1. Appendix

Here we present a more detailed proof of Theorem 1 and 2.

1.1. Proof of Theorem 1

We prove a more general result:

Theorem 1. *Consider vectors $x_i \in \mathbb{R}^m$ for $i = 1, 2, ..., n$ and their partitions $V_1, V_2, \ldots, V_K$ with sizes $n_1, n_2, \ldots, n_K$. Take the SON optimization:*

$$\min_{\{u_i \in \mathbb{R}^m\}} \frac{1}{2}\sum_{i=1}^{n} \|x_i - u_i\|_2^2 + \lambda \sum_{i \neq j} \|u_i - u_j\|_2 \tag{1}$$

and its associated centroid optimization:

$$\min_{\{v_\alpha \in \mathbb{R}^m\}} \frac{1}{2}\sum_{i=1}^{K} \|v_\alpha - c_\alpha\|_2^2 n_\alpha + \lambda \sum_{\alpha \neq \beta} n_\alpha n_\beta \|c_\alpha - c_\alpha\|_2 \tag{2}$$

where

$$c_\alpha = \frac{\sum\limits_{i \in V_\alpha} x_i}{n_\alpha}$$

.

1. *Suppose that for every $\alpha \in [K]$,*
   $$\frac{\max\limits_{i,j \in V_\alpha} \|x_i - x_j\|}{n_\alpha} \leq \lambda.$$
   Then, $u_i = v_\alpha$ for $i \in V_\alpha$ is a global solution of the SON clustering.
2. *If all c_αs are distinct and $\frac{d}{2n\sqrt{K}} \geq \lambda$ where $d = \min\limits_{\alpha \neq \beta} \|c_\alpha - c_\beta\|$, then all centroids v_α are distinct.*
3. *If $\max\limits_{\alpha} \frac{\|c_\alpha - c\|}{n - n_\alpha} \geq \lambda$ where $c = \sum\limits_{i=1}^{n} x_i/n$, then at least two centroids v_α are distinct.*

Proof. Notice that the solution of the centroid optimization satisfies

$$c_\alpha - v_\alpha = \lambda \sum_\beta n_\beta z_{\alpha,\beta}$$

where $\|z_{\alpha,\beta}\| \leq 1$, $z_{\alpha,\beta} = -z_{\beta,\alpha}$ and whenever $v_\alpha \neq v_\beta$, the relation $z_{\alpha,\beta} = \frac{v_\alpha - v_\beta}{\|v_\alpha - v_\beta\|_2}$ holds. Now, for the solution $u_i = v_\alpha$ for $i \in V_\alpha$, define

$$z'_{ij} = \begin{cases} z_{\alpha,\beta} & \alpha \neq \beta \\ \frac{x_i - x_j}{\lambda n_\alpha} & \alpha = \beta \end{cases},$$

where $i \in V_\alpha$, $j \in V_\beta$. It is easy to see that $\|z'_{ij}\|_2 \leq 1$, $z'_{ij} = -z'_{ji}$ and whenever $u_i \neq u_j$, we have that $z'_{ij} = \frac{u_i - u_j}{\|u_i - u_j\|_2}$. Further for each i,

$$\lambda \sum_j z'_{i,j} = \lambda \sum_\beta z_{\alpha,\beta} n_\beta + \sum_{j \in V_\alpha} \frac{x_i - x_j}{n_\alpha}$$

$$= c_\alpha - v_\alpha + x_i - c_\alpha = x_i - v_\alpha = x_i - u_i$$

This shows that the local optimality conditions for the SON optimization holds and proves part a.

For part b, denote the solution of the centroid optimization by $v_\alpha(\lambda)$ and notice that the solution of SON consists of distinct elements $v_\alpha = c_\alpha$ and is continuous at $\lambda = 0$. Hence, v_αs remain distinct in an interval $\lambda \in [0,\ \lambda_1)$. Take λ_0 as the supremum of all possible λ_1s. Hence, the solution in $\lambda \in [0,\ \lambda_0)$ contains distinct element and at $\lambda = \lambda_0$ contains two equal elements (otherwise, one can extend $[0,\ \lambda_0)$ to some $[0,\ \lambda_0 + \epsilon)$, which is against λ being supremum). Now, notice that for $\lambda \in [0\ \lambda_0)$ the objective function is smooth at the optimal point. Hence, $v_\alpha(\lambda)$ is differentiable and satisfies

