

490 **A Additional details on simulation studies in Section 4.1**

491 In this section, we provide more details of our simulation setups and results for Gaussian mixtures in  
 492 Section 4.1.

493 We set  $m_1 = 40 \cdot r, m_2 = r, m_3 = 20 \cdot r$  which means that there are total  $81 \cdot r$  number of distributions  
 494 in  $G_1$ ,  $20 \cdot r$  distributions in  $G_2$ . The  $r$  is set to be 1, 2, 3, where we have  $n = 101, 202, 303$   
 495 respectively. The mean for the Gaussian distributions are shown in the table below. The entries of  
 496 covariance matrices for the Gaussian distributions are chosen to be  $O(10^{-3})$  for  $\mu_1, \mu_2$  and they are  
 497 chosen to be  $O(10^{-6})$  for  $\mu_3$  and  $\mu_4$ . Then we scale down the distribution with scaling parameter  
 498 equals 0.5. This ensures that with high probability, all the distributions will fall into the bounded  
 499 range  $[0, 1] \times [0, 1]$ .

500 The algorithm we use to get the barycenter is Frank-Wolfe algorithm with 200 iterations. And we use  
 501 Sinkhorn divergence to calculate the Wasserstein distance. The regularization parameters for both  
 502 algorithms are chosen to be  $10^{-3}$ . To approximate the true distribution, first we divide  $[0, 1] \times [0, 1]$   
 503 range into  $80 \times 80$  grids, then we randomly sample 600 samples each time and count the number of  
 504 times it falls into certain grid to approximate the distribution. The results show us that for each  $n$  and  
 505 each iterations among 50 repetitions, all the distributions in  $G_2^*$  will be assigned to same cluster, so it  
 506 will be reasonable to define that  $\mu_3$  is misclassified if any copy of them are in the same cluster of an  
 507 arbitrarily chosen  $\mu$  from  $G_2^*$ .

508 The arrangement of mean for Gaussian mixture models shown in Table 3 indicates that the distribu-  
 509 tions are set based on Example 3. Recall that in Section 4.1,  $\Delta_* := \max_{k=1,2} \max_{i,j \in G_k} W_2(\mu_i, \mu_j)$   
 510 and  $\Delta^* := \min_{i \in G_1, j \in G_2} W^2(\mu_i, \mu_j)$  are the maximum within-cluster distance and the minimum  
 511 between-cluster distance respectively. Table 5 shows that  $\Delta_* < \Delta^*$  on average and  $\Delta_* < \Delta^*$  for  
 512 around 80% among 50 repetitions. So we can expect Wasserstein SDP to correctly cluster all data  
 513 points in the Wasserstein space. From Table 4 we can observe that in our settings the time cost for  
 514 Wasserstein SDP and distance-based Wasserstein  $K$ -means is relatively lower than the time cost  
 515 for barycenter-based Wasserstein  $K$ -means. But we can see that as  $n$  increases, the time cost for  
 516 B-WKM grows almost linearly w.r.t.  $n$  while almost quadratically for W-SDP and D-WKM. Thus we  
 517 should expect relatively higher time cost for W-SDP and D-WKM when  $n$  is sufficiently large, where  
 518 we can consider several methods to bring down the time cost (e.g., subsampling-based method for  
 519 SDP from Zhuang et al. [2022]).

520 Computationally speaking, the calculations of Wasserstein distances and barycenters are usually  
 521 based on one-step discretization and one-step application of entropic regularization methods such  
 522 as Sinkhorn (Genevay et al. [2018], Janati et al. [2020]). Dvinskikh and Tiapkin [2020] shows  
 523 that the complexity of calculating barycenters should be  $O(d^2 n)$ , where  $d$  and  $n$  are respectively  
 524 the discretization size and the total number of distributions; while Le et al. [2021] gives a  $O(p^2)$   
 525 complexity algorithm for calculating the Wasserstein distance.

Table 3: Positions  $(x, y) \in \mathbb{R}^2$  of means for two-dimensional mixture of Gaussian distributions for the counter example in Section 4.1.

	$a_{1,1}$	$a_{1,2}$	$a_{2,1}$	$a_{2,2}$	$a_{3,1}$	$a_{3,2}$	$a_{4,1}$	$a_{4,2}$
$x$	0.75	0.25	0.75	0.25	0.9	0.9	1.3	1.3
$y$	1.15	0.85	0.85	1.15	0.85	1.15	0.75	1.25

Table 4: The time cost with standard deviation shown in parentheses for the counter example. TC: Time cost, W-SDP: Wasserstein SDP, D-WKM: Distance-based Wasserstein  $K$ -means, B-WKM: Barycenter-based Wasserstein  $K$ -means.

$n$	TC for W-SDP (SD)	TC for D-WKM	TC for B-WKM (SD)
101	14.50 (0.5873)	14.15 (0.5132)	181.1 (372.4)
202	56.94 (1.490)	54.98 (1.516)	341.0 (136.2)
303	128.4 (3.640)	123.9 (3.606)	549.2 (200.2)

Table 5: Estimated Wasserstein distances with standard deviation shown in parentheses and frequency of  $\Delta^* > \Delta_*$  for the counter example.

$n$	$\Delta_*$	$\Delta^*$	Frequency of $\Delta_* < \Delta^*$
101	0.1978 (0.0055)	0.2046 (0.0050)	0.8200
202	0.1990 (0.0058)	0.2050 (0.0051)	0.8200
303	0.1996 (0.0067)	0.2052 (0.0050)	0.7600

## 526 B Additional proofs

527 In this section, we will give detailed proofs for Lemma 4 and Theorem 8. For the proof of Theorem 8,  
528 we will first introduce the main part and put the rest proofs of corresponding lemmas at the end of  
529 this section to make it clear.

### 530 B.1 Proof of Lemma 4 in Section 2.1

531 Recall the settings as following

$$\begin{aligned} \mu_1 &= 0.5 \delta_{(x,y)} + 0.5 \delta_{(-x,-y)}, & \mu_2 &= 0.5 \delta_{(x,-y)} + 0.5 \delta_{(-x,y)}, \\ \mu_3 &= 0.5 \delta_{(x+\epsilon_1,y)} + 0.5 \delta_{(x+\epsilon_1,-y)}, & \text{and } \mu_4 &= 0.5 \delta_{(x+\epsilon_1+\epsilon_2,y)} + 0.5 \delta_{(x+\epsilon_1+\epsilon_2,-y)}, \end{aligned}$$

532 where  $\delta_{(x,y)}$  denotes the point mass measure at point  $(x, y) \in \mathbb{R}^2$ , and  $(x, y, \epsilon_1, \epsilon_2)$  are positive  
533 constants. **Lemma 4 (Configuration characterization).** If  $(x, y, \epsilon_1, \epsilon_2)$  satisfies

$$y^2 < \min\{x^2, 0.25 \Delta_{\epsilon_1, x}\} \quad \text{and} \quad \Delta_{\epsilon_1, x} < \epsilon_2^2 < \Delta_{\epsilon_1, x} + y^2,$$

534 where  $\Delta_{\epsilon_1, x} := \epsilon_1^2 + 2x^2 + 2x\epsilon_1$ , then for all sufficiently large  $m$  (number of copies of  $\mu_1$  and  $\mu_2$ ),

$$W_2(\mu_3, \mu_2^*) < W_2(\mu_3, \mu_1^*) \quad \text{and} \quad \underbrace{\max_{k=1,2} \max_{i,j \in G_k} W_2(\mu_i, \mu_j)}_{\text{largest within-cluster distance}} < \underbrace{\min_{i \in G_1, j \in G_2} W_2(\mu_i, \mu_j)}_{\text{least between-cluster distance}},$$

535 where  $\mu_k^*$  denotes the Wasserstein barycenter of cluster  $G_k$  for  $k = 1, 2$ .

