An Efficient and Provable Approach for Mixture Proportion Estimation Using Linear Independence Assumption Supplementary Materials

Xiyu Yu[1] Tongliang Liu[1] Mingming Gong[2,3] Kayhan Batmanghelich[2] Dacheng Tao[1]
[1]UBTECH Sydney AI Centre, SIT, FEIT, The University of Sydney, Australia
[2]Department of Biomedical Informatics, University of Pittsburgh
[3]Department of Philosophy, Carnegie Mellon University
{xiyu0300@uni., tongliang.liu@, dacheng.tao@}sydney.edu.au {mig73, kayhan}@pitt.edu

Abstract

In this supplementary material, we provide the detailed proof of Proposition 1 and Theorem 3. Proposition 1 shows us that the independence assumption is the weakest assumption of MPE problem. In Theorem 3, we prove a data-independent error bound for the estimates of the weights, which ensures that the proposed method can uniformly converge to the optimal solution, no matter of data forms.

A. Proof of Proposition 1

Without loss of generality, we consider the case of two component distributions; that is, the mixture $P = \lambda_1 P_1 + \lambda_2 P_2$, where $\lambda_1 \geq 0, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$.

Here comes the proof of Proposition 1.

(i) The irreducibility condition implies the independence assumption while the independence assumption does not imply the irreducibility condition.

Proof. $(i.1)$ The irreducibility condition implies the independence assumption.

We prove this by contradiction. Suppose that P_1 is irreducible to P_2 but not independent with P_2, then there exists no $\gamma \in (0, 1]$ such that

$$P_1 = \gamma P_2 + (1 - \gamma)Q,$$

where Q is a new distribution which is different with P_1 and P_2. There also exist non-zero $v_1, v_2 \in \mathbb{R}$ such that

$$v_1 P_1 + v_2 P_2 = 0.$$

Note that v_1 and v_2 should be both non-zero; otherwise, $P_2 = 0$ or $P_1 = 0$, which is not a distribution. Then we have

$$P_1 = -\frac{v_2}{v_1} P_2.$$

Take the integral of $x \in \mathcal{X}$ on both sides of above equation, we have,

$$1 = \int_x P_1 dx = \int_x -\frac{v_2}{v_1} P_2 dx.$$

Then $-\frac{v_2}{v_1} = 1$, and $P_1 = P_2$, which conflicts with the fact that P_1 and P_2 satisfy the irreducibility condition ($\gamma = 1$).

Note that, according to the proof, two distributions are independent with each other if and only if they are different distributions.

$(i.2)$ The independence assumption does not imply the irreducibility condition.

Here we give a counterexample. We assume that P_2 is a Gaussian distribution with mean value vector $\mathbf{0}$, and Q is another Gaussian distribution with a different mean value vector $\mathbf{1}$. Then we have P_2 and Q are independent with each other. Suppose that $P_1 = 0.5P_2 + 0.5Q$, then P_1 and P_2 are also independent, but P_1 is not irreducible to P_2. □

(ii) The anchor set condition implies the independence assumption while the independence assumption does not imply the anchor set condition.

Proof. $(ii.1)$ The anchor set condition implies the independence assumption.

We prove this by contradiction. It is assumed that P_1, P_2 satisfy the anchor set condition but are not independent. Then there exist non-zero v_1 and v_2 such that $v_1 P_1 + v_2 P_2 = 0$. Similar to the proof of $(i.1)$, we have $P_1 = P_2$, which conflicts with the fact that P_1 has a compact support set not shared with P_2.

$(ii.2)$ The independence assumption does not imply the irreducibility condition.

We find that, in the proof of $(i.1)$, the independence of two component distributions only requires them to be different. It is easy to find out two different distributions with the same

support set, such as, two Gaussian distributions with different mean vectors. But obviously, they do not satisfy the anchor set condition. □

We now complete the proof of Proposition 1.

Remark. Based on the Proposition 1, We find that the independence assumption is the weakest assumption among these assumptions. According to the published result, we actually have: The anchor set condition implies the mutual irreducibility condition, which further implies the independence assumption.

B. Proof of Theorem 3

To prove Theorem 3, we need to first introduce the McDiarmid's inequality [2].

