Supplementary Material for Kernel Normalized Cut: a Theoretical Revisit

Yoshikazu Terada [1 2] **Michio Yamamoto** [3 2]

Abstract

In this supplementary material, we provide the proofs of the theoretical results in the main paper, some fundamental properties, and the additional experiments.

At first, we remark the following: since $\exists c_U, c_L > 0;\ \forall x, y \in \mathcal{X};\ c_L < k(x,y) < c_U$,

$$\|\psi_h(x)\|^2_{\mathcal{H}_h} = \|h(x,\cdot)\|^2_{\mathcal{H}_h} = h(x,x) = \frac{k(x,x)}{d^2(x)} \leq \frac{c_U}{c_L^2} \quad (\forall x \in \mathcal{X}).$$

Let $D = \sqrt{c_U}/c_L$, and then the image measure of $\mathbb{P}$ by $\psi_h : \mathcal{X} \to \mathcal{H}_h$ is D-bounded, that is, $\mathrm{supp}(\mathbb{P}) \subset \mathcal{B}(0, D)$.

A. Proofs of the results in Section 3.1

A.1. Proof of Lemma 1

Proof. For any binary membership matrix U_n,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \frac{k(X_i,X_i)}{\hat{d}_n(X_i)} - \frac{1}{n}\sum_{m=1}^m\sum_{i=1}^n\sum_{j=1}^n u_{im}u_{jm}\frac{k(X_i,X_j)}{\sum_{i=1}^n u_{im}\hat{d}_n(X_i)} - \frac{1}{n}\sum_{i=1}^n \frac{k(X_i,X_i)}{d(X_i)} + \frac{1}{n}\sum_{m=1}^m\sum_{i=1}^n\sum_{j=1}^n u_{im}u_{jm}\frac{k(X_i,X_j)}{\sum_{i=1}^n u_{im}d(X_i)}\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n k(X_i,X_i)\left|\frac{1}{\hat{d}_n(X_i)} - \frac{1}{d(X_i)}\right| + \frac{1}{n}\sum_{m=1}^m\sum_{i=1}^n\sum_{j=1}^n u_{im}u_{jm}k(X_i,X_j)\left|\frac{1}{\sum_{i=1}^n u_{im}d(X_i)} - \frac{1}{\sum_{i=1}^n u_{im}\hat{d}_n(X_i)}\right| \\
&\leq \frac{c_U}{c_L^2}\|\hat{d}_n - d\|_\infty + \frac{c_U}{n}\sum_{m=1}^m n_m^2\left|\frac{\sum_{i=1}^n u_{im}\hat{d}_n(X_i) - \sum_{i=1}^n u_{im}d(X_i)}{\sum_{i=1}^n\sum_{j=1}^n u_{im}u_{jm}d(X_i)\hat{d}_n(X_j)}\right| \\
&\leq \frac{c_U}{c_L^2}\|\hat{d}_n - d\|_\infty + \frac{c_U}{c_L^2}\sum_{m=1}^m\frac{1}{n}\sum_{i=1}^n u_{im}\|\hat{d}_n - d\|_\infty = \frac{2c_U}{c_L^2}\|\hat{d}_n - d\|_\infty.
\end{aligned}$$

Next, we derive the nonasymptotic upper bound for $\|\hat{d}_n - d\|_\infty$. Let $\mathscr{D} := \{k(x,\cdot) \mid x \in \mathcal{X}\}$. We will denote by $\hat{\mathfrak{R}}_{\mathcal{X}_n}(\mathscr{G})$ the empirical Rademacher complexity of $\mathscr{G} \subset \{f : \mathcal{X} \to \mathbb{R}\}$, that is,

$$\hat{\mathfrak{R}}_{\mathcal{X}_n}(\mathscr{G}) := \mathbb{E}_\epsilon\left[\sup_{g\in\mathscr{G}}\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right],$$

where $\epsilon_1, \ldots, \epsilon_n$ is a sequence of independent random variables taking the values $+1$ or -1 with the same probability $1/2$

[1]Graduate School of Engineering Science, Osaka University, Osaka, Japan [2]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan [3]Graduate School of Environmental and Life Science, Okayama University, Okayama, Japan. Correspondence to: Yoshikazu Terada <terada@sigmath.es.osaka-u.ac.jp>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

and $\mathbb{E}_\epsilon[\cdot]$ is the expectation with respect to $\epsilon_1, \ldots, \epsilon_n$. By the reproducing property and Jensen's inequality, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_{\mathcal{X}_n}(\mathscr{D}) &= \mathbb{E}_\epsilon\left[\sup_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \epsilon_i k(x, X_i)\right] \leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{x\in\mathcal{X}} \sum_{i=1}^n \epsilon_i k(x, X_i)\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{x\in\mathcal{X}} \left\langle k(x,\cdot), \sum_{i=1}^n \epsilon_i k(X_i,\cdot)\right\rangle_{\mathcal{H}_k}\right] \leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{x\in\mathcal{X}} \|k(x,\cdot)\|_{\mathcal{H}_k}\left\|\sum_{i=1}^n \epsilon_i k(X_i,\cdot)\right\|_{\mathcal{H}_k}\right] \\
&\leq \frac{c_U}{n}\left\{\mathbb{E}_\epsilon\left[\sum_{i=1}^n\sum_{j=1}^n \epsilon_i\epsilon_j k(X_i, X_j)\right]\right\}^{1/2} \leq \frac{c_U^{3/2}}{\sqrt{n}}.
\end{aligned}$$

Thus, using McDiarmid's concentration inequality and symmetrization, for all $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\sup_{x\in\mathcal{X}}\left|\hat{d}_n(x) - d(x)\right| \leq 2\frac{c_U^{3/2}}{\sqrt{n}} + (c_U - c_L)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

□

A.2. Proof of Proposition 3

Proof. Let $\mathscr{F} := \{d(\cdot)\|\psi_h(\cdot) - \mu\|^2_{\mathcal{H}_h} \mid \|\mu\|^2_{\mathcal{H}_h} \leq D^2 \text{ and } \mu \in \mathcal{H}_h\}$ and $\mathscr{F}_M := \{\min_{1\leq m\leq M} d(\cdot)\|\psi_h(\cdot) - \mu_m\|^2_{\mathcal{H}_h} \mid \forall 1 \leq m \leq M;\ \|\mu_m\|^2_{\mathcal{H}_h} \leq D^2 \text{ and } \mu_m \in \mathcal{H}_h\}$. From the basic properties of $\hat{\mathfrak{R}}_{\mathcal{X}_n}$ (Bartlett & Mendelson, 2002), we have

$$\hat{\mathfrak{R}}_{\mathcal{X}_n}(\mathscr{F}_M) \leq M\hat{\mathfrak{R}}_{\mathcal{X}_n}(\mathscr{F}).$$

From the reproducing property, we obtain

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i d(X_i)\|\psi_h(X_i) - \mu\|^2_{\mathcal{H}_h} = \frac{1}{n}\sum_{i=1}^n \epsilon_i d(X_i)\left\{h(X_i, X_i) + \|\mu\|^2_{\mathcal{H}_h} - 2\mu(X_i)\right\}.$$

Hence, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_{\mathcal{X}_n}(\mathscr{F}) &= \mathbb{E}_\epsilon\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right] = \mathbb{E}_\epsilon\left[\sup_{\mu\in\mathscr{M}} \frac{1}{n}\sum_{i=1}^n \epsilon_i d(X_i)\|\psi_h(X_i) - \mu\|^2_{\mathcal{H}_h}\right] \\
&= \mathbb{E}_\epsilon\left[\sup_{\mu\in\mathscr{M}} \frac{1}{n}\sum_{i=1}^n \epsilon_i d(X_i)\left\{h(X_i, X_i) + \|\mu\|^2_{\mathcal{H}_h} - 2\mu(X_i)\right\}\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mu\in\mathscr{M}} \sum_{i=1}^n \epsilon_i d(X_i)\|\mu\|^2_{\mathcal{H}_h}\right] + 2\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mu\in\mathscr{M}}\left|\sum_{i=1}^n \epsilon_i d(X_i)\langle\psi_h(x_i), \mu\rangle_{\mathcal{H}_h}\right|\right],
\end{aligned}$$

where write $\mathscr{M} := \{\mu(\cdot) \mid \|\mu\|^2_{\mathcal{H}_h} \leq D^2\} \subset \mathcal{H}_h$. Here, we have

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mu\in\mathscr{M}} \sum_{i=1}^n \epsilon_i d(X_i)\|\mu\|^2_{\mathcal{H}_h}\right] \leq \frac{1}{n}\mathbb{E}_\epsilon\left[\left|\sum_{i=1}^n \epsilon_i d(X_i)\right| \sup_{\mu\in\mathscr{M}} \|\mu\|^2_{\mathcal{H}_h}\right] \\
&= \frac{D^2}{n}\mathbb{E}_\epsilon\left[\left|\sum_{i=1}^n \epsilon_i d(X_i)\right|\right] \leq \frac{D^2}{n}\left\{\sum_{i=1}^n\sum_{j=1}^n d(x_i)d(x_j)\mathbb{E}_\epsilon[\epsilon_i\epsilon_j]\right\}^{1/2} \leq \frac{D^2}{n}\sqrt{nc_U^2} = \frac{D^2 c_U}{\sqrt{n}}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mu\in\mathscr{M}}\left|\sum_{i=1}^n \epsilon_i d(X_i)\langle\psi_h(x_i),\mu\rangle_{\mathcal{H}_h}\right|\right] = \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mu\in\mathscr{M}}\left|\left\langle\sum_{i=1}^n \epsilon_i d(X_i)\psi_h(x_i),\mu\right\rangle_{\mathcal{H}_h}\right|\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mu\in\mathscr{M}}\|\mu\|_{\mathcal{H}_h}\left\|\sum_{i=1}^n \epsilon_i d(X_i)\psi_h(x_i)\right\|_{\mathcal{H}_h}\right] \leq \frac{D}{n}\mathbb{E}_\epsilon\left[\left\|\sum_{i=1}^n \epsilon_i d(X_i)\psi_h(x_i)\right\|_{\mathcal{H}_h}\right] \\
&\leq \frac{D}{n}\left\{\mathbb{E}_\epsilon\left[\left\|\sum_{i=1}^n \epsilon_i d(X_i)\psi_h(x_i)\right\|_{\mathcal{H}_h}^2\right]\right\}^{1/2} = \frac{D}{n}\left\{\sum_{i=1}^n d(X_i)^2 h(X_i,X_i)\right\}^{1/2} \\
&= \frac{D}{n}\left\{\sum_{i=1}^n k(X_i,X_i)\right\}^{1/2} \leq \frac{D}{n}\sqrt{nc_U} = \frac{D\sqrt{c_U}}{\sqrt{n}}
\end{aligned}$$

Thus, using McDiarmid's concentration inequality and symmetrization (cf. Bartlett & Mendelson (2002) and Theorem 3.1 of Mohri et al. (2012)), for all $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\sup_{\boldsymbol{\mu}\in\mathcal{B}(0,D)^M}|\mathrm{WKKM}_h(\boldsymbol{\mu}\mid P_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}\mid\mathbb{P})| \leq \frac{2M\sqrt{c_U}D(\sqrt{c_U}D+2)}{\sqrt{n}} + 2c_U D^2\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Finally, for any $\boldsymbol{\mu}^* \in \mathcal{M}$,

$$\begin{aligned}
\ell(\hat{\boldsymbol{\mu}}_n,\boldsymbol{\mu}^*) &= \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n\mid\mathbb{P}) - \mathrm{WKKM}_h(\boldsymbol{\mu}^*\mid\mathbb{P}) \\
&= \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n\mid\mathbb{P}) - \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n\mid\mathbb{P}_n) + \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n\mid\mathbb{P}_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}^*\mid\mathbb{P}_n) \\
&\quad + \mathrm{WKKM}_h(\boldsymbol{\mu}^*\mid\mathbb{P}_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}^*\mid\mathbb{P}) \\
&\leq 2\sup_{\boldsymbol{\mu}\in\mathcal{B}(0,D)^M}|\mathrm{WKKM}_h(\boldsymbol{\mu}\mid\mathbb{P}_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}\mid\mathbb{P})| + \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n\mid\mathbb{P}_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}^*\mid\mathbb{P}_n).
\end{aligned}$$

Let $U_n^* = (u_{im})_{n\times M}$ denote the following membership matrix:

$$u_{im} = \begin{cases} 1 & \text{if } \|\psi_h(X_i)-\mu_m^*\|_{\mathcal{H}_h} \leq \|\psi_h(X_i)-\mu_j^*\|_{\mathcal{H}_h} \text{ for all } j\neq m, \\ 0 & \text{otherwise.}\end{cases}$$