536 *Proof.* For any  $w_i \in \mathbb{R}^2, i = 1, 2, 3, 4$ , let  $\mu = 0.5 \delta_{w_1} + 0.5 \delta_{w_2}$ ,  $\nu = 0.5 \delta_{w_3} + 0.5 \delta_{w_4}$ . By  
537 definition of Wasserstein distance we can show that

$$W_2^2(\mu, \nu) = 0.5 \min\{\|w_1 - w_3\|^2 + \|w_2 - w_4\|^2, \|w_1 - w_4\|^2 + \|w_2 - w_3\|^2\}$$

538 Let  $\mu_0 = 0.5 \delta_{(x,0)} + 0.5 \delta_{(-x,0)}$ , by algebraic calculation it is direct to check

$$W_2(\mu_3, \mu_2^*) < W_2(\mu_3, \mu_0) \quad \text{and} \quad \underbrace{\max_{k=1,2} \max_{i,j \in G_k} W_2(\mu_i, \mu_j)}_{\text{largest within-cluster distance}} < \underbrace{\min_{i \in G_1, j \in G_2} W_2(\mu_i, \mu_j)}_{\text{least between-cluster distance}},$$

539 once plugging in the assumptions. So we only need to show that  $\forall \varepsilon, \exists M$ , s.t. when  $m > M$  we  
540 have  $W_2^2(\mu_3, \mu_1^*) \geq W_2^2(\mu_3, \mu_0) - \varepsilon$ . For notation simplicity, let  $v_x = (x, 0), v_{-x} = (-x, 0), v_1 =$   
541  $(x, y), v_2 = (-x, -y), v_3 = (x, -y), v_4 = (x, -y)$ . By definition we know there exist measures  
542  $\xi_i, i = 1, 2, 3, 4$ , s.t.

$$W_2^2(\mu_1^*, \mu_1) = \int \|v - v_1\|^2 d\xi_1(v) + \int \|v - v_2\|^2 d\xi_2(v),$$

543

$$W_2^2(\mu_1^*, \mu_2) = \int \|v - v_3\|^2 d\xi_3(v) + \int \|v - v_4\|^2 d\xi_4(v),$$

544 where  $\mu_1^* = \xi_1 + \xi_2 = \xi_3 + \xi_4$  with  $\xi_i(\mathbb{R}^2) = 0.5, \forall i$ . Furthermore, if we define  $\xi_{i,j} = \xi_i \cdot \xi_j / \mu_1^*, i \in$   
545  $\{1, 2\}, j \in \{3, 4\}$ , then  $\xi_i = \xi_{i,3} + \xi_{i,4}, \xi_j = \xi_{1,j} + \xi_{2,j}, i \in \{1, 2\}, j \in \{3, 4\}$ . Thus

$$\begin{aligned} W_2^2(\mu_1^*, \mu_1) + W_2^2(\mu_1^*, \mu_2) &= \sum_{i=1}^4 \int \|v - v_i\|^2 d\xi_i(v) \\ &= \sum_{i \in \{1,2\}, j \in \{3,4\}} \int \|v - v_i\|^2 + \|v - v_j\|^2 d\xi_{i,j}(v). \end{aligned}$$

546 Now suppose  $t = \|v - v_x\|$ , by algebraic calculation we can get

$$\|v - v_1\|^2 + \|v - v_3\|^2 = t^2 + 2y^2.$$

547 Choose  $T > 0$  s.t.  $T^2 < \min\{2x^2 - 2y^2, y^2\}$ , then we have

$$\begin{aligned} W_2^2(\mu_1^*, \mu_1) + W_2^2(\mu_1^*, \mu_2) &= \sum_{i \in \{1,2\}, j \in \{3,4\}} \int \|v - v_i\|^2 + \|v - v_j\|^2 d\xi_{i,j}(v) \\ &\leq \int_{B_T(v_x)} \|v - v_1\|^2 + \|v - v_3\|^2 d\xi_{1,3}(v) + \int_{B_T(v_{-x})} \|v - v_2\|^2 + \|v - v_4\|^2 d\xi_{2,4}(v) \\ &\quad + (T^2 + 2y^2)(1 - \xi_{1,3}(B_T(v_x)) - \xi_{2,4}(B_T(v_{-x}))) \\ &= \int_{B_T(v_x)} t_1(v)^2 d\xi_{1,3}(v) + \int_{B_T(v_{-x})} t_2(v)^2 d\xi_{2,4}(v) + 2y^2 + T^2(1 - \xi_{1,3}(B_T(v_x)) - \xi_{2,4}(B_T(v_{-x}))), \end{aligned}$$

548 where  $B_t(v)$  stands for the ball with radius  $t$  centered at  $v$ ,  $t_1(v) := \|v - v_x\|$ ,  $t_2(v) := \|v - v_{-x}\|$ .

549 On the other hand, by definition we know that

$$\begin{aligned} m \cdot W_2^2(\mu_1^*, \mu_1) + m \cdot W_2^2(\mu_1^*, \mu_2) + W_2^2(\mu_1^*, \mu_3) \\ \leq m \cdot W_2^2(\mu_0, \mu_1) + m \cdot W_2^2(\mu_0, \mu_2) + W_2^2(\mu_0, \mu_3) \\ = m \cdot (2y^2) + C, \end{aligned}$$

550 where  $C := W_2^2(\mu_0, \mu_3)$ . So we have  $W_2^2(\mu_1^*, \mu_1) + W_2^2(\mu_1^*, \mu_2) \leq 2y^2 + C/m$ . i.e.,

$$\int_{B_T(v_x)} t_1(v)^2 d\xi_{1,3}(v) + \int_{B_T(v_{-x})} t_2(v)^2 d\xi_{2,4}(v) + T^2(1 - \xi_{1,3}(B_T(v_x)) - \xi_{2,4}(B_T(v_{-x}))) \leq \frac{C}{m}.$$

551 So we have

$$\begin{aligned} \int_{B_T(v_x)} t_1(v)^2 d\xi_{1,3}(v) &\leq \frac{C}{m}, \quad \int_{B_T(v_{-x})} t_2(v)^2 d\xi_{2,4}(v) \leq \frac{C}{m}, \\ 0.5 - \xi_{1,3}(B_T(v_x)) &\leq \frac{C}{T^2 m}, \quad 0.5 - \xi_{2,4}(B_T(v_{-x})) \leq \frac{C}{T^2 m}. \end{aligned}$$

