x_i) - A_i(\boldsymbol{\theta}_{i1})). \qquad (2.4)$$

Solving that equation, we finally obtain the boundary as

$$\sum_{i=1}^d (\boldsymbol{\theta}_{i1}-\boldsymbol{\theta}_{i0})^{\mathrm{T}} \boldsymbol{T}_i(x_i) = \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i1}) - \sum_{i=1}^d A_i(\boldsymbol{\theta}_{i0}) + \ln\frac{1-\pi}{\pi}. \qquad (2.5)$$

(2) *Proof.* Suppose $P(X_i|Y=y) \sim \mathcal{N}(\mu_{iy}, \sigma_i^2)$, By solving $P(Y=0)P(\boldsymbol{X}|Y=0) = P(Y=1)P(\boldsymbol{X}|Y=1)$, we get

$$\pi\exp[-\frac{1}{2}\sum_{i=1}^d \frac{(x_i-\mu_{i1})^2}{\sigma_i^2}] = (1-\pi)\exp[-\frac{1}{2}\sum_{i=1}^d \frac{(x_i-\mu_{i0})^2}{\sigma_i^2}]. \qquad (2.6)$$

After some simple math, we obtain the equation as

$$\begin{aligned}\ln\frac{1-\pi}{\pi} &= \sum_{i=1}^{d}\left(\frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}-\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2}\right)\\ &= \sum_{i=1}^{d}\left(\frac{\mu_{i1}-\mu_{i0}}{\sigma_i^2}x_i+\frac{\mu_{i0}^2-\mu_{i1}^2}{2\sigma_i^2}\right).\end{aligned}\tag{2.7}$$

Therefore, the decision boundary

$$\sum_{i=1}^{d}\frac{\mu_{i1}-\mu_{i0}}{\sigma_i^2}x_i=\ln\frac{1-\pi}{\pi}-\sum_{i=1}^{d}\frac{\mu_{i0}^2-\mu_{i1}^2}{2\sigma_i^2}\tag{2.8}$$

is linear to the sample space $\boldsymbol{X}=(x_1,x_2,\cdots,x_d)$. □

3 [70pts] Theoretical Analysis of k-means Algorithm

给定样本集 $\mathcal{D}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$, k-means 聚类算法希望获得簇划分 $\mathcal{C}=\{C_1,C_2,\cdots,C_k\}$, 使得最小化欧式距离

$$J(\gamma,\mu_1,\ldots,\mu_k)=\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}||\mathbf{x}_i-\mu_j||^2\tag{3.1}$$

其中, $\mu_1,\ldots,\mu_k$ 为 k 个簇的中心 (means), $\gamma\in\mathbb{R}^{n\times k}$ 为指示矩阵 (indicator matrix) 定义如下：若 $\mathbf{x}_i$ 属于第 j 个簇, 则 $\gamma_{ij}=1$, 否则为 0.

则最经典的 k-means 聚类算法流程如算法1中所示 (与课本中描述稍有差别, 但实际上是等价的)。

Algorithm 1: k-means Algorithm

1 Initialize $\mu_1,\ldots,\mu_k$.

2 **repeat**

3 **Step 1**: Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^n$ by assigning each of them to its nearest cluster center.

$$\gamma_{ij}=\begin{cases}1, & ||\mathbf{x}_i-\mu_j||^2\le||\mathbf{x}_i-\mu_{j'}||^2,\forall j'\\ 0, & \text{otherwise}\end{cases}$$

4 **Step 2**: For each $j\in\{1,\cdots,k\}$, recompute μ_j using the updated γ to be the center of mass of all points in C_j:

$$\mu_j=\frac{\sum_{i=1}^n\gamma_{ij}\mathbf{x}_i}{\sum_{i=1}^n\gamma_{ij}}$$

5 **until** *the objective function J no longer changes*;

(1) **[10pts]** 试证明, 在算法1中, **Step 1** 和 **Step 2** 都会使目标函数 J 的值降低.