$$\delta = \left[\frac{\mathrm{d}v_\alpha}{\mathrm{d}\lambda}\right]_\alpha = H^{-1}\frac{\partial g}{\partial \lambda} \tag{3}$$

where $[.]_\alpha$ and $[.]_{\alpha,\beta}$ denote block vectors and block matrices respectively. Moreover, H and g are the Hessian and the gradient of the objective function at the optimal point. In other words,

$$H = \Big[n_\alpha \delta_{\alpha,\beta} I + \frac{I\|v_\alpha - v_\beta\|_2^2 - (v_\alpha - v_\beta)(v_\alpha - v_\beta)^\top}{\|v_\alpha - v_\beta\|_2^3} \lambda n_\alpha n_\beta\Big]_{\alpha,\beta}$$

and

$$\frac{\partial g}{\partial \lambda} = \left[\sum_\beta z_{\alpha,\beta} n_\alpha n_\beta\right]_\alpha$$

Hence,

$$\delta = \left[\delta_{\alpha,\beta} I + \frac{I\|v_\alpha - v_\beta\|_2^2 - (v_\alpha - v_\beta)(v_\alpha - v_\beta)^\top}{\|v_\alpha - v_\beta\|_2^3} \lambda n_\beta\right]_{\alpha,\beta}^{-1} \times \left[\sum_\beta z_{\alpha,\beta} n_\beta\right]_\alpha$$

Simple calculations show that $\|\delta\|_2 \leq n\sqrt{K}$. Hence,

$$\left\|\frac{\mathrm{d}v_\alpha}{\mathrm{d}\lambda}\right\|_2 \leq \|\delta\|_2 \leq \sqrt{K} n$$

This yields for $\lambda < \lambda_0$ to

$$\|v_\alpha(\lambda) - v_\beta(\lambda)\|_2 = \left\|c_\alpha - c_\beta + \int_0^\lambda \left(\frac{\mathrm{d}v_\alpha}{\mathrm{d}\lambda} - \frac{\mathrm{d}v_\beta}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda\right\|_2$$

$$\geq \|c_\alpha - c_\beta\|_2 - \int_0^\lambda \left\|\frac{\mathrm{d}v_\alpha}{\mathrm{d}\lambda} - \frac{\mathrm{d}v_\beta}{\mathrm{d}\lambda}\right\|_2 \mathrm{d}\lambda$$

$$\geq d - 2n\lambda\sqrt{K}$$

Since at $\lambda = \lambda_0$, we have that $v_\alpha = v_\beta$ for some $\alpha \neq \beta$, we get that $d - 2n\lambda_0\sqrt{K} \leq 0$ or $\lambda_0 \geq d/2n\sqrt{K}$. this proves part b.

For part c, Take a value of λ, where $v_1 = v_2 = \ldots = v_K$. It is simple to see that in this case $v_\alpha = c$. The optimality condition leads to

$$c - c_\alpha = \lambda \sum_{\beta \neq \alpha} z_{\alpha,\beta} n_\beta$$

Hence, $\|c - c_\alpha\|_2 \leq \lambda(n - n_\alpha)$. This proves part c. □

1.2. Proof of Theorem 2

Denote by $\mathbf{U}_k$ a matrix where the i^{th} column is the value of u_i at the k^{th} iteration. Define

$$\psi_\mu(\mathbf{U}) = \mathcal{E}\left(\mathbf{U}_{k+1} \mid \mathbf{U}_k = \mathbf{U}, \mu_k = \mu\right), \tag{4}$$

which by simple manipulations leads to

$$\psi_\mu(\mathbf{U}) = \mathbf{U} + \frac{1}{\binom{n}{2}} \sum_{i<j} \left(\mathbf{L}_{ij}(\Pi_{ij}^{(\mu)}(u_i, u_j)) - \mathbf{L}_{ij}(u_i, u_j)\right)$$

where u_i denotes the i^{th} column of $\mathbf{U}$ and $\mathbf{L}_{ij}(x, y)$ is a matrix where the i^{th} column is x, the j^{th} column is y and the rest are zero. Also, denote

$$\begin{aligned}\sigma_\mu^2(\mathbf{U}) &= \mathrm{Var}\left(\mathbf{U}_{k+1} \mid \mathbf{U}_k = \mathbf{U}, \mu_k = \mu\right) \\ &= \mathcal{E}\left(\|\mathbf{U}_{k+1}\|_2^2 \mid \mathbf{U}_k = \mathbf{U}, \mu_k = \mu\right) - \|\phi_\mu(\mathbf{U})\|_2^2\end{aligned} \tag{5}$$

