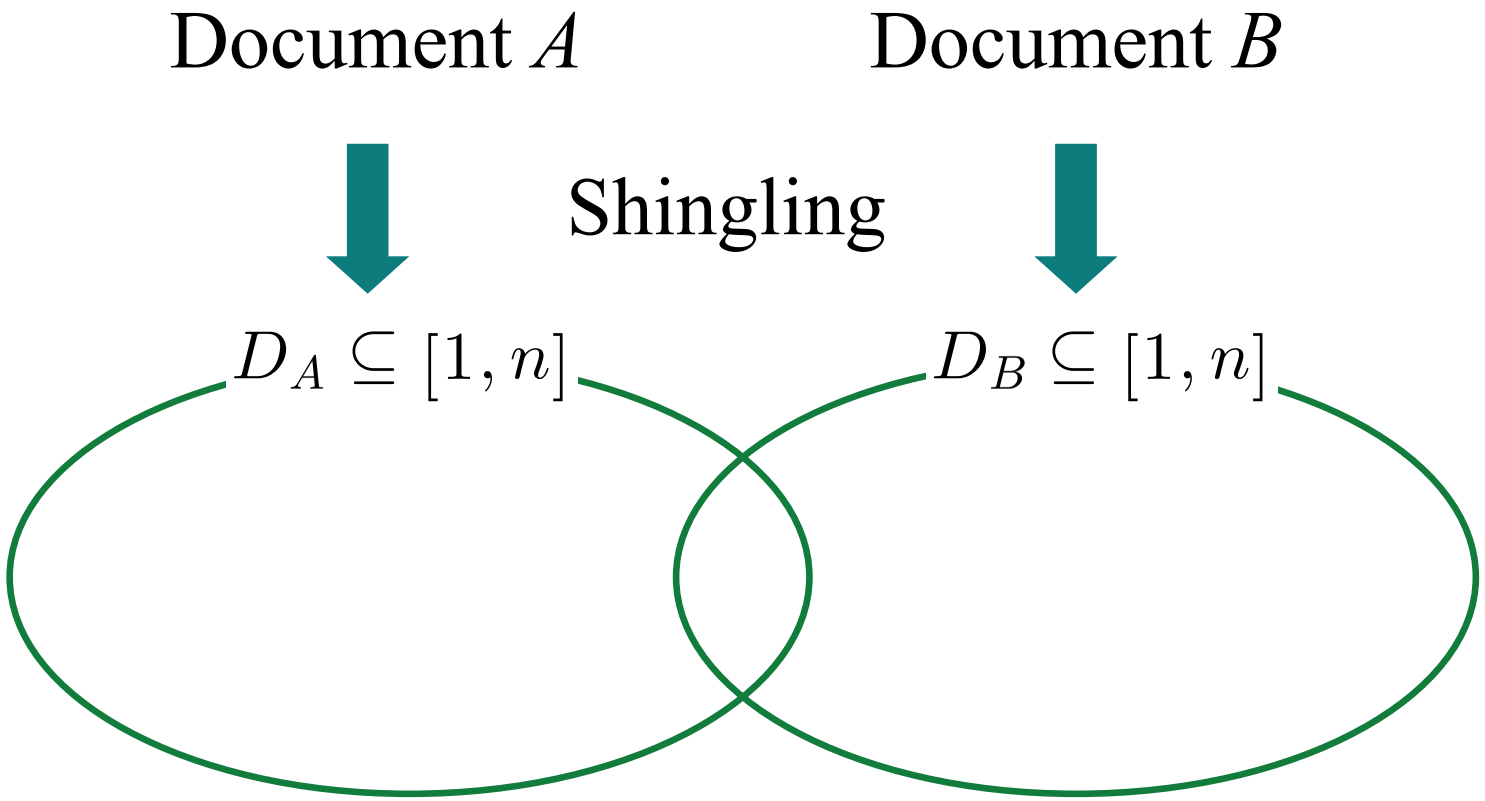
Improved Lower Bounds for Families of ε-Approximate k-Restricted Min-Wise Independent Permutations

Toshiya Itoh **Tatsuya Nagatani**

Tokyo Institute of Technology

Resemblance $r(A, B)$ of Documents A and B.



$$r(A, B) = \frac{\|D_A \cap D_B\|}{\|D_A \cup D_B\|}$$

Estimation of $r(A, B)$.