(2) **[10pts]** 试证明, 算法1会在有限步内停止。

(3) **[10pts]** 试证明, 目标函数 J 的最小值是关于 k 的非增函数, 其中 k 是聚类簇的数目。

(4) **[20pts]** 记 $\hat{\mathbf{x}}$ 为 n 个样本的中心点, 定义如下变量,

total deviation	$T(X) = \sum_{i=1}^{n} \lVert \mathbf{x}_i - \hat{\mathbf{x}} \rVert^2 / n$
intra-cluster deviation	$W_j(X) = \sum_{i=1}^{n} \gamma_{ij} \lVert \mathbf{x}_i - \mu_j \rVert^2 / \sum_{i=1}^{n} \gamma_{ij}$
inter-cluster deviation	$B(X) = \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} \lVert \mu_j - \hat{\mathbf{x}} \rVert^2$

试探究以上三个变量之间有什么样的等式关系? 基于此, 请证明, k-means 聚类算法可以认为是在最小化 intra-cluster deviation 的加权平均, 同时近似最大化 inter-cluster deviation.

(5) **[20pts]** 在公式(3.1)中, 我们使用 ℓ_2-范数来度量距离 (即欧式距离), 下面我们考虑使用 ℓ_1-范数来度量距离

$$J'(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} ||\mathbf{x}_i - \mu_j||_1 \tag{3.2}$$

- **[10pts]** 请仿效算法1(k-means-ℓ_2 算法), 给出新的算法 (命名为 k-means-ℓ_1 算法) 以优化公式3.2中的目标函数 J'.
- **[10pts]** 当样本集中存在少量异常点 (outliers) 时, 上述的 k-means-ℓ_2 和 k-means-ℓ_1 算法, 我们应该采用哪种算法? 即, 哪个算法具有更好的鲁棒性? 请说明理由。

Solution.

(1) *Proof.* In Step 1, a sample will be reassigned to another class if its previous nearest-class distance is not minimal. Specifically, for each sample $\boldsymbol{x}_i$, suppose its previous indicator vector was $\boldsymbol{\gamma}_i$ and is updated in Step 1 to be $\hat{\boldsymbol{\gamma}}_{\boldsymbol{i}}$. By the update criteria, we have

$$\sum_{j=1}^{k} \hat{\gamma}_{ij} \lVert \boldsymbol{x}_i - \boldsymbol{\mu}_j \rVert^2 = \sum_{\hat{\gamma}_{ij}=1} \lVert \boldsymbol{x}_i - \boldsymbol{\mu}_j \rVert^2 \le \sum_{\gamma_{ij}=1} \lVert \boldsymbol{x}_i - \boldsymbol{\mu}_j \rVert^2 = \sum_{j=1}^{k} \gamma_{ij} \lVert \boldsymbol{x}_i - \boldsymbol{\mu}_j \rVert^2. \tag{3.3}$$

Therefore, we get

$$\hat{J} = \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{\gamma}_{ij} \lVert \boldsymbol{x}_i - \boldsymbol{\mu}_j \rVert^2 \le \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \lVert \boldsymbol{x}_i - \boldsymbol{\mu}_j \rVert^2 = J, \tag{3.4}$$

so J is decreasing after Step 1.

In Step 2, suppose the updated cluster centers are $\hat{\boldsymbol{\mu}}_j (j = 1, 2, \cdots, k)$, while the previous ones were $\boldsymbol{\mu}_j$s. We define function $f_j(\boldsymbol{\mu}) = \sum_{i=1}^{n} \gamma_{ij} \lVert \boldsymbol{x}_i - \boldsymbol{\mu} \rVert^2$, and by setting the derivative $\frac{df_j(\boldsymbol{\mu})}{d\boldsymbol{\mu}} = \sum_{i=1}^{n} 2\gamma_{ij}(\boldsymbol{\mu} - \boldsymbol{x}_i)$ as 0, we obtain

$$\sum_{i=1}^{n} (\gamma_{ij} \tilde{\boldsymbol{\mu}} - \gamma_{ij} \boldsymbol{x}_i) = 0 \quad \Rightarrow \quad \tilde{\boldsymbol{\mu}} \sum_{i=1}^{n} \gamma_{ij} = \sum_{i=1}^{n} \gamma_{ij} \boldsymbol{x}_i \quad \Rightarrow \quad \tilde{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} \gamma_{ij} \boldsymbol{x}_i}{\sum_{i=1}^{n} \gamma_{ij}}, \tag{3.5}$$

which is exactly the expression in Step 2. Therefore, we have

$$\sum_{i=1}^{n}\gamma_{ij}\|\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_j\|^2 = f_j(\hat{\boldsymbol{\mu}}_j) = \min_{\boldsymbol{\mu}}\{f_j(\boldsymbol{\mu})\} \le f_j(\boldsymbol{\mu}_j) = \sum_{i=1}^{n}\gamma_{ij}\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2, \quad \forall j\in\{1,2,\cdots,k\}, \tag{3.6}$$