Theorem 1 (McDiarmid's Inequality)**.** *Let $X = \{X_1, X_2, \cdots, X_n\}$ be an i.i.d. sample and X^i be a new sample with the i-th example in X being replaced by an independent example X_i'. If there exist $b_1, b_2, \cdots, b_n > 0$ such that $f : \mathcal{X} \to \mathbb{R}$ satisifies,*

$$|f(X) - f(X^i)| \leq b_i, \forall i \in \{1, \cdots, n\}.$$

Then for any $X \in \mathcal{X}$ and $\epsilon > 0$, the following inequality holds,

$$P(\mathbb{E}[f(X)] - f(X) \geq \epsilon) \leq \exp(\frac{-2\epsilon^2}{\sum_{i=1}^n b_i^2}).$$

Sketch of Proof The error $D(\hat{\boldsymbol{\lambda}}) - D(\boldsymbol{\lambda}^*)$ is actually upper bounded by $2\sup_{\boldsymbol{\lambda}} |D(\boldsymbol{\lambda}) - \hat{D}(\boldsymbol{\lambda})|$, which is similiar to the relationship between the consistency and generalization error [5]. Thus, we need only to upper bound the later term. We define a function f,

$$f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda}) = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \psi(x_j) - \sum_{i=1}^c \frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\right] - \frac{1}{n}\sum_{j=1}^n \psi(x_j) + \sum_{i=1}^c \frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i).$$

It is observed that $2\sup_{\boldsymbol{\lambda}} |D(\boldsymbol{\lambda}) - \hat{D}(\boldsymbol{\lambda})|$ can be upper bounded by $\sup_{\boldsymbol{\lambda}} \|f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\|$, further be bounded by the Rademacher-like term $\mathbb{E}\sup_{\boldsymbol{\lambda}} \|f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\|^2$, which can be also bounded as in [1]. Then we can apply the McDiarmid's inequality [2] to obtain the final result. Note that the norm $\|\cdot\|$ denotes the l_2 norm for vectors and Frobenius norm for matrices.

Step one. We first upper bound the error $D(\hat{\boldsymbol{\lambda}}) - D(\boldsymbol{\lambda}^*)$ using the term $\sup_{\boldsymbol{\lambda}} \|f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\|$, where

$$f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda}) = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \psi(x_j) - \sum_{i=1}^c \lambda_i \frac{1}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\right] - \frac{1}{n}\sum_{j=1}^n \psi(x_j) + \sum_{i=1}^c \frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i).$$

Here comes the lemma,

Lemma 1. *We denote $\Delta \triangleq \{\boldsymbol{\lambda} | \lambda_i \geq 0, \forall i \in \{1, \cdots, c\}, \sum_{i=1}^c \lambda_i = 1\}$. Then we have*

$$D(\hat{\boldsymbol{\lambda}}) - D(\boldsymbol{\lambda}^*) \leq 4r \sup_{\boldsymbol{\lambda}\in\Delta} \|f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\|.$$

Proof. We observe that

$$\begin{aligned} D(\hat{\boldsymbol{\lambda}}) - D(\boldsymbol{\lambda}^*) &= D(\hat{\boldsymbol{\lambda}}) - \hat{D}(\hat{\boldsymbol{\lambda}}) \\ &\quad + \hat{D}(\hat{\boldsymbol{\lambda}}) - \hat{D}(\boldsymbol{\lambda}^*) + \hat{D}(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}^*) \\ &\leq D(\hat{\boldsymbol{\lambda}}) - \hat{D}(\hat{\boldsymbol{\lambda}}) + \hat{D}(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}^*) \\ &\leq 2\sup_{\boldsymbol{\lambda}\in\Delta} |D(\boldsymbol{\lambda}) - \hat{D}(\boldsymbol{\lambda})|. \end{aligned}$$

where the first inequality holds due to the fact that $\hat{\boldsymbol{\lambda}}$ minimizes $\hat{D}(\boldsymbol{\lambda})$, and thus $\hat{D}(\hat{\boldsymbol{\lambda}}) - \hat{D}(\boldsymbol{\lambda}^*) \leq 0$.