553 Now suppose  $v_{\epsilon_1} := (x + \epsilon_1, y)$ ,  $v_{-\epsilon_1} := (x + \epsilon_1, -y)$ , note that  $T^2 < y^2 < \epsilon_1^2 + y^2$  and  
554  $W_2^2(\mu_3, \mu_0) = 0.5\|v_x - v_{\epsilon_1}\|^2 + 0.5\|v_{-x} - v_{-\epsilon_1}\|^2$ . By definition of Wasserstein distance and  
555 symmetry we have

$$\begin{aligned} W_2^2(\mu_3, \mu_1^*) &\geq \int_{B_T(v_x)} (\|v_x - v_{\epsilon_1}\| - t_1(v))^2 d\xi_{1,3}(v) + \int_{B_T(v_{-x})} (\|v_{-x} - v_{-\epsilon_1}\| - t_2(v))^2 d\xi_{2,4}(v) \\ &\geq \|v_x - v_{\epsilon_1}\|^2 \xi_{1,3}(B_T(v_x)) + \|v_{-x} - v_{-\epsilon_1}\|^2 \xi_{2,4}(B_T(v_{-x})) \\ &\quad - 2\|v_x - v_{\epsilon_1}\| \int_{B_T(v_x)} t_1(v) d\xi_{1,3}(v) - 2\|v_{-x} - v_{-\epsilon_1}\| \int_{B_T(v_{-x})} t_2(v) d\xi_{2,4}(v) \\ &\geq W_2^2(\mu_3, \mu_0) - C_1^2 \cdot \frac{C}{T^2 m} - C_2^2 \cdot \frac{C}{T^2 m} \\ &\quad - 2C_1 \int_{B_T(v_x)} t_1(v) d\xi_{2,4}(v) - 2C_2 \int_{B_T(v_{-x})} t_2(v) d\xi_{2,4}(v), \end{aligned}$$

556 where  $C_1 = \|v_x - v_{\epsilon_1}\|$ ,  $C_2 = \|v_{-x} - v_{-\epsilon_1}\|$ . Set  $\forall \varepsilon > 0$ . Finally, by Hölder's inequality we have

$$\begin{aligned} W_2^2(\mu_3, \mu_1^*) &\geq W_2^2(\mu_3, \mu_0) - C_1^2 \cdot \frac{C}{T^2 m} - C_2^2 \cdot \frac{C}{T^2 m} \\ &\quad - 2C_1 \sqrt{\int_{B_T(v_x)} t_1^2(v) d\xi_{2,4}(v)} - 2C_2 \sqrt{\int_{B_T(v_{-x})} t_2^2(v) d\xi_{2,4}(v)} \\ &\geq W_2^2(\mu_3, \mu_0) - C_1^2 \cdot \frac{C}{T^2 m} - C_2^2 \cdot \frac{C}{T^2 m} - 2C_1 \sqrt{\frac{C}{m}} - 2C_2 \sqrt{\frac{C}{m}} \\ &\geq W_2^2(\mu_3, \mu_0) - \varepsilon, \end{aligned}$$

557 for large  $m$ , as desired. ■

558 **B.2 Proof of Theorem 8 in Section 3**

559 *Theorem 8 (Exact recovery for clustering Gaussians).* Let  $\Delta^2 := \min_{k \neq l} d^2(V^{(k)}, V^{(l)})$  denote  
 560 the minimal pairwise separation among clusters,  $\bar{n} := \max_{k \in [K]} n_k$  (and  $\underline{n} := \min_{k \in [K]} n_k$ ) the  
 561 maximum (minimum) cluster size, and  $m := \min_{k \neq l} \frac{2n_k n_l}{n_k + n_l}$  the minimal pairwise harmonic mean  
 562 of cluster sizes. Suppose the covariance matrix  $V_i$  of Gaussian distribution  $\nu_i = N(0, V_i)$  is  
 563 independently drawn from model (18) for  $i = 1, 2, \dots, n$ . Let  $\beta \in (0, 1)$ . If the separation  $\Delta^2$   
 564 satisfies

$$\Delta^2 > \bar{\Delta}^2 := \frac{C_1 t^2}{\min\{(1-\beta)^2, \beta^2\}} \mathcal{V} p^2 \log n,$$

then the SDP (17) achieves exact recovery with probability at least  $1 - C_2 n^{-1}$ , provided that

$$\underline{n} \geq C_3 \log^2 n, \quad t \leq C_4 \sqrt{\log n} / [(p + \log \bar{n}) \mathcal{V}^{1/2} T_v^{1/2}], \quad n/m \leq C_5 \log n,$$

565 where  $\mathcal{V} = \max_k \|V^{(k)}\|_{\text{op}}$ ,  $T_v = \max_k \text{Tr}[(V^{(k)})^{-1}]$ , and  $C_i, i = 1, 2, 3, 4, 5$  are constants.

566 *Lemma 10 (Dual argument for SDP (Section B in Chen and Yang [2021])).* The sufficient condition  
 567 for  $Z^* = \sum_{k \in [K]} \frac{1}{n_k} 1_{G_k} 1_{G_k}^T$  to be the unique solution of the SDP problem is to find  $(\lambda, \alpha, B)$  s.t.

$$(C_1) \quad B \succeq 0 \quad (B_{G_k G_k} = 0, B_{G_k G_l} > 0, \forall k \neq l),$$

$$(C_2) \quad W_n := \lambda Id + \frac{1}{2}(1\alpha^T + \alpha 1^T) - A - B \succeq 0,$$

$$(C_3) \quad \text{Tr}(W_n Z^*) = 0,$$

$$(C_4) \quad \text{Tr}(B Z^*) = 0,$$

568 which implies that

$$\alpha_{G_k} = \frac{2}{n_k} A_{G_k G_k} 1_{n_k} - \frac{\lambda}{n_k} 1_{n_k} - \frac{1}{n_k^2} (1_{n_k}^T A_{G_k G_k} 1_{n_k}).$$

569

$$\begin{aligned} [B_{G_l G_k} 1_{n_k}]_j &= -\frac{n_l + n_k}{2n_l} \lambda + \frac{n_k}{2} \left[ \frac{1}{n_l^2} \sum_{s,r \in G_l} d^2(V_s, V_r) - \frac{1}{n_k^2} \sum_{s,r \in G_k} d^2(V_s, V_r) \right] \\ &\quad + n_k \left[ \frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r) \right], \end{aligned}$$

570 for  $k \neq l, j \in G_l$ .

571 *Remarks.* It can be justified that if we can find  $(\lambda, B)$  satisfying above equations, then  $(C_3), (C_4)$   
 572 will hold automatically. Details can be found in Section B in Chen and Yang [2021].

573 Now we will proof the main theorem by two steps. First we will provide a lower bound for  
 574  $[B_{G_l G_k} 1_{n_k}]_j$ . Similar to the argument from Chen and Yang [2021], we want to set  $\lambda$  properly such  
 575 that  $(C_1)$  can hold. In the next step we will try to verify that the choice of  $(\lambda, \alpha, B)$  and the conditions  
 576 on the signals could actually imply  $(C_2)$ . And since number of clusters  $K$  is treated as fixed for most  
 577 practical settings, we will not emphasize  $K = O(1)$ .