$$\hat{J} = \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\|\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_{\boldsymbol{j}}\|^2 \le \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2 = J, \tag{3.7}$$

so J is decreasing after Step 2. □

(2) *Proof.* In Step 2, we set all $\boldsymbol{\mu}_j$ according to γ_{ij}, so given $\boldsymbol{\gamma}$, we can determine the J value (after some iteration) as follow:

$$J(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\left\|\boldsymbol{x}_i-\frac{\sum_{s=1}^{n}\gamma_{sj}\boldsymbol{x}_s}{\sum_{s=1}^{n}\gamma_{sj}}\right\|^2. \tag{3.8}$$

Because $\boldsymbol{\gamma}$, as the indicator matrix, have a finite number (namely, k^n) of possible assignments, $J(\boldsymbol{\gamma})$ also has a finite number ($\le k^n$) of possible values.

Now, if we define J_t as the J value after the tth iteration, in Exercise (1) we have shown that J_t is a non-increasing sequence. Hence, we can roughly run the k-means algorithm for k^n+1 iterations, and according to the Pigeonhole Principle, there must be at least two equal values in the sequence $J_1, J_2, \cdots, J_{k^n+1}$, and these two values must be consecutive since the sequence J_t is non-increasing. Thus, the algorithm must terminate in at most k^n+1 iterations. □

(3) *Proof.* We denote the minimum J value when dividing samples into k clusters as $J_{\min}^{(k)}$. Here, we show that for all integer $k(k\ge 1)$, the inequality $J_{\min}^{(k)} \ge J_{\min}^{(k+1)}$ always holds.

For any $k\ge 1$, suppose when $J^{(k)}$ reaches minimum, the cluster division is $\mathcal{C}^{(k)} = \{C_1, C_2, \cdots, C_k\}$, and the minimum value is given by

$$J_{\min}^{(k)} = \sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{I}(\boldsymbol{x}_i\in C_j)\cdot\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2. \tag{3.9}$$

Now, we randomly pick one cluster (for instance C_k, without loss of generality) such that $|C_k|\ge 2$. Note that if such cluster doesn't exist, which indicates that each cluster only has one sample, a $(k+1)$-clustering simply doesn't make sense: a trivial case! Thus, we can split one sample $\boldsymbol{x}_p$ out of C_k to be a new cluster $\hat{C}_{k+1} = \{\boldsymbol{x}_p\}$, so the new division $\hat{\mathcal{C}}^{(k+1)} = \{\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_{k+1}\}$ is a $(k+1)$-clustering, in which

$$\hat{C}_i = C_i, \quad \forall i\in\{1,2,\cdots,k-1\}, \tag{3.10}$$

$$\hat{C}_k = C_k\backslash\{\boldsymbol{x}_p\}, \quad \hat{C}_{k+1} = \{\boldsymbol{x}_p\}, \tag{3.11}$$

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i, \quad \forall i\in\{1,2,\cdots,k\}, \tag{3.12}$$

$$\hat{\boldsymbol{\mu}}_{k+1} = \boldsymbol{x}_p; \tag{3.13}$$

and its J value satisfies

$$\begin{aligned}
\hat{J}^{(k+1)} &= \sum_{i=1}^{n}\sum_{j=1}^{k+1}\mathbb{I}(\boldsymbol{x}_i \in \hat{C}_j)\cdot\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2 \\
&= \sum_{\boldsymbol{x}_i\in\hat{C}_k}\|\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k\|^2+\sum_{\boldsymbol{x}_i\in\hat{C}_{k+1}}\|\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_{k+1}\|^2+\sum_{j=1}^{k-1}\sum_{i=1}^{n}\mathbb{I}(\boldsymbol{x}_i\in\hat{C}_j)\cdot\|\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_j\|^2 \\
&= \sum_{\boldsymbol{x}_i\in\hat{C}_k}\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\|^2+\sum_{j=1}^{k-1}\sum_{i=1}^{n}\mathbb{I}(\boldsymbol{x}_i\in C_j)\cdot\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2 \\
&\le \sum_{i=1}^{n}\mathbb{I}(\boldsymbol{x}_i\in C_k)\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\|^2+\sum_{j=1}^{k-1}\sum_{i=1}^{n}\mathbb{I}(\boldsymbol{x}_i\in C_j)\cdot\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{I}(\boldsymbol{x}_i\in C_j)\cdot\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2 \\
&= J_{\min}^{(k)}. \qquad (3.14)
\end{aligned}$$