Since $\mathrm{Ncut}(\hat{U}_n\mid\mathbb{P}_n) - \mathrm{Ncut}(U_n^*\mid\mathbb{P}_n) \leq 0$, it follows that

$$\begin{aligned}
&\mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n\mid\mathbb{P}_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}^*\mid\mathbb{P}_n) \\
&= \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n\mid\mathbb{P}_n) - \mathrm{Ncut}(\hat{U}_n\mid\mathbb{P}_n) + \mathrm{Ncut}(\hat{U}_n\mid\mathbb{P}_n) - \mathrm{Ncut}(U_n^*\mid\mathbb{P}_n) + \mathrm{Ncut}(U_n^*\mid\mathbb{P}_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}^*\mid\mathbb{P}_n) \\
&\leq \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n\mid\mathbb{P}_n) - \mathrm{Ncut}(\hat{U}_n\mid\mathbb{P}_n) + \mathrm{Ncut}(U_n^*\mid\mathbb{P}_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}^*\mid\mathbb{P}_n) \leq 4D^2\|\hat{d}_n - d\|_\infty. \qquad \text{(A.1)}
\end{aligned}$$

When we have that $\|\psi_h(X_i)-\mu_m^*\|_{\mathcal{H}_h} = \|\psi_h(X_i)-\mu_j^*\|_{\mathcal{H}_h}$ for some i, there exists the set $A\subset\{1,\ldots,M\}$ of indexes such that $u_{im}^* = 1$ for $m\in A$. To get the crisp result, we set 1 only for $u_{im_1}^*$ where $m_1 = \min A$. Combining these results, we obtain

$$\ell(\hat{\boldsymbol{\mu}}_n,\boldsymbol{\mu}^*) \leq 2\sup_{\boldsymbol{\mu}\in\mathcal{B}(0,D)^M}|\mathrm{WKKM}_h(\boldsymbol{\mu}\mid\mathbb{P}_n) - \mathrm{WKKM}_h(\boldsymbol{\mu}\mid\mathbb{P})| + 4D^2\|\hat{d}_n - d\|_\infty.$$

Therefore, with probability at least $1-2\delta$,

$$\begin{aligned}
\ell(\hat{\boldsymbol{\mu}}_n,\boldsymbol{\mu}^*) &\leq \frac{4M\sqrt{c_U}D(\sqrt{c_U}D+2)}{\sqrt{n}} + 4c_U D^2\sqrt{\frac{\log(2/\delta)}{2n}} + 8D^2\frac{c_U^{3/2}}{\sqrt{n}} + 4D^2(c_U-c_L)\sqrt{\frac{\log(2/\delta)}{2n}} \\
&= \frac{4\sqrt{c_U}D\{\sqrt{c_U}MD + 2(M+Dc_U)\}}{\sqrt{n}} + 4D^2(2c_U-c_L)\sqrt{\frac{\log(2/\delta)}{2n}}.
\end{aligned}$$

□

B. Some fundamental properties about the weighted k-means

Note that the norm $\|\cdot\|_{\mathcal{H}_h}$ of RKHS $\mathcal{H}_h$ is strictly convex and smooth. For $f : \mathcal{X} \to \mathbb{R}$, let $\nabla_+ f(x, y)$ denote the one-sided directional derivative of f at $x \in \mathbb{R}^d$ in direction $y \in \mathbb{R}^d$:

$$\nabla_+ f(x, y) := \lim_{t \to 0+} \frac{f(x + ty) - f(x)}{t}.$$

For a distribution $\mathbb{P}$ on RKHS $\mathcal{H}_h$, the center of $\mathbb{P}$ is defined as a point $\mu \in \mathcal{H}_h$ such that

$$\mathbb{E}[\|X - \mu\|^2_{\mathcal{H}_h}] = \inf_{\eta \in \mathcal{H}_h} \mathbb{E}[\|X - \eta\|^2_{\mathcal{H}_h}].$$

Let $C(\mathbb{P})$ denote the set of all centers of $\mathbb{P}$.

The following properties of the weighted k-means on RKHS $\mathcal{H}_h$ are parallel with the results of the k-means in Graf & Luschgy (2000), Graf et al. (2007) and Levrard (2015). We can prove these results in much the same way as Graf & Luschgy (2000), Graf et al. (2007) and Levrard (2015). For the sake of completeness, we provide the proofs for all results.

Lemma B.1. *For $\mu \in \mathcal{H}_h$, we have $\mu \in C(\mathbb{P})$ if and only if*

$$\int d(x)\|\psi_h(x) - \mu\|_{\mathcal{H}_h} \nabla_+ \|\|_{\mathcal{H}_h}(\mu - \psi_h(x), \eta)\mathbb{P}(dx) \geq 0 \ \text{ for all } \ \eta \in \mathcal{H}_h.$$

More precisely , since $\|\|_{\mathcal{H}_h}$ is smooth,

$$\int_{\psi_h(x) \neq \mu} d(x)\{\mu - \psi_h(x)\}\mathbb{P}(dx) = 0.$$

Proof. Let $\psi(\mu) := \mathbb{E}[d(x)\|\psi_h(X) - \mu\|^2_{\mathcal{H}_h}]$. First, we will show the convexity of ψ. Note that

$$\begin{aligned}
&\|t\{\psi_h(X) - \mu\} + (1-t)\{\psi_h(X) - \eta\}\|^2_{\mathcal{H}_h} \\
&\leq t^2\|\psi_h(X) - \mu\|^2_{\mathcal{H}_h} + (1-t)^2\|\psi_h(X) - \eta\|^2_{\mathcal{H}_h} + 2t(1-t)\|\psi_h(X) - \mu\|_{\mathcal{H}_h}\|\psi_h(X) - \eta\|_{\mathcal{H}_h} \\
&\leq t\|\psi_h(X) - \mu\|^2_{\mathcal{H}_h} + (1-t)\|\psi_h(X) - \eta\|^2_{\mathcal{H}_h}.
\end{aligned}$$

For $t \in [0, 1]$,

$$\psi(t\mu + (1-t)\eta) = \mathbb{E}[d(x)\|t\{\psi_h(X) - \mu\} + (1-t)\{\psi_h(X) - \eta\}\|^2_{\mathcal{H}_h}] \leq t\psi(\mu) + (1-t)\psi(\eta).$$

Thus, we have

$$\mu \in C(\mathbb{P}) \ \rightleftharpoons \ \nabla_+ \psi(\mu, \eta) \geq 0.$$

We consider the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(u) = d(x)\|\mu + u\eta - \psi_h(x)\|^2_{\mathcal{H}_h}$. We have

$$\begin{aligned}
\frac{\psi(\mu + t\eta) - \psi(\mu)}{t} &= \int d(x) \frac{\|\mu + t\eta - \psi_h(x)\|^2_{\mathcal{H}_h} - \|\mu - \psi_h(x)\|^2_{\mathcal{H}_h}}{t} \mathbb{P}(dx) \\
&= \int \frac{g(t) - g(0)}{t} \mathbb{P}(dx).
\end{aligned}$$

For $t \in [0, 1]$,

$$\begin{aligned}
g(tu + (1-t)v) &= d(x)\|\mu + \{tu + (1-t)v\}\eta - \psi_h(x)\|^2_{\mathcal{H}_h} \\
&\leq td(x)\|\mu + u\eta - \psi_h(x)\|^2_{\mathcal{H}_h} + (1-t)d(x)\|\mu + v\eta - \psi_h(x)\|^2_{\mathcal{H}_h} \\
&= tg(u) + (1-t)g(v).
\end{aligned}$$

Since g is convex, Theorem 5.1.1 of Webster (1994) gives

$$g(0) - g(-1) \leq \frac{g(t) - g(0)}{t} \leq g(1) - g(0) \ \text{ for all } \ t \in (0, 1].$$

Thus, by Lebesgue's dominated convergence theorem,

$$\lim_{t\to 0+}\frac{\psi(\mu+t\eta)-\psi(\mu)}{t} = \int d(x)\lim_{t\to 0+}\frac{\|\mu+t\eta-\psi_h(x)\|_{\mathcal{H}_h}^2-\|\mu-\psi_h(x)\|_{\mathcal{H}_h}^2}{t}\mathbb{P}(dx)$$
$$= \int d(x)\|\mu-\psi_h(x)\|_{\mathcal{H}_h}\nabla_+\|\|_{\mathcal{H}_h}(\mu-\psi_h(x),\eta)\mathbb{P}(dx).$$

Since

$$\nabla_+\|\|_{\mathcal{H}_h}(0,\eta) = \lim_{t\to 0+}\frac{\|t\eta\|_{\mathcal{H}_h}}{t} = \|\eta\|_{\mathcal{H}_h},$$

we have

$$\int d(x)\|\mu-\psi_h(x)\|_{\mathcal{H}_h}\nabla_+\|\|_{\mathcal{H}_h}(\mu-\psi_h(x),\eta)\mathbb{P}(dx)$$
$$= \int_{\psi_h(x)\neq\mu} d(x)\|\mu-\psi_h(x)\|_{\mathcal{H}_h}\nabla_+\|\|_{\mathcal{H}_h}(\mu-\psi_h(x),\eta)\mathbb{P}(dx)$$
$$+\int_{\psi_h(x)=\mu} d(x)\|\mu-\psi_h(x)\|_{\mathcal{H}_h}\nabla_+\|\|_{\mathcal{H}_h}(\mu-\psi_h(x),\eta)\mathbb{P}(dx)$$
$$= \left\langle\int_{\psi_h(x)\neq\mu} d(x)\{\mu-\psi_h(x)\}\mathbb{P}(dx),\eta\right\rangle_{\mathcal{H}_h}.$$

□

B.1. Proof of Proposition 5

Proof. Note that the claim (ii) follows immediately from the claim (i) and Lemma B.1. First, we prove the claim (i). The proof is similar as the proof of Theorem 4.1 in Graf & Luschgy (2000). Write $\mathcal{U}^* := \{V_m(\boldsymbol{\mu}^*) \mid \mathbb{P}(V_m(\boldsymbol{\mu}^*)) > 0\}$. To obtain a contradiction, suppose $\#(\mathcal{U}^*) < M$. Here, note that $\mathcal{U}^*$ is also optimal. From $\#(\mathrm{supp}(\mathbb{P})) \geq M$, there exists $V_{m_0}^* \in \mathcal{U}^*$ such that $\mathbb{P}(V_{m_0}^* \setminus \{\mu_{m_0}^*\}) > 0$, where $\mu_{m_0}^*$ is the center of the Voronoi set $V_{m_0}^*$. Since $\mathcal{X}$ is separable and k is continuous, $\mathcal{H}_h$ is separable by Lemma 4.33 of Steinwart & Christmann (2008). Thus, the induced probability measure on $\mathcal{H}_h$ is tight and

$$\mathbb{P}(A) = \sup\{\mathbb{P}(K) \mid K \subset A,\ K : \text{compact}\}$$

for any Borel set A. Hence, for all δ satisfying $\mathbb{P}(V_{m_0}^* \setminus \{\mu_{m_0}^*\}) > \delta > 0$, there exists a compact set $K \subset A$ such that $\mathbb{P}(K) > \delta$. Let $H(a,b) := \{x \in \mathcal{H}_h \mid \|x-a\|_{\mathcal{H}_h} \leq \|x-b\|_{\mathcal{H}_h}\}$ and then $H^c(a,b)$ is open set. Since K is compact, there exists a finite set $B \subset K$ such that

$$K \subset \bigcup_{b\in B} H(\mu_{m_0}^*, b)^c.$$

It follows that there exists $b \in K$ such that

$$\mathbb{P}(H(\mu_{m_0}^*, b)^c \cap V_{m_0}^*) > 0.$$

Since

$$\int_{V_{m_0}^*\cap H(\mu_{m_0}^*,b)} d(x)\|\psi_h(x)-\mu_{m_0}^*\|_{\mathcal{H}_h}^2\mathbb{P}(dx) < \int_{V_{m_0}^*\cap H(\mu_{m_0}^*,b)} d(x)\|\psi_h(x)-b\|_{\mathcal{H}_h}^2\mathbb{P}(dx),$$

we have

$$\int d(x)\min_{\mu_m^*\in\mathcal{M}_M^*}\|\psi_h(x)-\mu_m^*\|_{\mathcal{H}_h}^2\mathbb{P}(dx) > \int d(x)\min_{\mu_m^*\in\mathcal{U}^*}\|\psi_h(x)-\mu_m^*\|_{\mathcal{H}_h}^2\mathbb{P}(dx),$$

which contradicts the optimality of $\mathcal{M}_M^*$ and (i) is proved.

