APPROXIMATE INFERENCE
FOR
DETERMINANTAL POINT PROCESSES

Jennifer Gillenwater
Advised by Ben Taskar and Emily Fox
Joint work with Alex Kulesza

OUTLINE

OUTLINE

Motivation & Background

OUTLINE

Motivation & Background

1. **Dimensionality Reduction**

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

2. **MAP Estimation**

GILLENWATER, KULESZA, AND TASKAR (NIPS 2012)

3. **Likelihood Maximization**

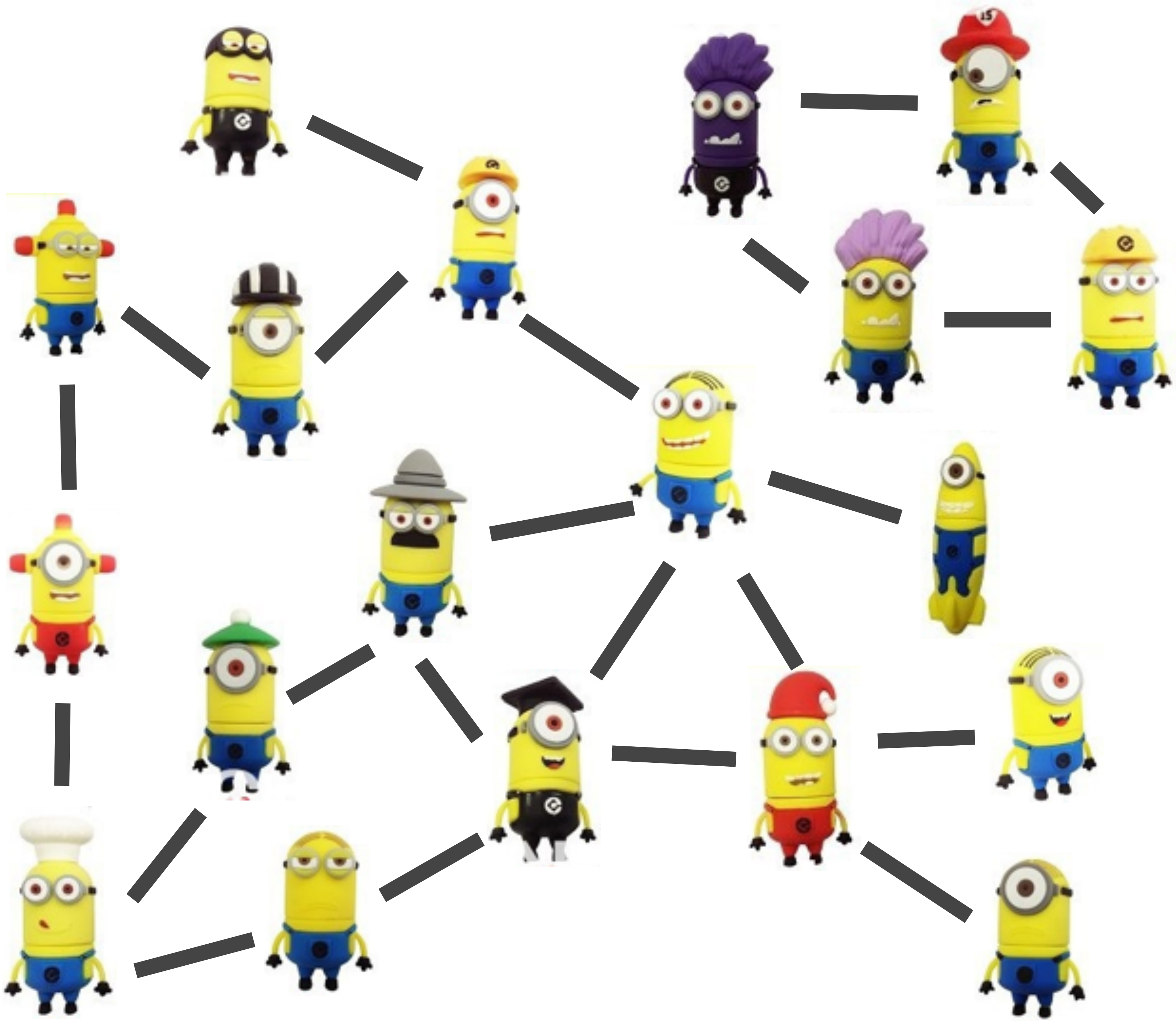
GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