We prove a more detailed theorem:

Theorem 2. *Starting from $\bar{\mathbf{U}}_0 = \mathbf{U}_0$ (the initialization of the algorithm), define the characteristic sequence $\{\bar{\mathbf{U}}_k\}_{k=0}^{\infty}$ by the following iteration:*

$$\bar{\mathbf{U}}_{k+1} = \psi_{\mu_k}(\bar{\mathbf{U}}_k)$$

1. *We have that*

$$\Pr\left(\sup_k \|\mathbf{U}_k - \bar{\mathbf{U}}_k\|_{\mathrm{F}}^2 + \sum_{l=k}^{\infty} \mu_l^2 > \lambda\right) \leq \frac{\sum_{k=0}^{\infty} \mu_k^2}{\lambda} \tag{6}$$

2. *Define $\tilde{\mathbf{U}}$ as the unique optimal solution of the SON optimization and suppose that $\{\mu_k\}$ is a non-increasing sequence.*

   (a) *There exists a positive sequence $h_n = O(\frac{1}{n})$, where n is the number of data points, such that*

$$\begin{aligned} R(\bar{\mathbf{U}}_k, \mu_k) \leq h_n \sum_{l=0}^{k-1} \mu_l^2 e^{-\frac{2}{n^2}\sum_{s=l+1}^{k-1}\frac{\mu_k}{1+\mu_k}} \\ + R(\mathbf{U}_0, \mu_0) e^{-\frac{2}{n^2}\sum_{s=0}^{k-1}\frac{\mu_k}{1+\mu_k}} \end{aligned} \tag{7}$$

   where

$$R(\mathbf{U}, \mu) = \frac{1}{2}\|\tilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2 + \mu\left(\Phi(\mathbf{U}) - \Phi(\tilde{\mathbf{U}})\right),$$

   (b) *There exists a universal constant a such that*

$$\begin{aligned} \|\bar{\mathbf{U}}_k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq a \sum_{l=0}^{k-1} \mu_l^2 e^{-\frac{2}{n^2}\sum_{s=l+1}^{k-1}\mu_s} \\ + \|\mathbf{U}_0 - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 e^{-\frac{2}{n^2}\sum_{s=0}^{k-1}\mu_s} \end{aligned}$$

3. *Assume that $\{\mu_k\}$ is non-increasing $\sum_0^\infty \mu_k = \infty$ and $\sum_0^\infty \mu_k^2 < \infty$. Then, the sequence $\mathbf{U}_k$ converges to $\tilde{\mathbf{U}}$ in the following strong probability sense:*

$$\forall \epsilon > 0; \lim_{k\to\infty} \Pr\left(\sup_{l\geq k} \|\mathbf{U}_l - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 > \epsilon\right) = 0 \tag{8}$$

4. *Take $\mu_k = \frac{\mu_1}{k^\alpha}$ for $k = 1, 2, \ldots$ and $\frac{2}{3} < \alpha < 1$. For sufficiently small values of $\epsilon > 0$ the relation*

$$\|\mathbf{U}_l - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 = O\left(\frac{1}{l^{3\alpha-2-\epsilon}}\right)$$

   holds with probability 1.

Proof. Denote by Ω_k the pair (i, j) which is selected in iteration k and $\Omega^k = (\Omega_0, \Omega_1, \ldots, \Omega_{k-1})$. Also, denote $\psi_\mu(\mathbf{U}, (i, j)) = \mathbf{U} + \mathbf{L}_{ij}(\Pi_{ij}^\mu(u_i, u_j)) - \mathbf{L}_{ij}(u_i, u_j)$. Then, the iterations can be written as

$$\begin{aligned}\mathbf{U}_{k+1} &= \psi_{\mu_k}(\mathbf{U}_k, \Omega_k) \\ \bar{\mathbf{U}}_{k+1} &= \mathcal{E}(\psi_{\mu_k}(\bar{\mathbf{U}}_k, \Omega) \mid \bar{\mathbf{U}}_k)\end{aligned} \tag{9}$$