We next prove the claim (iii). Choose a Voronoi partition $\{W_1(\boldsymbol{\mu}^*),\ldots,W_M(\boldsymbol{\mu}^*)\}$ such that $\mathring{V}_j(\boldsymbol{\mu}^*) \subset W_j(\boldsymbol{\mu}^*) \subset V_j(\boldsymbol{\mu}^*)$ $(j=1,\ldots,M)$. We abbreviate $W_j(\boldsymbol{\mu}^*)$ and $V_j(\boldsymbol{\mu}^*)$ to W_j and V_j, respectively. Then,

$$\mu_j \in C(\mathbb{P}(\cdot \mid W_j)) \cap C(\mathbb{P}(\cdot \mid V_j)).$$

From Lemma B.1, we obtain

$$\int_{W_i \setminus \{\mu\}} d(x)\{\mu_i - \psi_h(x)\}\mathbb{P}(dx) = 0 \quad \text{and} \quad \int_{V_i \setminus \{\mu\}} d(x)\{\mu_i - \psi_h(x)\}\mathbb{P}(dx) = 0.$$

Thus, we have

$$\int_{V_i \cap V_j} d(x)\{\mu_i - \psi_h(x)\}\mathbb{P}(dx) = 0.$$

Since $V_i \cap V_j \subset S(\mu_i, \mu_j) := \{x \in \mathcal{X} \mid \|\psi_h(x) - \mu_i\|_{\mathcal{H}_h} = \|\psi_h(x) - \mu_j\|_{\mathcal{H}_h}\}$, we have $\mu_j \in C(\mathbb{P}(\cdot \mid V_i \cap V_j))$. This contradicts with Theorem 2.4 in Graf & Luschgy (2000). □

As with Levrard (2015), let $p_{\min} := \inf_{\boldsymbol{\mu}^* \in \mathcal{M},\, 1 \le m \le M} \mathbb{P}(V_m(\boldsymbol{\mu}^*))$.

Proposition B.2. *Suppose $\mathbb{P}$ is D-bounded. Then, both B and $d_{\min}$ are positive.*

Proof. The proof of the strictly positiveness of B is given in Section 4.1 in Levrard (2015). Since $d_{\min} \ge c_L p_{\min}$, let us prove that $p_{\min} > 0$. For $m \in \mathbb{N}$, let $R_m^* = \inf_{\mathcal{M}_m} \mathrm{WKKM}_h(\mathcal{M}_M \mid \mathbb{P})$. Since the support of $\mathbb{P}$ contains more than M points, we have $R_{m+1}^* < R_m^*$. Suppose that $p_{\min} = \inf_{\boldsymbol{\mu} \in \mathcal{M}_m} \mathbb{P}(V_m(\boldsymbol{\mu})) = 0$, that is, $\forall \epsilon > 0;\ \exists \boldsymbol{\mu} \in \mathcal{M}_m;\ \mathbb{P}(V_1(\boldsymbol{\mu})) < \epsilon$. For all $\epsilon > 0$, we have

$$\begin{aligned}
R_M^* &= \sum_{m=1}^{M} \mathbb{E}\left[d(X)\|h(X,\cdot) - \mu_m^{(M)}\|_{\mathcal{H}_h}^2 \mathbb{1}_{V_m(\boldsymbol{\mu})}(X)\right] \\
&\ge \mathbb{E}\left[d(X)\|h(X,\cdot) - \mu_2^{(M)}\|_{\mathcal{H}_h}^2 \mathbb{1}_{V_1(\boldsymbol{\mu})}(X)\right] + \sum_{m=2}^{M} \mathbb{E}\left[d(X)\|h(X,\cdot) - \mu_m^{(M)}\|_{\mathcal{H}_h}^2 \mathbb{1}_{V_m(\boldsymbol{\mu})}(X)\right] \\
&\quad - \mathbb{E}\left[d(X)\|h(X,\cdot) - \mu_2^{(M)}\|_{\mathcal{H}_h}^2 \mathbb{1}_{V_1(\boldsymbol{\mu})}(X)\right] \\
&\ge R_{M-1}^* - 4\frac{c_U^2}{c_L^2}\epsilon.
\end{aligned}$$

This contradicts with $R_{m+1}^* < R_m^*$. Therefore, we obtain $p_{\min} > 0$. □

Lemma B.3. *Under the general assumption in Section 2, we have*

(i) *$\mathcal{B}(0,R)^M$ is weakly compact, for every $R \ge 0$,*

(ii) *$\boldsymbol{\mu} \mapsto \mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P})$ is weakly lower semi-continuous,*

(iii) *$\mathcal{M}$ is weakly compact.*

Proof. The claims (i, iii) are same as (i, iii) of Lemma 4.1 in Levrard (2015). Let $\{\boldsymbol{\mu}_n\}$ be a sequence in $\mathcal{B}(0,D)^M$ such that $\boldsymbol{\mu}_n$ converges weakly to $\boldsymbol{\mu} \in \mathcal{B}(0,D)^M$. From the proof of Lemma 4.1 in Levrard (2015), for a fixed $f \in \mathcal{H}_h$,

$$\min_{1 \le m \le M} \|f - \mu_m\|_{\mathcal{H}_h}^2 \le \liminf_{n \to \infty} \min_{1 \le m \le M} \|f - \mu_m^{(n)}\|_{\mathcal{H}_h}^2,$$

where $\boldsymbol{\mu}_n = (\mu_1^{(n)}, \ldots, \mu_1^{(n)})^T$. Thus, for given $x \in \mathcal{X}$,

$$\min_{1 \le m \le M} d(x)\|h(x,\cdot) - \mu_m\|_{\mathcal{H}_h}^2 \le \liminf_{n \to \infty} \min_{1 \le m \le M} d(x)\|h(x,\cdot) - \mu_m^{(n)}\|_{\mathcal{H}_h}^2.$$

From Fatou's lemma, we obtain

$$\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}) \le \liminf_{n \to \infty} \mathrm{WKKM}_h(\boldsymbol{\mu}_n \mid \mathbb{P}).$$

□

Proposition B.4. *Suppose that $\mathbb{P}$ satisfies a margin condition with radius r_0 described in Section 3.2. Then*

(i) *For every $\boldsymbol{\mu}^*$ in $\mathcal{M}$ and $\boldsymbol{\mu}$ in $\mathcal{B}(0,D)^M$, if $\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\| \le Br_0/(4\sqrt{2}D)$, then*

$$\ell(\boldsymbol{\mu},\boldsymbol{\mu}^*) \ge \frac{d_{\min}}{2}\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2.$$

(ii) *$\mathcal{M}$ is finite.*

(iii) *There exists $\epsilon > 0$ such that $\mathbb{P}$ is ϵ-separated.*

(iv) *For all $\boldsymbol{\mu} \in \mathcal{B}(0,D)^M$,*

$$\begin{aligned}&\frac{1}{16D^2\sigma^2}\mathrm{Var}\left(d(X)\min_{1\le m\le M}\|\psi_h(X)-\mu_m\|_{\mathcal{H}_h}^2 - d(X)\min_{1\le m\le M}\|\psi_h(X)-\mu_m^*(\boldsymbol{\mu})\|^2\right)\\ &\le \|\boldsymbol{\mu}-\boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2 \le \kappa_0\ell(\boldsymbol{\mu},\boldsymbol{\mu}^*),\end{aligned}$$

where $\sigma^2 = \mathbb{E}[d^2(X)]$, $\kappa_0 = 4MD^2\left(\epsilon^{-1} \vee 64D^2/(d_{\min}B^2r_0^2)\right)$, and $\boldsymbol{\mu}^*(\boldsymbol{\mu}) \in \arg\min_{\boldsymbol{\mu}^*\in\mathcal{M}}\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|$.

Proof. For $\boldsymbol{\mu} \in \mathcal{B}(0,D)^M$, $(W_1(\boldsymbol{\mu}),\dots,W_M(\boldsymbol{\mu}))$ be a Voronoi partition of $\boldsymbol{\mu}$. For $\boldsymbol{\mu}^* \in \mathcal{M}$,

$$\begin{aligned}&\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}) = \sum_{m=1}^M \mathbb{E}\left[d(X)\|\psi_h(X)-\mu_m\|_{\mathcal{H}_h}^2 \mathbb{1}_{W_m(\boldsymbol{\mu})}(X)\right]\\ &= \sum_{m=1}^M \mathbb{E}\left[d(X)\|\psi_h(X)-\mu_m\|_{\mathcal{H}_h}^2 \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right] + \sum_{m=1}^M \mathbb{E}\left[d(X)\|\psi_h(X)-\mu_m\|_{\mathcal{H}_h}^2 \{\mathbb{1}_{W_m(\boldsymbol{\mu})}(X) - \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\}\right].\end{aligned}$$

Since

$$\mu_m^* = \frac{\mathbb{E}\left[d(X)\psi_h(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right]}{\mathbb{E}\left[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right]}$$

we have

$$\begin{aligned}&\mathbb{E}\left[d(X)\langle\psi_h(X)-\mu_m^*, \mu_m^*-\mu_m\rangle_{\mathcal{H}_h}\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right] = \mathbb{E}\left[\langle d(X)\{\psi_h(X)-\mu_m^*\}\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X), \mu_m^*-\mu_m\rangle_{\mathcal{H}_h}\right]\\ &= \langle\mathbb{E}\left[d(X)\{\psi_h(X)-\mu_m^*\}\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right], \mu_m^*-\mu_m\rangle_{\mathcal{H}_h} = 0.\end{aligned}$$

We obtain

$$\mathbb{E}\left[d(X)\|\psi_h(X)-\mu_m\|_{\mathcal{H}_h}^2\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right] = \mathbb{E}\left[d(X)\|\psi_h(X)-\mu_m^*\|_{\mathcal{H}_h}^2\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right] + \|\mu_m^*-\mu_m\|_{\mathcal{H}_h}^2\mathbb{E}\left[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right],$$

and thus

$$\begin{aligned}\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}) = \mathrm{WKKM}_h(\boldsymbol{\mu}^* \mid \mathbb{P}) &+ \sum_{m=1}^M \|\mu_m^*-\mu_m\|_{\mathcal{H}_h}^2\mathbb{E}\left[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right]\\ &+ \sum_{m=1}^M \mathbb{E}\left[d(X)\|\psi_h(X)-\mu_m\|_{\mathcal{H}_h}^2\{\mathbb{1}_{W_m(\boldsymbol{\mu})}(X) - \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\}\right].\end{aligned}$$

Note that

$$\mathbb{1}_{W_m(\boldsymbol{\mu})}(x) - \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(x) = \begin{cases}+1 & \text{if } x \in W_m(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*)^c,\\ 0 & \text{if } x \in \{W_m(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*)\} \cap \{W_m(\boldsymbol{\mu})^c \cap V_m(\boldsymbol{\mu}^*)^c\},\\ -1 & \text{if } x \in W_m(\boldsymbol{\mu})^c \cap V_m(\boldsymbol{\mu}^*).\end{cases}$$

From $W_m(\boldsymbol{\mu})^c \cap V_m(\boldsymbol{\mu}^*) = V_m(\boldsymbol{\mu}^*) \cap \{\bigcup_{j \neq m} W_j(\boldsymbol{\mu})\}$ and $W_m(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*)^c = W_m(\boldsymbol{\mu}) \cap \{\bigcup_{j \neq m} V_j(\boldsymbol{\mu}^*)\}$, it follows that

$$
\begin{aligned}
&\sum_{m=1}^{M} d(X)\|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 \{\mathbb{1}_{W_m(\boldsymbol{\mu})}(X) - \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\} \\
&= \sum_{m=1}^{M} \sum_{j \neq m} d(X)\|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 \mathbb{1}_{W_m(\boldsymbol{\mu}) \cap V_j(\boldsymbol{\mu}^*)}(X) - \sum_{m=1}^{M} \sum_{j \neq m} d(X)\|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 \mathbb{1}_{W_j(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*)}(X) \\
&= \sum_{m=1}^{M} \sum_{j \neq m} d(X)\{\|\psi_h(X) - \mu_j\|_{\mathcal{H}_h}^2 - \|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2\} \mathbb{1}_{W_j(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*)}(X).
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
\ell(\boldsymbol{\mu}, \boldsymbol{\mu}^*) = &\sum_{m=1}^{M} \|\mu_m^* - \mu_m\|_{\mathcal{H}_h}^2 \mathbb{E}\left[d(X) \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right] \\
&+ \sum_{m=1}^{M} \sum_{j \neq m} \mathbb{E}\left[d(X)\{\|\psi_h(X) - \mu_j\|_{\mathcal{H}_h}^2 - \|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2\} \mathbb{1}_{W_j(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*)}(X)\right].
\end{aligned}
$$