Since $\hat{\mathcal{C}}^{(k+1)}$ is just one way to separate samples into $(k+1)$ clusters, its J value must be greater than the $(k+1)$-clustering minimum $J_{\min}^{(k+1)}$. Therefore, $J_{\min}^{(k)} \ge \hat{J}^{(k+1)} \ge J_{\min}^{(k+1)}$, so the minimum value of J is a non-increasing function of k. □

(4)

$$nT(\boldsymbol{X}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(\boldsymbol{X}) + nB(\boldsymbol{X}). \qquad (3.15)$$

Proof.

$$\begin{aligned}
nT(\boldsymbol{X}) &= \sum_{i=1}^{n}(\boldsymbol{x}_i-\hat{\boldsymbol{x}})^{\mathrm{T}}(\boldsymbol{x}_i-\hat{\boldsymbol{x}}) = \sum_{i=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i-2\hat{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x}_i+\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}}) \\
&= \sum_{i=1}^{n}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i-2\hat{\boldsymbol{x}}^{\mathrm{T}}\sum_{i=1}^{n}\boldsymbol{x}_i+n\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}} = \sum_{i=1}^{n}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i-n\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i-n\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2+\sum_{i=1}^{n}\sum_{j=1}^{k}2\gamma_{ij}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{x}_i-\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j-n\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(\boldsymbol{X})+\sum_{j=1}^{k}2\boldsymbol{\mu}_j^{\mathrm{T}}\sum_{i=1}^{n}\gamma_{ij}\boldsymbol{x}_i-\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j-n\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(\boldsymbol{X})+\sum_{j=1}^{k}2\boldsymbol{\mu}_j^{\mathrm{T}}\sum_{i=1}^{n}\gamma_{ij}\boldsymbol{\mu}_j-\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j-n\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(\boldsymbol{X})+\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j-n\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(\boldsymbol{X})+\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}(\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j+\hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}})-2\left(\sum_{i=1}^{n}\hat{\boldsymbol{x}}^{\mathrm{T}}\right)\hat{\boldsymbol{x}}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(\boldsymbol{X}) + \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}(\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j + \hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}}) - 2\left(\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\boldsymbol{\mu}_j^{\mathrm{T}}\right)\hat{\boldsymbol{x}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(\boldsymbol{X}) + \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}(\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j - 2\boldsymbol{\mu}_j^{\mathrm{T}}\hat{\boldsymbol{x}} + \hat{\boldsymbol{x}}^{\mathrm{T}}\hat{\boldsymbol{x}}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(\boldsymbol{X}) + nB(\boldsymbol{X}).
\end{aligned}
\tag{3.16}
$$

□

In k-means, the objective function

$$
\begin{aligned}
J(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\left\|\boldsymbol{x}_i - \frac{\sum_{s=1}^{n}\gamma_{sj}\boldsymbol{x}_s}{\sum_{s=1}^{n}\gamma_{sj}}\right\|^2 = \sum_{j=1}^{k}\min_{\boldsymbol{\mu}_j}\sum_{i=1}^{n}\gamma_{ij}\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2 = \sum_{j=1}^{k}\min_{\boldsymbol{\mu}_j}\sum_{i=1}^{n}\gamma_{ij}W_j(\boldsymbol{X}) \\
&= \min_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k}\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}\right)W_j(\boldsymbol{X}) \\
&= \min_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k}(nT(\boldsymbol{X}) - nB(\boldsymbol{X})) = nT(\boldsymbol{X}) - n\max_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k}B(\boldsymbol{X}).
\end{aligned}
\tag{3.17}
$$

Therefore, the algorithm can be considered as minimizing the weighted average of $W_j(\boldsymbol{X})$; since $T(\boldsymbol{X})$ is constant, the algorithm is also maximizing $B(\boldsymbol{X})$.

(5) The k-means-ℓ_1 algorithm is shown in Algo.(2). The ℓ_1 version is more robust for outliers, because when updating $\boldsymbol{\mu}$, we take the median instead of mean value, making outliers negligible in operation.