(1) Choose $\pi_1, \pi_2, \ldots, \pi_\ell \in S_n$ independently

(2) Define sketches S_A of A and S_B of B by

$$S_A = (\min\{\pi_1(D_A)\}, \min\{\pi_2(D_\mathrm{A})\}, \ldots, \min\{\pi_\ell(D_\mathrm{A})\}$$
$$= (s_{A,1}, s_{A,2}, \ldots, s_{A,\ell})$$
$$S_B = (\min\{\pi_1(D_B)\}, \min\{\pi_2(D_\mathrm{B})\}, \ldots, \min\{\pi_\ell(D_\mathrm{B})\}$$
$$= (s_{B,1}, s_{B,2}, \ldots, s_{B,\ell})$$

(3) Compute $\tilde{r}_\ell(A, B)$, estimation of $r(A, B)$, by

$$\tilde{r}_\ell(A, B) = \frac{\left\| \{ i \in [1, \ell] : \ s_{A,i} = s_{B,i} \} \right\|}{\ell}$$

$$\lim_{\ell \to \infty} \tilde{r}_\ell(A, B) = r(A, B)$$

$\mathcal{F} \subseteq S_n$ estimates $r(A, B) \Leftrightarrow \mathcal{F}$: k-restricted min-wise independent

Def. 1.1

$\mathcal{F} \subseteq S_n$: ε-Approximate k-Restricted Min-Wise Independent

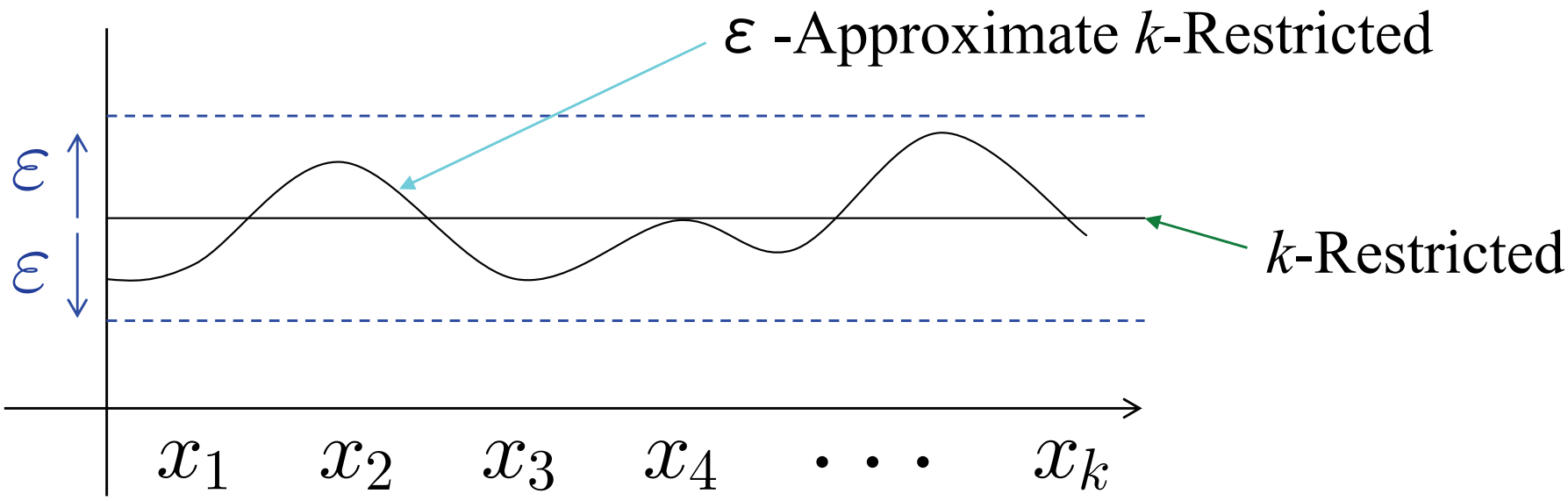
$$\forall X \subseteq [1, n] \quad \text{s. t.} \quad \|X\| \leq k \quad \forall x \in X$$

$$\left| \Pr_{\pi \in \mathcal{F}} [\min\{\pi(X)\} = \pi(x)] - \frac{1}{\|X\|} \right| \leq \frac{\varepsilon}{\|X\|},$$

where $\pi \in \mathcal{F}$ is chosen uniformly at random.