From (i) of Lemma 4.2 in Levrard (2015),

$$
\|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 - \|\psi_h(X) - \mu_j\|_{\mathcal{H}_h}^2 = 2\left\langle \mu_j - \mu_m, \psi_h(X) - \frac{\mu_j + \mu_m}{2} \right\rangle_{\mathcal{H}_h} \leq 8\sqrt{2}D\|\mu_j - \mu_m\|_{\mathcal{H}_h}.
$$

Using (ii) of Lemma 4.2 in Levrard (2015), we obtain

$$
\begin{aligned}
&\{\|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 - \|\psi_h(X) - \mu_j\|_{\mathcal{H}_h}^2\} \mathbb{1}_{W_j(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*)}(X) \\
&\leq 8\sqrt{2}D\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| \mathbb{1}_{W_j(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*) \cap N_{\boldsymbol{\mu}^*}(4\sqrt{2}D\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|/B)}(X).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sum_{m=1}^{M} d(X)\|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 \{\mathbb{1}_{W_m(\boldsymbol{\mu})}(X) - \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\} \\
&= \sum_{m=1}^{M} \sum_{j \neq m} d(X)\{\|\psi_h(X) - \mu_j\|_{\mathcal{H}_h}^2 - \|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2\} \mathbb{1}_{W_j(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*)}(X) \\
&\geq -8\sqrt{2}D\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| \sum_{m=1}^{M} \sum_{j \neq m} \mathbb{1}_{W_j(\boldsymbol{\mu}) \cap V_m(\boldsymbol{\mu}^*) \cap N_{\boldsymbol{\mu}^*}(4\sqrt{2}D\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|/B)}(X) \\
&\geq -8\sqrt{2}D\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| \mathbb{1}_{N_{\boldsymbol{\mu}^*}(4\sqrt{2}D\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|/B)}(X).
\end{aligned}
$$

Since $\mathbb{P}$ satisfies the margin condition,

$$
d\left(\frac{4\sqrt{2}D}{B}\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|\right) \leq \frac{B d_{\min}}{128 D^2} \frac{4\sqrt{2}D}{B}\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| = \frac{\sqrt{2} d_{\min}}{32D}\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|.
$$

Therefore, we obtain

$$
\begin{aligned}
\ell(\boldsymbol{\mu}, \boldsymbol{\mu}^*) &\geq d_{\min}\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2 - 8\sqrt{2}D\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| \frac{\sqrt{2} d_{\min}}{32D}\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| \\
&= \frac{d_{\min}}{2}\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2.
\end{aligned}
$$

Next, we prove the claim (ii). Conversely, suppose that $\mathcal{M}$ is not finite. From Lemma 4.1, there exists a sequence $\{\boldsymbol{\mu}_n\}$ of optimal codebooks and an optimal codebook $\boldsymbol{\mu}^*$ such that $\boldsymbol{\mu}_n \neq \boldsymbol{\mu}^*$ and $\boldsymbol{\mu}_n$ converges weakly to $\boldsymbol{\mu}^*$ as $n \to \infty$. If there exists $m \in \{1, \dots, M\}$ such that $\liminf_n \|\mu_m^{(n)}\|_{\mathcal{H}_h}^2 > \|\mu_m^*\|_{\mathcal{H}_h}^2$, $\liminf_n \|f - \mu_m^{(n)}\|_{\mathcal{H}_h}^2 > \|f - \mu_m^*\|_{\mathcal{H}_h}^2$ for arbitrary $f \in \mathcal{H}_h$. For $x \in \mathring{V}_m(\boldsymbol{\mu}^*)$ and $j \neq m$,

$$\liminf_{n\to\infty} \|\psi_h(x) - \mu_j^{(n)}\|_{\mathcal{H}_h}^2 \geq \|\psi_h(x) - \mu_j^*\|_{\mathcal{H}_h}^2 > \|\psi_h(x) - \mu_m^*\|_{\mathcal{H}_h}^2.$$

Thus, for every $x \in \mathcal{X}$,

$$\min_{1\leq m\leq M} d(x)\|\psi_h(x) - \mu_m^{(n)}\|_{\mathcal{H}_h}^2 > \min_{1\leq m\leq M} d(x)\|\psi_h(x) - \mu_m^*\|_{\mathcal{H}_h}^2.$$

Because of $\mathbb{P}(\mathring{V}_m(\boldsymbol{\mu}^*)) > 0$, it follows that

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{E}\left[\min_{1\leq m\leq M} d(x)\|\psi_h(x) - \mu_m^{(n)}\|_{\mathcal{H}_h}^2\right] &\geq \mathbb{E}\left[\liminf_{n\to\infty} \min_{1\leq m\leq M} d(x)\|\psi_h(x) - \mu_m^{(n)}\|_{\mathcal{H}_h}^2\right] \\ &> \mathbb{E}\left[\min_{1\leq m\leq M} d(x)\|\psi_h(x) - \mu_m^*\|_{\mathcal{H}_h}^2\right].\end{aligned}$$

This is impossible, and thus we have $\liminf_n \|\mu_m^{(n)}\|_{\mathcal{H}_h}^2 \leq \|\mu_m^*\|_{\mathcal{H}_h}^2$. We can take a subsequence $\{\check{\boldsymbol{\mu}}_n\}$ of $\{\boldsymbol{\mu}_n\}$ such that $\lim_{n\to\infty} \|\check{\mu}_m^{(n)}\|_{\mathcal{H}_h} = \|\mu_m^*\|_{\mathcal{H}_h}$ for every $m \in \{1, \dots, M\}$. From the uniform convexity of Hilbert spaces and the Radon-Riesz property, it follows that $\check{\boldsymbol{\mu}}_n \to \boldsymbol{\mu}^*$ as $n \to \infty$. For a large enough n, we have $0 < \|\check{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*\| \leq Br_0/(4\sqrt{2}M)$ and thus

$$0 = \ell(\bar{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) \geq \frac{d_{\min}}{2}\|\bar{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*\|,$$

a contradiction. Therefore, $\mathcal{M}$ is finite.

We now turn to the claim (iii). Let $\tilde{\boldsymbol{\mu}}$ be a local (not global) minimizer of $\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P})$. If $\mathbb{P}(V_m(\tilde{\boldsymbol{\mu}})) = 0$ for some $m \in \{1, \dots, M\}$, then $\mathrm{WKKM}_h(\tilde{\boldsymbol{\mu}} \mid \mathbb{P}) \geq R_{M-1}^* > R_M^*$ and the claim (iii) is proved. Suppose that $\mathbb{P}(V_m(\tilde{\boldsymbol{\mu}})) > 0$ for every $m \in \{1, \dots, M\}$. From Lemma B.5, we have $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\| \geq Br/(4\sqrt{2}D)$ for all $\boldsymbol{\mu}^* \in \mathcal{M}$. Thus, by Lemma B.6, there exists $\boldsymbol{\mu} \in \mathcal{B}(0, D + Br/(4\sqrt{2}D))^M \setminus \mathcal{B}^o(\mathcal{M}, Br/(4\sqrt{2}D))$ such that $\mathrm{WKKM}_h(\tilde{\boldsymbol{\mu}} \mid \mathbb{P}) \geq \mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}) > R_M^*$.

Finally, we prove the claim (iv). For all $x \in \mathcal{X}$,

$$\begin{aligned}&\min_{1\leq m\leq M} \|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 - \min_{1\leq m\leq M} \|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 \\ &= \min_{1\leq m\leq M} \left\{\|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 + \|\mu_m - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 + 2\langle\psi_h(X) - \mu_m^*(\boldsymbol{\mu}), \mu_m^*(\boldsymbol{\mu}) - \mu_m\rangle_{\mathcal{H}_h}\right\} \\ &\quad - \min_{1\leq m\leq M} \|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 \\ &= \min_{1\leq m\leq M} \left\{\|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 + \langle\psi_h(X) - \mu_m + \psi_h(X) - \mu_m^*(\boldsymbol{\mu}), \mu_m^*(\boldsymbol{\mu}) - \mu_m\rangle_{\mathcal{H}_h}\right\} \\ &\quad - \min_{1\leq m\leq M} \|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 \\ &\leq \min_{1\leq m\leq M} \left\{\|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 + 4D\|\mu_m^*(\boldsymbol{\mu}) - \mu_m\|_{\mathcal{H}_h}\right\} - \min_{1\leq m\leq M} \|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 \\ &\leq 4D \max_{1\leq m\leq M} \|\mu_m - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}\end{aligned}$$

Similarly, we have

$$\min_{1\leq m\leq M} \|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 - \min_{1\leq m\leq M} \|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 \geq -4D \max_{1\leq m\leq M} \|\mu_m - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}.$$

Hence,

$$\begin{aligned}
&\frac{1}{16D^2\sigma^2}\mathrm{Var}\left(d(X)\min_{1\le m\le M}\|\psi_h(X)-\mu_m\|_{\mathcal{H}_h}^2 - d(X)\min_{1\le m\le M}\|\psi_h(X)-\mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2\right)\\
&\le \frac{1}{16D^2\sigma^2}\mathbb{E}\left[d(X)^2\left\{\min_{1\le m\le M}\|\psi_h(X)-\mu_m\|_{\mathcal{H}_h}^2 - \min_{1\le m\le M}\|\psi_h(X)-\mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2\right\}^2\right]\\
&\le \frac{1}{16D^2\sigma^2}\mathbb{E}[d(X)^2]16D^2\max_{1\le m\le M}\|\mu_m-\mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2\\
&\le \|\boldsymbol{\mu}-\boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2.
\end{aligned}$$

From the claim (i), if $\|\boldsymbol{\mu}-\boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2 \le Br_0/(4\sqrt{2}D)$, then $\ell(\boldsymbol{\mu},\boldsymbol{\mu}^*) \ge d_{\min}\|\boldsymbol{\mu}-\boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2/2$. Suppose that $\|\boldsymbol{\mu}-\boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2 \ge Br_0/(4\sqrt{2}D)$. From Lemma B.6, there exists $\boldsymbol{\mu}'$ such that $\ell(\boldsymbol{\mu},\boldsymbol{\mu}^*) \ge \ell(\boldsymbol{\mu}',\boldsymbol{\mu}^*)$. Since $\boldsymbol{\mu}' \notin \mathcal{M}$ or $\|\boldsymbol{\mu}-\boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2 = Br_0/(4\sqrt{2}D)$, we have

$$\ell(\boldsymbol{\mu},\boldsymbol{\mu}^*) \ge \ell(\boldsymbol{\mu}',\boldsymbol{\mu}^*) \ge \epsilon \wedge \frac{d_{\min}}{2}\frac{B^2r_0^2}{32D^2} \ge \left(\epsilon \wedge \frac{B^2r_0^2d_{\min}}{64D^2}\right)\frac{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2}{4MD^2}.$$

Here, we note that

$$\left(\epsilon \wedge \frac{B^2r_0^2d_{\min}}{64D^2}\right)\frac{1}{4MD^2} \le \frac{4D^2D^2d_{\min}}{64D^2}\frac{1}{MD^2} = \frac{d_{\min}}{16}\frac{1}{M} \le \frac{d_{\min}}{16} \le \frac{d_{\min}}{2}.$$

The claim (iv) is proved. □

This lemma is a straightforward extension of Lemma 4.3 in Levrard (2015) for the weighted k-means.

Lemma B.5. *Let $\boldsymbol{\mu}$ be a element in $\mathcal{B}(\mathcal{M}, Br_0/(4\sqrt{2}M))$. If $\boldsymbol{\mu}$ satisfies the centroid condition, that is, for $m = 1,\dots,M$,*

$$\mu_m = \frac{\mathbb{E}\left[d(X)\psi_h(X)\mathbb{1}_{W_m(\boldsymbol{\mu})}(X)\right]}{\mathbb{E}\left[d(X)\mathbb{1}_{W_m(\boldsymbol{\mu})}(X)\right]}$$

then $\boldsymbol{\mu}$ should be a element of $\mathcal{M}$.