MOTIVATION & BACKGROUND

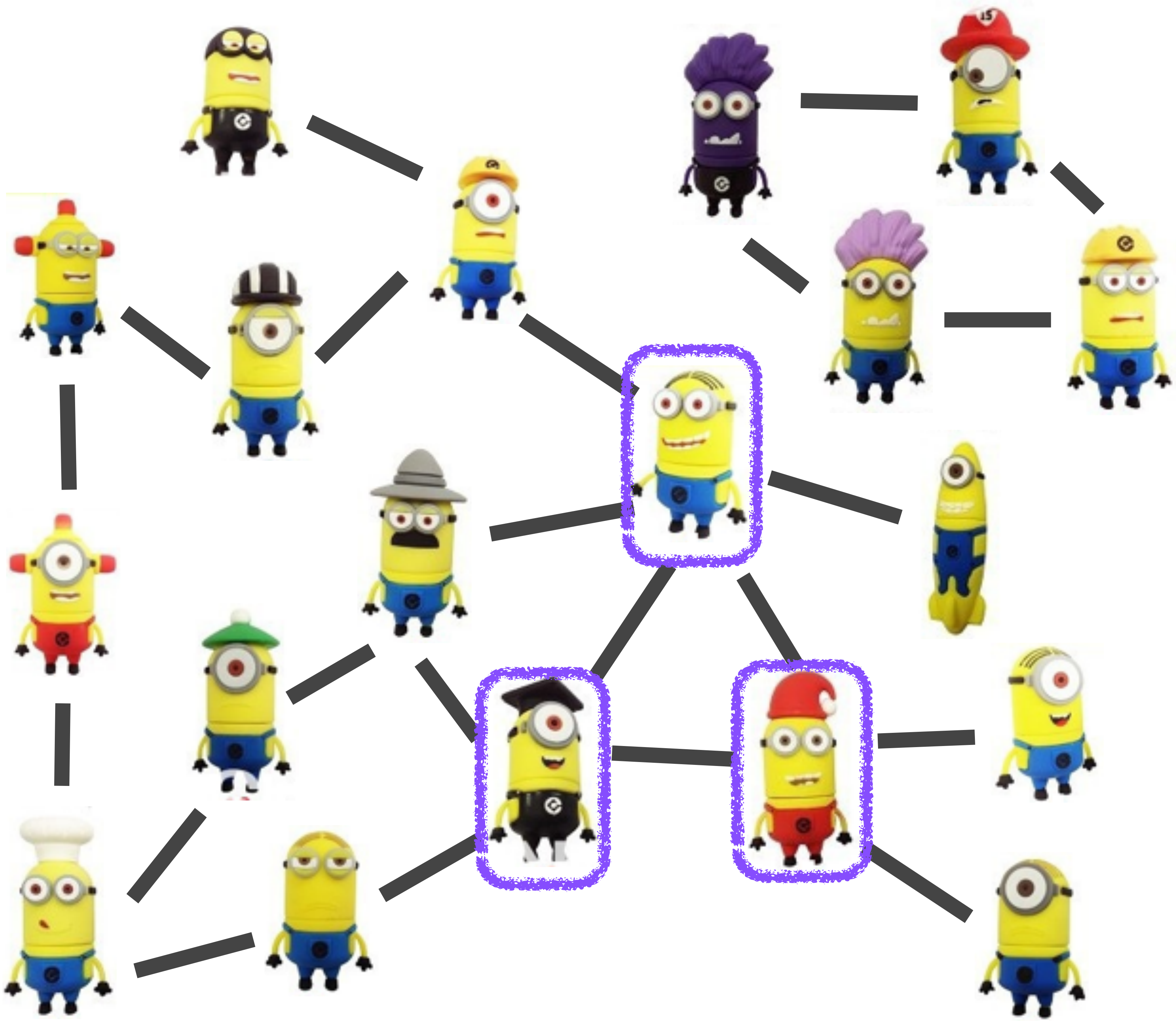
SOCIAL NETWORK MARKETING

SOCIAL NETWORK MARKETING

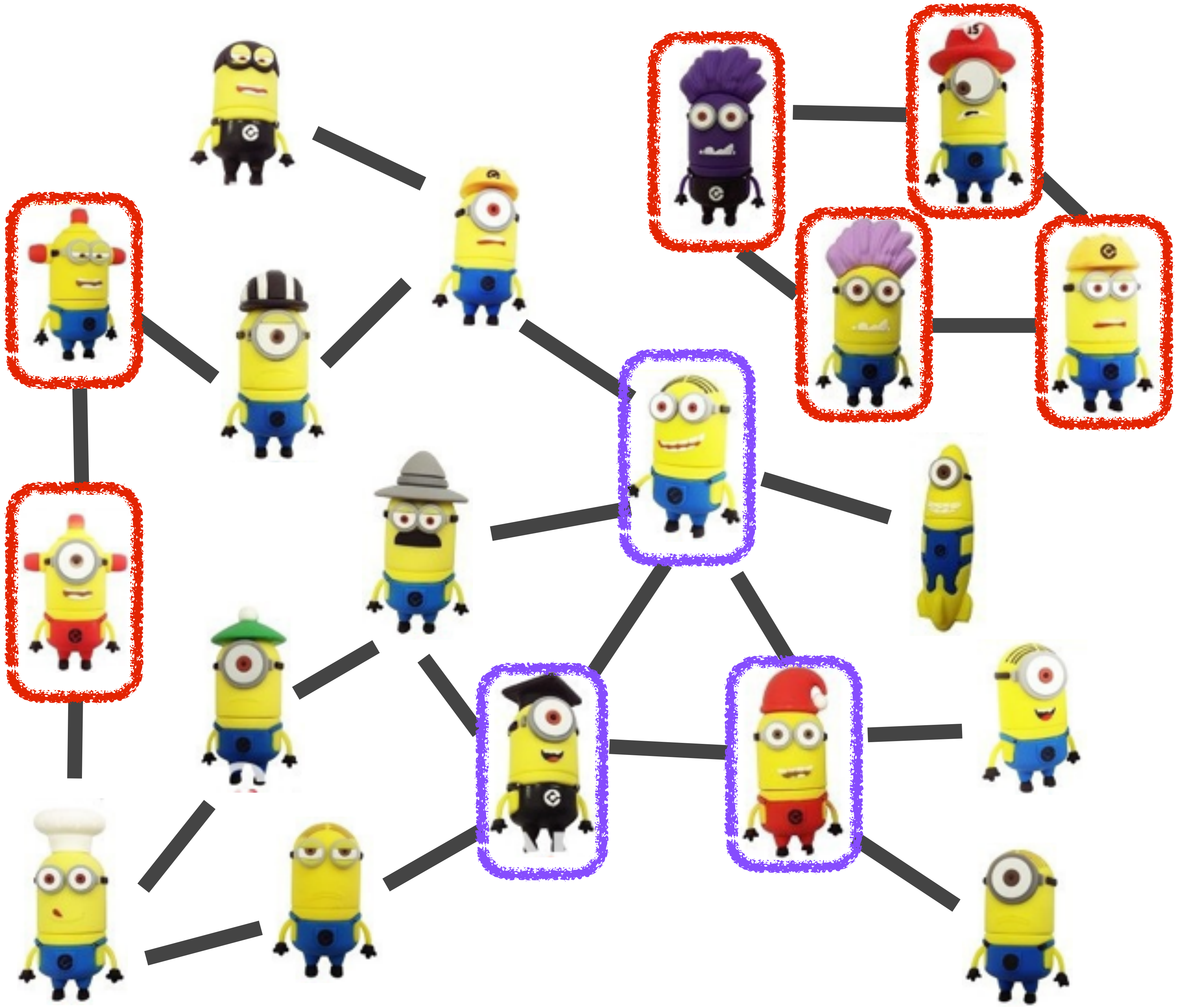
SOCIAL NETWORK MARKETING



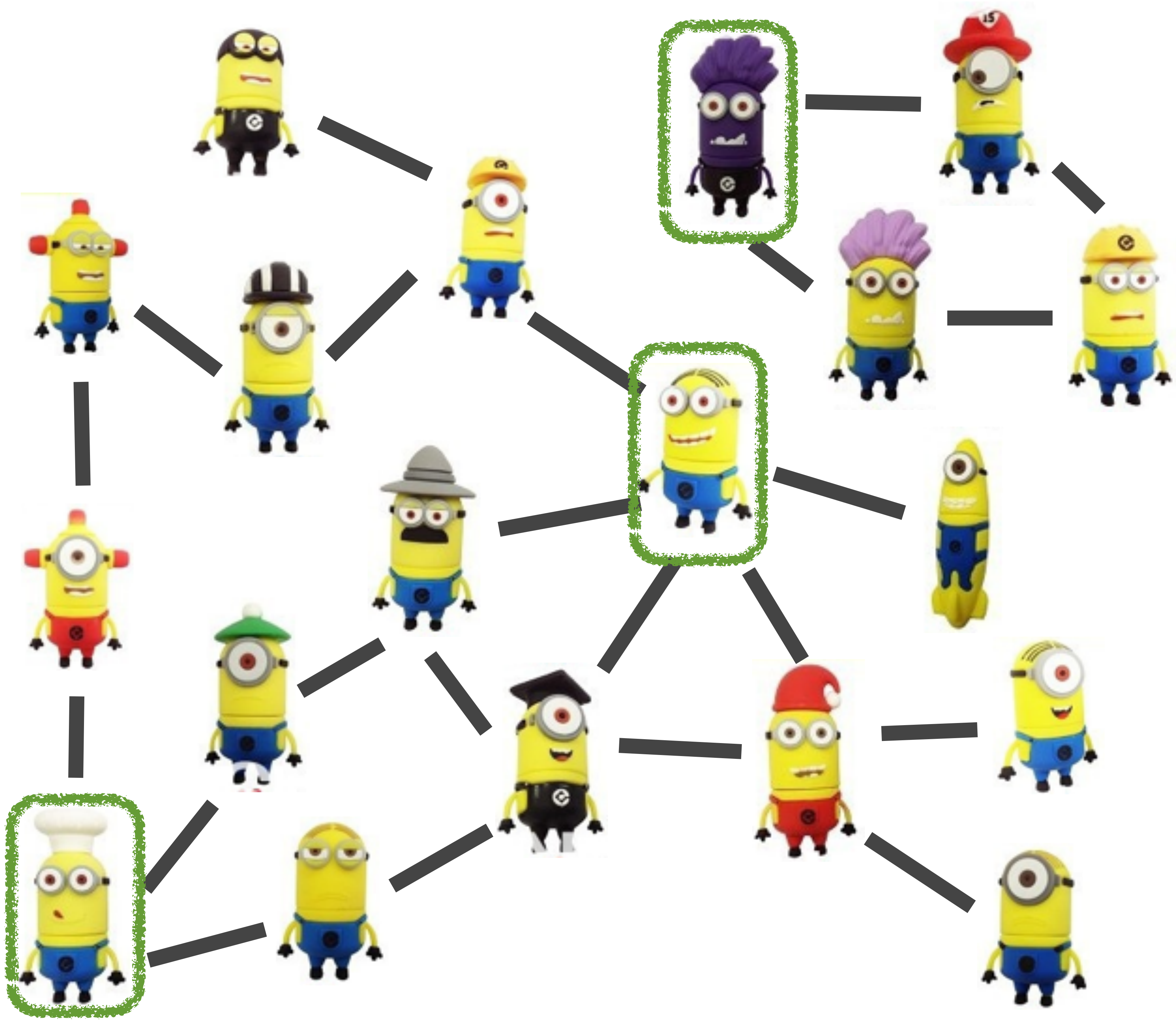
SOCIAL NETWORK MARKETING



SOCIAL NETWORK MARKETING



SOCIAL NETWORK MARKETING



VIDEO TRACKING

VIDEO TRACKING

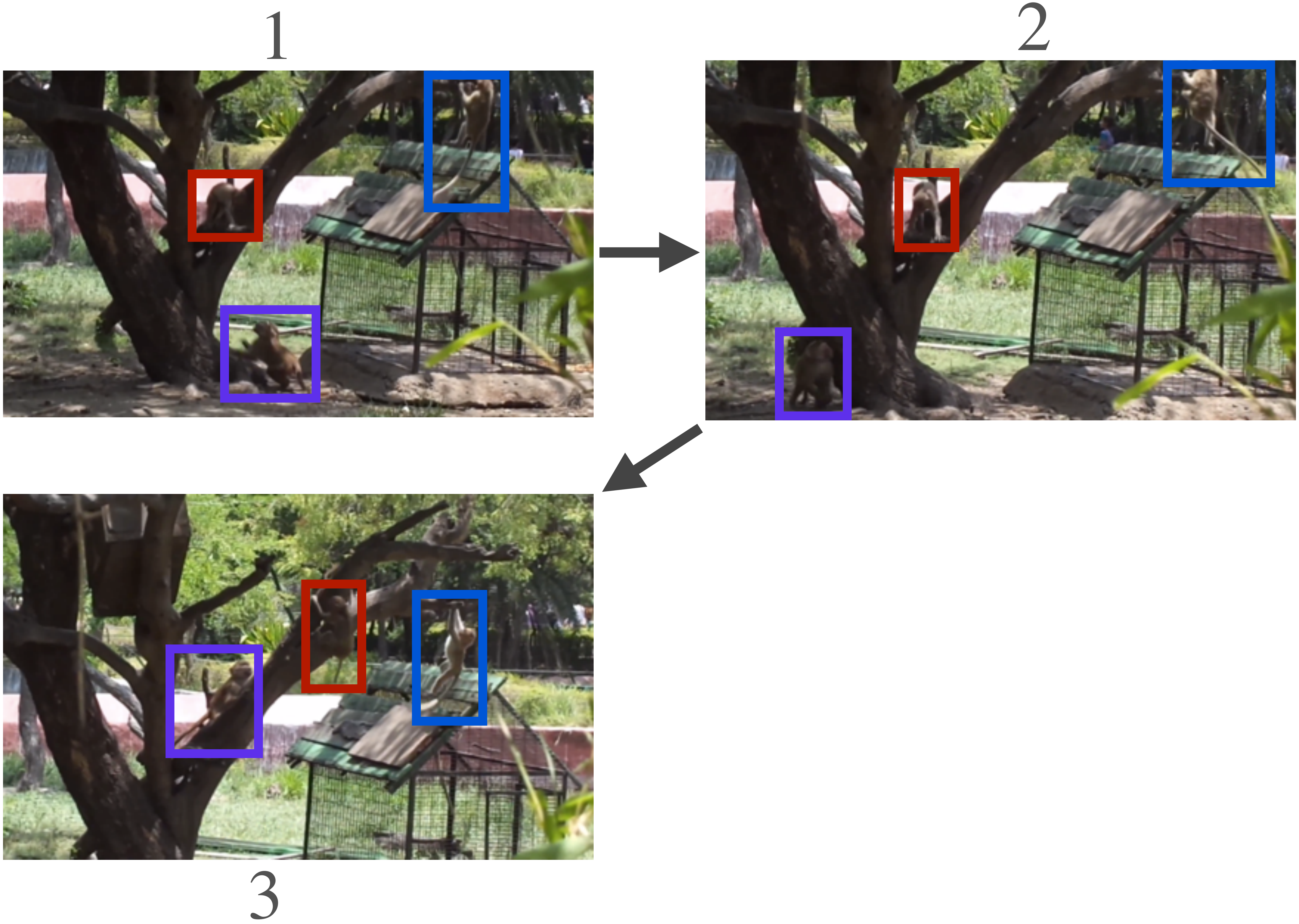
1

VIDEO TRACKING

1

2

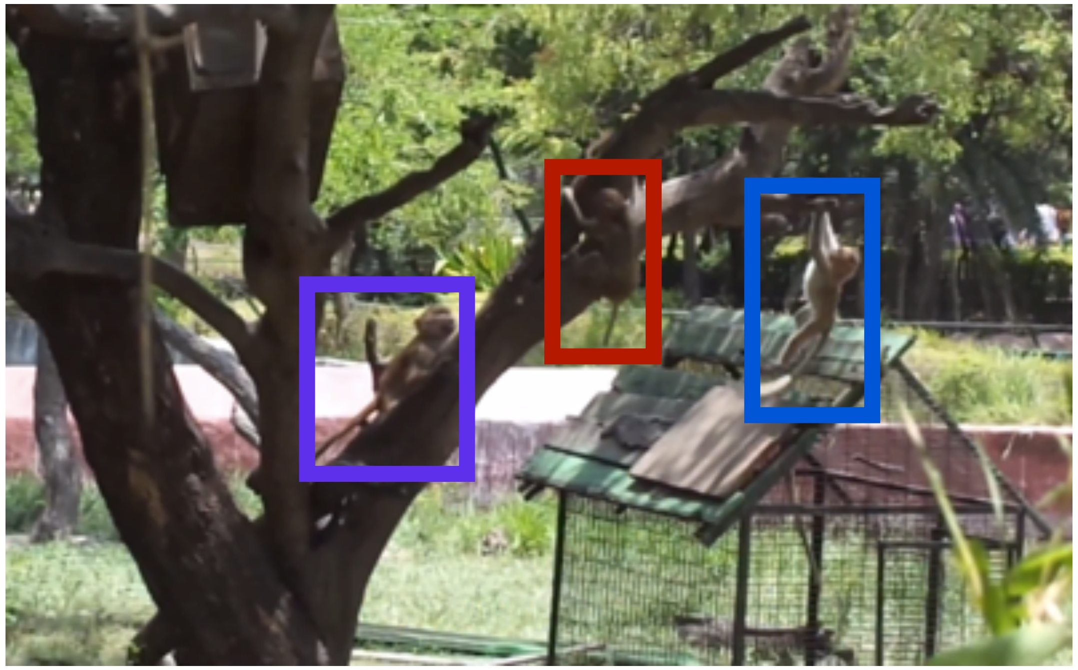
VIDEO TRACKING



VIDEO TRACKING

1

2

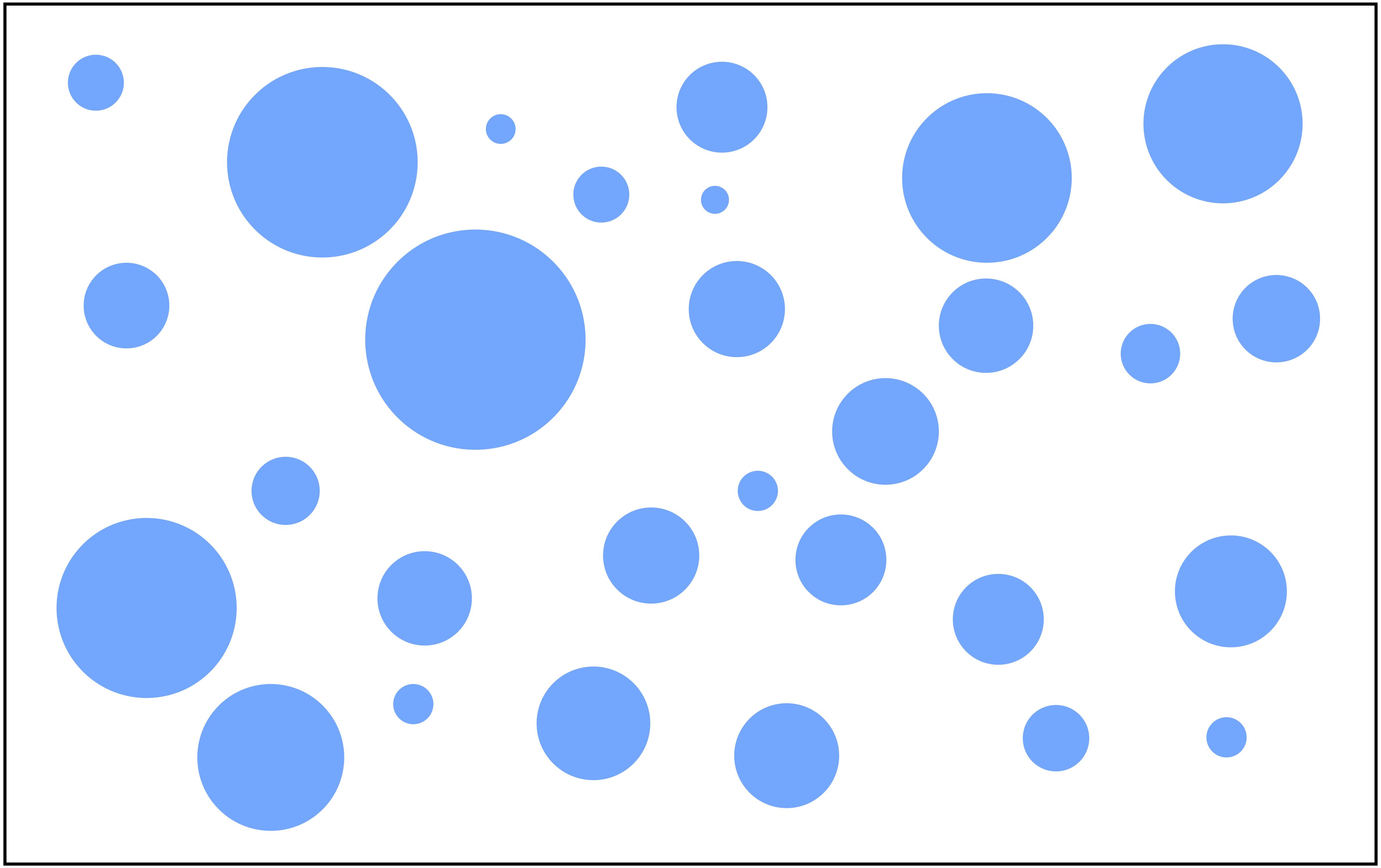


3

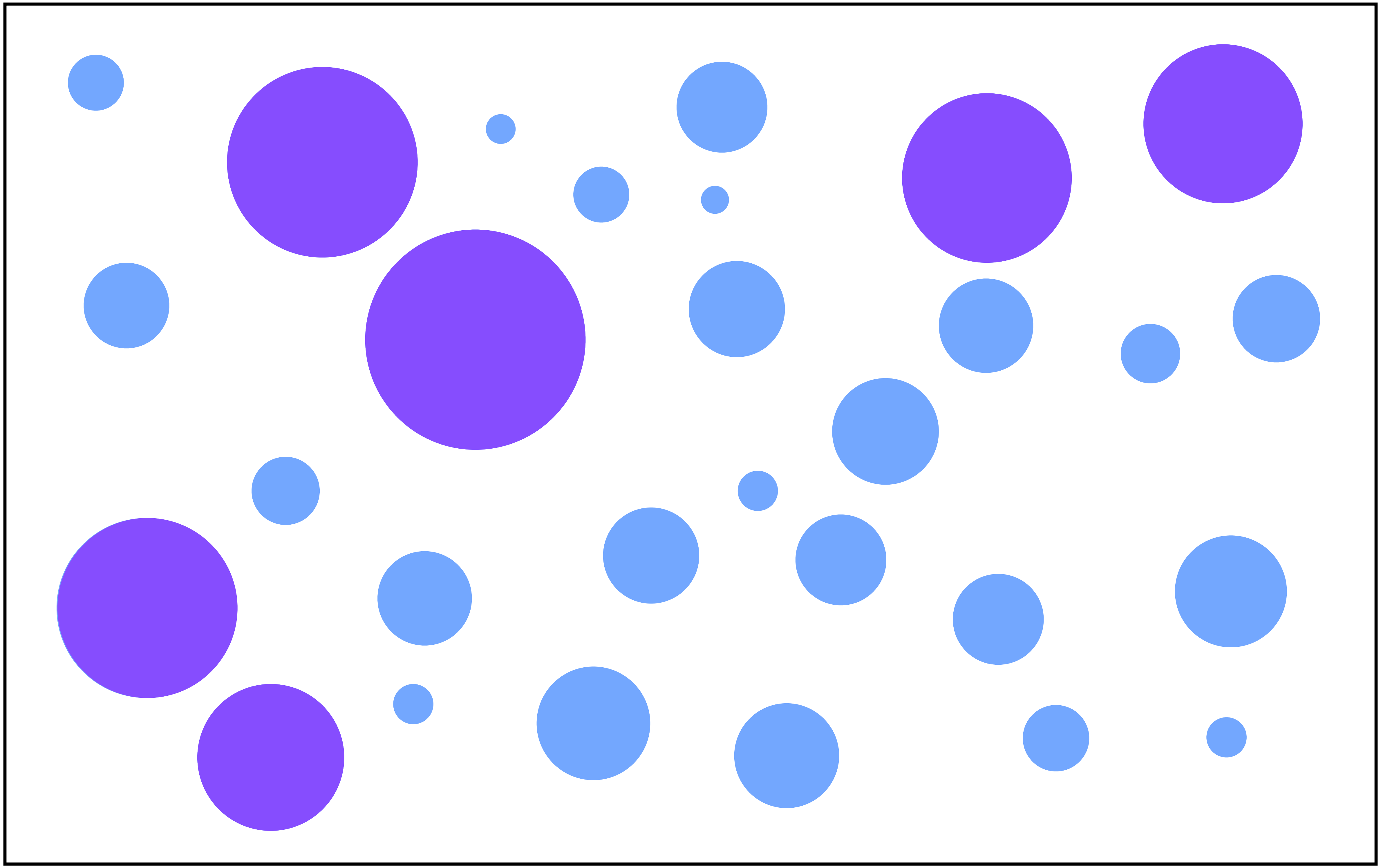
4

SUBSET SELECTION

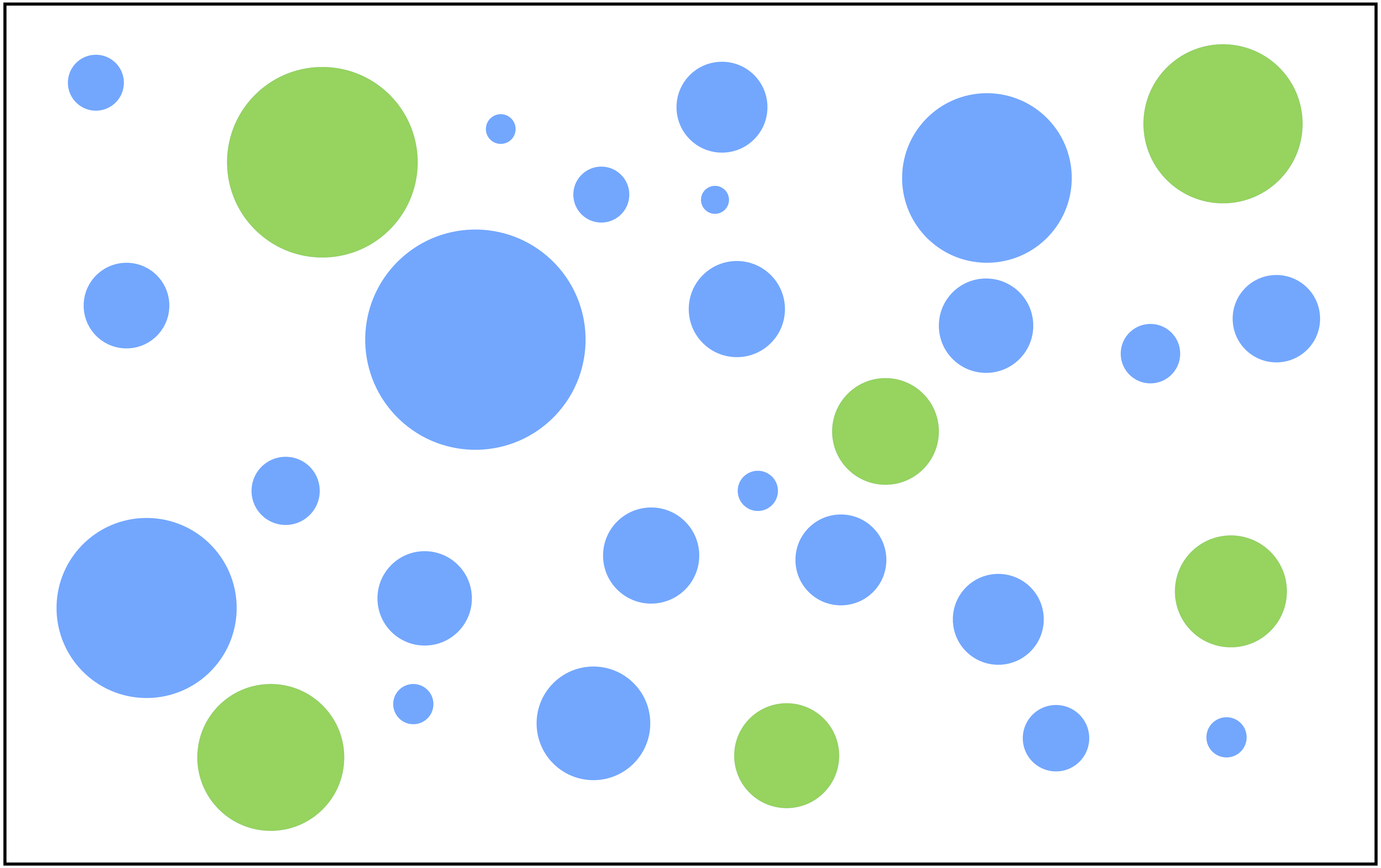
SUBSET SELECTION



SUBSET SELECTION

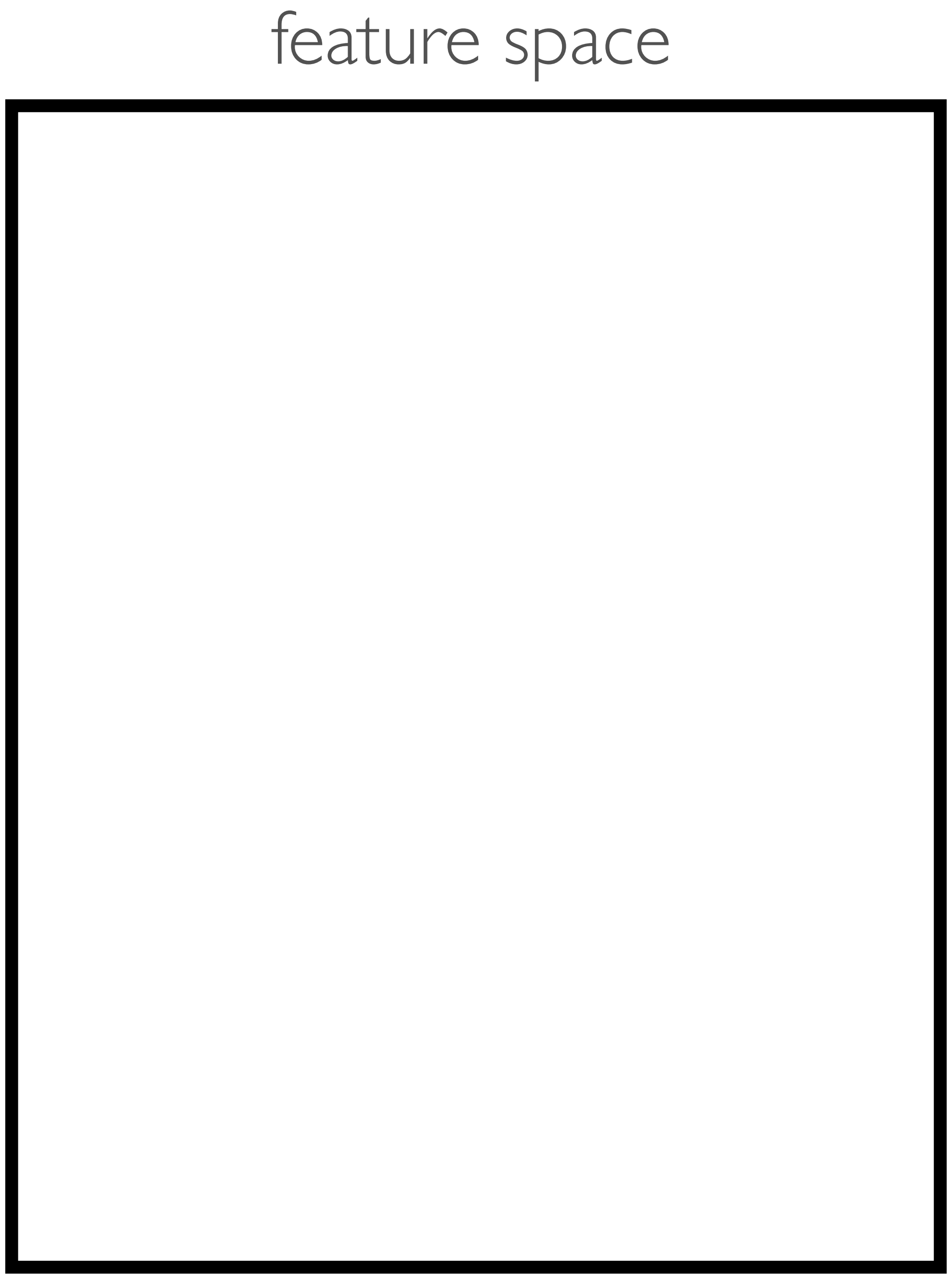


SUBSET SELECTION

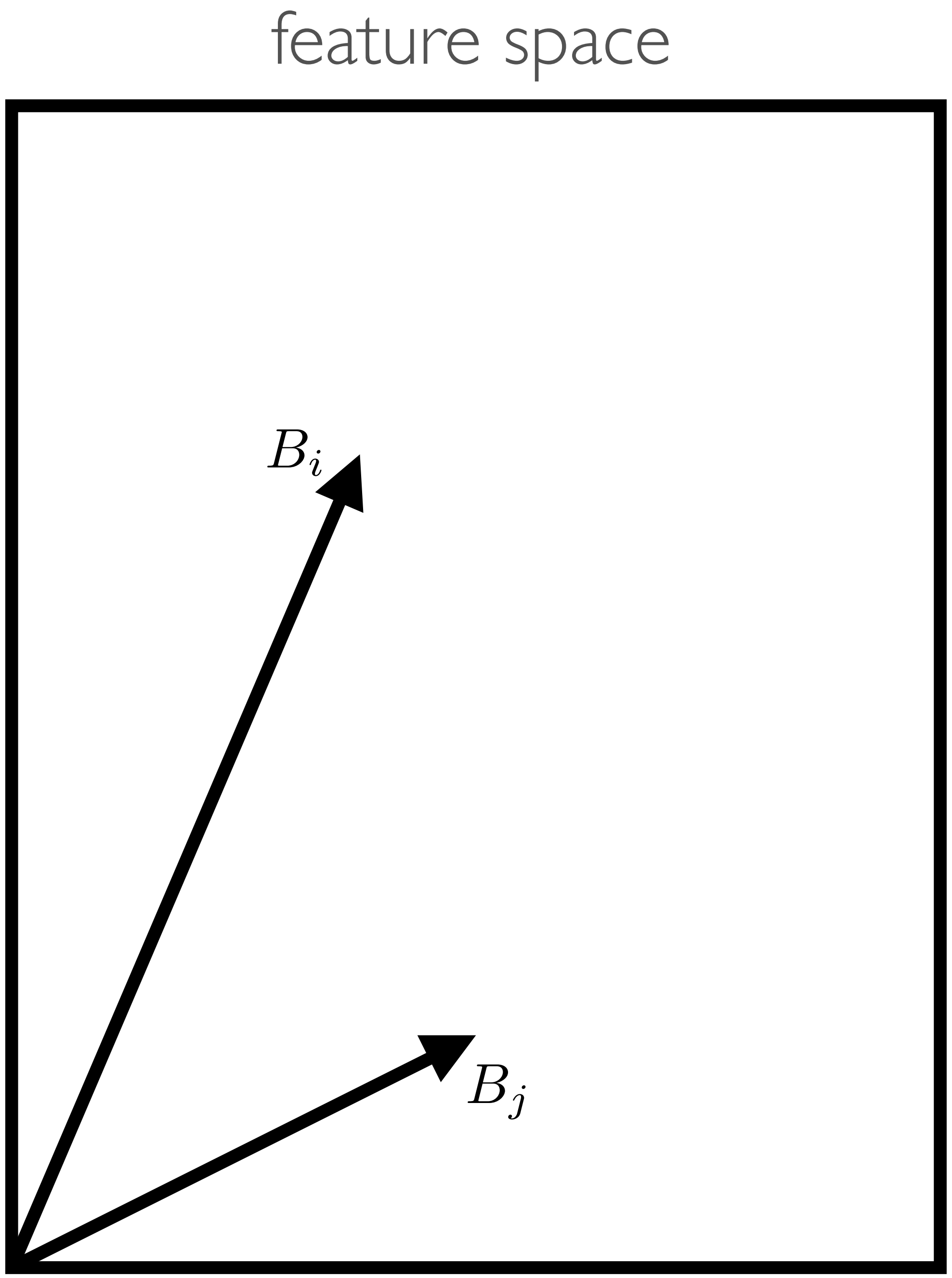


AREA AS SET-GOODNESS

AREA AS SET-GOODNESS

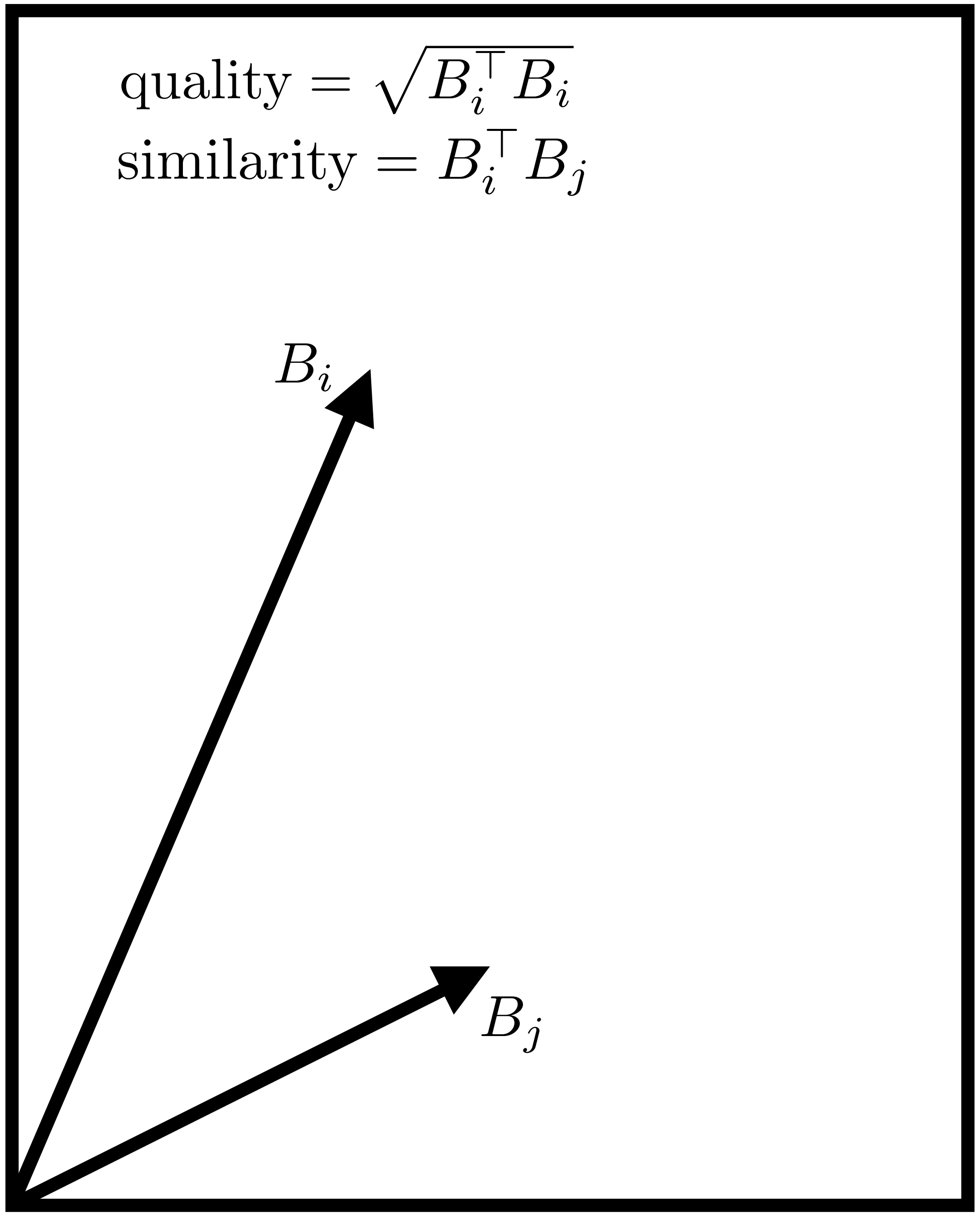


AREA AS SET-GOODNESS



AREA AS SET-GOODNESS

feature space

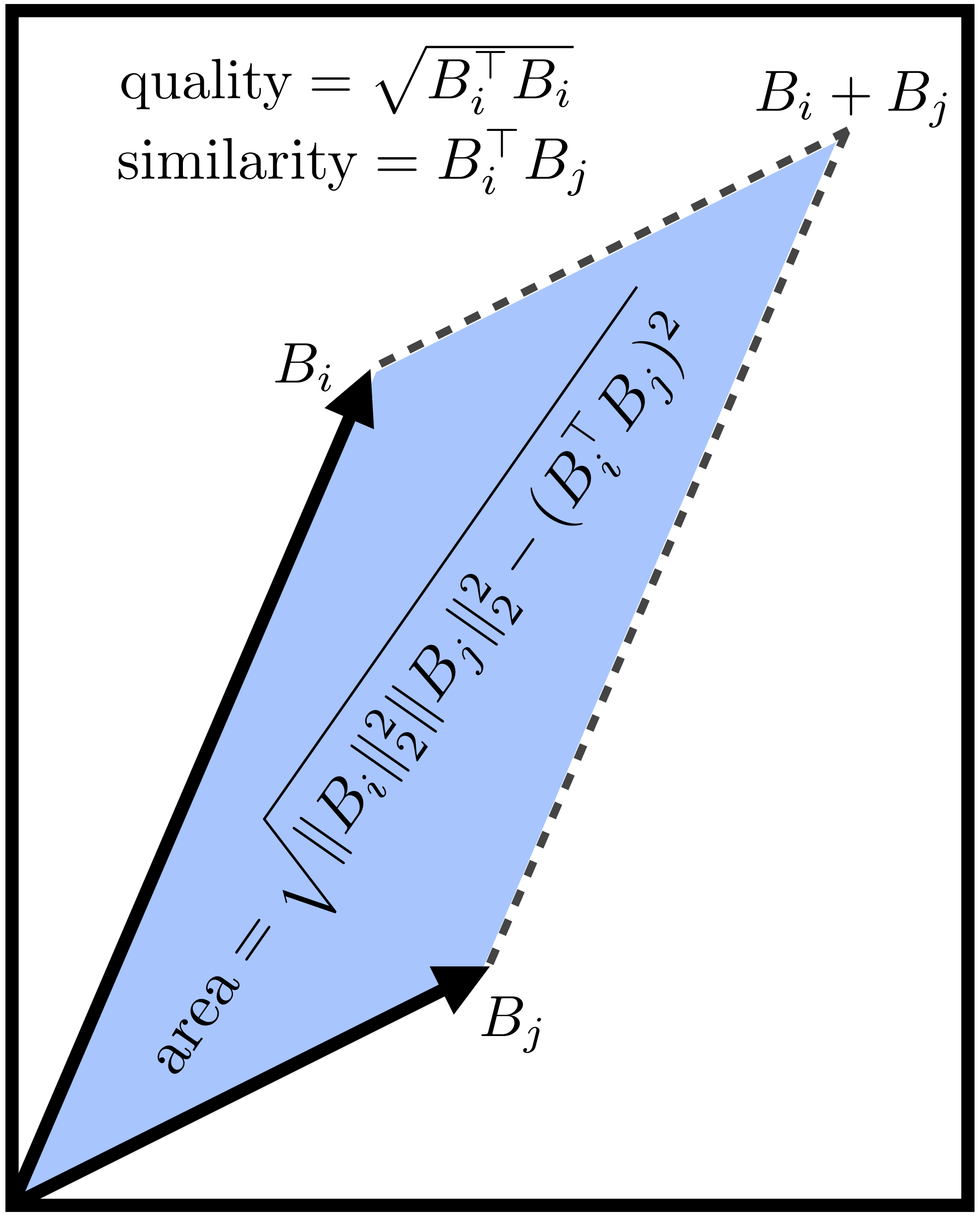


quality $= \sqrt{B_i^\top B_i}$

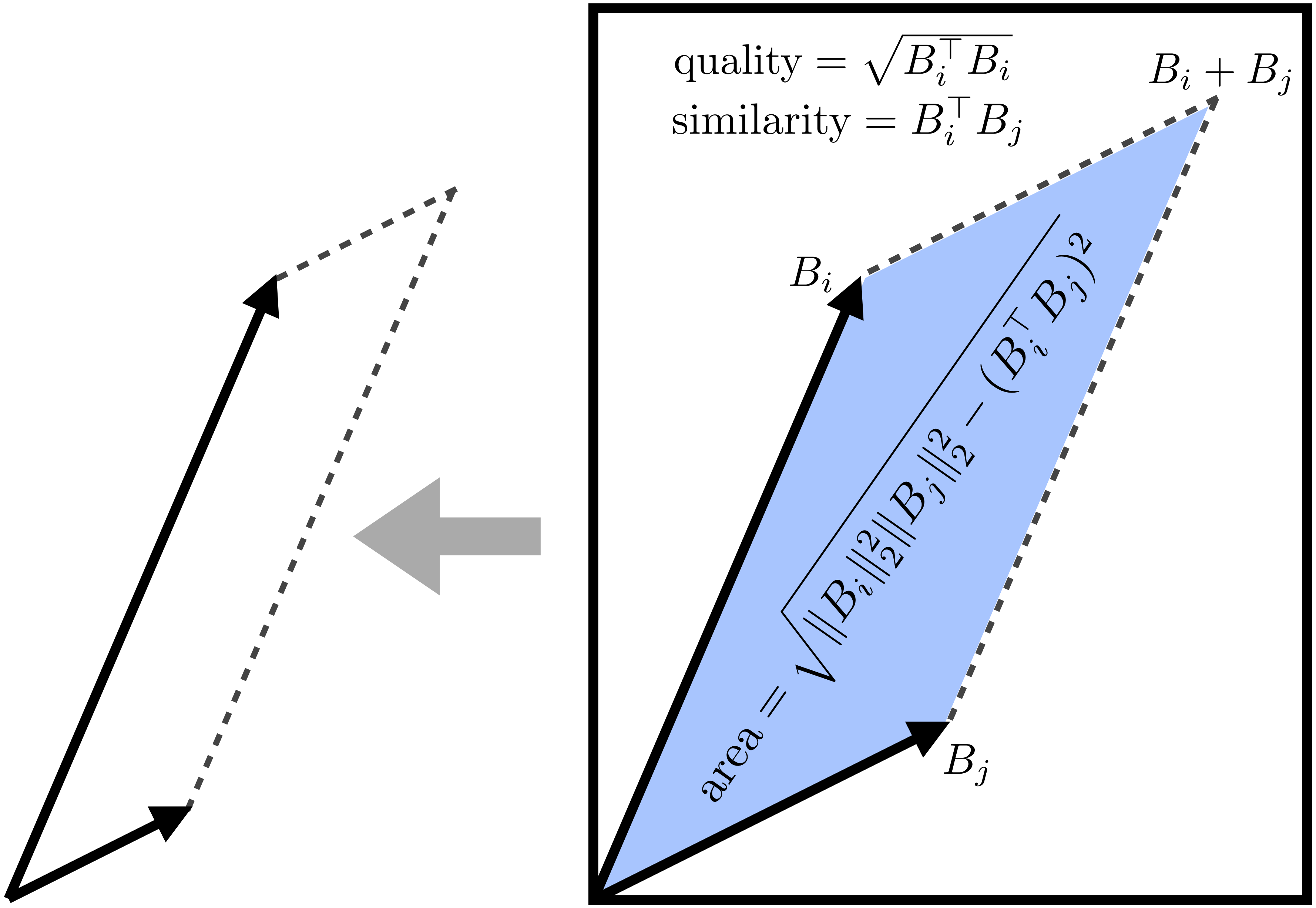
similarity $= B_i^\top B_j$

AREA AS SET-GOODNESS

feature space



AREA AS SET-GOODNESS

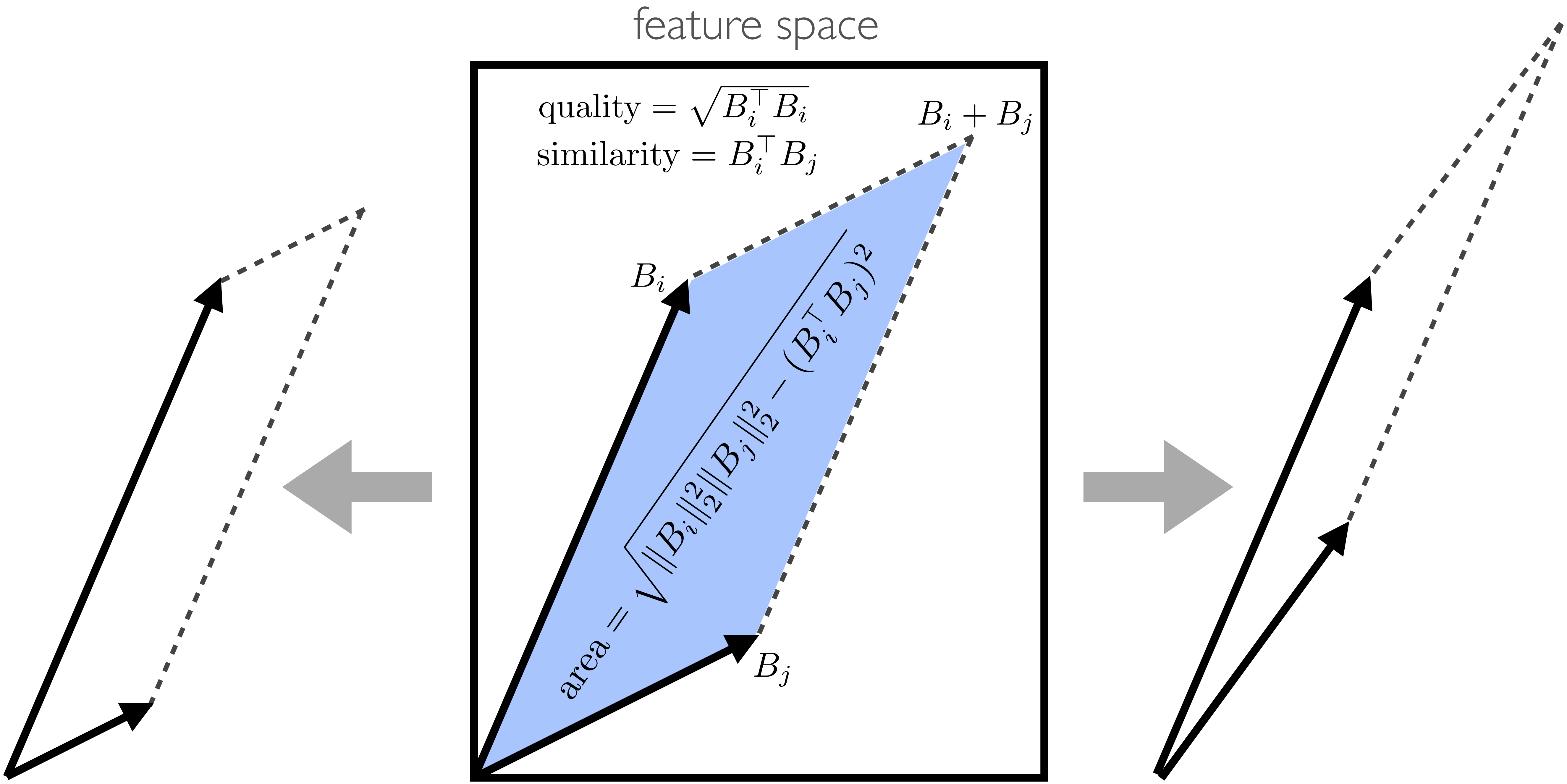


feature space

quality $= \sqrt{B_i^\top B_i}$

similarity $= B_i^\top B_j$

AREA AS SET-GOODNESS