$\varepsilon = 0 \Rightarrow \mathcal{F}$: k-Restricted Min-Wise Independent

$k = n \Rightarrow \mathcal{F}$: ε -Approximate Min-Wise Independent



Known Results (1)

	Upper Bound	Lower Bound
Min-Wise	$\ell\mathrm{cm}(n, n-1, \ldots, 1)$	$\ell\mathrm{cm}(n, n-1, \ldots, 1)$
k-Restricted Min-Wise	$O\left(n^k e^k\right)$	$\Omega\left(\binom{n-1}{\lfloor (k-1)/2 \rfloor}\right)$

Known Results (2)

ε -Approximate *k*-Restricted

Upper Bounds	$2^{4k+O(k)}k^2 \log\log(n/\varepsilon)$	(constructive)
	$O\left(\frac{k^2}{\varepsilon^2}\log\left(\frac{n}{\varepsilon}\right)\right)$	(nonconstructive)
Lower Bounds	$\Omega\left(k^2\left(1-\sqrt{8\varepsilon}\right)\right)$	(uniform)
	$\Omega\left(\min\left\{k2^{k/2}\log(\frac{n}{\varepsilon}),\ \frac{\log(1/\varepsilon)(\log n-\log\log(1/\varepsilon))}{\varepsilon^{1/3}}\right\}\right)$	(biased)
	$\Omega\left(k\sqrt{\frac{1}{\varepsilon}\log\left(\frac{n}{k}\right)}\right)$	(uniform)

$\mathcal{F} = \{\pi_1, \pi_2, \ldots, \pi_d\} \subseteq S_n$: ε -Approx. k-Restricted Min-Wise

$$s = k/3, \quad L = n/s, \quad N = L - 1$$

$$[1, n] = \{1, 2, \ldots, n\}$$

$$= \{\underbrace{1, \ldots, h, \ldots, s}_{X_0}, \underbrace{s+1, \ldots, 2s}_{X_1}, \ldots, \underbrace{(L-1)s+1, \ldots, L_s}_{X_N}\}$$

$$\forall h \in [1, s] \quad U_h = \begin{matrix} \pi_1 & \cdots & \pi_j & \cdots & \pi_d \end{matrix} \begin{matrix} \\ \\ \\ \\ \end{matrix}$$

$$\begin{bmatrix} & \vdots & \\ & u_{ij}^h & \cdots \\ & \vdots & \end{bmatrix} \begin{matrix} X_0 \cup X_1 \\ \vdots \\ X_0 \cup X_i \\ \vdots \\ X_0 \cup X_N \end{matrix}$$

$$u_{ij}^h = \begin{cases} 1/\sqrt{d} & \min\{\pi_j(X_0 \cup X_i)\} = \pi_j(h) \\ 0 & \text{otherwise} \end{cases}$$

$$X_0 = \{1, 2, 3, 4\}, X_1 = \{5, 6, 7, 8\}, X_2 = \{9, 10, 11, 12\}, \ldots$$

$$U_2 = \begin{array}{c} \begin{matrix} \pi_1 & \cdots & \pi_j & \cdots & \pi_d \end{matrix} \\ \begin{bmatrix} & & u_{1j}^2 & & \\ & & u_{2j}^2 & & \\ & & \vdots & & \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix} \end{array} \begin{array}{l} X_0 \cup X_1 = \{1, 2, 3, 4, 5, 6, 7, 8\} \\ X_0 \cup X_1 = \{1, 2, 3, 4, 9, 10, 11, 12\} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{array}$$

$$u_{1j}^2 = \begin{cases} 1/\sqrt{d} & \min\{\pi_j(\{1, 2, 3, 4, 5, 6, 7, 8\})\} = \pi_j(2) \\ 0 & \text{otherwise} \end{cases}$$

$$u_{2j}^2 = \begin{cases} 1/\sqrt{d} & \min\{\pi_j(\{1, 2, 3, 4, 9, 10, 11, 12\})\} = \pi_j(2) \\ 0 & \text{otherwise} \end{cases}$$

$$\forall h \in [1, s] \qquad V_h = (v_{ij}^h) = U_h U_h^T = \begin{bmatrix} \frac{\delta_{11}^h}{2s} & \frac{\delta_{12}^h}{3s} & \frac{\delta_{13}^h}{3s} & \cdots & \frac{\delta_{1N}^h}{3s} \\ \frac{\delta_{12}^h}{3s} & \frac{\delta_{22}^h}{2s} & \frac{\delta_{23}^h}{3s} & \cdots & \frac{\delta_{2N}^h}{3s} \\ \frac{\delta_{13}^h}{3s} & \frac{\delta_{23}^h}{3s} & \frac{\delta_{33}^h}{2s} & \cdots & \frac{\delta_{3N}^h}{3s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\delta_{1N}^h}{3s} & \frac{\delta_{2N}^h}{3s} & \frac{\delta_{3N}^h}{3s} & \cdots & \frac{\delta_{NN}^h}{2s} \end{bmatrix}$$