Algorithm 2: k-means-ℓ_1 Algorithm

1 Initialize $\mu_1, \ldots, \mu_k$.

2 **repeat**

3 **Step 1**: Decide the class memberships of $\{\boldsymbol{x}_i\}_{i=1}^n$ by assigning each of them to its nearest cluster center.

$$
\gamma_{ij} = \begin{cases} 1, & \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_1 \le \|\boldsymbol{x}_i - \boldsymbol{\mu}_{j'}\|_1, \forall j' \\ 0, & \text{otherwise} \end{cases}
$$

4 **Step 2**: For each $j \in \{1, \cdots, k\}$, recompute μ_j using the updated γ to be the center of mass of all points in C_j:

$$
\mu_j = \mathrm{med}(\{\boldsymbol{x}_i | \gamma_{ij} = 1\}),
$$

in which $\{\boldsymbol{x}_i | \gamma_{ij} = 1\}$ means the set of samples in cluster C_j, and $\mathrm{med}(\cdot)$ means taking the median for each dimension.

5 **until** *the objective function J no longer changes*;

4 [50pts] Kernel, Optimization and Learning

给定样本集 $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, $\mathcal{F} = \{\Phi_1 \cdots, \mathbf{\Phi}_d\}$ 为非线性映射族。考虑如下的优化问题

$$\min_{\mathbf{w}, \mu \in \Delta_q} \quad \frac{1}{2} \sum_{k=1}^{d} \frac{1}{\mu_k} \|\mathbf{w}_k\|_2^2 + C \sum_{i=1}^{m} \max \left\{ 0, 1 - y_i \left(\sum_{k=1}^{d} \mathbf{w}_k \cdot \mathbf{\Phi}_k(\mathbf{x}_i) \right) \right\} \tag{4.1}$$

其中, $\Delta_q = \{\boldsymbol{\mu} | \mu_k \geq 0, k = 1, \cdots, d; \|\boldsymbol{\mu}\|_q = 1\}$.

(1) **[30pts]** 请证明, 下面的问题4.2是优化问题4.1的对偶问题。

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha} \end{matrix} \right\|_p \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{aligned} \tag{4.2}$$

其中, p 和 q 满足共轭关系, 即 $\frac{1}{p} + \frac{1}{q} = 1$. 同时, $\mathbf{Y} = \mathrm{diag}([y_1, \cdots, y_m])$, $\mathbf{K}_k$ 是由 $\mathbf{\Phi}_k$ 定义的核函数 (kernel).

(2) **[20pts]** 考虑在优化问题4.2中, 当 $p = 1$ 时, 试化简该问题。

Solution.

(1) *Proof.* Define the slack variables $\xi_i \geq 0 (i = 1, \cdots m)$, so the problem is equivalent as

$$\begin{aligned} \min_{\boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\xi}} \quad & \frac{1}{2} \sum_{k=1}^{d} \frac{1}{\mu_k} \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{w}_k + C \sum_{i=1}^{m} \xi_i, \\ \text{s.t.} \quad & \sum_{k=1}^{d} \mu_k^q = 1; \quad \mu_k \geq 0, \quad k = 1, 2, \cdots, d; \\ & y_i \left(\sum_{k=1}^{d} \boldsymbol{w}_k^{\mathrm{T}} \cdot \mathbf{\Phi}_k(\boldsymbol{x_i}) \right) \geq 1 - \xi_i, \quad i = 1, 2, \cdots m; \\ & \xi_i \geq 0, \quad i = 1, 2, \cdots m. \end{aligned} \tag{4.3}$$

The Lagrange function is given by

$$\begin{aligned} L(\boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\xi}, \lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \quad = \quad & \frac{1}{2} \sum_{k=1}^{d} \frac{1}{\mu_k} \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{w}_k + C \sum_{i=1}^{m} \xi_i + \lambda (\sum_{k=1}^{d} \mu_k^q - 1) + \\ + \quad & \sum_{i=1}^{m} \alpha_i [1 - \xi_i - y_i (\sum_{k=1}^{d} \boldsymbol{w}_k^{\mathrm{T}} \cdot \mathbf{\Phi}_k(\boldsymbol{x_i}))] + \sum_{i=1}^{m} \beta_i (-\xi_i) + \sum_{k=1} \gamma_k (-\mu_k), \end{aligned} \tag{4.4}$$

where λ, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ are Lagrange multipliers and satisfy $\alpha_i \geq 0$, $\beta_i \geq 0$ $(i = 1, 2, \cdots, m)$. Setting the partial derivatives $\frac{\partial L}{\partial \boldsymbol{w}_k}$, $\frac{\partial L}{\partial \mu_k}$ and $\frac{\partial L}{\partial \xi_i}$ to 0, we get