Proof. Let $\tilde{\boldsymbol{\mu}}$ be a local minimum codebook in $\mathcal{B}(\mathcal{M}, Br_0/(4\sqrt{2}M)) \cap \mathcal{M}^c$. Then, there exists $\boldsymbol{\mu}^* \in \mathcal{M}$ such that $\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\| \le Br_0/(4\sqrt{2}M)$. For $m = 1,\dots,M$,

$$\begin{aligned}
&\mathbb{E}\left[d(X)\psi_h(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)\right] - \mathbb{E}\left[d(X)\psi_h(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right]\\
&= \mathbb{E}\left[d(X)\psi_h(X)\{\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X) - \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\}\right]\\
&= \mathbb{E}\left[d(X)\psi_h(X)\{\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})\cap V_m(\boldsymbol{\mu}^*)^c}(X) - \mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})^c\cap V_m(\boldsymbol{\mu}^*)}(X)\}\right]\\
&= \sum_{j\ne m}\mathbb{E}\left[d(X)\psi_h(X)\{\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})\cap V_j(\boldsymbol{\mu}^*)}(X) - \mathbb{1}_{V_j(\tilde{\boldsymbol{\mu}})\cap V_m(\boldsymbol{\mu}^*)}(X)\}\right].
\end{aligned}$$

From the centroid condition of $\tilde{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}^*$, we have

$$\begin{aligned}
&\sum_{m=1}^M \left\|\mathbb{E}[d(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)]\tilde{\mu}_m - \mathbb{E}[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)]\mu_m^*\right\|_{\mathcal{H}_h}\\
&= \sum_{m=1}^M \left\|\mathbb{E}\left[d(X)\psi_h(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)\right] - \mathbb{E}\left[d(X)\psi_h(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\right]\right\|_{\mathcal{H}_h}\\
&\le \sum_{m=1}^M\sum_{j\ne m}\left\{\left\|\mathbb{E}\left[d(X)\psi_h(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})\cap V_j(\boldsymbol{\mu}^*)}(X)\right]\right\|_{\mathcal{H}_h} + \left\|\mathbb{E}\left[d(X)\psi_h(X)\mathbb{1}_{V_j(\tilde{\boldsymbol{\mu}})\cap V_m(\boldsymbol{\mu}^*)}(X)\right]\right\|_{\mathcal{H}_h}\right\}\\
&\le 2D\sum_{m=1}^M\sum_{j\ne m}\mathbb{E}\left[d(X)\mathbb{1}_{V_j(\tilde{\boldsymbol{\mu}})\cap V_m(\boldsymbol{\mu}^*)}(X)\right].
\end{aligned}$$

From (ii) of Lemma 4.2 in Levrard (2015) and the margin condition of $\mathbb{P}$, we have

$$\sum_{m=1}^{M}\sum_{j\neq m}\mathbb{E}\left[d(X)\mathbb{1}_{V_j(\tilde{\boldsymbol{\mu}})\cap V_m(\boldsymbol{\mu}^*)}(X)\right] \leq \mathbb{E}\left[d(X)\mathbb{1}_{N_{\boldsymbol{\mu}^*}((4\sqrt{2}D/B)\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\|)}(X)\right] \leq \frac{Bd_{\min}}{128D^2}\frac{4\sqrt{2}D}{B}\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\| = \frac{d_{\min}}{16\sqrt{2}D}\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\|.$$

This gives

$$\sum_{m=1}^{M}\left\|\mathbb{E}[d(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)]\tilde{\mu}_m - \mathbb{E}[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)]\mu_m^*\right\|_{\mathcal{H}_h} \leq \frac{d_{\min}}{8\sqrt{2}}\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\|.$$

By the triangle inequality,

$$\begin{aligned}
&\left\|\mathbb{E}[d(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)]\tilde{\mu}_m - \mathbb{E}[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)]\mu_m^*\right\|_{\mathcal{H}_h}\\
&= \left\|\mathbb{E}[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)](\tilde{\mu}_m-\mu_m^*) - \tilde{\mu}_m\mathbb{E}\left[d(X)\{\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)-\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)\}\right]\right\|_{\mathcal{H}_h}\\
&\geq \mathbb{E}[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)]\,\|\tilde{\mu}_m-\mu_m^*\|_{\mathcal{H}_h} - \|\tilde{\mu}_m\|_{\mathcal{H}_h}\left|\mathbb{E}\left[d(X)\{\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)-\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)\}\right]\right|.
\end{aligned}$$

If $r_0 \geq 2D$, then

$$\mathbb{E}[d(X)] = q(2D) \leq \frac{Bd_{\min}}{128D^2}2D \leq \frac{4D^2d_{\min}}{128D^2} = \frac{1}{32}d_{\min}.$$

which is impossible. Hence, we have $r_0 < 2D$. Since $B \leq 2D$ and $r_0 < 2D$, we have

$$\|\tilde{\mu}_m\|_{\mathcal{H}_h} \leq \|\mu_m^*\|_{\mathcal{H}_h} + \|\tilde{\mu}_m-\mu_m^*\|_{\mathcal{H}_h} \leq \|\mu_m^*\|_{\mathcal{H}_h} + \|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\| \leq D + \frac{Br_0}{4\sqrt{2}D} \leq 2D.$$

Moreover,

$$\begin{aligned}
&\sum_{m=1}^{M}\left|\mathbb{E}\left[d(X)\{\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)-\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\}\right]\right|\\
&\leq \sum_{m=1}^{M}\left\{\mathbb{E}\left[d(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})\cap V_m(\boldsymbol{\mu}^*)^c}(X)\right] + \mathbb{E}\left[d(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})^c\cap V_m(\boldsymbol{\mu}^*)}(X)\right]\right\}\\
&= 2\sum_{m=1}^{M}\sum_{j\neq m}\mathbb{E}\left[d(X)\mathbb{1}_{V_j(\tilde{\boldsymbol{\mu}})\cap V_m(\boldsymbol{\mu}^*)}(X)\right]\\
&\leq 2\mathbb{E}\left[d(X)\mathbb{1}_{N_{\boldsymbol{\mu}^*}((4\sqrt{2}D/B)\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\|)}(X)\right] \leq 2\frac{Bd_{\min}}{128D^2}\frac{4\sqrt{2}D}{B}\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\| = \frac{d_{\min}}{8\sqrt{2}D}\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\|.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
&\sum_{m=1}^{M}\left\|\mathbb{E}[d(X)\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)]\tilde{\mu}_m - \mathbb{E}[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)]\mu_m^*\right\|_{\mathcal{H}_h}\\
&\geq \sum_{m=1}^{M}\mathbb{E}[d(X)\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)]\,\|\tilde{\mu}_m-\mu_m^*\|_{\mathcal{H}_h} - \sum_{m=1}^{M}\|\tilde{\boldsymbol{\mu}}\|\left|\mathbb{E}\left[d(X)\{\mathbb{1}_{V_m(\tilde{\boldsymbol{\mu}})}(X)-\mathbb{1}_{V_m(\boldsymbol{\mu}^*)}(X)\}\right]\right|\\
&\geq d_{\min}\sum_{m=1}^{M}\|\tilde{\mu}_m-\mu_m^*\|_{\mathcal{H}_h} - 2D\frac{d_{\min}}{8\sqrt{2}D}\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\|\\
&\geq \left(1-\frac{1}{4\sqrt{2}}\right)d_{\min}\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\| = \left(\frac{8\sqrt{2}-2}{8\sqrt{2}}\right)d_{\min}\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^*\|.
\end{aligned}$$

This contradicts the upper bound. □

Lemma B.6. *For any $r > 0$, there exists $\boldsymbol{\mu}_r \in \mathcal{B}(0, D+r)^M \setminus \mathcal{B}^o(\mathcal{M}, r)$ such that*

$$\inf_{\boldsymbol{\mu} \in \mathcal{H}_h^M \setminus \mathcal{B}^o(\mathcal{M},r)} \mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}) = \mathrm{WKKM}_h(\boldsymbol{\mu}_r \mid \mathbb{P}).$$

Proof. Choose a codebook $\boldsymbol{\mu}$ such that $\mu_m \notin \mathcal{B}(0, D+r)$ for some m. Let $s : \mathcal{H}_h \to \mathcal{H}_h$ denote the projection onto the closed convex set $\mathcal{B}(0, D+r)$. Let $\bar{\boldsymbol{\mu}} = (\bar{\mu}_1, \ldots, \bar{\mu}_M)$ be given by $\bar{\mu}_m = \mu_m$ if $\mu_m \in \mathcal{B}(0, D+r)$ and $\bar{\mu}_m = s(\mu_m)$ if $\mu_m \notin \mathcal{B}(0, D+r)$. From Theorem 5.2 in Brezis (2011), it follows that, for any $f \in \mathcal{B}(0, D+r)$,

$$\|f - \mu_m\|_{\mathcal{H}_h}^2 \geq \|f - s(\mu_m)\|_{\mathcal{H}_h}^2 + \|s(\mu_m) - \mu_m\|_{\mathcal{H}_h}^2.$$

By the D-boundedness of $\mathbb{P}$, we have

$$\mathrm{WKKM}_h(\bar{\boldsymbol{\mu}} \mid \mathbb{P}) \leq \mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}).$$

This gives

$$\inf_{\mathcal{H}_h^M \setminus \mathcal{B}^o(\mathcal{M},r)} \mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}) = \inf_{\mathcal{B}(0,D+r)^M \setminus \mathcal{B}^o(\mathcal{M},r)} \mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}).$$

From the weakly compactness of $\mathcal{B}(0, D+r)^M \setminus \mathcal{B}^o(\mathcal{M}, r)$, there exists $\boldsymbol{\mu}_r$ that achieves the above infimum. □

C. Proofs of Theorem 6

Proof. Assume that $\mathbb{P}$ satisfies a margin condition with $r_0 > 0$, and we will denote by ϵ the separation constant of $\mathbb{P}$.

Let $\zeta(\boldsymbol{\mu}, x) = d(x) \min_{1 \leq m \leq M} \|\psi_h(x) - \mu_m\|_{\mathcal{H}_h}^2$, and let

$$\mathcal{F} = \{\zeta(\boldsymbol{\mu}, \cdot) - \zeta(\boldsymbol{\mu}^*(\boldsymbol{\mu}), \cdot) \mid \boldsymbol{\mu} \in \mathcal{B}(0, D)^M\}.$$

Since $\sup_{x \in \mathcal{X}} d(x) \leq c_U$ and

$$\left| \min_{1 \leq m \leq M} \|\psi_h(X) - \mu_m\|_{\mathcal{H}_h}^2 - \min_{1 \leq m \leq M} \|\psi_h(X) - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h}^2 \right| \leq 4D \max_{1 \leq m \leq M} \|\mu_m - \mu_m^*(\boldsymbol{\mu})\|_{\mathcal{H}_h},$$

it follows that, for all $f \in \mathcal{F}$,

$$\|f\|_\infty \leq 8 c_U D^2, \quad \mathrm{Var}(f) \leq 16 D^2 \sigma^2 \|\boldsymbol{\mu} - \boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2 =: \omega(f).$$

Let us denote by Pf and $P_n f$ the integration of the function f with respect to $\mathbb{P}$ and $\mathbb{P}_n$, respectively. Let, if it exists, $\Phi(r)$ be an sub-root function such that

$$\mathbb{E}\left[\sup_{w(f) \leq r} |(P - P_n) f| \right] \leq \Phi(r) \quad (\forall r > 0).$$

From Theorem 4.1 in Levrard (2015), we see that, for all $x > 0$, with probability larger than $1 - \exp(-x)$,

$$Pf - P_n f \leq \frac{1}{K} \left\{ \omega(f) + r^* + \frac{9K^2 + 16K \sup_{f \in \mathcal{F}} \|f\|_\infty}{4n} x \right\} \quad (\forall f \in \mathcal{F}),$$

where $K > 0$ is a constant, and r^* be the unique solution of $\Phi(r) = r/(24K)$.