(matrix of size $N \times N$)

Proposition 2.1

(i) $\forall i \in [1, N]$ $\qquad \frac{1-\varepsilon}{2s} \leq v_{ii}^h \leq \frac{1+\varepsilon}{2s}$

(ii) $\forall i, j \in [1, N](i \neq j)$ $\qquad \frac{1-\varepsilon}{3s} \leq v_{ij}^h \leq \frac{1+\varepsilon}{3s}$

Proof of Proposition 2.1

$$X_0 = \{1,2,3,4\}, X_1 = \{5,6,7,8\}, X_2 = \{9,10,11,12\}, \ldots$$

$$V_2 = (v_{ij}^2) = U_2 U_2^T$$

$$v_{11}^2 = (u_{11}^2, u_{12}^2, \ldots, u_{1d}^2) \cdot (u_{11}^2, u_{12}^2, \ldots, u_{1d}^2)^T$$

$$= \Pr_{\pi \in \mathcal{F}}[\min\{\pi(\{\underbrace{1,2,3,4}_{X_0}, \underbrace{5,6,7,8}_{X_1}\})\} = \pi(2)] = \frac{1 \pm \varepsilon}{8}$$

$$v_{12}^2 = (u_{11}^2, u_{12}^2, \ldots, u_{1d}^2) \cdot (u_{21}^2, u_{22}^2, \ldots, u_{2d}^2)^T$$

$$= \Pr_{\pi \in \mathcal{F}}[\min\{\pi(\{\underbrace{1,2,3,4}_{X_0}, \underbrace{5,6,7,8}_{X_1}, \underbrace{9,10,11,12}_{X_2}\})\} = \pi(2)] = \frac{1 \pm \varepsilon}{12}$$

$$V = \overbrace{\begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_s \end{bmatrix}}^{\|\mathcal{F}\|} \left[U_1^T, U_2^T, \ldots, U_s^T\right] = \begin{bmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_s \end{bmatrix}$$

Proposition 2.2

$\mathcal{F} \subseteq S_n$: ε-Approximate k-Restricted Min-Wise

$$\begin{aligned} \|\mathcal{F}\| &\geq \operatorname{rank}(V) \\ &= \operatorname{rank}(V_1) + \operatorname{rank}(V_2) + \cdots + \operatorname{rank}(V_s). \end{aligned}$$

$$A = \begin{bmatrix} a_{11} & a & a & \cdots & a \\ a & a_{22} & a & \cdots & a \\ a & a & a_{33} & \cdots & a \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & a & a & \cdots & a_{tt} \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \overset{t}{\longleftrightarrow} \quad \Big\updownarrow t$$

Proposition 2.3

(C1) $\forall i, j \in [1, t] (i \neq j) \quad a_{ij} = a > 0$

(C2) $\min\{a_{11}, a_{22}, \ldots, a_{tt}\} > a$

$$\Rightarrow A : \text{nonsingular} \equiv \text{rank}(A) = t$$

Proof of Proposition 2.3

$$A = \begin{bmatrix} a_{11} & a & \cdots & a \\ a & a_{12} & \cdots & a \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & a_{tt} \end{bmatrix} = \begin{bmatrix} a & a & \cdots & a \\ a & a & \cdots & a \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & a \end{bmatrix} + \begin{bmatrix} a_{11}-a & 0 & \cdots & 0 \\ 0 & a_{22}-a & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{tt}-a \end{bmatrix}$$

$$= \underbrace{a(1,1,\ldots,1)\cdot(1,1,\ldots,1)^T}_{\text{positive semidefinite}} + \underbrace{\mathrm{diag}(a_{11}-a, a_{22}-a, \ldots, a_{tt}-a)}_{\text{positive definite}}$$

(C2) $\min\{a_{11}, a_{22}, \ldots, a_{tt}\} > a$

$$\Rightarrow a_{11}-a>0, a_{22}-a>0, \ldots, a_{tt}-a \quad >0$$

A : nonsingular $\equiv \mathrm{rank}(A) = t$

Generalized Ramsey Number

$\mathcal{C}_m = \{c_1, c_2, \ldots, c_m\}$ a set of m colors

$\chi : E \to \mathcal{C}_m$ edge coloring of $K_\ell = (V, E)$

$R(t_1, t_2, \ldots, t_m)$

$$\min \ell \text{ s.t. } \exists i \subseteq [1, m] \quad \exists K_{t_i} = (V_i, E_i) \subseteq K_\ell \quad \forall e \in E_i \quad \chi(e) = c_i$$