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{w}_k} &= \frac{1}{\mu_k}\boldsymbol{w}_k - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{\Phi}_k(\boldsymbol{x_i}) = 0, && (4.5)\\
\frac{\partial L}{\partial \mu_k} &= -\frac{1}{2\mu_k^2}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + \lambda q \mu_k^{q-1} - \gamma_k = 0, && (4.6)\\
\frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0, && (4.7)
\end{aligned}$$

or equivalently,

$$\begin{aligned}
\boldsymbol{w}_k = \mu_k \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{\Phi}_k(\boldsymbol{x_i}), &\qquad \forall k \in \{1, 2, \cdots, d\}; && (4.8)\\
\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k = 2\lambda q \mu_k^{q+1} - 2\mu_k^2 \gamma_k, &\qquad \forall k \in \{1, 2, \cdots, d\}; && (4.9)\\
\alpha_i + \beta_i = C, &\qquad \forall i \in \{1, 2, \cdots, m\}. && (4.10)
\end{aligned}$$

For the sake of simplicity, we define the variables $M_k (k = 1, 2, \cdots, d)$ as

$$M_k = \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_i)\boldsymbol{\Phi}_k(\boldsymbol{x}_j). \tag{4.11}$$

It is easy to prove that $M_k = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_k\boldsymbol{Y}\boldsymbol{\alpha}$. Specifically, by its definition Eq.(4.11),

$$\begin{aligned}
M_k &= (\alpha_1 y_1, \cdots, \alpha_m y_m) \begin{pmatrix} \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_1)\boldsymbol{\Phi}_k(\boldsymbol{x}_1) & \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_1)\boldsymbol{\Phi}_k(\boldsymbol{x}_2) & \cdots & \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_1)\boldsymbol{\Phi}_k(\boldsymbol{x}_m) \\ \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_2)\boldsymbol{\Phi}_k(\boldsymbol{x}_1) & \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_2)\boldsymbol{\Phi}_k(\boldsymbol{x}_2) & \cdots & \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_2)\boldsymbol{\Phi}_k(\boldsymbol{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_m)\boldsymbol{\Phi}_k(\boldsymbol{x}_1) & \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_m)\boldsymbol{\Phi}_k(\boldsymbol{x}_2) & \cdots & \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_m)\boldsymbol{\Phi}_k(\boldsymbol{x}_m) \end{pmatrix} \begin{pmatrix} \alpha_1 y_1 \\ \alpha_2 y_2 \\ \vdots \\ \alpha_m y_m \end{pmatrix} \\
&= (\alpha_1, \alpha_2, \cdots, \alpha_m) \begin{pmatrix} y_1 & 0 & \cdots & 0 \\ 0 & y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_m \end{pmatrix} \boldsymbol{K}_k \begin{pmatrix} y_1 & 0 & \cdots & 0 \\ 0 & y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_m \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} \\
&= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_k\boldsymbol{Y}\boldsymbol{\alpha}, && (4.12)
\end{aligned}$$

where $\boldsymbol{Y} = \mathrm{diag}(y_1, \cdots, y_m)$ and $\boldsymbol{K}_k$ is the kernel matrix of mapping $\boldsymbol{\Phi}_k(\cdot)$.

Now, we will derive the dual problem. The Lagrange function Eq.(4.4) could be rewritten as

$$\begin{aligned}
L(\boldsymbol{\alpha}) &= \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + \sum_{i=1}^{m}(C-\alpha_i-\beta_i)\xi_i + \lambda(\sum_{k=1}^{d}\mu_k^q - 1) - \sum_{k=1}^{d}\gamma_k\mu_k \\
&\quad + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i y_i(\sum_{k=1}^{d}\boldsymbol{w}_k^{\mathrm{T}}\cdot\boldsymbol{\Phi}_k(\boldsymbol{x_i})) - \sum_{k=1}^{d}\gamma_k\mu_k \\
&= \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i y_i(\sum_{k=1}^{d}\boldsymbol{w}_k^{\mathrm{T}}\cdot\boldsymbol{\Phi}_k(\boldsymbol{x_i})) - \sum_{k=1}^{d}\gamma_k\mu_k \quad (by\ Eq.(4.10)) \\
&= \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + \sum_{i=1}^{m}\alpha_i - \sum_{k=1}^{d}\boldsymbol{w}_k^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{\Phi}_k(\boldsymbol{x_i}) - \sum_{k=1}^{d}\gamma_k\mu_k \\
&= \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + \sum_{i=1}^{m}\alpha_i - \sum_{k=1}^{d}\frac{1}{\mu_k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k - \sum_{k=1}^{d}\gamma_k\mu_k \quad (by\ Eq.(4.8)) \\
&= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k - \sum_{k=1}^{d}\gamma_k\mu_k \\
&= \sum_{i=1}^{m}\alpha_i - \sum_{k=1}^{d}\lambda q\mu_k^q \quad (by\ Eq.(4.9)) \\
&= \sum_{i=1}^{m}\alpha_i - \lambda q. \qquad (4.13)
\end{aligned}$$