$$\text{quality} = \sqrt{B_i^\top B_i}$$

$$\text{similarity} = B_i^\top B_j$$

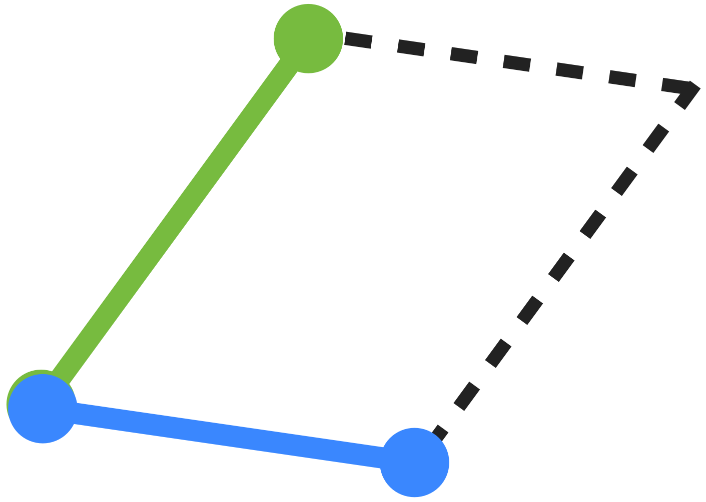
$$\text{area} = \sqrt{\|B_i\|_2^2 \|B_j\|_2^2 - (B_i^\top B_j)^2}$$

VOLUME AS SET-GOODNESS

$$\text{area} = \sqrt{\|B_i\|_2^2 \|B_j\|_2^2 - (B_i^\top B_j)^2}$$

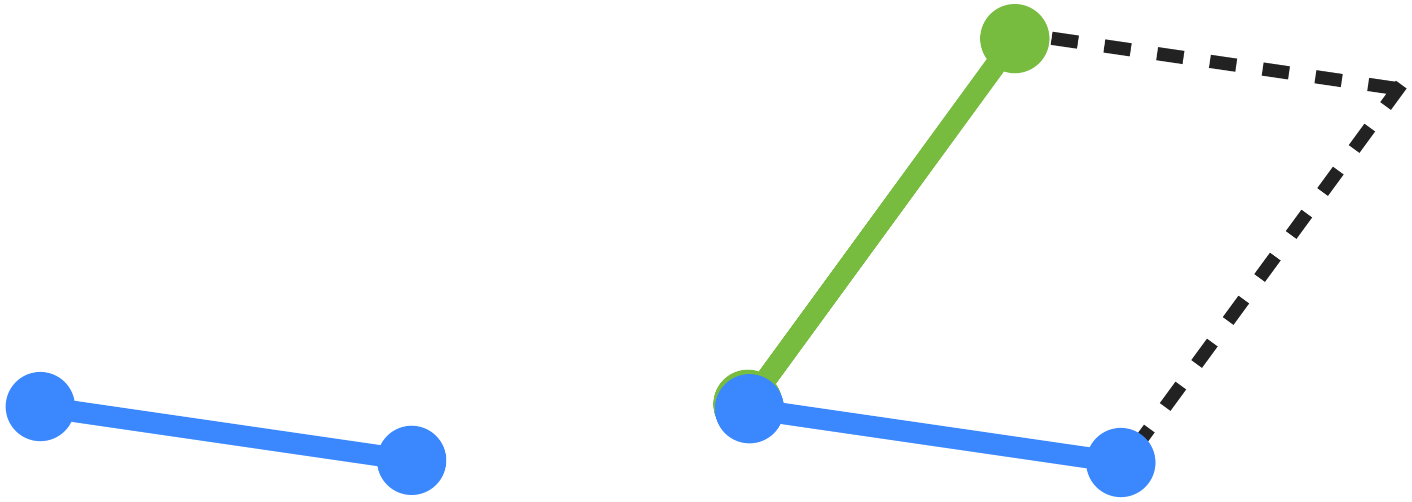
VOLUME AS SET-GOODNESS

$$\text{area} = \sqrt{\|B_i\|_2^2 \|B_j\|_2^2 - (B_i^\top B_j)^2}$$



VOLUME AS SET-GOODNESS

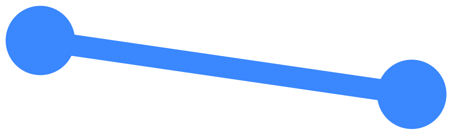
$$\text{area} = \sqrt{\|B_i\|_2^2 \|B_j\|_2^2 - (B_i^\top B_j)^2}$$

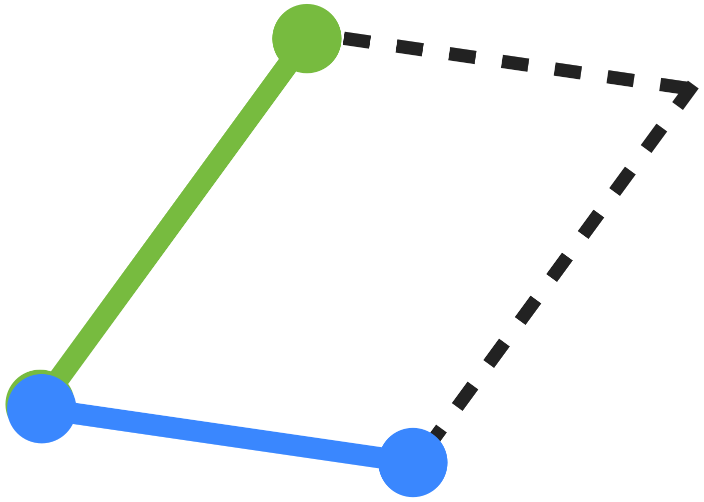


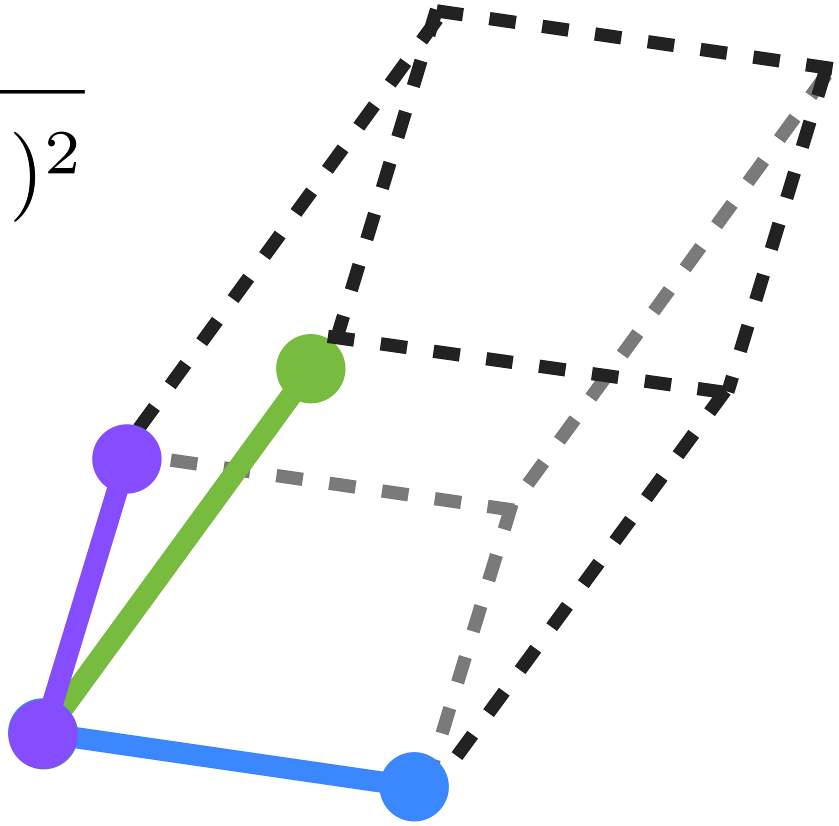
$$\text{length} = \|B_i\|_2$$

VOLUME AS SET-GOODNESS

$$\text{area} = \sqrt{\|B_i\|_2^2 \|B_j\|_2^2 - (B_i^\top B_j)^2}$$



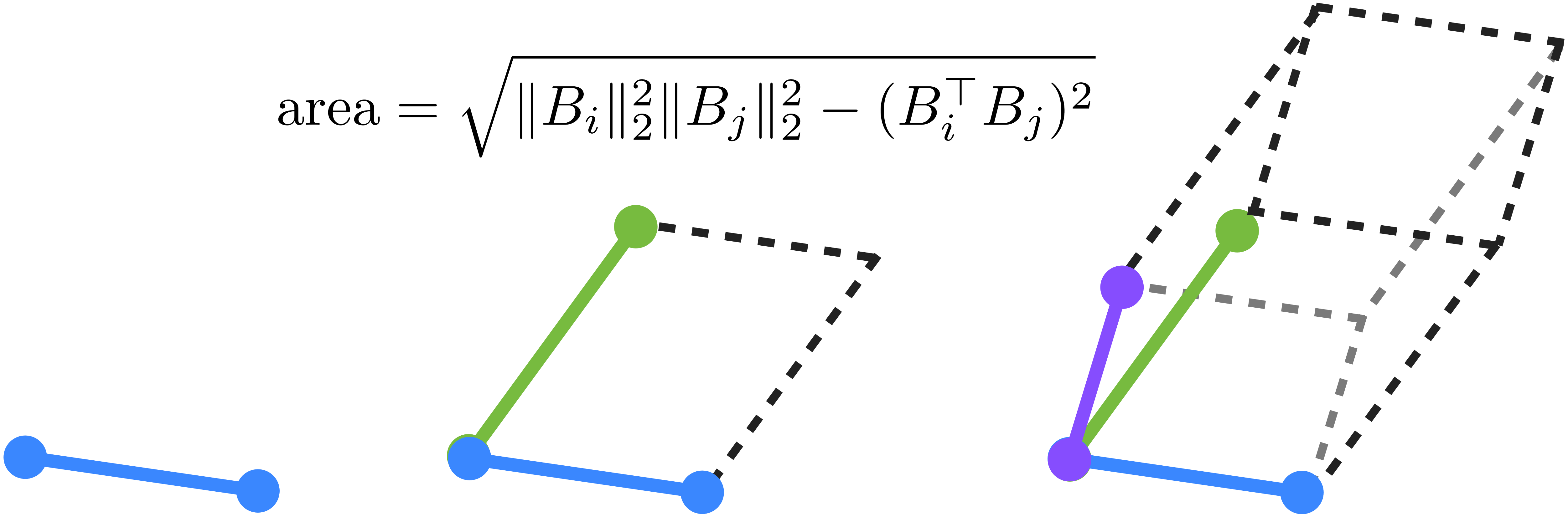




$$\text{length} = \|B_i\|_2$$

$$\text{volume} = \text{base} \times \text{height}$$

VOLUME AS SET-GOODNESS



$\text{area} = \sqrt{\|B_i\|_2^2 \|B_j\|_2^2 - (B_i^\top B_j)^2}$

$\text{length} = \|B_i\|_2$

$\text{volume} = \text{base} \times \text{height}$

$$\begin{aligned}\text{vol}(B) &= \text{height} \times \text{base} \\ &= ||B_1||_2 \text{vol}(\text{proj}_{\perp B_1}(B_{2:N}))\end{aligned}$$

AREA AS A DET

$$\text{area} = \sqrt{\|B_i\|_2^2\|B_j\|_2^2 - (B_i^\top B_j)^2}$$

AREA AS A DET