Define $\boldsymbol{\Delta}_k = \mathbf{U}_k - \bar{\mathbf{U}}_k$ and $\boldsymbol{\eta}_k = \psi_{\mu_k}(\bar{\mathbf{U}}_k, \Omega_k) - \mathcal{E}(\psi_{\mu_k}(\bar{\mathbf{U}}_k, \Omega) \mid \bar{\mathbf{U}}_k)$. Also, denote $\mathcal{U} = \{\bar{\mathbf{U}}_k\}_{k=0}^{\infty}$. Notice that the sequence $\{\boldsymbol{\eta}_k\}_{k=0}^{\infty}$ consists of zero-mean independent elements. Subtracting the two iterations in (9) gives us:

$$\boldsymbol{\Delta}_{k+1} = \psi_{\mu_k}(\mathbf{U}_k, \Omega_k) - \psi_{\mu_k}(\bar{\mathbf{U}}_k, \Omega_k) + \boldsymbol{\eta}_k \tag{10}$$

It is simple to see that $\Pi_{ij}^\mu(u_i, u_j)$ is a contraction map for any μ, i, j. Then, it is simple to deduce that $\psi_\mu(\mathbf{U}, \Omega)$ is a contraction map for any Ω and μ. As a result, we obtain from (10) that

$$\mathcal{E}\left(\|\boldsymbol{\Delta}_{k+1} - \boldsymbol{\eta}_k\|_{\mathrm{F}}^2 \mid \Omega^k\right) \leq \|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2,$$

which can also be written as

$$\mathcal{E}\left(\|\boldsymbol{\Delta}_{k+1}\|_{\mathrm{F}}^2 \mid \Omega^k\right) \leq$$
$$\|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2 + 2\mathcal{E}\left(\langle \psi_{\mu_k}(\mathbf{U}_k, \Omega_k), \boldsymbol{\eta}_k\rangle \mid \Omega^k\right) - \mathcal{E}\|\boldsymbol{\eta}_k\|_{\mathrm{F}}^2$$

Now, it is simple to see that $\|\psi_\mu(\mathbf{U}, \Omega) - \mathbf{U}\| \leq \sqrt{2}\mu$. Furthermore, $\mathbf{U}_k$ only depends on $\Omega_0, \Omega_1, \ldots, \Omega_{k-1}$, while $\boldsymbol{\eta}_k$ is a function of Ω_k. Hence, $\mathbf{U}_k$ and $\boldsymbol{\eta}_k$ are independent and $\mathcal{E}(\langle \mathbf{U}_k, \boldsymbol{\eta}_k\rangle \mid \Omega^k) = \mathbf{0}$ This leads to

$$\mathcal{E}\left(\|\boldsymbol{\Delta}_{k+1}\|_{\mathrm{F}}^2 \mid \Omega^k\right) \leq$$
$$\|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2 + 2\mathcal{E}\left(\langle \psi_{\mu_k}(\mathbf{U}_k, \Omega_k) - \mathbf{U}_k, \boldsymbol{\eta}_k\rangle \mid \Omega^k\right) - \mathcal{E}\|\boldsymbol{\eta}_k\|_{\mathrm{F}}^2$$
$$\leq \|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2 + 2\sqrt{2}\mu_k\sqrt{\mathcal{E}(\|\boldsymbol{\eta}_k\|_2^2)} - \mathcal{E}\|\boldsymbol{\eta}_k\|_{\mathrm{F}}^2$$

Notice that $\mathcal{E}(\|\boldsymbol{\eta}_l\|_2^2) = \sigma_{\mu_l}^2(\bar{\mathbf{U}}_l)$ and

$$\|\mathbf{U}_{k+1} - \mathbf{U}_k\|_2 = \|\psi_{\mu_k}(U_k, \Omega_k) - \mathbf{U}_k\|_2 \leq \sqrt{2}\mu_k$$

which leads to

$$\sigma_\mu^2(\mathbf{U}) \leq 2\mu^2.$$

We conclude that

$$\mathcal{E}\left(\|\boldsymbol{\Delta}_{k+1}\|_{\mathrm{F}}^2 \mid \Omega^k\right) \leq \|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2 + 4\mu_k^2$$

Define $s_k = \sum_{l=k}^{\infty} \mu_l^2$. We observe that $\|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2 + s_k$ is a supermartingale. Hence, from the suprmartingale version of the Doob's inequality we obtain that

$$\Pr\left(\sup_k \|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2 + s_k > \lambda\right) \leq \frac{\mathcal{E}\|\boldsymbol{\Delta}_0\|_{\mathrm{F}}^2 + s_0}{\lambda} = \frac{\sum_{k=0}^{\infty} \mu_k^2}{\lambda}$$

This proves part (1).