In fact, it is easy to see that μ_k can never equal 0, because otherwise, the $\frac{1}{\mu_k}$ term in Eq.(4.1) does not make sense (making the objective function value go to $+\infty$). Thus, the optimal solution could never lie on the boundary of constraint $\mu_k \geq 0$. According to the KKT condition $\gamma_k\mu_k = 0$, we have $\gamma_k = 0$ holds for all $k \in \{1, 2, \cdots, d\}$.

Now, by plugging Eq.(4.8) into Eq.(4.9), we get

$$\begin{aligned}
\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k &= \left(\mu_k\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{\Phi}_k(\boldsymbol{x_i})\right)^{\mathrm{T}}\left(\mu_k\sum_{j=1}^{m}\alpha_j y_j\boldsymbol{\Phi}_k(\boldsymbol{x_j})\right) = \mu_k^2\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_i)\boldsymbol{\Phi}_k(\boldsymbol{x}_j) \\
&= \mu_k^2 M_k = 2\lambda q\mu_k^{q+1} - 2\mu_k^2\gamma_k. \qquad (4.14)
\end{aligned}$$

So, we have $M_k = 2\lambda q\mu_k^{q-1} - 2\gamma_k = 2\lambda q\mu_k^{q-1}$, and therefore, for the conjugate number $p = \frac{q}{q-1}$, we have

$$\sum_{k=1}^{d}M_k^p = \sum_{k=1}^{d}(2\lambda q)^p\mu_k^q = (2\lambda q)^p\sum_{k=1}^{d}\mu_k^q = (2\lambda q)^p. \qquad (4.15)$$

This just indicates that

$$2\lambda q = \left(\sum_{k=1}^{d}M_k^p\right)^{\frac{1}{p}} = \left\|\begin{matrix}M_1\\ \vdots\\ M_d\end{matrix}\right\|_p = \left\|\begin{matrix}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_1\boldsymbol{Y}\boldsymbol{\alpha}\\ \vdots\\ \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_d\boldsymbol{Y}\boldsymbol{\alpha}\end{matrix}\right\|_p. \qquad (4.16)$$

Thus, according to Eq.(4.16), the dual problem's objective function could be

$$\Gamma(\boldsymbol{\alpha}) = 2L(\boldsymbol{\alpha}) = 2\sum_{i=1}^{m}\alpha_i - 2\lambda q = 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_1\boldsymbol{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_d\boldsymbol{Y}\boldsymbol{\alpha} \end{matrix} \right\|_p . \tag{4.17}$$

The corresponding KKT conditions are: for $\forall i \in \{1, 2, \cdots, m\}$, $\forall k \in \{1, 2, \cdots, d\}$

$$\begin{cases} \alpha_i \geq 0, \quad \beta_i \geq 0, \\ y_i(\sum_{k=1}^{d} \boldsymbol{w}_k \cdot \boldsymbol{\Phi}_k(\boldsymbol{x_i})) \geq 1 - \xi_i, \\ \alpha_i(y_i(\sum_{k=1}^{d} \boldsymbol{w}_k \cdot \boldsymbol{\Phi}_k(\boldsymbol{x_i})) - 1 + \xi_i) = 0, \\ \xi_i \geq 0, \quad \beta_i\xi_i = 0, \\ \mu_k \geq 0, \quad \gamma_k\mu_k = 0. \end{cases} \tag{4.18}$$

Since $\alpha_i + \beta_i = C$, $\alpha_i \geq 0$, $\beta_i \geq 0$, we conclude that $\mathbf{0} \leq \boldsymbol{\alpha} \leq \boldsymbol{C}$. Consequently, we achieve the dual problem as

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_1\boldsymbol{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_d\boldsymbol{Y}\boldsymbol{\alpha} \end{matrix} \right\|_p , \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \boldsymbol{C}. \end{aligned} \tag{4.19}$$

$\square$

(2) Taking $p = 1$, we claim that the optimization problem would transform into a classic kernelized soft margin SVM (and without intercept term).