$$\text{area} = \sqrt{\|B_i\|_2^2 \|B_j\|_2^2 - (B_i^\top B_j)^2}$$

$$= \det \begin{pmatrix} ||B_i||_2^2 & B_i^\top B_j \\ B_i^\top B_j & ||B_j||_2^2 \end{pmatrix}^{\frac{1}{2}}$$

AREA AS A DET

$$\text{area} = \sqrt{\|B_i\|_2^2 \|B_j\|_2^2 - (B_i^\top B_j)^2}$$

$$= \det \begin{pmatrix} ||B_i||_2^2 & B_i^\top B_j \\ B_i^\top B_j & ||B_j||_2^2 \end{pmatrix}^{\frac{1}{2}}$$

$$= \det \left(\begin{bmatrix} \text{---} \; B_i \; \text{---} \\ \text{---} \; B_j \; \text{---} \end{bmatrix} \begin{bmatrix} | & | \\ B_i & B_j \\ | & | \end{bmatrix} \right)^{\frac{1}{2}}$$

VOLUME AS A DET

$$\mathrm{vol}(B_{\{i,j\}}) = \det\left(\begin{pmatrix} - & B_i & - \\ - & B_j & - \end{pmatrix} \begin{pmatrix} | & | \\ B_i & B_j \\ | & | \end{pmatrix} \right)^{\frac{1}{2}}$$

VOLUME AS A DET

$$\mathrm{vol}(B_{\{i,j\}}) = \det\left(\begin{bmatrix} \text{—}\; B_i \;\text{—} \\ \text{—}\; B_j \;\text{—} \end{bmatrix}\begin{bmatrix} | & | \\ B_i & B_j \\ | & | \end{bmatrix}\right)^{\frac{1}{2}}$$

$$\mathrm{vol}(B) = \det\left(\begin{bmatrix} \text{—}\; B_1 \;\text{—} \\ \vdots \\ \text{—}\; B_N \;\text{—} \end{bmatrix}\begin{bmatrix} | & & | \\ B_1 & \cdots & B_N \\ | & & | \end{bmatrix}\right)^{\frac{1}{2}}$$

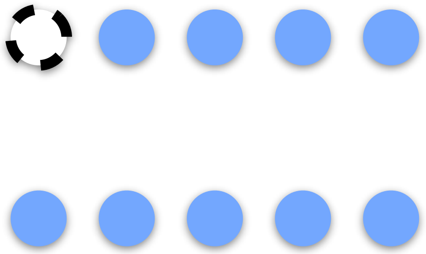
$$\mathrm{vol}(B)^2 = \det(B^\top B) = \det(L)$$