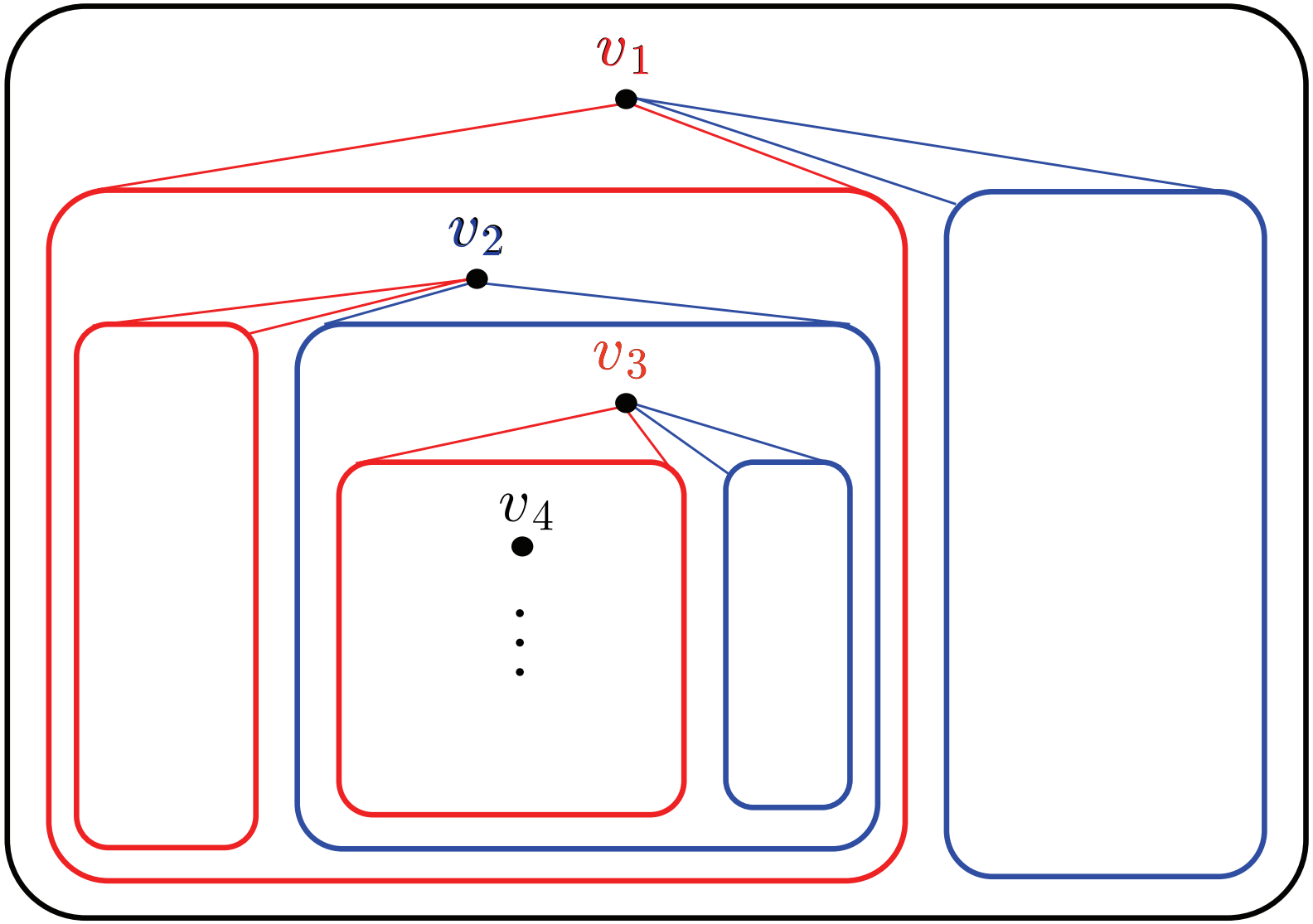
$$t_1 = t_2 = \cdots = t_m = t \qquad R_m(t) \equiv R(t, t, \ldots, t)$$