Recall that the kernel matrices $\boldsymbol{K}_k$ are positive semi-definite, so $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_k\boldsymbol{Y}\boldsymbol{\alpha} \geq 0$. We obtain the objective function, given $p = 1$, in Eq.(4.2) as

$$\begin{aligned} \Gamma(\boldsymbol{\alpha}) &= 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_1\boldsymbol{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_d\boldsymbol{Y}\boldsymbol{\alpha} \end{matrix} \right\|_1 \\ &= 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \sum_{k=1}^{d} \left| \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{K}_k\boldsymbol{Y}\boldsymbol{\alpha} \right| \\ &= 2\sum_{i=1}^{m}\alpha_i - \sum_{k=1}^{d}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{\Phi}_k^{\mathrm{T}}(\boldsymbol{x}_i)\boldsymbol{\Phi}_k(\boldsymbol{x}_j) \\ &= 2\sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \sum_{k=1}^{d}\boldsymbol{K}_k(\boldsymbol{x}_i, \boldsymbol{x}_j). \end{aligned} \tag{4.20}$$

Therefore, if we define the linear combination $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{d}\boldsymbol{K}_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, which is also a valid kernel function, the dual problem Eq.(4.2) is equivalent as

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j), \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \boldsymbol{C}. \end{aligned} \tag{4.21}$$

Notice that Eq.(4.21) is exactly the same as the dual problem of the kernelized soft margin SVM, except that the intercept term b in the boundary expression $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$ is dropped (so one constraint is deleted). ln conclusion, setting $p=1$, the dual problem would decay into a classic kernelized soft margin SVM (without intercept term).

Remark. In fact, if we take the conjugate of 1 as ∞, and start from the primal problem's prospective, we can get the same consistent result.

Given $p=1$, $\frac{1}{p}+\frac{1}{q}=1$, we get $q=\infty$, so the primal problem constrains that $\|\boldsymbol{\mu}\|_\infty=1$, i.e. $\max_k\{\mu_k\}=1$. Thus, the objective function in Eq.(4.1) satisfies

$$
\begin{aligned}
f(\boldsymbol{W},\boldsymbol{\mu}) &= \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\boldsymbol{w}_k\|_2^2 + C\sum_{i=1}^{m}\max\left\{0, 1-y_i\left(\sum_{k=1}^{d}\boldsymbol{w}_k\cdot\boldsymbol{\Phi}_k(\boldsymbol{x}_i)\right)\right\} \\
&\geq \frac{1}{2}\sum_{k=1}^{d}\|\boldsymbol{w}_k\|_2^2 + C\sum_{i=1}^{m}\max\left\{0, 1-y_i\left(\sum_{k=1}^{d}\boldsymbol{w}_k\cdot\boldsymbol{\Phi}_k(\boldsymbol{x}_i)\right)\right\} \\
&= f(\boldsymbol{W},\mathbf{1}).
\end{aligned}
\tag{4.22}
$$

Therefore, the function $f(\boldsymbol{W},\boldsymbol{\mu})$ must reach its minimum at $\boldsymbol{\mu}=\mathbf{1}$. Plug $\boldsymbol{\mu}=\mathbf{1}$ into Eq.(4.3), we get the primal problem as

$$
\begin{aligned}
\min_{\boldsymbol{W},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{k=1}^{d}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + C\sum_{i=1}^{m}\xi_i, \\
\text{s.t.} \quad & y_i\left(\sum_{k=1}^{d}\boldsymbol{w}_k^{\mathrm{T}}\cdot\boldsymbol{\Phi}_k(\boldsymbol{x}_i)\right) \geq 1-\xi_i, \quad i=1,2,\cdots m; \\
& \xi_i \geq 0, \quad i=1,2,\cdots m,
\end{aligned}
\tag{4.23}
$$

which is obviously the classic form of a kernelized soft margin SVM (without intercept term b).

Reference

[1] Wikipedia contributors. "Exponential family." Wikipedia, The Free Encyclopedia, 16 May. 2017. Available at: `https://en.wikipedia.org/wiki/Exponential_family#Bayesian_estimation:_conjugate_distributions` [Accessed 11 Jun. 2017]