According to the definition of f, we have

$$\begin{aligned} &\|f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\| \\ &\leq \|\mathbb{E}\frac{1}{n}\sum_{j=1}^n \psi(x_j)\| + \|\mathbb{E}\sum_{i=1}^c \frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\| \\ &+ \|\frac{1}{n}\sum_{j=1}^n \psi(x_j)\| + \|\sum_{i=1}^c \frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\| \\ &\leq 4r. \end{aligned}$$

Denote

$$g(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda}) = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \psi(x_j) - \sum_{i=1}^c \frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\right] + \frac{1}{n}\sum_{j=1}^n \psi(x_j) - \sum_{i=1}^c \frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i).$$

We also have

$$\|g(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\| \leq 4r.$$

Further, we have

$$\begin{aligned} |D(\boldsymbol{\lambda}) - \hat{D}(\boldsymbol{\lambda})| &= \left|g(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda}) f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\right| \\ &\leq \|g(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\| \|f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\| \\ &\leq 4r\|f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\| \\ &\leq 4r \sup_{\boldsymbol{\lambda}\in\Delta} \|f(\mathbf{x}^M, \mathbf{x}^C, \boldsymbol{\lambda})\|, \end{aligned} \tag{1}$$

where the first inequality holds due to the Cauchy-Schwarz inequality. □

Step two. We try to upper bound the term $\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|$. This can be done by first upper bounding the Rademacher-like term $\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2$.

Lemma 2. *If the kernel is characteristic, and be upper bounded by $\|\psi(x)\|\leq r$ for all $x\in\mathcal{X}$. Then we have*

$$\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2\leq 8r^2(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{n_0}}),$$

where $n_0=\min(n_1,\cdots,n_c)$.

Proof. Note that

$$\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2\leq 4r\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|.$$

Here the expectation is taken on the i.i.d. samples $\mathbf{x}^M$ and $\mathbf{x}^C$. Now we introduce the "ghost" samples $\mathbf{x'}^M$ and $\mathbf{x'}^C$ [3] which is an independent sample. Then

$$\begin{aligned}
&\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|\\
&=\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\psi(x_j)-\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\right]\\
&-\frac{1}{n}\sum_{j=1}^{n}\psi(x_j)+\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\|\\
&=\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|\mathbb{E}_{\mathbf{x'}^M,\mathbf{x'}^C}\left[\frac{1}{n}\sum_{j=1}^{n}\psi(x'_j)-\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j'^i)\right]\\
&-\frac{1}{n}\sum_{j=1}^{n}\psi(x_j)+\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\|\\
&\leq\mathbb{E}_{\mathbf{x}^M,\mathbf{x}^C,\mathbf{x'}^M,\mathbf{x'}^C}\sup_{\boldsymbol{\lambda}\in\Delta}\|\frac{1}{n}\sum_{j=1}^{n}\psi(x'_j)-\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j'^i)\\
&-\frac{1}{n}\sum_{j=1}^{n}\psi(x_j)+\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\psi(x_j^i)\|\\
&=\mathbb{E}_{\mathbf{x}^M,\mathbf{x}^C,\mathbf{x'}^M,\mathbf{x'}^C}\sup_{\boldsymbol{\lambda}\in\Delta}\|\frac{1}{n}\sum_{j=1}^{n}\left(\psi(x'_j)-\psi(x_j)\right)\\
&-\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\left(\psi(x_j'^i)-\psi(x_j^i)\right)\|,
\end{aligned}$$

where the first inequality holds due to the Jensen's inequality. The norm and $\sup(\cdot)$ are convex.