For part (2) from the definition of the proximal operator, there exists a vector $\boldsymbol{\zeta} \in \partial\phi_\Omega(\psi_\mu(\mathbf{U}, \Omega))$ such that $\psi_\mu(\mathbf{U}, \Omega) = \mathbf{U} - \mu\boldsymbol{\zeta}$. We conclude that

$$\phi_\Omega(\tilde{\mathbf{U}}) - \phi_\Omega(\psi_\mu(\mathbf{U}, \Omega)) \geq$$
$$\frac{1}{\mu}\langle \mathbf{U} - \psi_\mu(\mathbf{U}, \Omega), \tilde{\mathbf{U}} - \psi_\mu(\mathbf{U}, \Omega)\rangle =$$
$$\frac{1}{2\mu}\left(\|\tilde{\mathbf{U}} - \psi_\mu(\mathbf{U}, \Omega)\|_{\mathrm{F}}^2 - \|\tilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2 + \|\mathbf{U} - \psi_\mu(\mathbf{U}, \Omega)\|_{\mathrm{F}}^2\right)$$

Hence,

$$\Phi(\tilde{\mathbf{U}}) - \sum_\Omega \phi_\Omega(\psi_\mu(\mathbf{U}, \Omega))$$
$$\geq \tfrac{n(n-1)}{4\mu}\left(\mathcal{E}\|\tilde{\mathbf{U}} - \psi_\mu(\mathbf{U}, \Omega)\|_{\mathrm{F}}^2 - \|\tilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2\right)$$
$$\geq \tfrac{n(n-1)}{4\mu}\left(\|\tilde{\mathbf{U}} - \psi_\mu(\mathbf{U})\|_{\mathrm{F}}^2 - \|\tilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2\right) \quad (11)$$

where the last inequality is obtained by Jensen's inequality. Notice that

$$\sum_\Omega \phi_\Omega(\psi_\mu(\mathbf{U}, \Omega)) =$$
$$\sum_{\Omega,\Omega'} \phi_{\Omega'}(\psi_\mu(\mathbf{U}, \Omega)) - \sum_{\Omega\neq\Omega'} \phi_{\Omega'}(\psi_\mu(\mathbf{U}, \Omega))$$
$$\geq \frac{n(n-1)}{2}\Phi(\psi_\mu(\mathbf{U})) - \sum_{\Omega\neq\Omega'} \phi_{\Omega'}(\psi_\mu(\mathbf{U}, \Omega))$$
$$= \Phi(\mathbf{U}) + \frac{n(n-1)}{2}\left(\Phi(\psi_\mu(\mathbf{U})) - \Phi(\mathbf{U})\right)$$
$$- \sum_{\Omega\neq\Omega'}\left(\phi_{\Omega'}(\psi_\mu(\mathbf{U}, \Omega)) - \phi_{\Omega'}(\mathbf{U})\right)$$

Now, notice that $\phi_{\Omega'}(\psi_\mu(\mathbf{U}, \Omega)) - \phi_{\Omega'}(\mathbf{U}) = 0$ when Ω and Ω' do not overlap. Also, there exists a constant a such that $|\phi_{\Omega'}(\psi_\mu(\mathbf{U}, \Omega)) - \phi_{\Omega'}(\mathbf{U})| < a\mu$. We conclude that

$$\sum_\Omega \phi_\Omega(\psi_\mu(\mathbf{U}, \Omega)) \geq$$
$$\Phi(\mathbf{U}) + \frac{n(n-1)}{2}\left(\Phi(\psi_\mu(\mathbf{U})) - \Phi(\mathbf{U})\right) - 2(n-2)a\mu$$