578 **B.2.1 Proof of main result.**

579 *Step 1 (Construct  $(\lambda, B)$ ).* Recall  $[B_{G_l G_k} 1_{n_k}]_j = -\frac{n_l + n_k}{2n_l} \lambda + n_k L$ , where  $L$  equals

$$\frac{1}{2} \left[ \frac{1}{n_l^2} \sum_{s,r \in G_l} d^2(V_s, V_r) - \frac{1}{n_k^2} \sum_{s,r \in G_k} d^2(V_s, V_r) \right] + \left[ \frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r) \right].$$

580 For  $L$  defined above, by Lemma 14, we have

$$L \geq d^2(V^{(l)}, V^{(k)}) - d(V^{(l)}, V^{(k)}) K_1 - K_2,$$

581 w.p. at least  $(1 - c/n^2)$ , where

$$\begin{aligned} K_1 &= C\sqrt{\log nt}\mathcal{V}^{1/2} + Ct^2(p + \log \bar{n})\mathcal{V}T v^{1/2}, \\ K_2 &= Ct^2 p^2 \log n\mathcal{V}, \end{aligned}$$

582 for some constant  $C, c$ . Now we chose  $\beta \in (0, 1)$  and let  $m := \min_{k \neq l} \frac{2n_k n_l}{n_k + n_l}$ . If we suppose

$$\Delta \geq Ctp\sqrt{\log n}\mathcal{V}^{1/2}/(1 - \beta),$$

583 for some constant  $C$ , then we have

$$(1 - \beta)d^2(V^{(l)}, V^{(k)}) - d(V^{(l)}, V^{(k)})K_1 - K_2 \geq 0, \forall k \neq l,$$

584 which implies that

$$L \geq \beta d^2(V^{(l)}, V^{(k)}).$$

585 Define for  $k \neq l$ ,

$$c_j^{(k,l)} := [B_{G_l G_k} \mathbf{1}_{n_k}]_j, \quad j \in G_l,$$

586

$$r_i^{(k,l)} := [1_{n_l}^T B_{G_l G_k}]_i, \quad i \in G_k,$$

587

$$t^{(k,l)} := 1_{n_l}^T B_{G_l G_k} \mathbf{1}_{n_k},$$

588

$$(B_{G_l G_k}^\#)_{ij} := r_i^{(k,l)} c_j^{(k,l)} / t^{(k,l)}.$$

589 And define  $(B_{G_l G_l}^\#)_{ij} := 0, \forall l$ . By setting  $\lambda = \frac{\beta}{4} m \Delta^2$ , further we have

$$c_j^{(k,l)} \geq \frac{\beta}{2} n_k d^2(V^{(l)}, V^{(k)}), r_i^{(k,l)} \geq \frac{\beta}{2} n_l d^2(V^{(l)}, V^{(k)}), t^{(k,l)} \geq \frac{\beta}{2} n_l n_k d^2(V^{(l)}, V^{(k)}),$$

590 which implies that  $(B_{G_l G_k}^\#)_{ij} > 0, \forall i \in G_k, j \in G_l$ . And  $[B_{G_l G_k} \mathbf{1}_{n_k}]_j = [B_{G_l G_k}^\# \mathbf{1}_{n_k}]_j$ , which  
 591 means we can construct  $B^\#$  based on  $[B_{G_l G_k} \mathbf{1}_{n_k}]_j$  with  $[B_{G_l G_k} \mathbf{1}_{n_k}]_j = [B_{G_l G_k}^\# \mathbf{1}_{n_k}]_j$ . So essentially,  
 592 they are the same in the sense that we only care about they quantity through  $[B_{G_l G_k} \mathbf{1}_{n_k}]_j$ . And  
 593 thus for notation simplicity, we will use the symbol  $B$  instead of  $B^\#$ .

**Step 2 (Verify the condition for  $W_n$  in  $(C_2)$ ).** Next we would like to find sufficient condition for  $(C_2)$ , i.e.,

$$v^T W_n v \geq 0, \forall v \in \Gamma_K := \text{span}\{\mathbf{1}_{G_k} : k \in [K]\}^\perp, \|v\| = 1.$$

594 Note that  $v^T W_n v = \lambda - v^T A v - v^T B v \geq \lambda - v^T B v$ . And by definition as well as simple calculation  
 595 we have

$$v^T B v = \sum_{k=1}^K \sum_{l \neq k} \frac{1}{t^{(k,l)}} \left( \sum_{i \in G_k} v_i r_i^{(k,l)} \right) \left( \sum_{j \in G_l} v_j c_j^{(k,l)} \right),$$

596

$$\sum_{j \in G_l} v_j c_j^{(k,l)} = n_k \sum_{j \in G_l} \left( \frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r) \right) v_j.$$

597 Further note that

$$\frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r) = d^2(V^{(l)}, V^{(k)}) + E_j^{(k,l)},$$

598 where

$$\begin{aligned} E_j^{(k,l)} &= \left[ \frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - d^2(V_j, V^{(k)}) \right] + \left[ d^2(V_j, V^{(k)}) - d^2(V^{(l)}, V^{(k)}) \right] \\ &\quad - \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r). \end{aligned}$$

599 Then by triangle inequality and throwing away the last term of  $E_j^{(k,l)}$ , we have

$$\sum_{j \in G_l} v_j c_j^{(k,l)} = n_k \sum_{j \in G_l} E_j^{(k,l)} v_j \leq n_k \sum_{j \in G_l} (E_{1,j}^{(k,l)} + E_{2,j}^{(k,l)}) |v_j|,$$

600 where

$$E_{1,j}^{(k,l)} = \frac{1}{n_k} \sum_{r \in G_k} d^2(V^{(k)}, V_r) + \left[ \frac{2}{n_k} \sum_{r \in G_k} d(V^{(k)}, V_r) d(V_j, V^{(k)}) \right],$$

601

$$E_{2,j}^{(k,l)} = d^2(V^{(l)}, V_j) + 2d(V^{(l)}, V_j) d(V^{(l)}, V^{(k)}).$$

602 If we set  $\tilde{E}_{h,j}^{(k,l)} = E_{h,j}^{(k,l)} / d(V^{(l)}, V^{(k)})$ ,  $h = 1, 2$ , then the inequality can be written as

$$\sum_{j \in G_l} v_j c_j^{(k,l)} \leq n_k d(V^{(l)}, V^{(k)}) \sum_{j \in G_l} (\tilde{E}_{1,j}^{(k,l)} + \tilde{E}_{2,j}^{(k,l)}) |v_j|.$$

603 By Lemma 15 we know

$$\sum_{j \in G_l} \tilde{E}_{1,j}^{(k,l)} |v_j| \leq C \mathcal{V}^{1/2} p t \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2},$$

604 w.p.  $\geq 1 - cn^{-2}$ . And by Lemma 16 we have

$$\sum_{j \in G_l} \tilde{E}_{2,j}^{(k,l)} |v_j| \leq C t \mathcal{V}^{1/2} p (\sqrt{n_l} + \log^2(n)) \left( \sum_{j \in G_l} v_j^2 \right)^{1/2},$$

605 w.p.  $\geq 1 - cn^{-1}$ . Now if we assume  $\min_k n_k \geq C \log^2 n$  and notice that  $t^{(k,l)} \geq$   
 606  $\frac{\beta}{2} n_l n_k d^2(V^{(l)}, V^{(k)})$ , then further we can get

$$\begin{aligned} v^T B v &\leq \sum_{k,l} \frac{n_k n_l}{t^{(k,l)}} \sqrt{n_l} \sqrt{n_k} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2} \left( \sum_{i \in G_k} v_i^2 \right)^{1/2} C t^2 \mathcal{V} p^2 \\ &\leq \frac{C t^2}{\beta} \left( \sum_l \sum_{j \in G_l} v_j^2 \right)^{1/2} \left( \sum_l n_l \right)^{1/2} \left( \sum_k \sum_{i \in G_k} v_i^2 \right)^{1/2} \left( \sum_k n_k \right)^{1/2} \mathcal{V} p^2 \\ &= \frac{C t^2}{\beta} p^2 n \mathcal{V}, \end{aligned}$$