By Proposition C.1, let

$$\Phi(\delta) = \frac{4\sqrt{\pi M} + \sqrt{2 \log(|\bar{\mathcal{M}}|)}}{\sqrt{n}} \sqrt{\delta},$$

and for some positive constant $K > 0$, let us denote by δ^* the solution of the equation $\Phi(\delta) = \delta/(24K)$. The solution δ^* can be written by

$$\delta^* = \frac{576 K^2}{n} \left(4\sqrt{\pi M} + \sqrt{2 \log(|\bar{\mathcal{M}}|)} \right)^2 \leq C \frac{K^2 \{M + \log(|\bar{\mathcal{M}}|)\}}{n} =: \frac{K^2 \Xi}{n},$$

where $C = 18432\pi$ and $\Xi = C\{M + \log(|\bar{\mathcal{M}}|)\}$. Combining Theorem 4.1 in Levrard (2015) with this fact, with probability larger than $1 - \exp(-x)$,

$$(P - P_n)(\zeta(\boldsymbol{\mu}, \cdot) - \zeta(\boldsymbol{\mu}^*(\boldsymbol{\mu}))) \leq K^{-1} 16 D^2 \sigma^2 \|\boldsymbol{\mu} - \boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2 + \frac{K\Xi}{n} + \frac{9K + 128 c_U D^2}{4n} x.$$

Here, we remark that $\kappa_0 \ell(\boldsymbol{\mu}, \boldsymbol{\mu}^*) \geq \|\boldsymbol{\mu} - \boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2$. Choosing $K = 32 D^2 \sigma^2 \kappa_0$, with probability larger than $1 - \exp(-x)$

$$\begin{aligned}
\ell(\bar{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) &\leq 2 \frac{32 D^2 \sigma^2 \kappa_0 \Xi}{n} + 32 \frac{D^2 \sigma^2 9\kappa_0 + 4 c_U D^2}{2n} x \\
&\leq C_0 \sigma^2 \kappa_0 \frac{D^2 \{M + \log(|\bar{\mathcal{M}}|)\}}{n} + (9 c_U \kappa_0 + 4) \frac{16 D^2 c_U}{n} x,
\end{aligned}$$

which proves the theorem. For $\hat{\boldsymbol{\mu}}$, by the inequality (A.1), we can immediately get the last statement. □

Proposition C.1. *For all $\delta > 0$,*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F},\, \omega(f) \leq \delta} |(P - P_n) f| \right] \leq \frac{4\sqrt{\pi M} + \sqrt{2 \log(|\bar{\mathcal{M}}|)}}{\sqrt{n}} \sqrt{\delta}.$$

Proof. Note that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|\boldsymbol{\mu} - \boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2 \leq \delta/(16 D^2 \sigma^2)} |(P - P_n)(\zeta(\boldsymbol{\mu}, \cdot) - \zeta(\boldsymbol{\mu}^*(\boldsymbol{\mu}), \cdot))| \right] \\
&\leq \mathbb{E}\left[\sup_{\boldsymbol{\mu}^* \in \mathcal{M}} \sup_{\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2 \leq \delta/(16 D^2 \sigma^2)} |(P - P_n)(\zeta(\boldsymbol{\mu}, \cdot) - \zeta(\boldsymbol{\mu}^*, \cdot))| \right].
\end{aligned}$$

For $\boldsymbol{\mu}^* \in \bar{\mathcal{M}}$, write

$$Y_{\boldsymbol{\mu}^*} = \sup_{\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|^2 \leq \delta/(16 D^2 \sigma^2)} |(P - P_n)(\zeta(\boldsymbol{\mu}, \cdot) - \zeta(\boldsymbol{\mu}^*, \cdot))|.$$

We can see that

$$\mathbb{E}\left[\sup_{\boldsymbol{\mu}^* \in \mathcal{M}} Y_{\boldsymbol{\mu}^*} \right] = \mathbb{E}\left[\sup_{\boldsymbol{\mu}^* \in \mathcal{M}} (Y_{\boldsymbol{\mu}^*} - \mathbb{E}[Y_{\boldsymbol{\mu}^*}] + \mathbb{E}[Y_{\boldsymbol{\mu}^*}]) \right] \leq \mathbb{E}\left[\sup_{\boldsymbol{\mu}^* \in \mathcal{M}} (Y_{\boldsymbol{\mu}^*} - \mathbb{E}[Y_{\boldsymbol{\mu}^*}]) \right] + \sup_{\boldsymbol{\mu}^* \in \mathcal{M}} \mathbb{E}[Y_{\boldsymbol{\mu}^*}].$$

Here, for $\boldsymbol{\mu}^* \in \bar{\mathcal{M}}$, when $\|\boldsymbol{\mu} - \boldsymbol{\mu}^*(\boldsymbol{\mu})\|^2 \leq \delta/(16 D^2 \sigma^2)$,

$$\|\zeta(\boldsymbol{\mu}, \cdot) - \zeta(\boldsymbol{\mu}^*, \cdot)\|_\infty \leq 4 D c_U \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| \leq 4 D c_U \frac{\sqrt{\delta}}{4 D \sigma} \leq \sqrt{\delta}.$$

Note that $Y_{\boldsymbol{\mu}^*}$ depends on $X_1, \ldots, X_n$, and we write it $\eta(X_1, \ldots, X_n)$. For all $(x_1, \ldots, x_n),\ (y_1, \ldots, y_n) \in \mathcal{X}^n$ and all $i \in \{1, \ldots, n\}$,

$$|\eta(x_1, \ldots, x_i, \ldots, x_n) - \eta(x_1, \ldots, y_i, \ldots, x_n)| \leq \frac{2\sqrt{\delta}}{n}.$$

Thus, from the bounded difference inequality (Theorem 5.1 in Massart (2007)), it follows that, for all $\lambda > 0$,

$$\log\left(\mathbb{E}\left[\exp\{\lambda (Y_{\boldsymbol{\mu}^*} - \mathbb{E}[Y_{\boldsymbol{\mu}^*}])\} \right] \right) \leq \frac{\lambda^2 \delta}{2n}.$$

Hence, for all $\boldsymbol{\mu}^* \in \bar{\mathcal{M}}$ and all $x > 0$,

$$\mathbb{P}\left(Y_{\boldsymbol{\mu}^*} - \mathbb{E}[Y_{\boldsymbol{\mu}^*}] \geq x \right) \leq \exp\left(-\frac{2nx^2}{4\delta} \right)$$

and

$$\mathbb{P}\left(\mathbb{E}[Y_{\boldsymbol{\mu}^*}] - Y_{\boldsymbol{\mu}^*} \geq x \right) \leq \exp\left(-\frac{2nx^2}{4\delta} \right).$$

From the discussion in Section 6.1.1 of Massart (2007), we obtain

$$\mathbb{E}\left[\sup_{\boldsymbol{\mu}^* \in \bar{\mathcal{M}}} (Y_{\boldsymbol{\mu}^*} - \mathbb{E}[Y_{\boldsymbol{\mu}^*}])\right] \leq \sqrt{\frac{2\log(|\bar{\mathcal{M}}|)\delta}{n}}.$$

Thus, from Lemma C.2,

$$\mathbb{E}\left[\sup_{\boldsymbol{\mu}^* \in \bar{\mathcal{M}}} Y_{\boldsymbol{\mu}^*}\right] \leq \sqrt{\frac{2\log(|\bar{\mathcal{M}}|)\delta}{n}} + \sup_{\boldsymbol{\mu}^* \in \bar{\mathcal{M}}} \mathbb{E}[Y_{\boldsymbol{\mu}^*}] \leq \frac{\sqrt{2\log(|\bar{\mathcal{M}}|)} + 4\sqrt{\pi M}}{\sqrt{n}}\sqrt{\delta},$$

which completes the proof. □

Lemma C.2. *For a fixed $\boldsymbol{\mu}^* \in \bar{\mathcal{M}}$,*

$$\mathbb{E}\left[\sup_{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2 \leq \delta/(16D^2\sigma^2)} |(P - P_n)(\zeta(\boldsymbol{\mu}, \cdot) - \zeta(\boldsymbol{\mu}^*, \cdot))|\right] \leq \frac{4\sqrt{\pi M}\sqrt{\delta}}{\sqrt{n}}.$$

Proof. Fix $\boldsymbol{\mu}^* \in \bar{\mathcal{M}}$ and let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher sequence. By the symmetrization principle,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2 \leq \delta/(16D^2\sigma^2)} |(P - P_n)(\zeta(\boldsymbol{\mu}, \cdot) - \zeta(\boldsymbol{\mu}^*, \cdot))|\right] \\
&\leq 2\mathbb{E}_{X,\epsilon}\left[\sup_{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2 \leq \delta/(16D^2\sigma^2)} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}^*, X_i)\}\right|\right].
\end{aligned}$$

Let $g_1, \ldots, g_n$ be independent standard Gaussian variables. From Lemma 4.5 in Ledoux & Talagrand (1991), we have

$$\begin{aligned}
&\mathbb{E}_{X,\epsilon}\left[\sup_{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2 \leq \delta/(16D^2\sigma^2)} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}^*, X_i)\}\right|\right] \\
&\leq \sqrt{\frac{\pi}{2}}\mathbb{E}_{X,g}\left[\sup_{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2 \leq \delta/(16D^2\sigma^2)} \left|\frac{1}{n}\sum_{i=1}^n g_i\{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}^*, X_i)\}\right|\right] \\
&= \sqrt{\frac{\pi}{2}}\mathbb{E}_{X,g}\left[\sup_{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2 \leq \delta/(16D^2\sigma^2),\ s=\pm 1} \frac{s}{n}\sum_{i=1}^n g_i\{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}^*, X_i)\}\right].
\end{aligned}$$

From Slepian's lemma (see, e.g., Theorem 3.14 in Massart (2007)), for fixed $X_1, \ldots, X_n$ and for fixed optimal codebook $\boldsymbol{\mu}^*$, we consider the following Gaussian process $Z_{\boldsymbol{\mu},s}$:

$$Z_{\boldsymbol{\mu},s} = s\sum_{i=1}^n g_i\{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}^*, X_i)\}, \quad \boldsymbol{\mu} \in \mathcal{V}(\delta) := \mathcal{B}(\boldsymbol{\mu}^*, \sqrt{\delta}/(4D\sigma)), \quad s \in \{-1, +1\}.$$

For $i \in \{1, \ldots, n\}$ and $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{V}(\delta)$, it follows that

$$\begin{aligned}
\{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}', X_i)\}^2 &\leq d(X_i)^2 \max_{1\leq m\leq M}\{\|\psi_h(X_i) - \mu_m\|_{\mathcal{H}_h}^2 - \|\psi_h(X_i) - \mu'_m\|_{\mathcal{H}_h}^2\}^2 \\
&\leq d(X_i)^2 \max_{1\leq m\leq M}\left[2\left(\|\mu_m\|_{\mathcal{H}_h}^2 - \|\mu'_m\|_{\mathcal{H}_h}^2\right)^2 + 8\langle \mu_m - \mu'_m, \psi_h(X_i)\rangle_{\mathcal{H}_h}^2\right].
\end{aligned}$$

and, for $M \geq 2$,

$$
\begin{aligned}
&\{\zeta(\boldsymbol{\mu}, X_i) + \zeta(\boldsymbol{\mu}', X_i) - 2\zeta(\boldsymbol{\mu}^*, X_i)\}^2 \\
&\leq d(X_i)^2 \big\{ \min_{1\leq m\leq M} [\|\psi_h(X_i) - \mu_m\|_{\mathcal{H}_h}^2 - \|\psi_h(X_i) - \mu_m^*\|_{\mathcal{H}_h}^2] + \min_{1\leq m\leq M} [\|\psi_h(X_i) - \mu_m'\|_{\mathcal{H}_h}^2 - \|\psi_h(X_i) - \mu_m^*\|_{\mathcal{H}_h}^2] \big\}^2 \\
&\leq d(X_i)^2 \min_{1\leq m\leq M} \big\{ \|\psi_h(X_i) - \mu_m\|_{\mathcal{H}_h}^2 - \|\psi_h(X_i) - \mu_m^*\|_{\mathcal{H}_h}^2 + \|\psi_h(X_i) - \mu_m'\|_{\mathcal{H}_h}^2 - \|\psi_h(X_i) - \mu_m^*\|_{\mathcal{H}_h}^2 \big\}^2 \\
&\leq \frac{2}{M} d(X_i)^2 \sum_{m=1}^{M} \big\{ \big(\|\mu_m\|_{\mathcal{H}_h}^2 - \|\mu_m^*\|_{\mathcal{H}_h}^2 - 2\langle \mu_m - \mu_m^*, \psi_h(X_i)\rangle_{\mathcal{H}_h} \big)^2 + \big(\|\mu_m'\|_{\mathcal{H}_h}^2 - \|\mu_m^*\|_{\mathcal{H}_h}^2 - 2\langle \mu_m' - \mu_m^*, \psi_h(X_i)\rangle_{\mathcal{H}_h} \big)^2 \big\} \\
&\leq d(X_i)^2 \sum_{m=1}^{M} \big\{ 2\left(\|\mu_m\|_{\mathcal{H}_h}^2 - \|\mu_m^*\|_{\mathcal{H}_h}^2 \right)^2 + 8\langle \mu_m - \mu_m^*, \psi_h(X_i)\rangle_{\mathcal{H}_h}^2 + 2\left(\|\mu_m'\|_{\mathcal{H}_h}^2 - \|\mu_m^*\|_{\mathcal{H}_h}^2 \right)^2 + 8\langle \mu_m' - \mu_m^*, \psi_h(X_i)\rangle_{\mathcal{H}_h}^2 \big\}.
\end{aligned}
$$

Let ξ_{im} $(i = 1, \ldots, n; m = 1, \ldots, M)$ and ξ_m' $(m = 1, \ldots, M)$ be independent standard Gaussian variables. We will consider the Gaussian process $X_{\boldsymbol{\mu}, s}$ defined as follows:

$$
X_{\boldsymbol{\mu},s} = s \left\{ 2\sqrt{2} \sum_{i=1}^{n} d(X_i) \sum_{m=1}^{M} \langle \mu_m - \mu_m^*, \psi_h(X_i)\rangle_{\mathcal{H}_h} \xi_{im} + \sqrt{2 \sum_{i=1}^{n} d(X_i)^2} \sum_{m=1}^{M} \left(\|\mu_m\|_{\mathcal{H}_h}^2 - \|\mu_m^*\|_{\mathcal{H}_h}^2 \right) \xi_m' \right\}.
$$