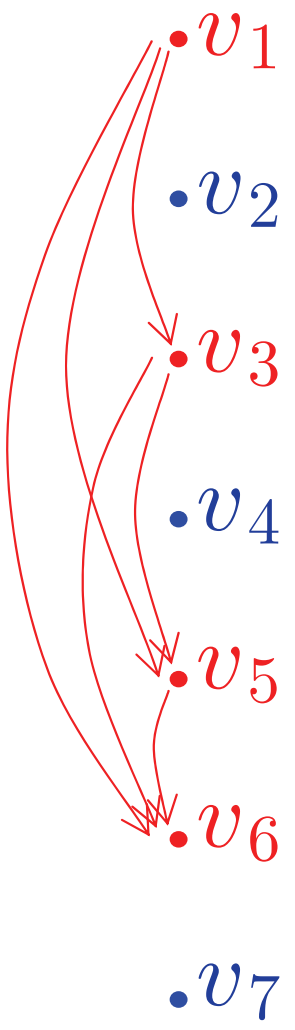
Lemma 3.1

$$\forall m \geq 2 \ \forall t \geq 1 \qquad R_m(t) \leq m^{mt-(m-1)}$$

Proof of Lemma 3.1

$$m = 2, \quad t = 4 \qquad n = 2^{2 \cdot 4 - (2-1)} = 2^7$$





Analysis for $\mathrm{rank}(V_h)$

$$\|\mathcal{F}\| \geq \mathrm{rank}(V_1) + \mathrm{rank}(V_2) + \cdots + \mathrm{rank}(V_s)$$

Lemma 3.2

$\forall 0 < \varepsilon < \frac{1}{5} \quad \forall k \geq 3$

$\mathcal{F} \subseteq S_n$: ε -Approximate k-Restricted Min-Wise

$$\|\mathcal{F}\| < \frac{k}{2\varepsilon} m \quad (m \geq 1)$$

$$\Rightarrow \forall h \in [1, s] \quad \mathrm{rank}(V_h) \begin{cases} = N & m = 1 \\ \geq \left\lfloor \dfrac{\log(3n/k)}{m \log m} \right\rfloor & m \geq 2 \end{cases}$$

Proof of Lemma 3.2

$\forall h \in [1, s] \quad \forall i, j \in [1, N]\ (i \neq j)$

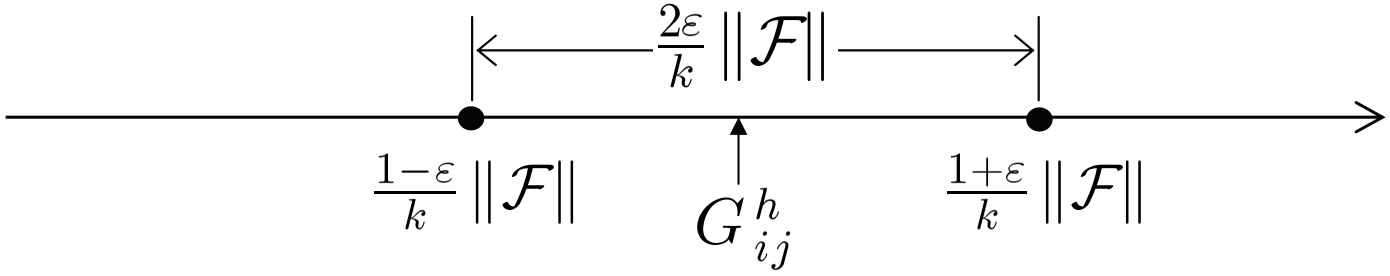
$$\mathcal{G}_{ij}^{h} = \{\pi \in \mathcal{F} : \ \min\{\pi(X_0 \cup X_i \cup X_j)\} = \pi(h)\}$$

$$G_{ij}^{h} = \left\| \mathcal{G}_{ij}^{h} \right\| : \ \text{integer}$$

$\mathcal{F} \subseteq S_n$: ε-Approximate k-Restricted Min-Wise

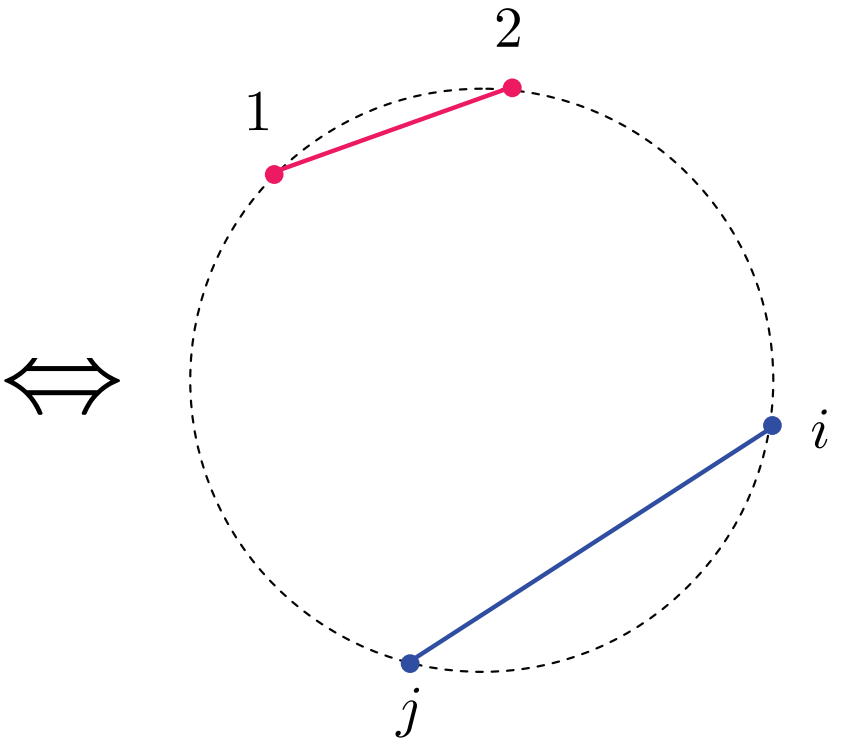
$$\Rightarrow \frac{1-\varepsilon}{3s} \leq \frac{G_{ij}^{h}}{\|\mathcal{F}\|} \leq \frac{1+\varepsilon}{3s}$$