COMPLEX STATISTICS

COMPLEX STATISTICS

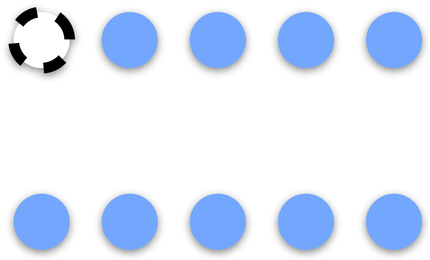
● ● ● ● ●

COMPLEX STATISTICS



COMPLEX STATISTICS

$$\Sigma$$



COMPLEX STATISTICS

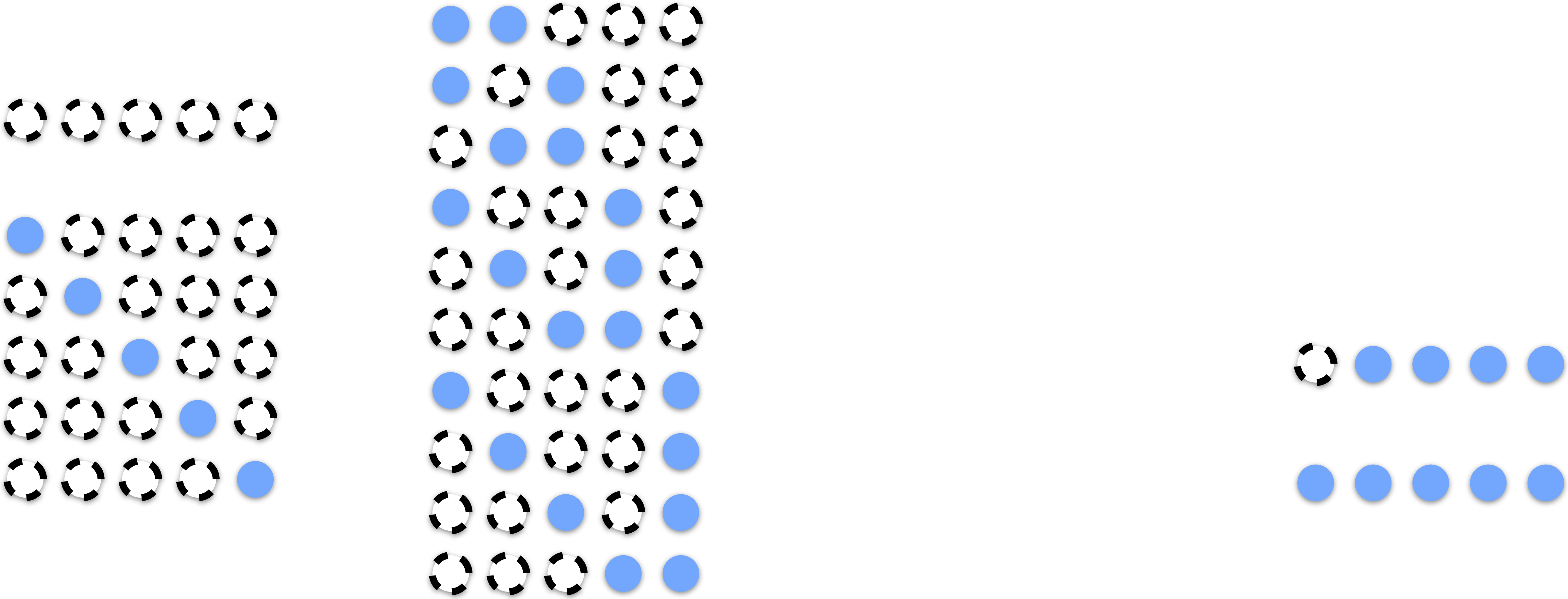
$$\Sigma$$

COMPLEX STATISTICS

$$\Sigma$$

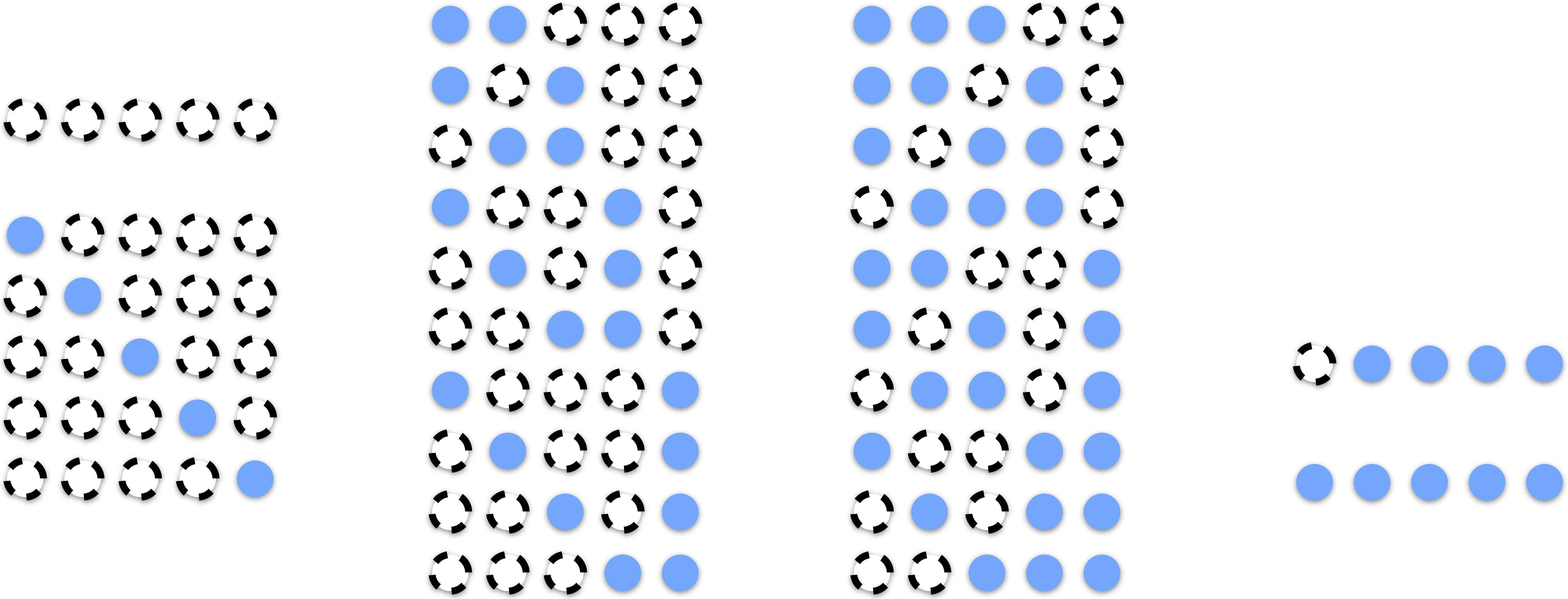
COMPLEX STATISTICS

$$\Sigma$$



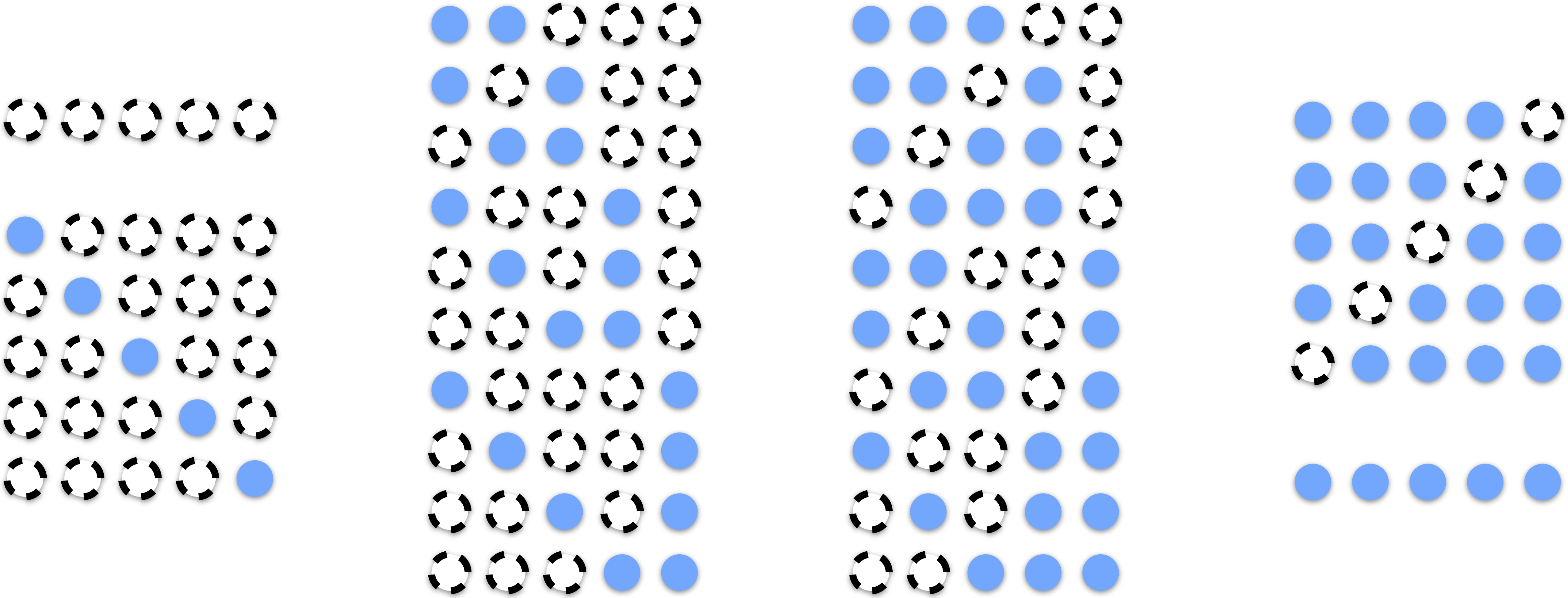
COMPLEX STATISTICS

$$\Sigma$$



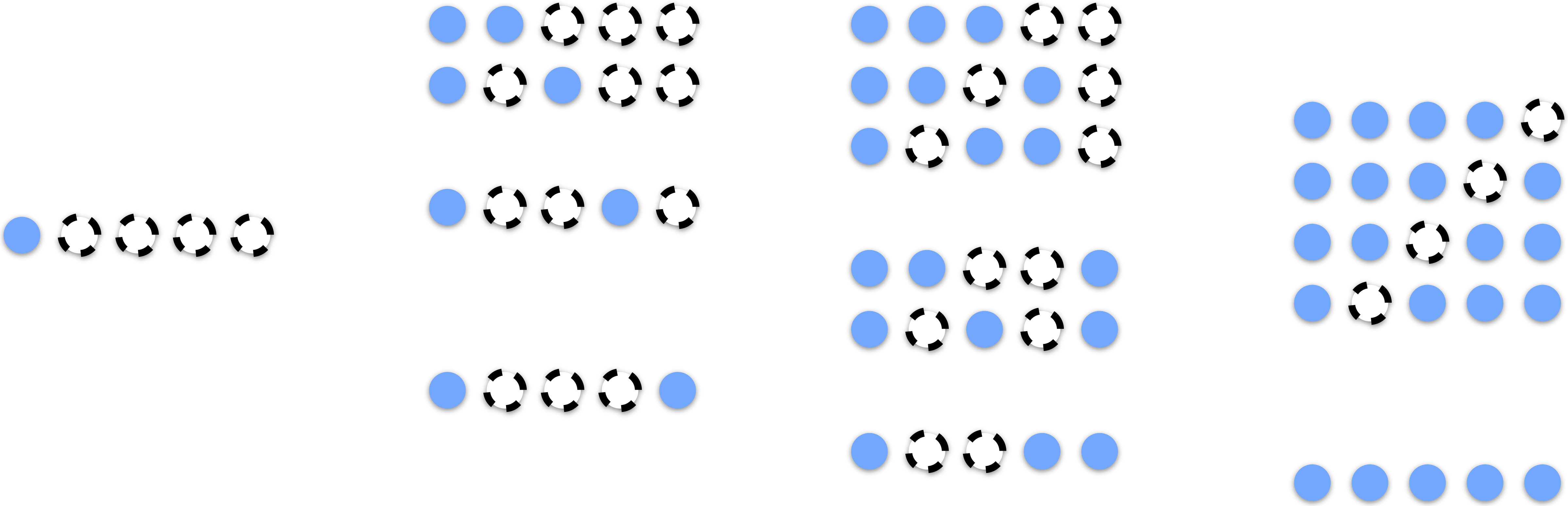
COMPLEX STATISTICS

$$\Sigma$$



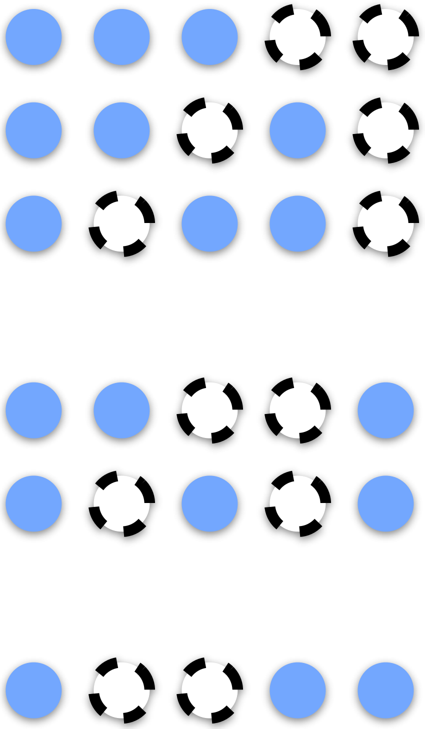
COMPLEX STATISTICS

$$\Sigma$$



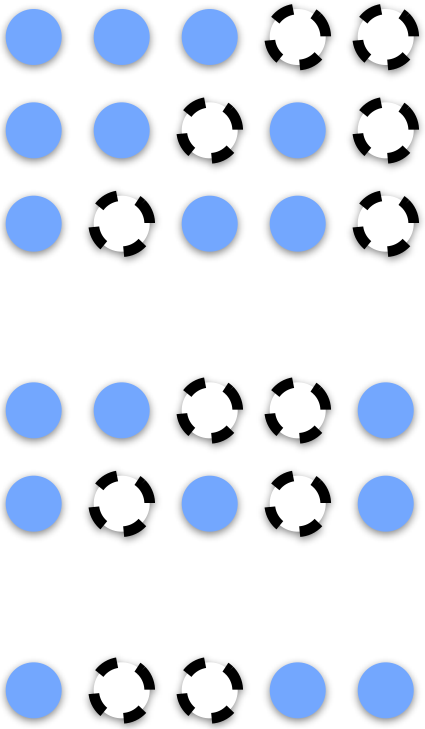
COMPLEX STATISTICS

$$\Sigma$$



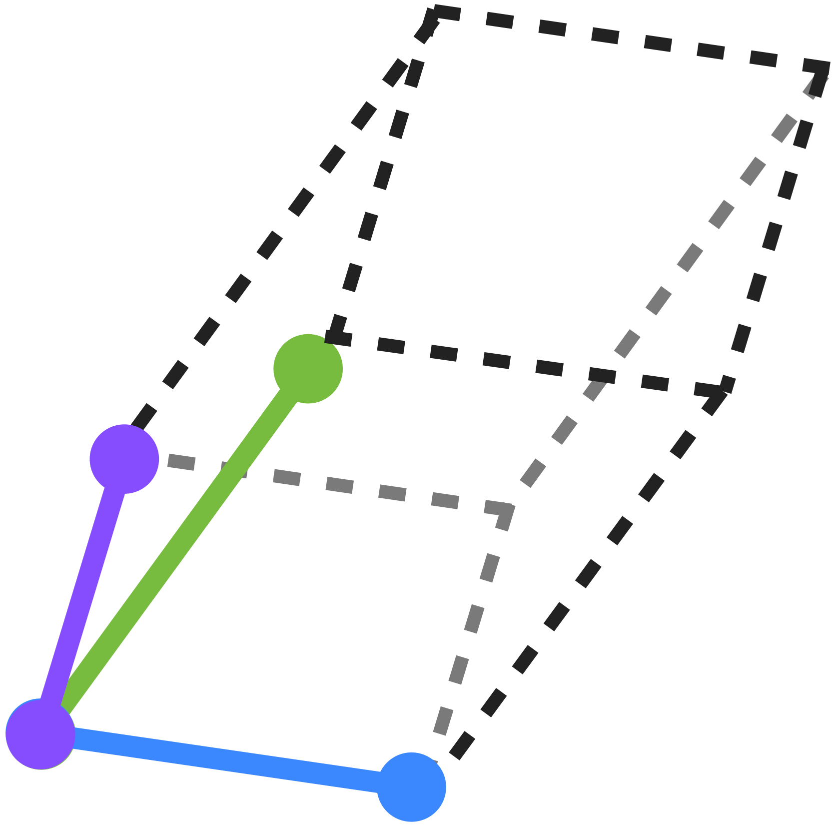
COMPLEX STATISTICS

$$N \text{ items} \implies 2^N \text{ sets}$$



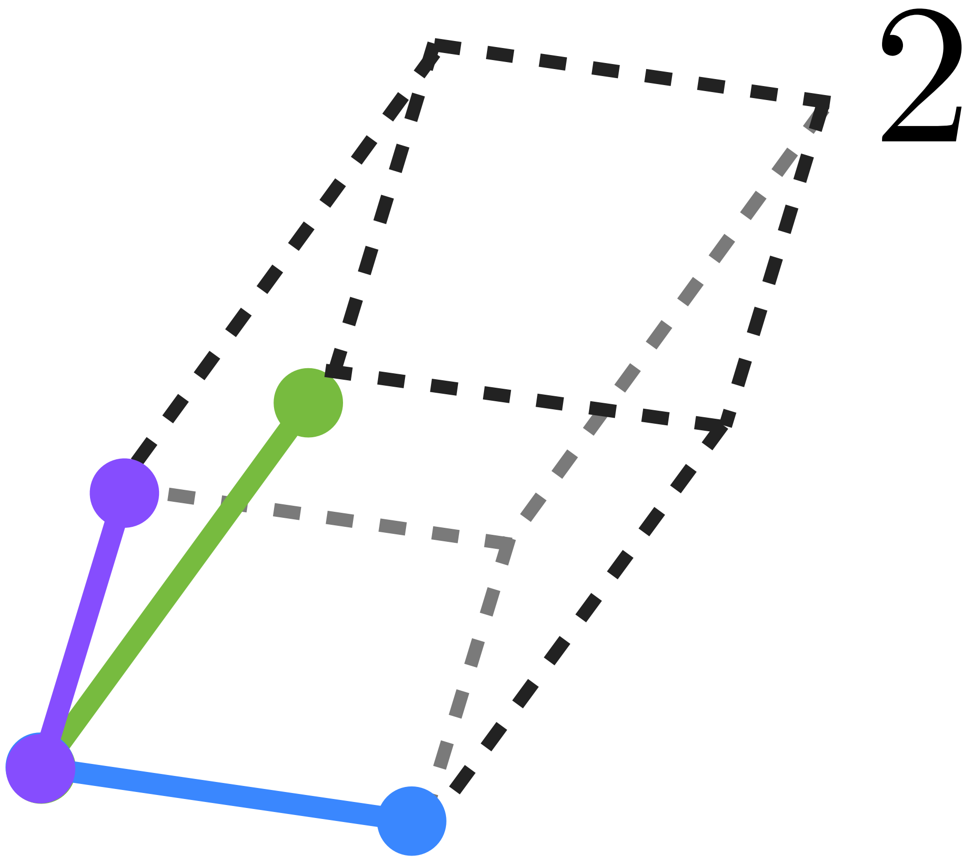
EFFICIENT COMPUTATION

EFFICIENT COMPUTATION



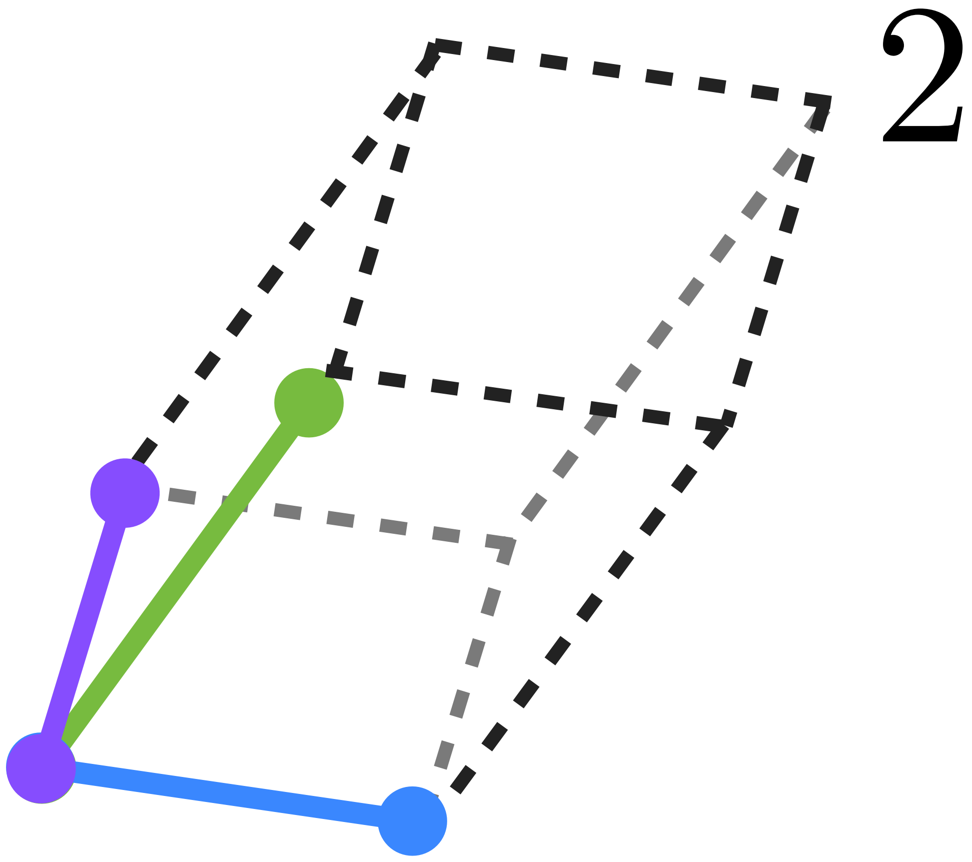
$$\det^{\frac{1}{2}}$$

EFFICIENT COMPUTATION



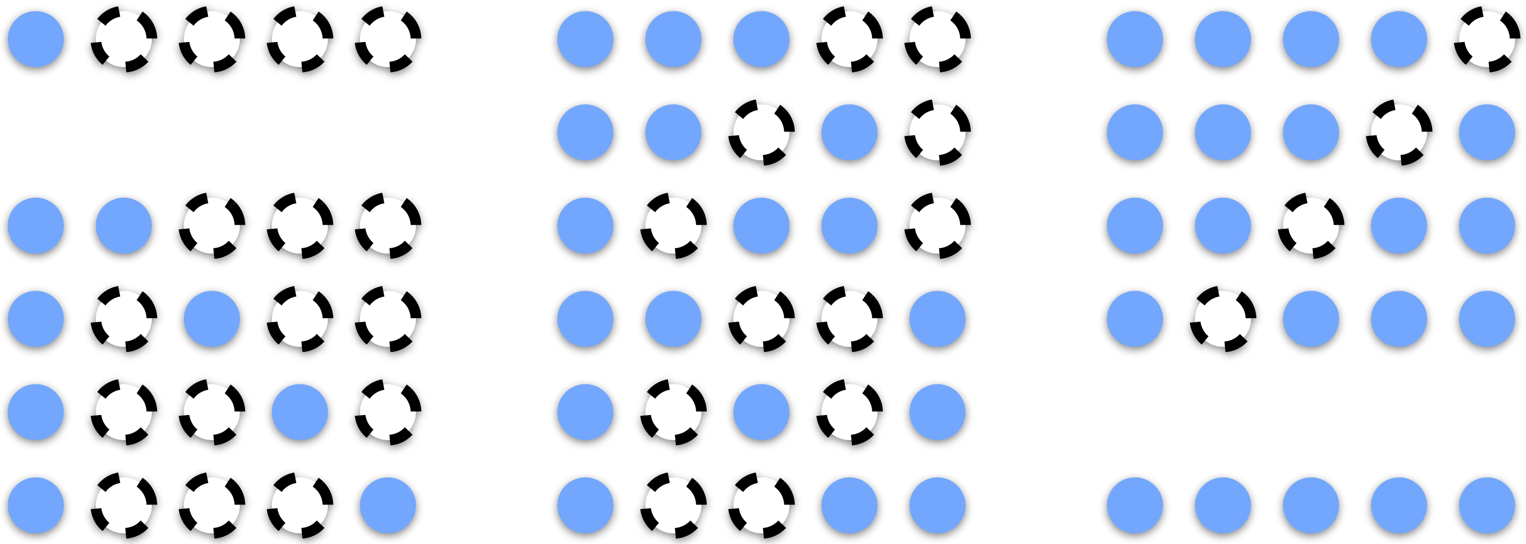
$$\det$$

EFFICIENT COMPUTATION

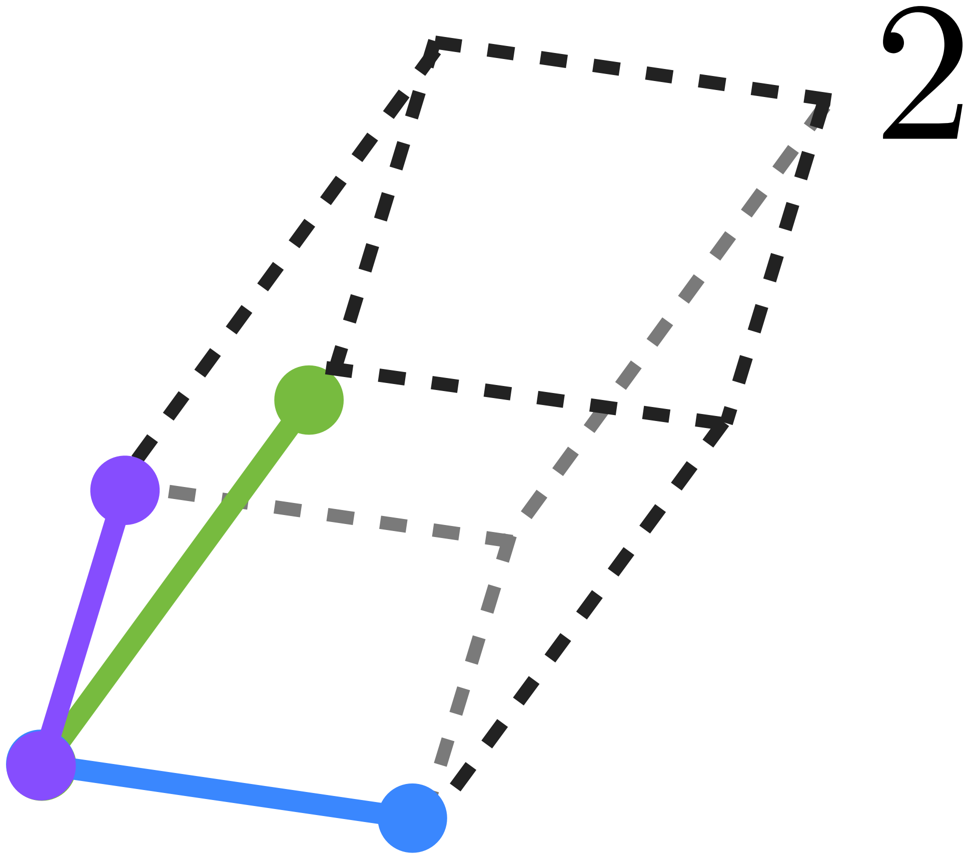


det

$$\Sigma$$

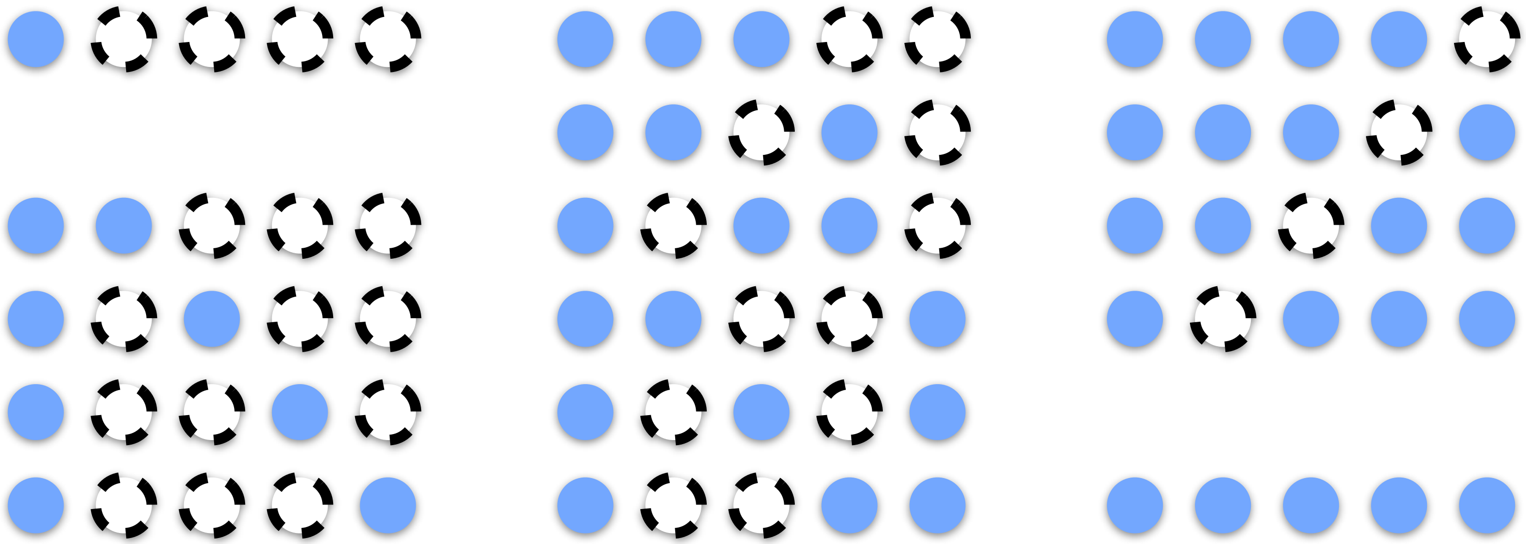


EFFICIENT COMPUTATION



det

$O(N^3)$

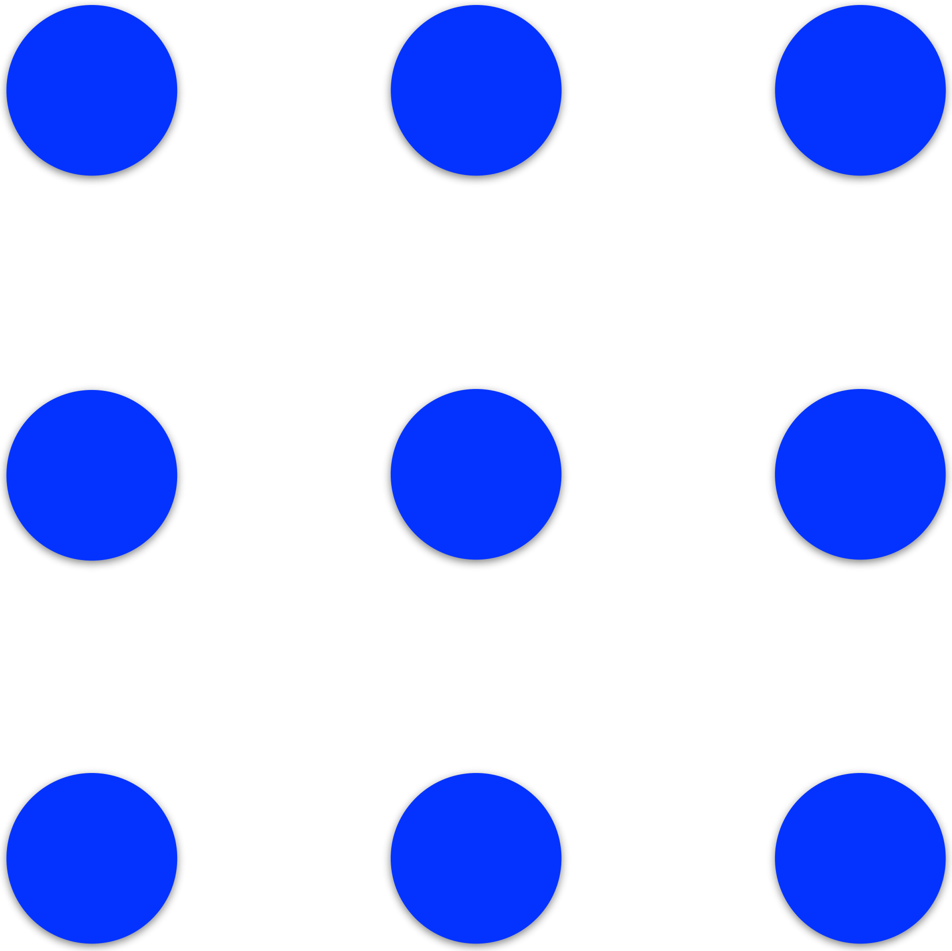


POINT PROCESSES

$$\mathcal{Y} = \{1, \ldots, N\}$$

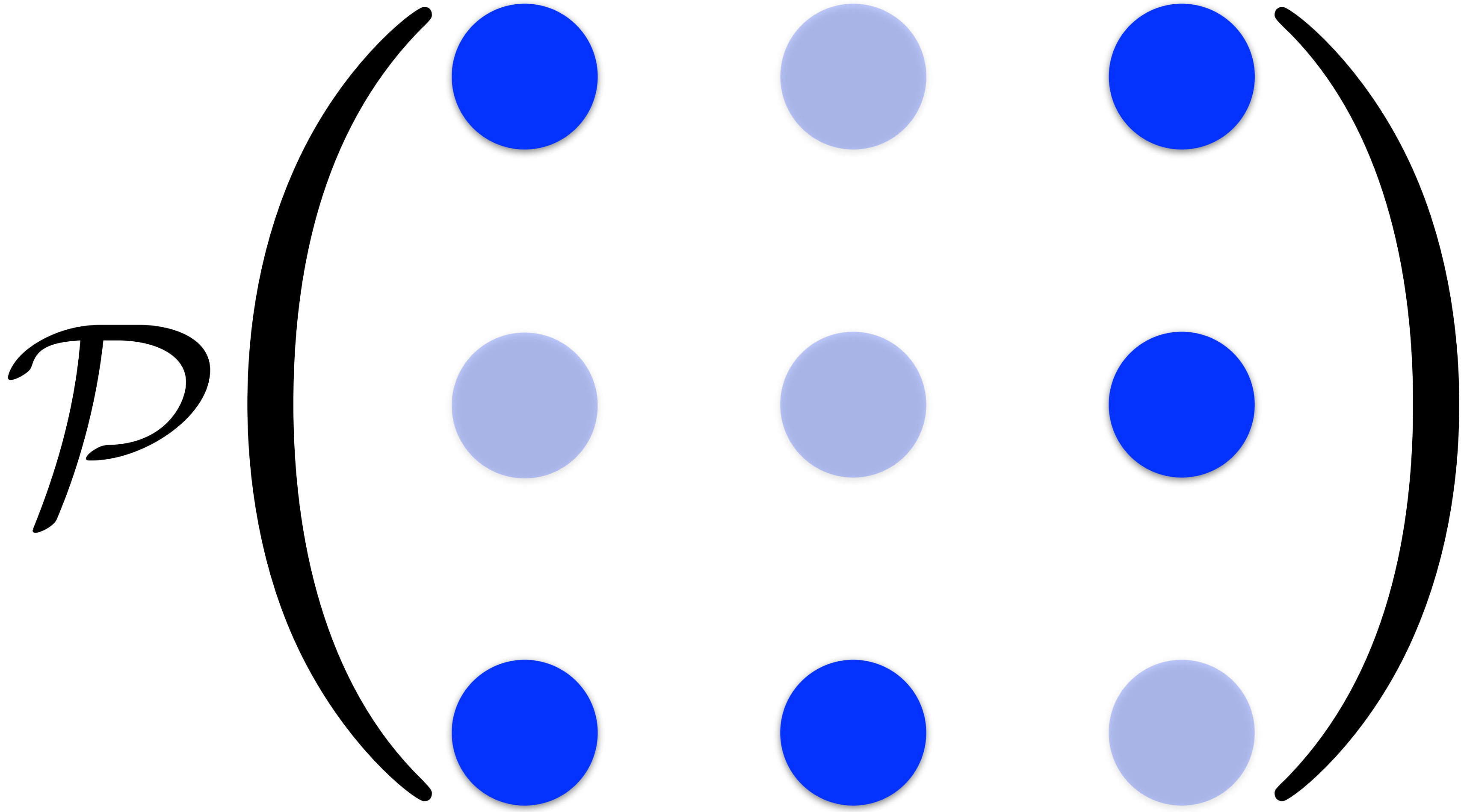
POINT PROCESSES

$$\mathcal{Y} = \{1, \ldots, N\}$$



POINT PROCESSES

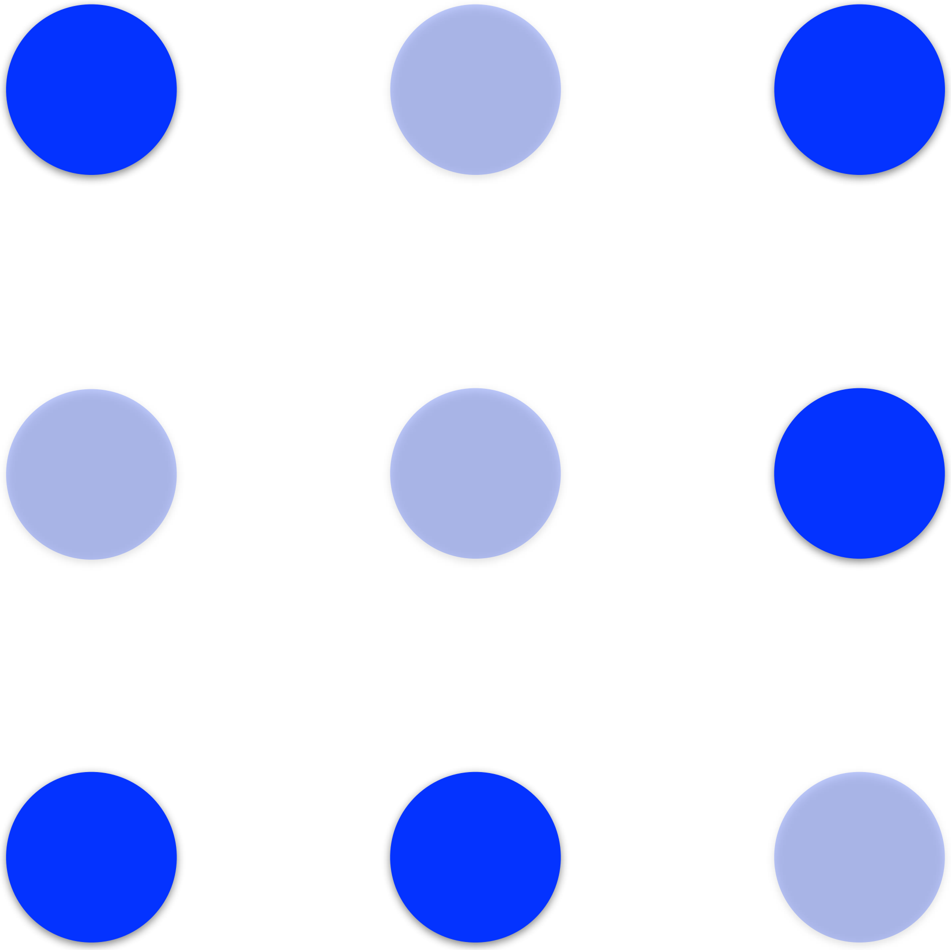
$$\mathcal{Y} = \{1, \ldots, N\}$$