Define $h_n = 8(n-2)a/n(n-1) = O(\frac{1}{n})$. Replacing this result in (11) and performing straightforward calculations leads to

$$h_n\mu^2 \geq \tfrac{2\mu}{n^2}\left(\Phi(\mathbf{U}) - \Phi(\tilde{\mathbf{U}})\right)$$
$$+\mu\left(\Phi(\psi_\mu(\mathbf{U})) - \Phi(\mathbf{U})\right)$$
$$+\tfrac{1}{2}\left(\|\tilde{\mathbf{U}} - \psi_\mu(\mathbf{U})\|_{\mathrm{F}}^2 - \|\tilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2\right) \quad (12)$$

Now, we introduce the recursion to (12). We introduce $R_k = R(\bar{\mathbf{U}}_k, \mu_k)$ and use monotonicity of μ_k to conclude that:

$$h_n\mu_k^2 \geq \frac{2\mu_k}{n^2}\left(\Phi(\bar{\mathbf{U}}_k) - \Phi(\tilde{\mathbf{U}})\right) + R_{k+1} - R_k$$

Finally, we use the fact that $\Phi(.)$ is a $1-$strongly convex function which leads to $\Phi(\mathbf{U}) - \Phi(\tilde{\mathbf{U}}) \geq \frac{1}{2}\|\tilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2$, and conclude that

$$\Phi(\mathbf{U}) - \Phi(\tilde{\mathbf{U}}) \geq \frac{R(\mathbf{U}, \mu)}{1+\mu}$$

This yields to

$$R_{k+1} - h_n\mu_k^2 \leq \left(1 - \frac{\frac{2\mu_k}{n^2}}{1+\mu_k}\right)R_k \leq e^{-\frac{\frac{2\mu_k}{n^2}}{1+\mu_k}}R_k$$

where the last equality holds because $1-x \leq e^{-x}$ for every positive x. It is now simple to see by induction that

$$R_k \leq h_n\sum_{l=0}^{k-1}\mu_l^2 e^{-\frac{2}{n^2}\sum_{s=l+1}^{k-1}\frac{\mu_s}{1+\mu_s}} + R_0 e^{-\frac{2}{n^2}\sum_{s=0}^{k-1}\frac{\mu_s}{1+\mu_s}} \quad (13)$$

which proves part (2a).

For part (2b), we observe from (11) that

$$\Phi(\tilde{\mathbf{U}}) - \Phi(\mathbf{U}) + \frac{n(n-1)}{2} a\mu \geq$$

$$\frac{n(n-1)}{4\mu}\left(\|\tilde{\mathbf{U}} - \psi_\mu(\mathbf{U})\|_{\mathrm{F}}^2 - \|\tilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2\right)$$

which with the similar argument to above leads to

$$\frac{1}{2}\|\bar{\mathbf{U}}_{k+1} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq \left(1 - \frac{2\mu_k}{n^2}\right)\frac{1}{2}\|\bar{\mathbf{U}}_k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 + a\mu_k^2$$

$$\leq \frac{1}{2}\|\bar{\mathbf{U}}_k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 e^{-\frac{2\mu_k}{n^2}} + a\mu_k^2$$

We conclude part (2b).

For part (3,4), define $\mathcal{U}^k = \{\bar{\mathbf{U}}_l^k\}_{l=0}^\infty$ as the sequence obtained by starting from $\bar{\mathbf{U}}_0^k = \mathbf{U}_k$ and applying

$$\bar{\mathbf{U}}_{l+1}^k = \psi_{\mu_{l+k}}(\bar{\mathbf{U}}_l^k)$$

Take arbitrary (non-zero) positive numbers ϵ, δ. Take λ such that $\lambda \geq \frac{2}{\delta}\sum\limits_{l=0}^{\infty}\mu_l^2$. Define

$$\Phi_{\max} = \max_{\|\mathbf{U} - \tilde{\mathbf{U}}\| \leq \lambda} \Phi(\mathbf{U})$$

Define l_0, k such that $\sum\limits_{l=k}^{\infty}\mu_l^2 < \epsilon\delta/8$ and

$$\forall l > l_0;\ h_n \sum_{t=0}^{l-1}\mu_{t+k}^2 e^{-\frac{2}{n^2}\sum\limits_{s=t+1}^{l-1}\frac{\mu_{s+k}}{1+\mu_{s+k}}} +$$

$$(\lambda + \mu_k\Phi_{\max})e^{-\frac{2}{n^2}\sum\limits_{s=0}^{l-1}\frac{\mu_{s+k}}{1+\mu_{s+k}}} < \frac{\epsilon}{8}$$