607 where the second inequality comes from Cauchy-Schwartz inequality. So by assuming

$$\Delta^2 \geq \frac{C t^2}{\beta^2} \mathcal{V} \cdot p^2 n / m,$$

for some constant  $C$ , we have

$$v^T W_n v \geq \lambda - v^T B v \geq \frac{\beta}{4} m \Delta^2 - \frac{C t^2}{\beta} p^2 n \mathcal{V} > 0.$$

608 Or it is sufficient to assume

$$\Delta^2 \geq \frac{C t^2}{\beta^2} \mathcal{V} \cdot p^2 \log n,$$

609 if  $n/m = o(\log n)$ . To sum up, if we assume

$$\Delta^2 \geq \frac{C t^2}{\min\{1 - \beta, \beta\}^2} \mathcal{V} \cdot p^2 \log n,$$

610 then w.p.  $\geq 1 - C/n$ , we have  $(C_1) - (C_4)$  hold by the construction of  $(\lambda, B)$ . Finally by Lemma 10  
 611 we know the solution of SDP  $Z^*$  exists uniquely, which is

$$Z^* = \sum_{k \in [K]} \frac{1}{n_k} 1_{G_k} 1_{G_k}^T$$

612 as desired. ■

613 *Remarks.* In our theorem, we focus on the relation between minimum cluster distance  $\bar{\Delta}$  with number  
614 of distributions  $n$ , which should be tight enough in the sense that  $\bar{\Delta} \asymp \sqrt{\log n}$ . This is the same order  
615 for the cut-off of exact recovery of SDP for Euclidean case from Chen and Yang [2021].

On the other hand, one sufficient condition for  $V_i, i = 1, \dots, n$  to be psd is  $1 - t \max_i \|X_i\|_{op} > 0$ ,  
which will hold w.p.  $\geq 1 - c/n^2$  if  $t \leq C/[\sqrt{p} + \sqrt{\log n}]$  for some constant  $C, c$ . Recall from our  
assumption,

$$t \leq c\sqrt{\log n}/[(p + \log \bar{n})\mathcal{V}^{1/2}T_v^{1/2}] \leq c\sqrt{\log n}/(p + \log \bar{n}),$$

616 for some constant, since  $T_v = \max_k \text{Tr}((V^{(k)})^{-1}) \geq p/\min_k \|V^{(k)}\|_{op}$ . This indicates that our  
617 bound for  $t$  guarantees  $V_i$  to be psd w.p.  $\geq 1 - c/n^2$ . And our bound should be tight as  $n \asymp \bar{n}$ . One  
618 may apply triangle inequality directly to Lemma 14 to get the upper bound of  $t$  with less order in  $p$ ,  
619 which is of less concern in our theorem, where we put more emphasis on the order in  $n$ .

## 620 B.2.2 Proofs of lemmas.

621 Before proving Lemma 14, let us first look at the Taylor expansion for psd matrix.

622 *Lemma 11 (Taylor expansion for psd matrix* (Theorem 1.1 in Del Moral and Niclas [2017])). The  
623 square root function  $\varphi : Q \in \mathcal{S}_r^+ \mapsto Q^{1/2}$  is Fréchet differentiable at any order on  $\mathcal{S}_r^+$  with the first  
624 order derivative given for any  $(A, H) \in \mathcal{S}_r^+ \times \mathcal{S}_r$  by the formula

$$\nabla \varphi(A) \cdot H = \int_0^\infty e^{-t\varphi(A)} H e^{-t\varphi(A)} dt,$$

625 where  $\mathcal{S}_r^+, \mathcal{S}_r$  are the positive semi-definite matrix and symmetric matrix respectively. The higher  
626 order derivatives are defined inductively for any  $n \geq 2$  by

$$\nabla^n \varphi(A) \cdot H = -\nabla \varphi(A) \cdot \left[ \sum_{p+q=n-2 \& p, q \geq 0} \frac{n!}{(p+1)!(q+1)!} [\nabla^{p+1} \varphi(A) \cdot H][\nabla^{q+1} \varphi(A) \cdot H] \right].$$

627 Again from the same paper, we have the Taylor expansion for  $\varphi(A)$ :

$$\varphi(A + H) = \varphi(A) + \sum_{1 \leq k \leq n} \frac{1}{k!} \nabla^k \varphi(A) \cdot H + \bar{\nabla}^{n+1} \varphi[A, H],$$

628 with

$$\bar{\nabla}^{n+1} \varphi[A, H] := \frac{1}{n!} \int_0^1 (1 - \epsilon)^n \nabla^{n+1} \varphi(A + \epsilon H) \cdot H d\epsilon.$$

629 *Corollary 12 (Decomposition of Wasserstein distance for Gaussians).* If we choose  $n = 1$  in  
630 Lemma 11, we have for  $k \neq l, j \in G_i^*$ , and under the assumptions in the Theorem, the following  
631 expansion holds.

$$\begin{aligned} & d^2(V_j, V^{(k)}) - d^2(V_j, V^{(l)}) \\ &= d^2(V^{(l)}, V^{(k)}) + \left\langle \mathcal{A}(V^{(l)}, V^{(k)}), t(X_j V^{(l)} + V^{(l)} X_j) + t^2 X_j V^{(l)} X_j \right\rangle \\ & - d^2(V_j, V^{(l)}) - \Delta_0, \end{aligned}$$

632

$$\begin{aligned} & \frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - d^2(V_j, V^{(k)}) \\ &= \left\langle \mathcal{A}(V_j, V^{(k)}), \frac{1}{n_k} \sum_{r \in G_k} t(X_r V^{(k)} + V^{(k)} X_r) + t^2 X_r V^{(k)} X_r \right\rangle - \Delta_1, \end{aligned}$$

633 where  $\mathcal{A}(U, V) := Id - U^{1/2}(U^{1/2}VU^{1/2})^{-1/2}U^{1/2}$ , for  $U, V : \text{psd}$ . And  $\Delta_0 \leq 0, \Delta_1 \leq 0$ , which  
634 are extra terms (high order terms in Lemma 11).

635 *Proof.* By definition we know  $d^2(V, U) = W_2^2(\nu, \mu)$ , where  $\nu \sim N(0, V), \mu \sim N(0, U)$ . Thus

$$d^2(V, U) = \text{Tr}(V) + \text{Tr}(U) - 2\text{Tr}[\sqrt{V^{1/2}UV^{1/2}}].$$

636 So we have

$$\begin{aligned} & d^2(V_j, V^{(k)}) - d^2(V^{(k)}, V^{(l)}) \\ &= \text{Tr}[V_j - V^{(l)}] - 2\text{Tr} \left[ \sqrt{(V^{(k)})^{1/2} V_j (V^{(k)})^{1/2}} - \sqrt{(V^{(k)})^{1/2} V^{(l)} (V^{(k)})^{1/2}} \right]. \end{aligned}$$

637 On the other hand, by definition we know  $V_j = (I + tX_j)V^{(l)}(I + tX_j) = V^{(l)} + t(X_j V^{(l)} +$   
 638  $V^{(l)} X_j) + t^2 X_j V^{(l)} X_j$ . Then by Lemma 11 and note the second order remainder term is always  
 639 negative semi-definite, we can directly get the results by first order Taylor expansion. ■

**Lemma 13 (Norm for operator A).** We conclude that for any  $U, V$ : psd, we have

$$\|\mathcal{A}(U, V) \cdot V^{1/2}\|_F^2 = \|V^{1/2} \cdot \mathcal{A}(U, V)\|_F^2 = d^2(U, V).$$