Let $\sigma_1,\cdots,\sigma_n,\ \sigma_1^i,\cdots,\sigma_{n_i}^i,\forall i\in\{1,\cdots,c\}$ be the independent random variables such that $P(\sigma=1)=P(\sigma=-1)=1/2$, which are also known as Rademacher variables. Due to the fact that $\mathbf{x'}^M$ and $\mathbf{x'}^C$ are the i.i.d. copies of $\mathbf{x}^M$ and $\mathbf{x}^C$, then the random variable $\frac{1}{n}\sum_{j=1}^{n}\left(\psi(x'_j)-\psi(x_j)\right)-\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\left(\psi(x_j'^i)-\psi(x_j^i)\right)$ is symmetric, and it has the same distribution with $\frac{1}{n}\sum_{j=1}^{n}\sigma_j\left(\psi(x'_j)-\psi(x_j)\right)-\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\sigma_j^i\left(\psi(x_j'^i)-\psi(x_j^i)\right)$. Thus, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}^M,\mathbf{x}^C,\mathbf{x'}^M,\mathbf{x'}^C}\sup_{\boldsymbol{\lambda}\in\Delta}\|\frac{1}{n}\sum_{j=1}^{n}\left(\psi(x'_j)-\psi(x_j)\right)\\
&-\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\left(\psi(x_j'^i)-\psi(x_j^i)\right)\|\\
&=\mathbb{E}_{\mathbf{x}^M,\mathbf{x}^C,\mathbf{x'}^M,\mathbf{x'}^C,\sigma}\sup_{\boldsymbol{\lambda}\in\Delta}\|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\left(\psi(x'_j)-\psi(x_j)\right)\\
&-\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\sigma_j^i\left(\psi(x_j'^i)-\psi(x_j^i)\right)\|\\
&\leq\mathbb{E}_{\mathbf{x}^M,\mathbf{x'}^M,\sigma}\sup_{\boldsymbol{\lambda}\in\Delta}\|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\left(\psi(x'_j)-\psi(x_j)\right)\|\\
&+\mathbb{E}_{\mathbf{x}^C,\mathbf{x'}^C,\sigma}\sup_{\boldsymbol{\lambda}\in\Delta}\|\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\sigma_j^i\left(\psi(x_j'^i)-\psi(x_j^i)\right)\|\\
&\leq 2\mathbb{E}_{\mathbf{x}^M,\sigma}\|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\psi(x_j)\|\\
&+2\mathbb{E}_{\mathbf{x}^C,\sigma}\sup_{\boldsymbol{\lambda}\in\Delta}\|\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\sigma_j^i\psi(x_j^i)\|.
\end{aligned}$$

Now we bound the term $\mathbb{E}_{\mathbf{x}^M,\sigma}\|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\psi(x_j)\|$ and $\mathbb{E}_{\mathbf{x}^C,\sigma}\sup_{\boldsymbol{\lambda}\in\Delta}\|\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\sigma_j^i\psi(x_j^i)\|$, respectively.

We have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}^C,\sigma}\sup_{\boldsymbol{\lambda}\in\Delta}\|\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\sigma_j^i\psi(x_j^i)\|\\
&\leq r\mathbb{E}_{\sigma}\sup_{\boldsymbol{\lambda}\in\Delta}\|\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\sum_{j=1}^{n_i}\sigma_j^i\|\\
&\leq r\sup_{\boldsymbol{\lambda}\in\Delta}\sum_{i=1}^{c}\frac{\lambda_i}{n_i}\mathbb{E}_{\sigma}\|\sum_{j=1}^{n_i}\sigma_j^i\|\\
&\leq\frac{r}{\sqrt{n_0}}.
\end{aligned}\tag{2}$$

where $n_0=\min(n_1,\cdots,n_c)$; and the first inequality holds due to the Talagrand Contraction Lemma [4].

Similarly, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}^M,\sigma}\|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\psi(x_j)\|\\&\leq\frac{r}{n}\mathbb{E}_\sigma\sqrt{\left(\sum_{j=1}^{n}\sigma_j\right)^2}\\&\leq\frac{r}{\sqrt{n}}.\end{aligned}\tag{3}$$

where the first inequality holds due to the Talagrand Contraction Lemma [4].

The combining above results, we have our conclusion

$$\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2\leq 8r^2(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{n_0}}).$$

□

Step three. New we can finally conclude the proof. Denote $\mathbf{x}^{iC}$ as the design matrix of the sample drawn from P_i. Let $\mathbf{x}^{Mp}$ be a new sample drawn from the mixture with the p-th example in $\mathbf{x}^M$ being replaced by an independent example x'_p, where $p\in\{1,\cdots,n\}$, and $\mathbf{x}^{iCq}$ be a new sample drawn from the component P_i with the q-th example in $\mathbf{x}^{iC}$ being replaced by an independent example x'^i_q, where $q\in\{1,\cdots,n_i\}$, for all $i\in\{1,\cdots,c\}$.