It is simple to see that such a choice exists because of the conditions in part (3). Now, we define two outcomes H_1 and H_2:

$$H_1 : \forall k \geq 0;\ \|\mathbf{U}_k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq \lambda$$

$$H_2 : \forall l \geq 0; \|\bar{\mathbf{U}}_l^k - \mathbf{U}_{l+k}\| \leq \frac{\epsilon}{4}$$

Notice that from part (1) we have that $\Pr(H_1^c)$ and $\Pr(H_2^c)$ are less than $\delta/2$. Furthermore, under $H_1 \cap H_2$ we have that:

$$\forall l > l_0;\ \|\mathbf{U}_{l+k} - \tilde{\mathbf{U}}\|_2^2 \leq 2(\|\mathbf{U}_{l+k} - \bar{\mathbf{U}}_l^k\|_{\mathrm{F}}^2 + \|\bar{\mathbf{U}}_l^k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2)$$

$$\leq 2(\frac{\epsilon}{4} + \frac{\epsilon}{4}) = \epsilon$$

This is because according to part (2),

$$\|\bar{\mathbf{U}}_l^k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq 2R(\bar{\mathbf{U}}_l^k, \mu_{l+k}) \leq$$

$$2h_n\sum_{t=0}^{l-1}\mu_{t+k}^2 e^{-\frac{2}{n^2}\sum\limits_{s=t+1}^{l-1}\frac{\mu_{s+k}}{1+\mu_{s+k}}}$$

$$+2R(\mathbf{U}_k, \mu_k)e^{-\frac{2}{n^2}\sum\limits_{s=0}^{l-1}\frac{\mu_{s+k}}{1+\mu_{s+k}}} \leq \frac{\epsilon}{4}$$

where we used H_1 to conclude that $R(\mathbf{U}_k, \mu_k) \leq \lambda + \Phi_{\max}\mu_k$. We conclude that

$$\Pr(\sup_{l > l_0 + k}\|\mathbf{U}_l - \tilde{\mathbf{U}}\|_2^2 > \epsilon) \leq \Pr(H_1^c) + \Pr(H_2^c) \leq \delta$$

which proves part (3).

For part (4), define $k_r = r^\gamma$, $\lambda_r = r^{-\beta}$, where $\gamma = \frac{1-\frac{\epsilon}{2}}{1-\alpha}$, $\beta < \gamma(2\alpha - 1) - 1$, and the outcomes:

$$Q_r : \sup_{l \geq 0}\|\mathbf{U}_{l+k_r} - \bar{\mathbf{U}}_l^{k_r}\|_{\mathrm{F}}^2 > \lambda_r.$$

By part (1), we have that

$$\sum_{r=1}^{\infty}\Pr(Q_r) < \infty.$$

Hence by Borel-Cantelli lemma, $Q_{r_0}^c, Q_{r_0+1}^c, Q_{r_0+2}^c, \ldots$ simultaneously hold for some r_0 with probability 1. For simplicity and without loss of generality, we assume that $r_0 = 0$ as it does not affect the asymptotic rate. Then for any $r > 0$, we have that

$$\sup_{l \geq 0}\|\mathbf{U}_{l+k_r} - \bar{\mathbf{U}}_l^{k_r}\|_{\mathrm{F}}^2 \leq \lambda_r$$

In particular,

$$\|\mathbf{U}_{k_{r+1}} - \bar{\mathbf{U}}_{l_r}^{k_r}\|_{\mathrm{F}}^2 \leq \lambda_r$$

where $l_r = k_{r+1} - k_r$. From part (2b), we conclude that

$$\|\bar{\mathbf{U}}_{l_r}^{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq A\sum_{t=0}^{l_r-1}\frac{1}{(t+k_r)^{2\alpha}}e^{-2a\sum\limits_{s=t+1}^{l_r-1}\frac{1}{(s+k_r)^\alpha}}$$

$$+\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 e^{-2a\sum\limits_{s=0}^{l_r-1}\frac{1}{(s+k_r)^\alpha}}$$

where we introduce $\mu_1 = bn^2$ and $A = 4an^4b^2$ for simplicity. This leads to