640

641 *Proof.* Suppose we have the SVD

$$U^{1/2} V^{1/2} = Q_1^T \Sigma Q_2,$$

642 then we have

$$\begin{aligned} \mathcal{A}(U, V) \cdot V^{1/2} &= (I - U^{1/2} (U^{1/2} V U^{1/2})^{-1/2}) V^{1/2} \\ &= V^{1/2} - U^{1/2} Q_1^T Q_2, \end{aligned}$$

643 which implies that

$$\begin{aligned} \|\mathcal{A}(U, V) \cdot V^{1/2}\|_F^2 &= \text{Tr}(V) + \text{Tr}(U) - 2\text{Tr}(V^{1/2} U^{1/2} Q_1^T Q_2) \\ &= \text{Tr}(V) + \text{Tr}(U) - 2\text{Tr}(Q_2^T \Sigma Q_2) \\ &= \text{Tr}(V) + \text{Tr}(U) - 2\text{Tr}(\sqrt{U^{1/2} V U^{1/2}}). \end{aligned}$$

644

645 **Lemma 14 (Lower bound for L).** Recall that  $L$  equals

$$\frac{1}{2} \left[ \frac{1}{n_l^2} \sum_{s,r \in G_l} d^2(V_s, V_r) - \frac{1}{n_k^2} \sum_{s,r \in G_k} d^2(V_s, V_r) \right] + \left[ \frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r) \right],$$

646 we have

$$L \geq d^2(V^{(l)}, V^{(k)}) - d(V^{(l)}, V^{(k)}) K_1 - K_2,$$

647 w.p. at least  $(1 - c/n^2)$ , where

$$\begin{aligned} K_1 &= C\sqrt{\log nt} \mathcal{V}^{1/2} + Ct^2(p + \log \bar{n}) \mathcal{V} T v^{1/2}, \\ K_2 &= Ct^2 p^2 \log n \mathcal{V}, \end{aligned}$$

648 for some constant  $C, c$ .

649 *Proof.* First note that we can decompose the term into three terms:

$$\frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r) = U_1 - U_2 + U_3,$$

650 where

$$\begin{aligned} U_1 &:= \frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - d^2(V_j, V^{(k)}), \\ U_2 &:= \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r) - d^2(V_j, V^{(l)}) \\ U_3 &:= d^2(V_j, V^{(k)}) - d^2(V_j, V^{(l)}). \end{aligned}$$

651 If we further define  $U_0 := \frac{1}{2} \left[ \frac{1}{n_l^2} \sum_{s,r \in G_l} d^2(V_s, V_r) - \frac{1}{n_k^2} \sum_{s,r \in G_k} d^2(V_s, V_r) \right]$ , then we have

$$L = U_0 + U_1 - U_2 + U_3.$$

652 From Corollary 12 we know  $U_1$  and  $U_2$  can be lower bounded by throwing out the remainders  $\Delta_1, \Delta_2$ ,  
653 i.e.,

$$\begin{aligned} U_1 &= \frac{1}{n_k} \sum_{r \in G_k} d^2(V_j, V_r) - d^2(V_j, V^{(k)}) \\ &\geq \left\langle \mathcal{A}(V_j, V^{(k)}), \frac{1}{n_k} \sum_{r \in G_k} t(X_r V^{(k)} + V^{(k)} X_r) + t^2 X_r V^{(k)} X_r \right\rangle, \end{aligned}$$

$$\begin{aligned} U_3 &= d^2(V_j, V^{(k)}) - d^2(V_j, V^{(l)}) \\ &\geq d^2(V^{(l)}, V^{(k)}) + \left\langle \mathcal{A}(V^{(l)}, V^{(k)}), t(X_j V^{(l)} + V^{(l)} X_j) + t^2 X_j V^{(l)} X_j \right\rangle \\ &\quad - d^2(V_j, V^{(l)}). \end{aligned}$$

654 As for the  $U_0$  and  $U_3$ , we choose to use triangle inequality to get a rough bound, i.e., by noting  
655  $d(V_j, V_r) \leq d(V_j, V^{(l)}) + d(V_r, V^{(l)})$ , we have

$$\begin{aligned} U_2 &= \frac{1}{n_l} \sum_{r \in G_l} d^2(V_j, V_r) - d^2(V_j, V^{(l)}) \\ &\leq \frac{1}{n_l} \sum_{r \in G_l} d^2(V^{(l)}, V_r) + \frac{2}{n_l} d(V^{(l)}, V_r) \sum_{r \in G_l} d(V_j, V^{(l)}). \end{aligned}$$

656 And

$$\begin{aligned} U_0 &= \frac{1}{2} \left[ \frac{1}{n_l^2} \sum_{s,r \in G_l} d^2(V_s, V_r) - \frac{1}{n_k^2} \sum_{s,r \in G_k} d^2(V_s, V_r) \right] \\ &\geq -\frac{1}{2} \frac{1}{n_l^2} \sum_{s,r \in G_k} (d(V_s, V^{(k)}) + d(V_r, V^{(k)}))^2 \\ &\geq -\frac{2}{n_l} \sum_{r \in G_k} d^2(V^{(k)}, V_r). \end{aligned}$$

657 For the RHS of the inequality for  $U_1$ , it can be divided into two parts.

$$Z_1^1 := \left\langle \mathcal{A}(V_j, V^{(k)}), \frac{1}{n_k} \sum_{r \in G_k} t(X_r V^{(k)} + V^{(k)} X_r) \right\rangle$$

658 and

$$Z_2^1 := \left\langle \mathcal{A}(V_j, V^{(k)}), t^2 \frac{1}{n_k} \sum_{r \in G_k} X_r V^{(k)} X_r \right\rangle.$$

659 the first part is a Gaussian distribution whose variance can be bounded by  
660  $c_1 t^2 \|\mathcal{A}(V_j, V^{(k)}) V^{(k)}\|_F^2 / n_k$ , for some constant  $c_1$ . By Gaussian tail bound  $P(|N(0, 1)| >$   
661  $u) \leq e^{-u^2/2}, \forall u > 0$  and Lemma 13, we have

$$\begin{aligned} |Z_1^1| &\leq c_2 t \sqrt{\log n} \|V^{(k)}\|^{1/2} / \sqrt{n_k} \cdot d(V_j, V^{(k)}) \\ &\leq c_2 t \sqrt{\log n} \mathcal{V}^{1/2} \cdot d(V_j, V^{(k)}), \end{aligned}$$

662 w.p.  $\geq 1 - c_3/n^2$ , for some constant  $c_2, c_3$ . On the other hand,

$$\begin{aligned}
|Z_2^1| &= t^2 \left\langle \mathcal{A}(V_j, V^{(k)})(V^{(k)})^{1/2}, \frac{1}{n_k} \sum_{r \in G_k} X_r V^{(k)} X_r (V^{(k)})^{-1/2} \right\rangle \\
&\leq t^2 \left\| \frac{1}{n_k} \sum_{r \in G_k} X_r V^{(k)} X_r (V^{(k)})^{-1/2} \right\|_F \cdot d(V_j, V^{(k)}) \\
&\leq t^2 \frac{1}{n_k} \sum_{r \in G_k} \|X_r\|^2 \|V^{(k)}\| \left\| (V^{(k)})^{-1/2} \right\|_F \cdot d(V_j, V^{(k)}) \\
&\leq t^2 \max_{r \in G_k} \|X_r\|^2 \mathcal{V} T v^{1/2} \cdot d(V_j, V^{(k)}) \\
&\leq c_4 t^2 (p + \log n) \mathcal{V} T v^{1/2} \cdot d(V_j, V^{(k)}),
\end{aligned}$$

663 w.p.  $\geq 1 - c_5/n^2$ , for some constant  $c_4, c_5$ . The last inequality can be implied from union bound and  
664 Corollary 4.4.8 in Vershynin [2018]:

$$\|X_r\| \leq C(\sqrt{p} + u), \quad \text{w.p.} \geq 1 - 4e^{-u^2}.$$

665 Now by combining  $Z_1^1, Z_2^1$  we have

$$\begin{aligned}
U_1 &\geq Z_1^1 + Z_2^1 \\
&\geq - \left[ c_2 t \sqrt{\log n} \mathcal{V}^{1/2} + c_4 t^2 (p + \log n) \mathcal{V} T v^{1/2} \right] \cdot d(V_j, V^{(k)}),
\end{aligned}$$

666 w.p.  $\geq 1 - (c_3 + c_5)/n^2$ .