$$\Rightarrow \frac{1-\varepsilon}{k} \|\mathcal{F}\| \leq G_{ij}^{h} \leq \frac{1+\varepsilon}{k} \|\mathcal{F}\|$$



- $\frac{2\varepsilon}{k}\|\mathcal{F}\| < m \Rightarrow G_{ij}^{h}$ takes at most m (integer) values
- m (integer) values $\equiv$ $\mathcal{C}_m = \{c_1, c_2, \ldots, c_m\}$: a set of m colors
- V_h is symmetric

$$V_h = \begin{array}{c} \begin{array}{cccccccc} & 1 & 2 & \cdots & i & & j & \cdots & N \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ i \\ \\ j \\ \vdots \\ N \end{array} \left[\begin{array}{cccccccc} & v_{12} & & & & & & \\ v_{12} & & & & & & & \\ & & & & & & & \\ & & & & & v_{ij} & & \\ & & & & & & & \\ & & & & v_{ij} & & & \\ & & & & & & & \\ & & & & & & & \end{array}\right] \end{array}$$



Case 1 : $m = 1$

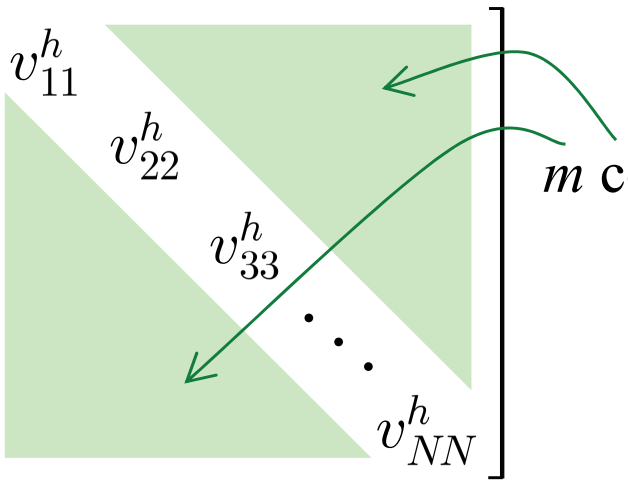
$$V_h = (v_{ij}^h) = \begin{bmatrix} v_{11}^h & v & v & \cdots & v \\ v & v_{22}^h & v & \cdots & v \\ v & v & v_{33}^h & \cdots & v \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v & v & v & \cdots & v_{NN}^h \end{bmatrix} \begin{matrix} \leftarrow N \rightarrow \\ \\ \\ \\ \\ \updownarrow N \end{matrix}$$

- $\forall\, i, j \in [1, N] (i \neq j) \quad v_{ij}^h = v > 0$
- $0 < \varepsilon < \frac{1}{5} \Rightarrow \min\{v_{11}^h, v_{22}^h, \ldots, v_{NN}^h\} > v$
- Proposition 2.3 $\Rightarrow V_h$: nonsingular

$$\Rightarrow \operatorname{rank}(V_h) = N$$

Case 2 : $m \geq 2$

$$V_h = (v_{ij}^h) = \begin{bmatrix} v_{11}^h & & & \\ & v_{22}^h & & \\ & & v_{33}^h & \\ & & & \ddots \\ & & & & v_{NN}^h \end{bmatrix}$$



$$\frac{3n}{k} - 1 = N \geq m^{mt-(m-1)} \geq R_m(t)$$

Lemma 3.1

$$\forall \chi : E \to \mathcal{C}_m \text{ of } K_N = (V, E)$$

$$\exists i \in [1, m] \quad \exists K_t = (V_t, E_t) \quad \text{s.t.} \quad \forall e \in E_t \ \chi(e) = c_i$$