$$\|\mathbf{U}_{k_{r+1}} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq 2\lambda_r + A\sum_{t=0}^{l_r-1}\frac{1}{(t+k_r)^{2\alpha}}e^{-2b\sum\limits_{s=t+1}^{l_r-1}\frac{1}{(s+k_r)^\alpha}}$$

$$+2\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 e^{-2b\sum\limits_{s=0}^{l_r-1}\frac{1}{(s+k_r)^\alpha}}$$

$$\leq Le^{Lk_r^{1-\alpha} - Lk_{r+1}^{1-\alpha}}\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 +$$

$$2\lambda_r + A\sum_{t=0}^{l_r}\frac{1}{(t+k_r)^{2\alpha}}e^{L(k_r+t)^{1-\alpha} - Lk_{r+1}^{1-\alpha}}$$

where L denotes "some suitable constant" which may vary in difference occurrences. Notice that

$$\sum_{t=0}^{l_r} \frac{1}{(t+k_r)^{2\alpha}} e^{L(k_r+t)^{1-\alpha} - Lk_{r+1}^{1-\alpha}} = \sum_{t=k_r}^{k_{r+1}} \frac{1}{t^{2\alpha}} e^{Lt^{1-\alpha} - Lk_{r+1}^{1-\alpha}}$$

$$\leq L \sum_{t=k_r}^{k_{r+1} - Lk_{r+1}^{\alpha}(1+\rho\log(k_{r+1}))} \frac{1}{t^{2\alpha}} e^{-L\rho\log(k_{r+1})}$$

$$+ \sum_{t=k_{r+1} - Lk_{r+1}^{\alpha}(1+L\rho\log(k_{r+1}))}^{k_{r+1}} \frac{1}{t^{2\alpha}}$$

$$\leq L\left(\frac{1}{(k_{r+1} - Lk_{r+1}^{\alpha}(1+\rho\log(k_{r+1})))^{2\alpha-1}} - \frac{1}{k_{r+1}^{2\alpha-1}}\right)$$

$$+ L\frac{e^{-L\rho\log(k_{r+1})}}{k_r^{2\alpha-1}} \leq \frac{L\log(k_{r+1})}{k_{r+1}^{\alpha}} \leq \frac{L\log r}{r^{\gamma\alpha}} < \frac{L}{r^{\beta}}$$

where ρ is a sufficiently large constant and we use the fact that $\gamma\alpha > \gamma(2\alpha-1) - 1 > \beta$. Moreover,

$$k_r^{1-\alpha} - k_r^{1-\alpha} = r^{\gamma(1-\alpha)} - (r+1)^{\gamma(1-\alpha)} \leq -Lr^{\gamma(1-\alpha)-1}$$

We conclude that

$$\|\mathbf{U}_{k_{r+1}} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq \frac{L}{r^{\beta}} + Le^{-Lr^{\gamma(1-\alpha)-1}} \|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2$$

which leads to

$$\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq$$

$$L\left(\sum_{s=1}^{r-1} \frac{1}{s^{\beta}} e^{-L\sum_{t=s+1}^{r-1} t^{\gamma(1-\alpha)-1}} + e^{-L\sum_{t=0}^{r-1} t^{\gamma(1-\alpha)-1}}\right)$$

$$\leq L\left(\sum_{s=1}^{r-1} \frac{1}{s^{\beta}} e^{L\left(s^{\gamma(1-\alpha)} - r^{\gamma(1-\alpha)}\right)} + e^{-Lr^{\gamma(1-\alpha)}}\right)$$

With a similar approach to the above, we observe that

$$\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq \frac{L\log r}{r^{\beta-\frac{\epsilon}{2}}} \leq \frac{L}{r^{\beta-\epsilon}}$$

Take $k_r < l \leq k_{r+1}$. We observe that

$$\|\mathbf{U}_l - \tilde{\mathbf{U}}\|_2^2 \leq 2(\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_2^2 + \|\mathbf{U}_{k_r} - \mathbf{U}_l\|_2^2)$$

$$\leq 2\lambda_r + \frac{L}{r^{\beta-\epsilon}} \leq \frac{L}{r^{\beta-\epsilon}} \leq \frac{L}{l^{\frac{\beta-\epsilon}{\gamma}}}$$

By taking $\beta = \gamma(2\alpha-1) - 1$, we obtain part (4). □