667

668 For  $U_0$ , we have

$$\begin{aligned}
U_0 &\geq -\frac{2}{n_k} \sum_{r \in G_k} d^2(V^{(k)}, V_r) \\
&= -\frac{2t^2}{n_k} \sum_{r \in G_k} \text{Tr}(X_r V^{(k)} X_r) \\
&\geq -2t^2 \mathcal{V} \frac{1}{n_k} \sum_{r \in G_k} \text{Tr}(X_r^2) \\
&\geq -c_6 t^2 \mathcal{V} p^2,
\end{aligned}$$

669 w.p.  $\geq 1 - c_7/n^2$  for some constant  $c_6, c_7$ . The equation is a direct result by definition of Wasserstein  
670 distance for Gaussians:

$$d^2(V^{(k)}, V_r) = \text{Tr}(V^{(k)}) + \text{Tr}(V_r) - 2\text{Tr}(\sqrt{(V^{(k)})^{1/2} V_r (V^{(k)})^{1/2}}).$$

671 Note here

$$\begin{aligned}
\sqrt{(V^{(k)})^{1/2} V_r (V^{(k)})^{1/2}} &= \sqrt{(V^{(k)})^{1/2} (I + tX_r) V^{(k)} (I + tX_r) (V^{(k)})^{1/2}} \\
&= (V^{(k)})^{1/2} (I + tX_r) (V^{(k)})^{1/2}.
\end{aligned}$$

672 The last inequality can be derived through Bernstein's inequality (Theorem 2.8.2) by noting that  
673  $\text{Tr}(X_r^2)$  is sub-exponential with mean  $\mathbb{E}(\text{Tr}(X_r^2)) = p^2$ . Similar to the argument for  $U_0, U_1$ , after we  
674 apply high-dimensional bound for sub-Gaussian or sub-exponential distributions we can get bound  
675 for  $U_2, U_3$ :

$$\begin{aligned}
U_2 &\leq \frac{1}{n_l} \sum_{r \in G_l} d^2(V^{(l)}, V_r) + \frac{2}{n_l} \sum_{r \in G_l} d(V_r, V^{(l)}) d(V^{(l)}, V_j) \\
&\leq c_8 t^2 \mathcal{V} p^2 \log n,
\end{aligned}$$

676 w.p.  $\geq 1 - c_9/n^2$ , for some constant  $c_8, c_9$ .

$$\begin{aligned} U_3 &\geq d^2(V^{(l)}, V^{(k)}) + \left\langle \mathcal{A}(V^{(l)}, V^{(k)}), t(X_j V^{(l)} + V^{(l)} X_j) + t^2 X_j V^{(l)} X_j \right\rangle \\ &\quad - d^2(V_j, V^{(l)}) \\ &\geq d^2(V^{(l)}, V^{(k)}) - \left[ c_2 t \sqrt{\log n} \mathcal{V}^{1/2} + c_4 t^2 (p + \log \bar{n}) \mathcal{V} T v^{1/2} \right] \cdot d(V^{(l)}, V^{(k)}) \\ &\quad - c_{10} t^2 (p + \log n) p \mathcal{V}, \end{aligned}$$

677 w.p.  $\geq 1 - c_{11}/n^2$ , for some constant  $c_{10}, c_{11}$ . Lastly, by noting  $d(V_j, V^{(k)}) \leq d(V^{(l)}, V^{(k)}) +$   
678  $d(V_j, V^{(l)})$  in  $U_1$ , and combine them together we have

$$\begin{aligned} L &= U_0 + U_1 - U_2 + U_3 \\ &\geq d^2(V^{(l)}, V^{(k)}) - d(V^{(l)}, V^{(k)}) K_1 - K_2, \end{aligned}$$

679 w.p. at least  $(1 - c/n^2)$ , where

$$\begin{aligned} K_1 &= C \sqrt{\log n} t \mathcal{V}^{1/2} + C t^2 (p + \log \bar{n}) \mathcal{V} T v^{1/2}, \\ K_2 &= C t^2 p^2 \log n \mathcal{V}, \end{aligned}$$

680 for some constant  $C, c$ . ■

681 **Lemma 15** ( $\tilde{E}_{1,j}^{(k,l)}$  upper bound). Suppose  $v \in \Gamma_K := \text{span}\{1_{G_k} : k \in [K]\}^\perp$ ,  $\|v\| = 1$ . Let

$$E_{1,j}^{(k,l)} = \frac{1}{n_k} \sum_{r \in G_k} d^2(V^{(k)}, V_r) + \left[ \frac{2}{n_k} \sum_{r \in G_k} d(V^{(k)}, V_r) d(V_j, V^{(k)}) \right],$$

682 and  $\tilde{E}_{1,j}^{(k,l)} = E_{1,j}^{(k,l)} / d(V^{(l)}, V^{(k)})$ . Then w.p.  $\geq 1 - n^{-2}$ , we have

$$\sum_{j \in G_l} \tilde{E}_{1,j}^{(k,l)} |v_j| \leq C \mathcal{V}^{1/2} p t \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2},$$

683 *Proof.* Note  $E(\text{Tr}(X_r^2)) = p^2$ ,  $E(\sqrt{\text{Tr}(X_r^2)}) \leq \sqrt{E(\text{Tr}(X_r^2))} = p$  by Jensen's inequality. From  
684 high-dimension bound for sub-exponential and sub-Gaussian (Hoeffding's inequality and Bernstein's  
685 inequality) we have that w.p.  $\geq 1 - c/n^2$ ,

$$\begin{aligned} \frac{1}{n_k} \sum_{r \in G_k} d^2(V^{(k)}, V_r) &= \frac{1}{n_k} \sum_{r \in G_k} \text{Tr}(X_r^2 V^{(k)}) \leq C \mathcal{V} p^2 t^2, \\ \frac{1}{n_k} \sum_{r \in G_k} d(V^{(k)}, V_r) &= \frac{1}{n_k} \sum_{r \in G_k} \sqrt{\text{Tr}(X_r^2 V^{(k)})} \leq C \mathcal{V}^{1/2} p t, \end{aligned}$$

687 for some constants  $C, c$ . Suppose that  $d(V^{(l)}, V^{(k)}) \geq C_0 t \mathcal{V}^{1/2} \sqrt{\log n} p$ , for some fixed constant  
688  $C_0$ . Then we have w.p.  $\geq 1 - c/n^2$

$$d(V_j, V^{(k)}) \leq d(V_j, V^{(l)}) + d(V^{(l)}, V^{(k)}) \leq C t p \sqrt{\log n} \mathcal{V}^{1/2} + d(V^{(l)}, V^{(k)}) \leq C d(V^{(l)}, V^{(k)}),$$