For $s, s' \in \{-1, +1\}$,

$$
Z_{\boldsymbol{\mu},s} - Z_{\boldsymbol{\mu}',s'} = \begin{cases} s\left[\sum_{i=1}^n g_i \{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}', X_i)\}\right] & \text{if } ss' = +1, \\ s\left[\sum_{i=1}^n g_i \{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}^*, X_i) + \zeta(\boldsymbol{\mu}', X_i) - \zeta(\boldsymbol{\mu}^*, X_i)\}\right] & \text{if } ss' = -1. \end{cases}
$$

Here, we note that, if $ss' = 1$,

$$
\begin{aligned}
\mathrm{Var}(Z_{\boldsymbol{\mu},s} - Z_{\boldsymbol{\mu}',s'}) &= \mathrm{Var}\left(\sum_{i=1}^{n} g_i \{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}', X_i)\} \right) = \mathbb{E}\left[\sum_{i=1}^{n} g_i^2 \{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}', X_i)\}^2 \right] \\
&\leq \sum_{i=1}^{n} d(X_i)^2 \max_{1\leq m\leq M} \left[2\left(\|\mu_m\|_{\mathcal{H}_h}^2 - \|\mu_m'\|_{\mathcal{H}_h}^2 \right)^2 + 8\langle \mu_m - \mu_m', \psi_h(X_i)\rangle_{\mathcal{H}_h}^2 . \right] \\
&\leq \sum_{i=1}^{n} \sum_{m=1}^{M} d(X_i)^2 \left[2\left(\|\mu_m\|_{\mathcal{H}_h}^2 - \|\mu_m'\|_{\mathcal{H}_h}^2 \right)^2 + 8\langle \mu_m - \mu_m', \psi_h(X_i)\rangle_{\mathcal{H}_h}^2 . \right] \\
&= \mathrm{Var}(X_{\boldsymbol{\mu},s} - X_{\boldsymbol{\mu}',s'}),
\end{aligned}
$$

and that, if $ss' = -1$,

$$
\begin{aligned}
\mathrm{Var}(Z_{\boldsymbol{\mu},s} - Z_{\boldsymbol{\mu}',s'}) &= \mathrm{Var}\left(\sum_{i=1}^{n} g_i \{\zeta(\boldsymbol{\mu}, X_i) - \zeta(\boldsymbol{\mu}^*, X_i) + \zeta(\boldsymbol{\mu}', X_i) - \zeta(\boldsymbol{\mu}^*, X_i)\} \right) \\
&\leq \sum_{i=1}^{n} d(X_i)^2 \sum_{m=1}^{M} \big\{ 2\left(\|\mu_m\|_{\mathcal{H}_h}^2 - \|\mu_m^*\|_{\mathcal{H}_h}^2 \right)^2 + 8\langle \mu_m - \mu_m^*, \psi_h(X_i)\rangle_{\mathcal{H}_h}^2 \\
&\qquad + 2\left(\|\mu_m'\|_{\mathcal{H}_h}^2 - \|\mu_m^*\|_{\mathcal{H}_h}^2 \right)^2 + 8\langle \mu_m' - \mu_m^*, \psi_h(X_i)\rangle_{\mathcal{H}_h}^2 \big\} = \mathrm{Var}(X_{\boldsymbol{\mu},s} - X_{\boldsymbol{\mu}',s'}).
\end{aligned}
$$

Therefore, by Slepian's lemma, it follows that

$$
\begin{aligned}
\mathbb{E}_g \left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta),\ s=\pm 1} Z_{\boldsymbol{\mu},s} \right] &\leq 2\mathbb{E}_{\xi,\xi'} \left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta),\ s=\pm 1} X_{\boldsymbol{\mu},s} \right] \\
&\leq 4\sqrt{2}\mathbb{E}_{\xi} \left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta),\ s=\pm 1} s \sum_{i=1}^{n} \sum_{m=1}^{M} d(X_i) \langle \mu_m - \mu_m^*, \psi_h(X_i)\rangle_{\mathcal{H}_h} \xi_{im} \right] \\
&\quad + 2\sqrt{2 \sum_{i=1}^{n} d(X_i)^2} \mathbb{E}_{\xi'} \left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta),\ s=\pm 1} s \sum_{m=1}^{M} \left(\|\mu_m\|_{\mathcal{H}_h}^2 - \|\mu_m^*\|_{\mathcal{H}_h}^2 \right) \xi_m' \right].
\end{aligned}
$$

By Cauchy-Schwarz inequality and Jensen's inequality, the first term is bounded as follows:

$$\mathbb{E}_\xi\left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta),\ s=\pm1} s\sum_{i=1}^n\sum_{m=1}^M d(X_i)\langle\mu_m-\mu_m^*,\psi_h(X_i)\rangle_{\mathcal{H}_h}\xi_{im}\right]$$
$$\leq \mathbb{E}_\xi\left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta)}\sum_{m=1}^M\|\mu_m-\mu_m^*\|_{\mathcal{H}_h}\left\|\sum_{i=1}^n d(X_i)\psi_h(X_i)\xi_{im}\right\|_{\mathcal{H}_h}\right] \leq \mathbb{E}_\xi\left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta)}\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|\sqrt{\sum_{m=1}^M\left\|\sum_{i=1}^n d(X_i)\psi_h(X_i)\xi_{im}\right\|_{\mathcal{H}_h}^2}\right]$$
$$\leq \frac{\sqrt{\delta}}{4D\sigma}\sqrt{\sum_{m=1}^M\mathbb{E}_\xi\left[\left\|\sum_{i=1}^n d(X_i)\psi_h(X_i)\xi_{im}\right\|_{\mathcal{H}_h}^2\right]} = \frac{\sqrt{M\delta}}{4D\sigma}\sqrt{\sum_{i=1}^n d(X_i)^2\|\psi_h(X_i)\|_{\mathcal{H}_h}^2}.$$

Jensen's inequality gives

$$\mathbb{E}_X\left[\sqrt{\sum_{i=1}^n d(X_i)^2\|\psi_h(X_i)\|_{\mathcal{H}_h}^2}\right] \leq \sqrt{\sum_{i=1}^n\mathbb{E}_X\left[d(X_i)^2\|\psi_h(X_i)\|_{\mathcal{H}_h}^2\right]} \leq \sqrt{n}D\sigma.$$

Since

$$\sum_{m=1}^M\left(\|\mu_m\|_{\mathcal{H}_h}^2-\|\mu_m^*\|_{\mathcal{H}_h}^2\right)^2 = \sum_{m=1}^M\left(\|\mu_m-\mu_m^*\|_{\mathcal{H}_h}^2+2\langle\mu_m-\mu_m^*,\mu_m^*\rangle_{\mathcal{H}_h}\right)^2 = \sum_{m=1}^M\langle\mu_m-\mu_m^*,\mu_m+\mu_m^*\rangle_{\mathcal{H}_h}^2$$
$$\leq \sum_{m=1}^M\|\mu_m-\mu_m^*\|_{\mathcal{H}_h}^2\|\mu_m+\mu_m^*\|_{\mathcal{H}_h}^2 \leq 4D^2\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2 \leq \frac{\delta}{4\sigma^2},$$

the second term can be bounded by

$$\mathbb{E}_{\xi'}\left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta),\ s=\pm1}\sum_{m=1}^M\left(\|\mu_m\|_{\mathcal{H}_h}^2-\|\mu_m^*\|_{\mathcal{H}_h}^2\right)(s\xi'_m)\right] \leq \mathbb{E}_{\xi'}\left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta),\ s=\pm1}\sqrt{\sum_{m=1}^M\left(\|\mu_m\|_{\mathcal{H}_h}^2-\|\mu_m^*\|_{\mathcal{H}_h}^2\right)^2}\sqrt{\sum_{m=1}^M(s\xi'_m)^2}\right]$$
$$\leq \frac{\sqrt{\delta}}{2\sigma}\mathbb{E}_{\xi'}\left[\sqrt{\sum_{m=1}^M(\xi'_m)^2}\right] \leq \frac{\sqrt{M\delta}}{2\sigma}.$$

Note that

$$\mathbb{E}_X\left[\sqrt{\sum_{i=1}^n d(X_i)^2}\right] \leq \sqrt{\sum_{i=1}^n\mathbb{E}_X\left[d(X_i)^2\right]} \leq \sqrt{n}\sigma.$$

Combining these, we obtain

$$\mathbb{E}_{X,g}\left[\sup_{\boldsymbol{\mu}\in\mathcal{V}(\delta),\ s=\pm1} Z_{\boldsymbol{\mu},s}\right] \leq 4\sqrt{2}\frac{\sqrt{nM\delta}}{4}+2\sqrt{2}\frac{\sqrt{nM\delta}}{2} = 2\sqrt{2nM}\sqrt{\delta}$$

and thus

$$\mathbb{E}\left[\sup_{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2\leq\delta/(16D^2\sigma^2)}|(P-P_n)(\zeta(\boldsymbol{\mu},\cdot)-\zeta(\boldsymbol{\mu}^*,\cdot))|\right] \leq \frac{1}{n}2\sqrt{\frac{\pi}{2}}2\sqrt{2nM}\sqrt{\delta} = \frac{4\sqrt{\pi M}\sqrt{\delta}}{\sqrt{n}}.$$

□

D. Proof of Theorem 7

Proof. For a partition $\mathscr{P}_M = \{W_1, \ldots, W_M\}$ of the data space $\mathcal{X}$ such that $\mathbb{P}(W_m) > 0$ $(m = 1, \ldots, M)$, let

$$\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathscr{P}_M, \mathbb{P}) := \sum_{m=1}^{M} \int_{W_m} d(x) \|\psi_h(x) - \mu_m\|_{\mathcal{H}_h}^2 \mathbb{P}(dx).$$

The optimal solution $\boldsymbol{\mu}(\mathscr{P}_M) := \{\mu_1(W_1), \ldots, \mu_M(W_M)\}$ of $\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathscr{P}_M, \mathbb{P})$ is given by

$$\mu_m(W_m) := \frac{\int_{W_m} d(y)\psi_h(y)\mathbb{P}(dy)}{\int_{W_m} d(y)\mathbb{P}(dy)} \quad (m = 1, \ldots, M).$$

By the Riesz representation theorem, we can easily see that $\mu_m(W_m) \in \mathcal{H}_h$ for $m = 1, \ldots, M$. Using the reproducing property, we have

$$\begin{aligned}
&\|\psi(x) - \mu_m(W_m)\|_{\mathcal{H}_h}^2 \\
&= \left\langle \frac{k(x,\cdot)}{d(x)d(\cdot)}, \frac{k(x,\cdot)}{d(x)d(\cdot)} \right\rangle_{\mathcal{H}_h} + \frac{1}{d(W_m)^2} \left\langle \int_{W_m} d(y) \frac{k(\cdot,y)}{d(\cdot)d(y)} \mathbb{P}(dy), \int_{W_m} d(z) \frac{k(\cdot,z)}{d(\cdot)d(z)} \mathbb{P}(dz) \right\rangle_{\mathcal{H}_h} \\
&\quad - \frac{2}{d(W_m)} \left\langle \frac{k(x,\cdot)}{d(x)d(\cdot)}, \int_{W_m} d(y) \frac{k(\cdot,y)}{d(\cdot)d(y)} \mathbb{P}(dy) \right\rangle_{\mathcal{H}_h} \\
&= \left\langle \frac{k(x,\cdot)}{d(x)d(\cdot)}, \frac{k(x,\cdot)}{d(x)d(\cdot)} \right\rangle_{\mathcal{H}_h} + \frac{1}{d(W_m)^2} \int_{W_m} \int_{W_m} d(y)d(z) \left\langle \frac{k(\cdot,y)}{d(\cdot)d(y)}, \frac{k(\cdot,z)}{d(\cdot)d(z)} \right\rangle_{\mathcal{H}_h} \mathbb{P}(dy)\mathbb{P}(dz) \\
&\quad - \frac{2}{d(W_m)} d(y) \int_{W_m} \left\langle \frac{k(x,\cdot)}{d(x)d(\cdot)}, \frac{k(\cdot,y)}{d(\cdot)d(y)} \right\rangle_{\mathcal{H}_h} \mathbb{P}(dy) \\
&= \frac{k(x,x)}{d(x)d(x)} + \frac{1}{d(W_m)^2} \int_{W_m} \int_{W_m} k(y,z)\mathbb{P}(dy)\mathbb{P}(dz) - \frac{2}{d(W_m)} \int_{W_m} \frac{k(x,y)}{d(x)} \mathbb{P}(dy).
\end{aligned}$$