POINT PROCESSES

$$\mathcal{Y} = \{1, \ldots, N\}$$

$\mathcal{P}\Bigg($



$\Bigg) = 0.2$

DETERMINANTAL

DETERMINANTAL

$$\mathcal{P}(\{2,3,5\}) \propto$$

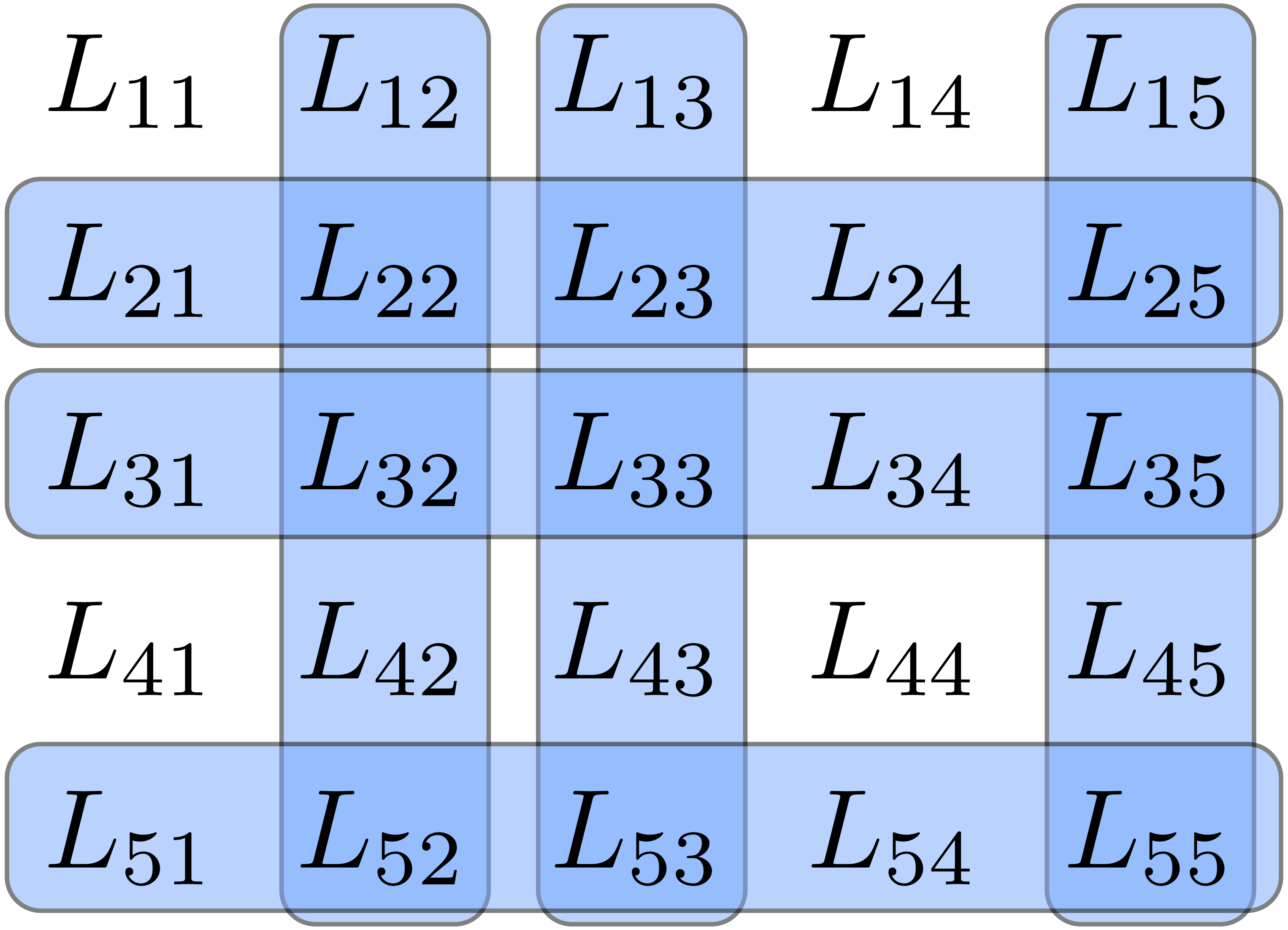
DETERMINANTAL

$$\mathcal{P}(\{2, 3, 5\}) \propto$$

$$\begin{matrix} L_{11} & L_{12} & L_{13} & L_{14} & L_{15} \\ L_{21} & L_{22} & L_{23} & L_{24} & L_{25} \\ L_{31} & L_{32} & L_{33} & L_{34} & L_{35} \\ L_{41} & L_{42} & L_{43} & L_{44} & L_{45} \\ L_{51} & L_{52} & L_{53} & L_{54} & L_{55} \end{matrix}$$

DETERMINANTAL

$$\mathcal{P}(\{2, 3, 5\}) \propto$$



$$\begin{matrix} L_{11} & L_{12} & L_{13} & L_{14} & L_{15} \\ L_{21} & L_{22} & L_{23} & L_{24} & L_{25} \\ L_{31} & L_{32} & L_{33} & L_{34} & L_{35} \\ L_{41} & L_{42} & L_{43} & L_{44} & L_{45} \\ L_{51} & L_{52} & L_{53} & L_{54} & L_{55} \end{matrix}$$

DETERMINANTAL

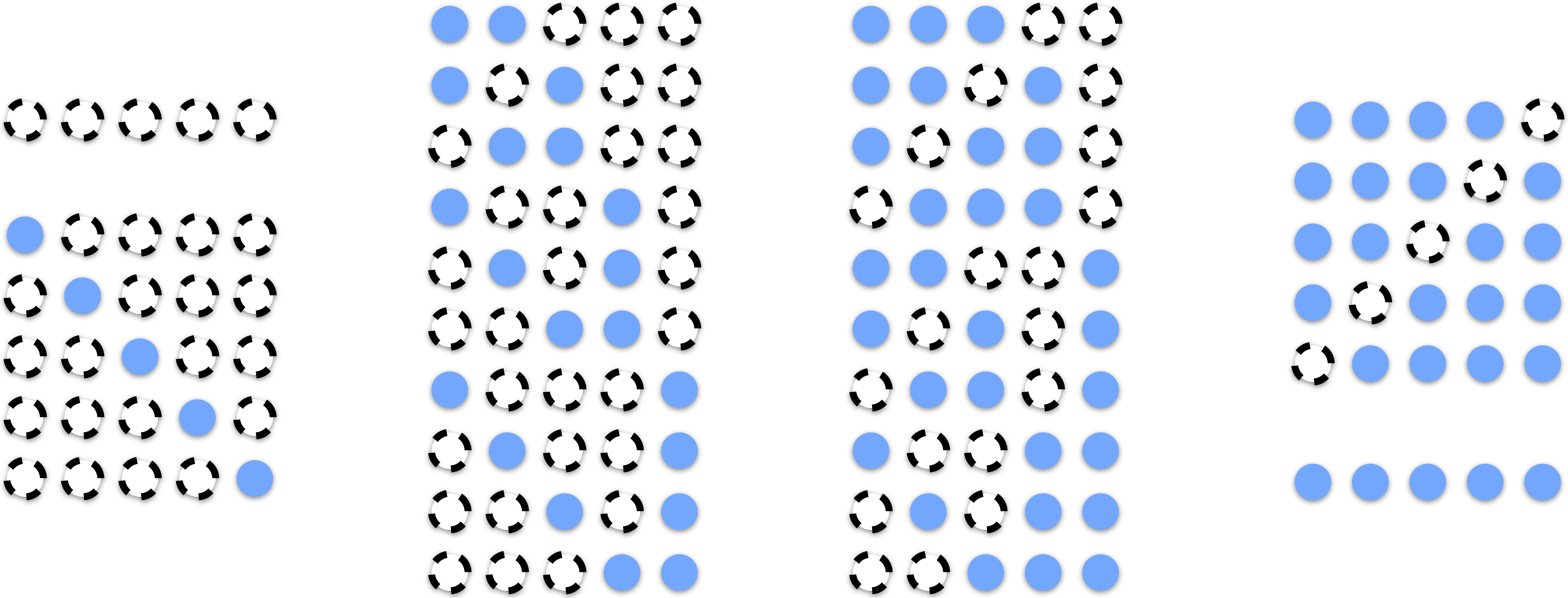
$$\mathcal{P}(\{2,3,5\}) \propto \qquad \begin{matrix} L_{22} & L_{23} & L_{25} \\ L_{32} & L_{33} & L_{35} \\ L_{52} & L_{53} & L_{55} \end{matrix}$$

DETERMINANTAL

$$\mathcal{P}(\{2,3,5\}) \propto \det\begin{pmatrix} L_{22} & L_{23} & L_{25} \\ L_{32} & L_{33} & L_{35} \\ L_{52} & L_{53} & L_{55} \end{pmatrix}$$

DETERMINANTAL

$$\mathcal{P}(\{2,3,5\}) = \det\begin{pmatrix} \boxed{\begin{matrix} L_{22} & L_{23} & L_{25} \\ L_{32} & L_{33} & L_{35} \\ L_{52} & L_{53} & L_{55} \end{matrix}} \end{pmatrix}$$