689 for some large constant  $C$ . So we have w.p.  $\geq 1 - c/n^2$

$$\sum_{j \in G_l} \tilde{E}_{1,j}^{(k,l)} |v_j| \leq C \mathcal{V}^{1/2} p t \sum_{j \in G_l} |v_j| \leq C \mathcal{V}^{1/2} p t \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2},$$

690 where  $\mathcal{V} = \max_k \|V^{(k)}\|$ , for some large constant  $C$ . ■

691 **Lemma 16** ( $\tilde{E}_{2,j}^{(k,l)}$  upper bound). Suppose  $v \in \Gamma_K := \text{span}\{1_{G_k} : k \in [K]\}^\perp$ ,  $\|v\| = 1$ . Let

$$E_{2,j}^{(k,l)} = d^2(V^{(l)}, V_j) + 2d(V^{(l)}, V_j) d(V^{(l)}, V^{(k)}).$$

692 and  $\tilde{E}_{2,j}^{(k,l)} = E_{2,j}^{(k,l)} / d(V^{(l)}, V^{(k)})$ . Then w.p.  $\geq 1 - n^{-1}$ , we have

$$\sum_{j \in G_l} \tilde{E}_{2,j}^{(k,l)} |v_j| \leq C t \mathcal{V}^{1/2} p (\sqrt{n_l} + \log^2(n)) \left( \sum_{j \in G_l} v_j^2 \right)^{1/2},$$

693 *Proof.* First we make the following claim:

694 *Claim 17.* Following the above setting, w.p.  $\geq 1 - cn^{-1}$ , we have

$$\sum_{j \in G_l} d^2(V^{(l)}, V_j) |v_j| \leq Ct^2 \mathcal{V} p^2 (\sqrt{n_l} + \log(n)^2) \left( \sum_{j \in G_l} v_j^2 \right)^{1/2}, \quad (20)$$

695

$$\sum_{j \in G_l} d(V^{(l)}, V_j) |v_j| \leq Ct \mathcal{V}^{1/2} p \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2}, \quad (21)$$

696 for some large constant  $C$ .

697 If the claim holds, by plugging in the lower bound for  $\Delta$  in the assumption, we have

$$\begin{aligned} \sum_{j \in G_l} \tilde{E}_{2,j}^{(k,l)} |v_j| &\leq \frac{Ct^2 \mathcal{V} p^2 (\sqrt{n_l} + \log(n)^2)}{C_0 t \mathcal{V}^{1/2} p \sqrt{\log n}} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2} + Ct \mathcal{V}^{1/2} p \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2} \\ &\leq Ct \mathcal{V}^{1/2} p (\sqrt{n_l} + \log^2(n)) \left( \sum_{j \in G_l} v_j^2 \right)^{1/2} \end{aligned}$$

698 *Proof of the claim.* First we look at (21):

$$\sum_{j \in G_l} d(V^{(l)}, V_j) |v_j| \leq t \mathcal{V}^{1/2} \sum_{j \in G_l} \sqrt{\text{Tr}(X_j^2)} |v_j|.$$

699 By Theorem 2.6.3 (General Hoeffding's inequality) in Vershynin [2018] we have w.p.  $\geq 1 - c/n^2$ ,

$$\sum_{j \in G_l} \sqrt{\text{Tr}(X_j^2)} |v_j| \leq p \sum_{j \in G_l} |v_j| + Cp \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2},$$

700 for some constant  $C$ . i.e., w.p.  $\geq 1 - c/n^2$ ,

$$\sum_{j \in G_l} d(V^{(l)}, V_j) |v_j| \leq Ct \mathcal{V}^{1/2} p \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2},$$

for some constant  $C$ . Next we will show (20). First note that  $d^2(V^{(l)}, V_j) \leq t^2 \mathcal{V} \text{Tr}(X_j^2)$ , let

$$G_1(v) = \left| \sum_{j \in G_l} [\text{Tr}(X_j^2) - \mathbb{E} \text{Tr}(X_j^2)] |v_j| \right|,$$

701 then

$$\sum_{j \in G_l} d^2(V^{(l)}, V_j) |v_j| \leq t^2 \mathcal{V} G_1(v) + t^2 \mathcal{V} p^2 \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2}.$$

702 W.O.L.G., we may assume  $v \in \mathbb{V} := \{v \in \Gamma_K : \|v\| = 1\}$ ,  $\|G_1\|_{\mathbb{V}} := \sup_{v \in \mathbb{V}} |G_1(v)|$ . Then by  
703 Theorem 4 in Adamczak [2008] we know

$$\mathbb{P}(\|G_1\|_{\mathbb{V}} \geq 2\mathbb{E}\|G_1\|_{\mathbb{V}} + s) \leq \exp\left(-\frac{s^2}{3\tau_1^2}\right) + 3 \exp\left(-\frac{s}{3\|M_1\|_{\psi_1}}\right),$$

704 where

$$\tau_1^2 = \sup_{v \in \mathbb{V}} \sum_{j \in G_l} v_j^2 \mathbb{E}[\text{Tr}(X_j^2) - \mathbb{E} \text{Tr}(X_j^2)]^2 \leq \mathbb{E}[\text{Tr}(X_j^2)]^2 \leq p^4,$$

705

$$M_1 = \max_{j \in G_l, v \in \mathbb{V}} |v_j [\text{Tr}(X_j^2) - \mathbb{E}\text{Tr}(X_j^2)]| \leq \max_{j \in G_l} |[\text{Tr}(X_j^2) - \mathbb{E}\text{Tr}(X_j^2)]|.$$

706 By maximal inequality (Lemma 2.2.2 in van der Vaart and Wellner [1996]) we have

$$\|M_1\|_{\psi_1} \leq C \log(n_l) \max_{j \in G_l} \|[\text{Tr}(X_j^2) - \mathbb{E}\text{Tr}(X_j^2)]\|_{\psi_1} \leq C \log(n_l) p^2.$$

707 So by choosing  $s = C \log^2(n) p^2$ , we have w.p.  $\geq 1 - c/n$ ,

$$G_1(v) \leq 2\mathbb{E}\|G_1\|_{\mathbb{V}} + C \log^2(n) p^2,$$

708 for some  $C, c$ . On the hand,

$$\begin{aligned} \mathbb{E}\|G_1\|_{\mathbb{V}} &= \mathbb{E} \left| \sum_{j \in G_l} [\text{Tr}(X_j^2) - \mathbb{E}\text{Tr}(X_j^2)] v_j \right| \\ &\leq \sum_{j \in G_l} \mathbb{E} |\text{Tr}(X_j^2) - \mathbb{E}\text{Tr}(X_j^2)| |v_j| \\ &\leq 2\mathbb{E} |\text{Tr}(X_1^2)| \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2} \\ &= 2p^2 \sqrt{n_l} \left( \sum_{j \in G_l} v_j^2 \right)^{1/2}. \end{aligned}$$

709 So w.p.  $\geq 1 - cn^{-1}$ , we have

$$\sum_{j \in G_l} d^2(V^{(l)}, V_j) |v_j| \leq Ct^2 \mathcal{V} p^2 (\sqrt{n_l} + \log(n)^2) \left( \sum_{j \in G_l} v_j^2 \right)^{1/2}.$$

710

■