$0 < \varepsilon < \frac{1}{5} +$ Proposition 2.3 $\Rightarrow \operatorname{rank}(V_h) \geq \left\lfloor \frac{\log(3n/k)}{m \log m} \right\rfloor$

Main Result

Theorem 4.1

$\forall 0 < \varepsilon < \frac{1}{5}\ \forall k \geq 3 \quad \mathcal{F} \subseteq S_n$: ε-Approx. k-Restricted Min-Wise

$$\|\mathcal{F}\| = \Omega\left(k\sqrt{\tfrac{1}{\varepsilon}\log(n/k)}\right).$$

Proof of Theorem 4.1

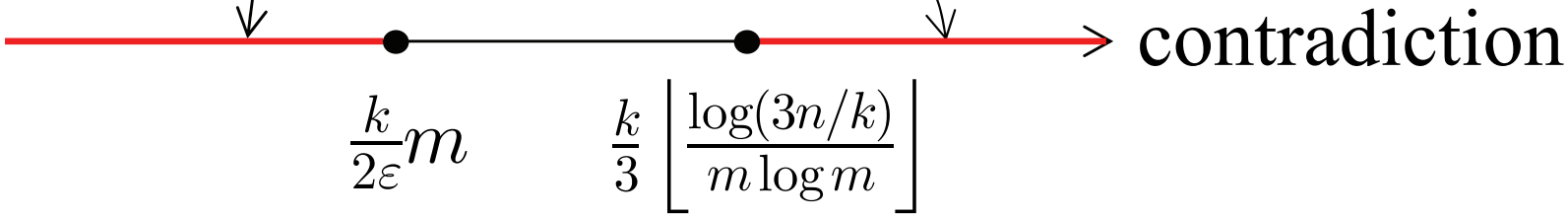
$$\forall m \geq 1 \quad \tfrac{2\varepsilon}{k}\|\mathcal{F}\| < m \ (\Leftrightarrow\ \|\mathcal{F}\| < \tfrac{k}{2\varepsilon}m)$$

Proposition 2.2 + Lemma 3.2

$$\|\mathcal{F}\| \geq \operatorname{rank}(V_1) + \operatorname{rank}(V_2) + \cdots + \operatorname{rank}(V_s)$$

$$\geq s\left\lfloor\frac{\log(3n/k)}{m\log m}\right\rfloor = \frac{k}{3}\left\lfloor\frac{\log(3n/k)}{m\log m}\right\rfloor$$

$$\forall m \geq 1 \quad \left[\|\mathcal{F}\| < \frac{k}{2\varepsilon} m \Rightarrow \|\mathcal{F}\| \geq \frac{k}{3} \left\lfloor \frac{\log(3n/k)}{m \log m} \right\rfloor \right]$$



$$\forall m \geq 1 \quad \left[\frac{k}{2\varepsilon} m \leq \frac{k}{3} \left\lfloor \frac{\log(3n/k)}{m \log m} \right\rfloor \Rightarrow \|\mathcal{F}\| \geq \frac{k}{2\varepsilon} m \right]$$

$$\|\mathcal{F}\| = \Omega\left(k \sqrt{\frac{1}{\varepsilon} \log(n/k)} \right)$$

Discussions

(LB1) $\|\mathcal{F}\| = \Omega(k^2(1-\sqrt{8\varepsilon}))$

(LB2) $\|\mathcal{F}\| = \Omega\left(\min\left\{ k2^{k/2}\log\left(\frac{n}{\varepsilon}\right), \frac{\log(1/\varepsilon)(\log n - \log\log(1/\varepsilon))}{\varepsilon^{1/3}} \right\}\right)$

(Ours) $\|\mathcal{F}\| = \Omega\left(k\sqrt{\frac{1}{\varepsilon}\log\left(n/k\right)}\right)$

(1) $n \gg k \Rightarrow$ (Ours) $>$ (LB1)

(2) $k < \frac{2}{3}\log(1/\varepsilon) \Rightarrow$ (Ours) $> k2^{k/2}\log(\frac{n}{\varepsilon})$

(3) $k \geq \frac{2}{3}\log(1/\varepsilon) \Rightarrow$ (Ours) $> \frac{\log(1/\varepsilon)(\log n - \log\log(1/\varepsilon))}{\varepsilon^{1/3}}$

$\Rightarrow$ (Ours) $>$ (LB2)


Thank you