Thus, we obtain

$$\mathrm{WKKM}_h(\mathscr{P}_M \mid \mathbb{P}) = \int_{\mathcal{X}} \frac{k(x,x)}{d(x)} \mathbb{P}(dx) - \sum_{m=1}^{M} \frac{1}{d(W_m)} \int_{W_m} \int_{W_m} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy). \tag{D.1}$$

Here, we note that

$$\begin{aligned}
d(W_m) &= \int_{W_m} \int_{\mathcal{X}} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy) \\
&= \int_{W_m} \int_{W_m} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy) + \int_{W_m} \int_{\mathcal{X}\setminus W_m} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\mathrm{WKKM}_h(\mathscr{C}_M \mid \mathbb{P}) &= \int_{\mathcal{X}} \frac{k(x,x)}{d(x)} \mathbb{P}(dx) - \sum_{m=1}^{M} \frac{1}{d(W_m)} \left\{ d(W_m) - \int_{W_m} \int_{\mathcal{X}\setminus W_m} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy) \right\} \\
&= \int_{\mathcal{X}} \frac{k(x,x)}{d(x)} \mathbb{P}(dx) - M + \sum_{m=1}^{M} \frac{1}{d(W_m)} \int_{W_m} \int_{\mathcal{X}\setminus W_m} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy) \\
&= Const. + \mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}),
\end{aligned}$$

where $Const.$ does not depend on $\mathscr{P}_M$. From this, we can conclude that the weighted kernel k-means $\mathrm{WKKM}_h(\mathscr{P}_M \mid \mathbb{P})$ is equivalent to the normalized cut $\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P})$. □

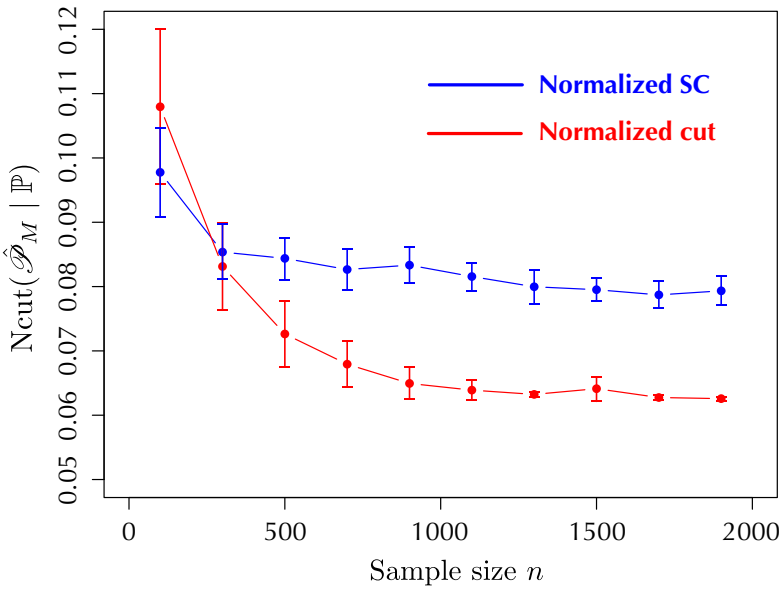

Figure 1. Convergence of $\mathrm{Ncut}(\hat{\mathscr{P}}_M \mid \mathbb{P})$ for both NSC and Ncut at each sample size. Each point is the mean of 20 replicates $\pm$ 2 times standard error.

E. Additional numerical experiments

In this section, we provide more detailed numerical experiments related to the essential difference between normalized spectral clustering (NSC) and normalized cut (Ncut). First, we consider the two moon data example in Section 5 of the main paper. Here, we employed the same setting in Section 5. For each sample size, we generated independently 40 datasets and applied normalized spectral clustering and normalized cut. We approximately evaluated the loss of the population level Ncut for each estimated partitioning using the Monte-Carlo simulation as follows:

$$\mathrm{Ncut}(\hat{\mathscr{P}}_M \mid \mathbb{P}) = \sum_{m=1}^{M} \frac{1}{d(\hat{W}_m)} \int_{\hat{W}_m} \int_{\mathcal{X}\setminus\hat{W}_m} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy) \approx \frac{1}{N^2} \sum_{m=1}^{M} \frac{1}{\tilde{d}_N(\hat{W}_m)} \sum_{X_i^* \in \hat{W}_m} \sum_{X_j^* \notin \hat{W}_m} k(X_i^*, X_j^*),$$

where $\hat{\mathscr{P}}_M = \{\hat{W}_1, \ldots, \hat{W}_M\}$ be an estimated partition of $\mathcal{X} = \mathbb{R}^2$ from data $\mathcal{X}_n$, $\mathcal{X}_N^* = (X_1^*, \ldots, X_N^*)$ and $\mathcal{X}_N^\dagger = (X_1^\dagger, \ldots, X_N^\dagger)$ are two independently generated i.i.d. sample from $\mathbb{P}$, and $\tilde{d}_N(\hat{W}_m) = \sum_{X_i^* \in \hat{W}_m} \sum_{j=1}^{N} k(X_i^*, X_j^\dagger)/N^2$. Here, we set $N = 10^4$, and we used 10^4 random initial values to avoid local minima in each method. From the consistency results, the values of $\mathrm{Ncut}(\mathscr{P} \mid \mathbb{P})$ converge to the corresponding limit in both spectral clustering and normalized cut. Figure 1 shows the means of 40 loss values with $\pm 2\hat{\sigma}/\sqrt{40}$ at each sample size, where $\hat{\sigma}$ is a standard deviation of 40 loss values for both NSC and Ncut. From this figure, we can see that values of $\mathrm{Ncut}(\hat{\mathscr{P}}_M \mid \mathbb{P})$ for NSC and Ncut converge to around 0.08 and 0.06, respectively. Thus, it seems that NSC does not provide the optimal partition of $\mathrm{Ncut}(\mathscr{P} \mid \mathbb{P})$ even in the large sample limit in this setting. Here, we remark that we used the same tuning parameter for both NSC and Ncut.

Next, we provide more detail results of the image segmentation example in Section 5 of the main paper. There are two major spectral clustering algorithms (see von Luxburg (2007)). Here, we refer these algorithms proposed by Shi & Malik (2000) and Ng et al. (2002) as NSC-SM and NSC-NJW, respectively. We also use the Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ to compute similarities among the representing points in this application. For each image, we applied NSC-NJW, NSC-SM, and Ncut with several values of σ^2 in the same way as Section 5. Figure 2 and Figure 3 show the image segmentation results of the first and second images at several values of σ^2, respectively. From these results, we can see that, at the same values of σ^2, spectral clustering provides a different segmentation from Ncut. Moreover, it seems that spectral clustering requires a smaller value of σ^2 to provide a similar result to that of Ncut.

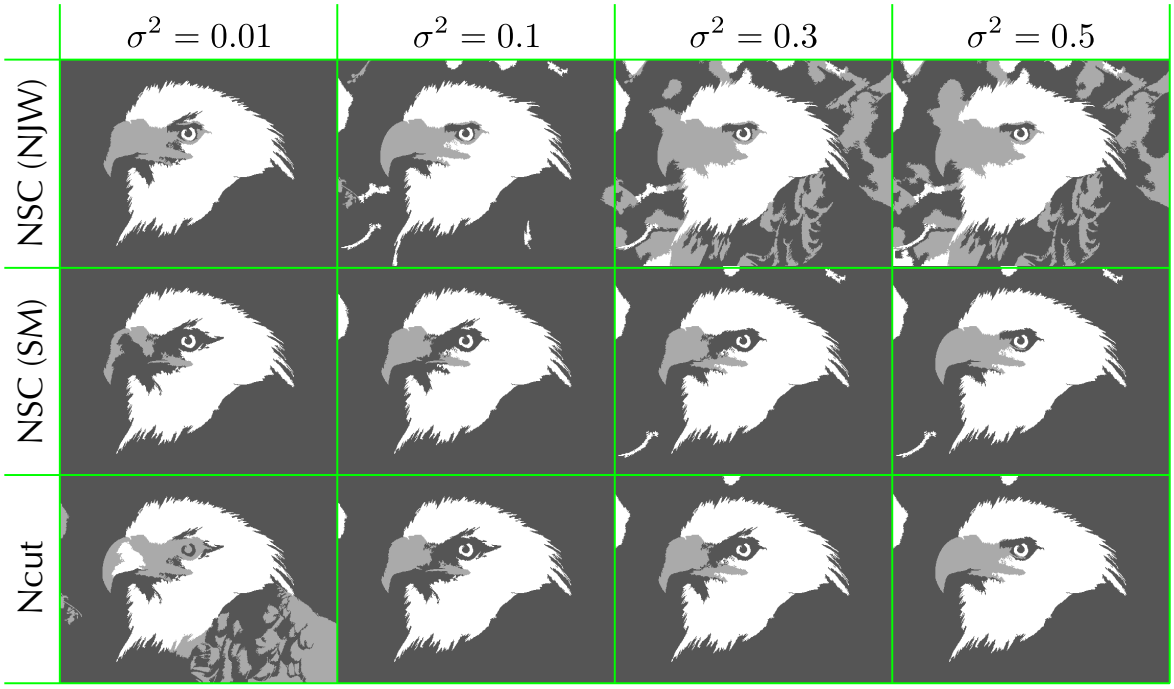

Figure 2. The image segmentation results of the first image at several values of the tuning parameter σ^2.

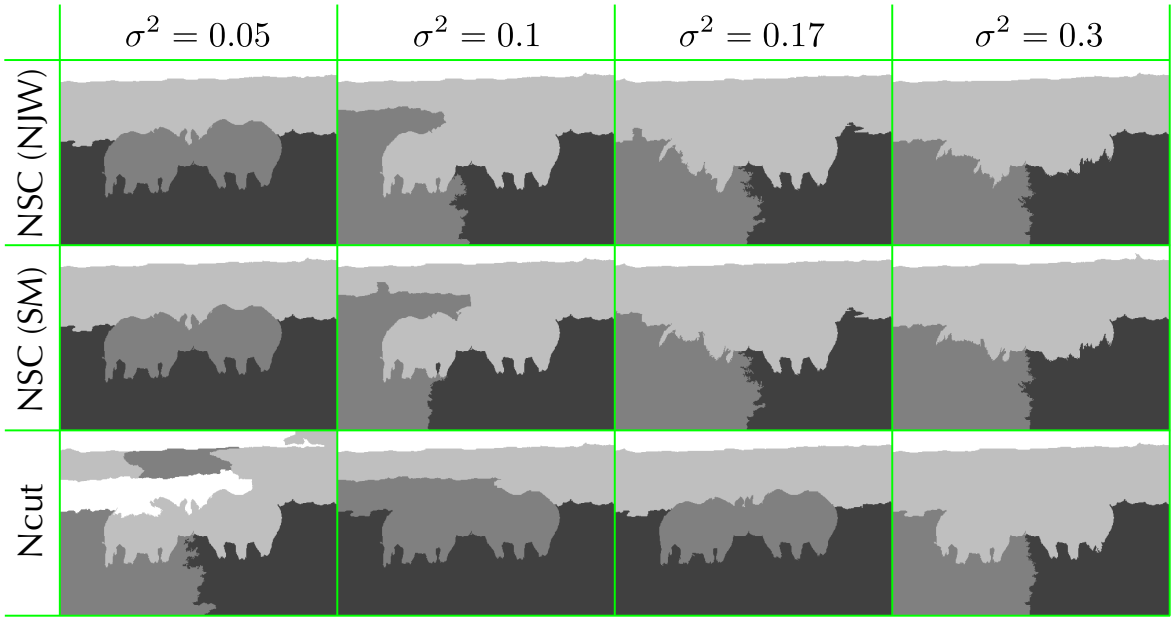

Figure 3. The image segmentation results of the second image at several values of the tuning parameter σ^2.

References

Bartlett, P. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Brezis, H. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer-Verlag, New York, 2011.

Graf, S. and Luschgy, H. *Foundations of Quantization for Probability Distributions*. Springer-Verlag, 2000.

Graf, S., Luschgy, H., and Pagès, G. Optimal quantizers for radon random vectors in a banach space. *Journal of Approximation Theory*, 144:27–53, 2007.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces*. Springer-Verlag, 1991.

Levrard, C. Nonasymptotic bounds for vector quantization in hilbert spaces. *Annals of Statistics*, 43:592–619, 2015.

Massart, P. *Concentration Inequalities and Model Selection*. Springer-Verlag, 2007.

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

Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: analysis and an algorithm. In *Proceedings of Advances in Neural Information Processing Systems 14*, pp. 849–856, 2002.

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

Steinwart, I. and Christmann, A. (eds.). *Support Vector Machines*. Springer-Verlag, 2008.

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

Webster, R. *Comvexity*. Oxford University Press, 1994.