DETERMINANTAL

$$\mathcal{P}(\{2,3,5\}) = \frac{\det\begin{pmatrix} L_{22} & L_{23} & L_{25} \\ L_{32} & L_{33} & L_{35} \\ L_{52} & L_{53} & L_{55} \end{pmatrix}}{\det(L+I)}$$

EFFICIENT INFERENCE

EFFICIENT INFERENCE

Normalizing: $\mathcal{P}_L(\mathbf{Y} = Y)$

EFFICIENT INFERENCE

Normalizing: $\mathcal{P}_L(\mathbf{Y} = Y)$

Marginalizing: $\mathcal{P}(Y \subseteq \mathbf{Y})$

EFFICIENT INFERENCE

Normalizing: $\mathcal{P}_L(\mathbf{Y} = Y)$

Marginalizing: $\mathcal{P}(Y \subseteq \mathbf{Y})$

Conditioning: $\mathcal{P}_L(\mathbf{Y} = B \mid A \subseteq \mathbf{Y})$

$\mathcal{P}_L(\mathbf{Y} = B \mid A \cap \mathbf{Y} = \emptyset)$

EFFICIENT INFERENCE

Normalizing: $\mathcal{P}_L(\mathbf{Y} = Y)$

Marginalizing: $\mathcal{P}(Y \subseteq \mathbf{Y})$

Conditioning: $\mathcal{P}_L(\mathbf{Y} = B \mid A \subseteq \mathbf{Y})$

$\mathcal{P}_L(\mathbf{Y} = B \mid A \cap \mathbf{Y} = \emptyset)$

Sampling: $\mathbf{Y} \sim \mathcal{P}_L$

EFFICIENT INFERENCE

Normalizing: $\mathcal{P}_L(\mathbf{Y} = Y)$

Marginalizing: $\mathcal{P}(Y \subseteq \mathbf{Y})$

Conditioning: $\mathcal{P}_L(\mathbf{Y} = B \mid A \subseteq \mathbf{Y})$

$\mathcal{P}_L(\mathbf{Y} = B \mid A \cap \mathbf{Y} = \emptyset)$

Sampling: $\mathbf{Y} \sim \mathcal{P}_L$

$O(N^3)$

1. DIMENSIONALITY REDUCTION

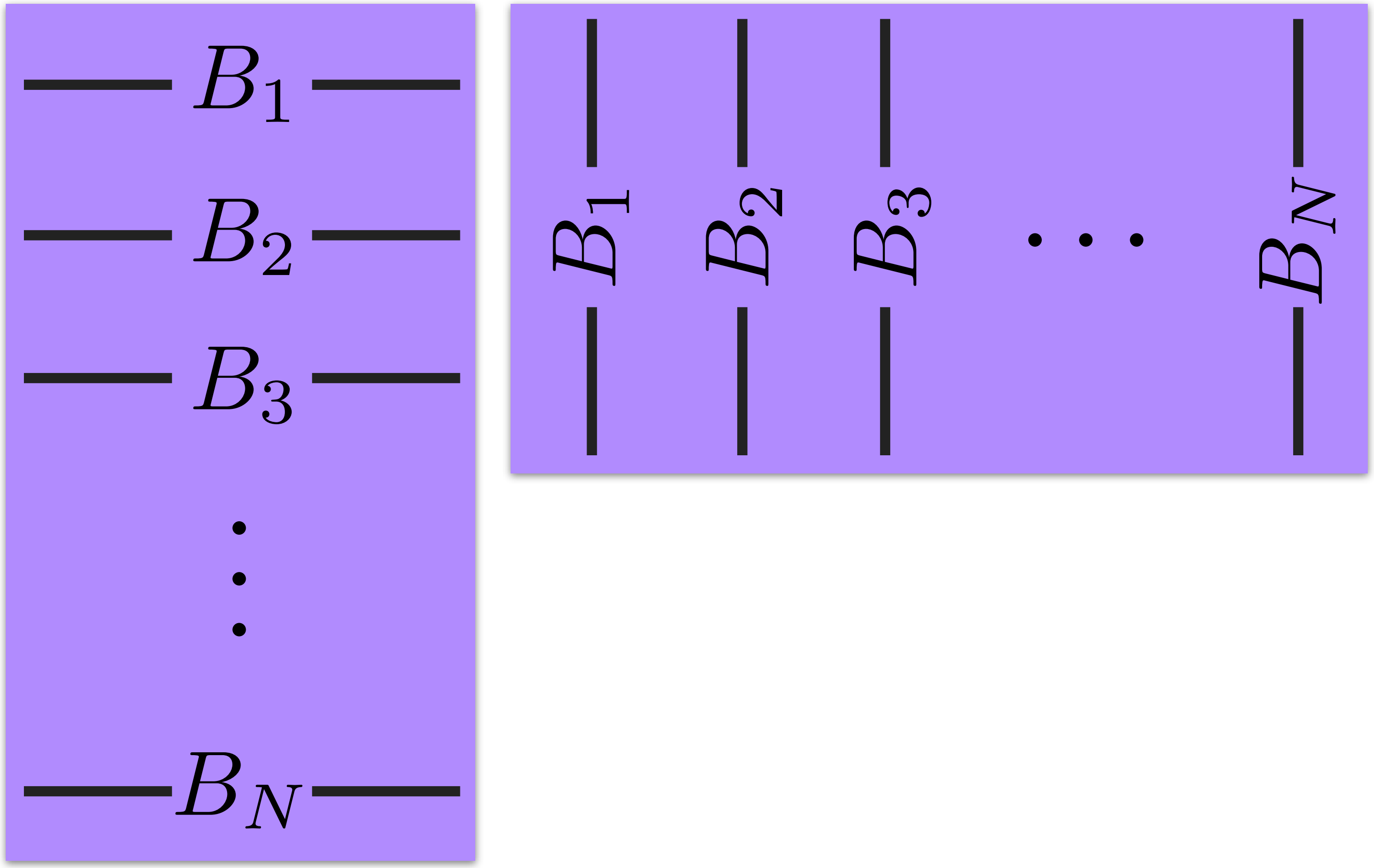
DUAL KERNEL

KULESZA AND TASKAR (NIPS 2010)