Then for any $p\in\{1,\cdots,n\}$, we have

$$\begin{aligned}&\left|\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^{Mp},\mathbf{x}^C,\boldsymbol{\lambda})\|^2-\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2\right|\\&\leq\sup_{\boldsymbol{\lambda}\in\Delta}\left|f(\mathbf{x}^{Mp},\mathbf{x}^C,\boldsymbol{\lambda})+f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\right|\\&\cdot\left|f(\mathbf{x}^{Mp},\mathbf{x}^C,\boldsymbol{\lambda})-f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\right|\\&\leq 8r\left|\frac{1}{n}\left(\psi(x'_p)-\psi(x_p)\right)\right|\\&\leq\frac{8r^2}{n}.\end{aligned}$$

Similarly, for any $q\in\{1,\cdots,n_i\}$ and $i\in\{1,\cdots,c\}$, we have

$$\begin{aligned}&\left|\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^{Mp},\mathbf{x}^{C\backslash i},\mathbf{x}^{iCq},\boldsymbol{\lambda})\|^2-\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2\right|\\&\leq\sup_{\boldsymbol{\lambda}\in\Delta}\left|f(\mathbf{x}^{Mp},\mathbf{x}^{C\backslash i},\mathbf{x}^{iCq},\boldsymbol{\lambda})+f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\right|\\&\cdot\left|f(\mathbf{x}^M,\mathbf{x}^{C\backslash i},\mathbf{x}^{iCq},\boldsymbol{\lambda})-f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\right|\\&\leq 8r\sup_{\boldsymbol{\lambda}\in\Delta}\left|\frac{\lambda_i}{n_i}\left(\psi(x'^i_q)-\psi(x^i_q)\right)\right|\\&\leq\frac{8r^2}{n_i}.\end{aligned}$$

Then employing the McDiarmid's inequality, we have

$$\begin{aligned}&P(\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2-\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2\geq\epsilon)\\&\leq\exp(\frac{-\epsilon^2}{64r^4(\frac{1}{n}+\sum_{i=1}^{c}\frac{1}{n_i})}).\end{aligned}$$

Let

$$\delta=\exp(\frac{-\epsilon^2}{64r^4(\frac{1}{n}+\sum_{i=1}^{c}\frac{1}{n_i})}).$$

Then for any $\delta>0$, with the probability at least $1-\delta$, we have

$$\begin{aligned}&\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|\\&\leq\sqrt{\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2+8r^2\sqrt{\frac{1}{2}(\frac{1}{n}+\sum_{i=1}^{c}\frac{1}{n_i})\log\frac{1}{\delta}}}.\end{aligned}$$

Then combining the results in Lemma 1 and 2, we have

$$\begin{aligned}&D(\hat{\boldsymbol{\lambda}})-D(\boldsymbol{\lambda}^*)\\&\leq 2\sup_{\boldsymbol{\lambda}\in\Delta}|D(\boldsymbol{\lambda})-\hat{D}(\boldsymbol{\lambda})|\\&\leq 4r\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|\\&\leq 4r\sqrt{\mathbb{E}\sup_{\boldsymbol{\lambda}\in\Delta}\|f(\mathbf{x}^M,\mathbf{x}^C,\boldsymbol{\lambda})\|^2+8r^2\sqrt{\frac{1}{2}(\frac{1}{n}+\sum_{i=1}^{c}\frac{1}{n_i})\log\frac{1}{\delta}}}\\&\leq 4r\sqrt{8r^2(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{n_0}})+8r^2\sqrt{\frac{1}{2}(\frac{1}{n}+\sum_{i=1}^{c}\frac{1}{n_i})\log\frac{1}{\delta}}}\\&=8\sqrt{2}r^2\sqrt{(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{n_0}})+\sqrt{\frac{1}{2}(\frac{1}{n}+\sum_{i=1}^{c}\frac{1}{n_i})\log\frac{1}{\delta}}}.\end{aligned}$$

Then we complete our proof of Theorem 3.

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002. 2

[2] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013. 2

[3] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004. 3

[4] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer Science & Business Media, 2013. 3, 4

[5] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning.* MIT press, 2012. 2