DUAL KERNEL

KULESZA AND TASKAR (NIPS 2010)

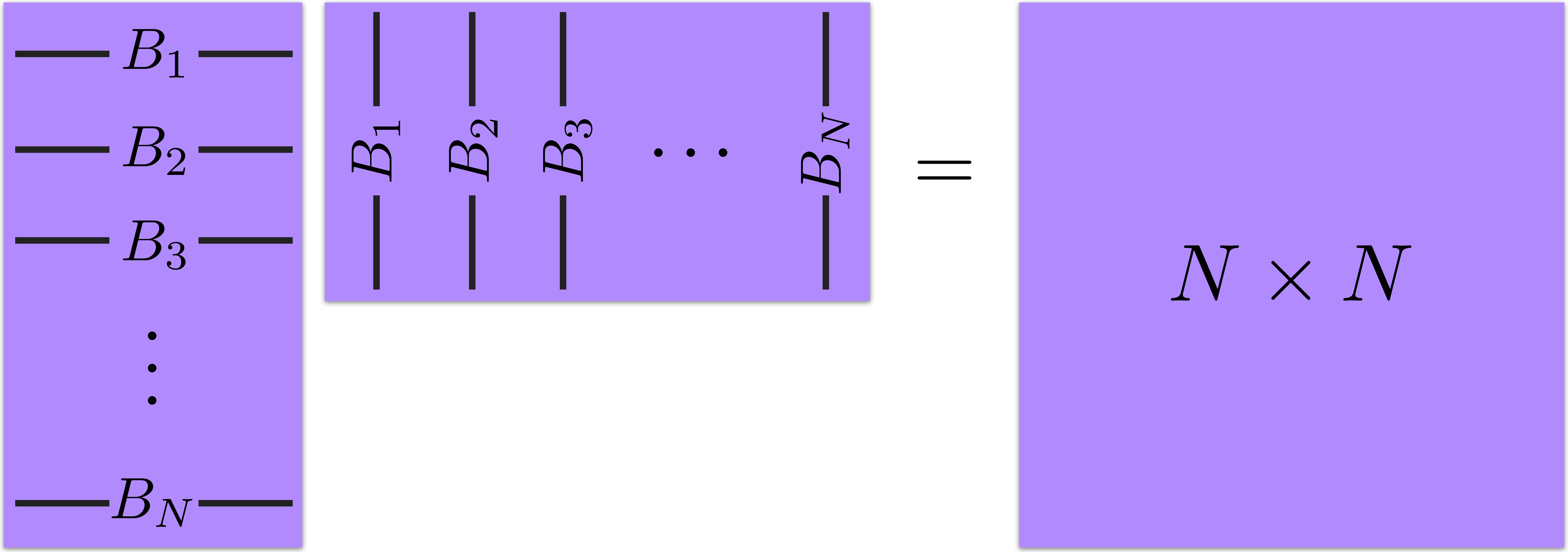
$$L$$



DUAL KERNEL

KULESZA AND TASKAR (NIPS 2010)

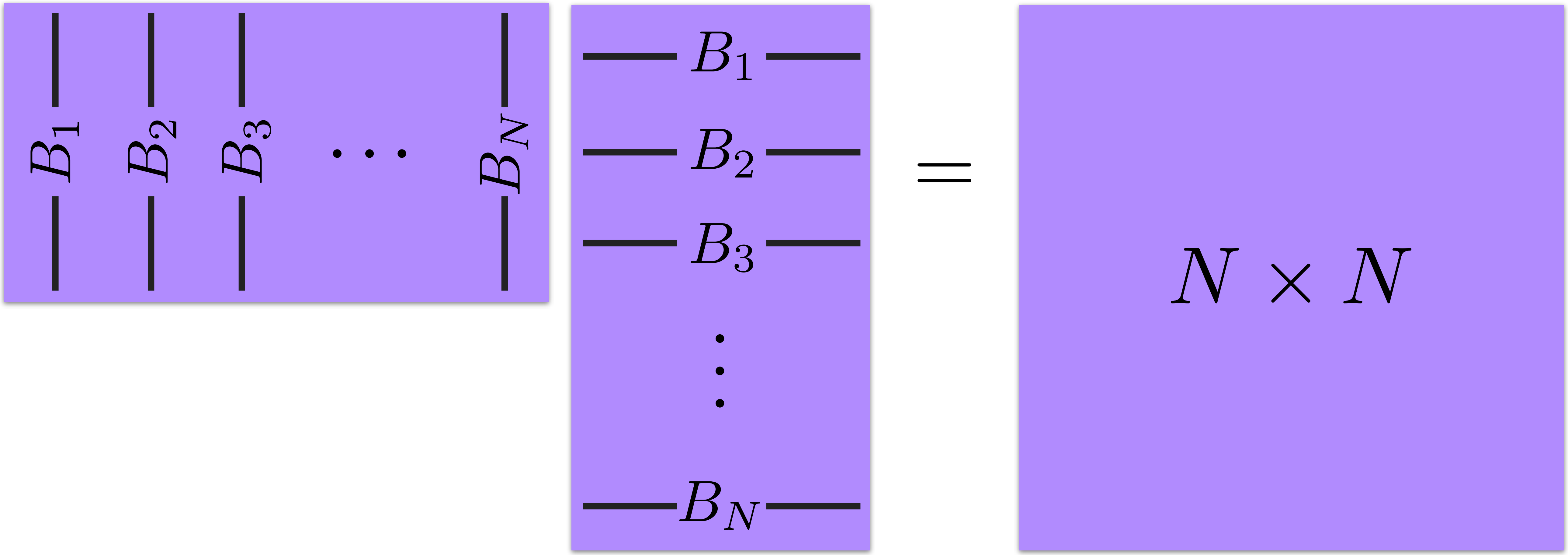
$$L$$



DUAL KERNEL

KULESZA AND TASKAR (NIPS 2010)

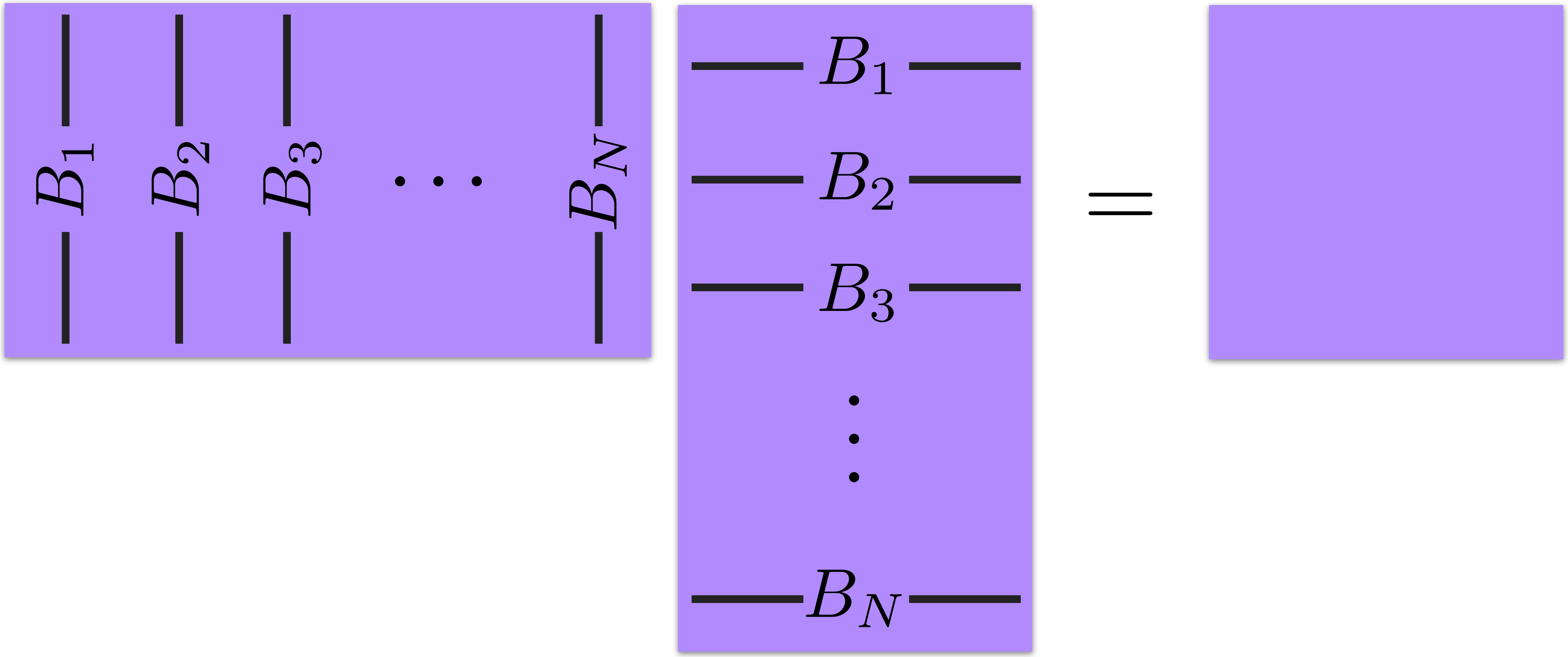
$$C$$



DUAL KERNEL

KULESZA AND TASKAR (NIPS 2010)

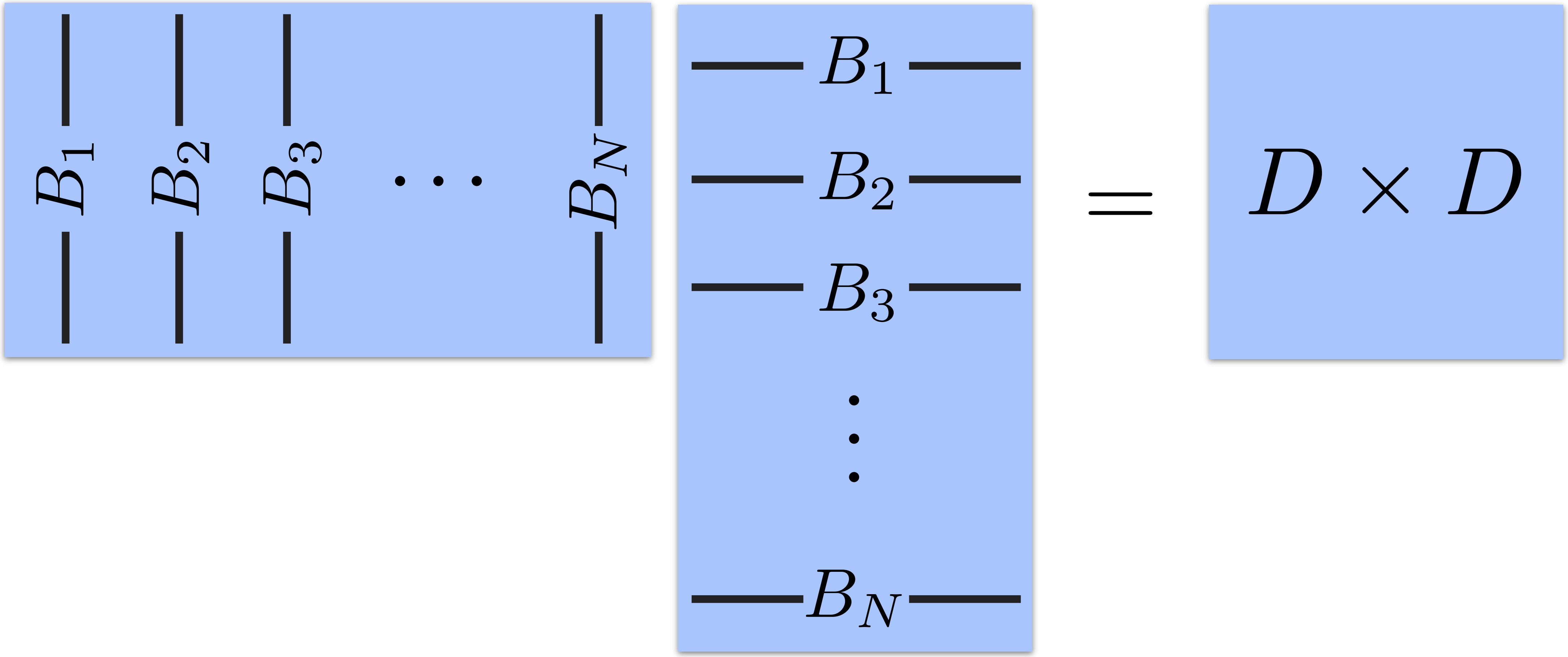
$$C$$



DUAL KERNEL

KULESZA AND TASKAR (NIPS 2010)

$$C$$



DUAL INFERENCE

DUAL INFERENCE

$L = V \Lambda V^\top$

$C = \hat{V} \Lambda \hat{V}^\top$

DUAL INFERENCE

$$L = V \Lambda V^\top \quad \overset{V = B^\top \hat{V} \Lambda^{-\frac{1}{2}}}{\Longleftrightarrow} \quad C = \hat{V} \Lambda \hat{V}^\top$$

DUAL INFERENCE

$L = V \Lambda V^\top$ $\quad V = B^\top \hat{V} \Lambda^{-\frac{1}{2}} \quad$ $C = \hat{V} \Lambda \hat{V}^\top$

Normalizing $\sum_Y \det(L_Y)$ $\quad O(D^3)$

DUAL INFERENCE

$$L = V \Lambda V^\top \quad \xLeftrightarrow{\;V = B^\top \hat{V} \Lambda^{-\frac{1}{2}}\;} \quad C = \hat{V} \Lambda \hat{V}^\top$$

Normalizing $\sum_Y \det(L_Y)$ $O(D^3)$

Marginalizing & Conditioning $O(D^3 + D^2 k^2)$

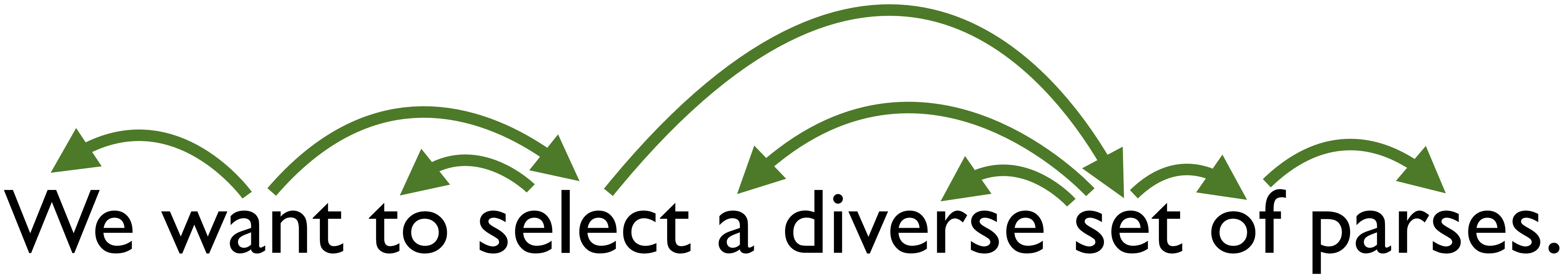
DUAL INFERENCE

$$L = V\Lambda V^\top \quad \xLeftrightarrow{\; V = B^\top \hat{V}\Lambda^{-\frac{1}{2}} \;} \quad C = \hat{V}\Lambda\hat{V}^\top$$

Normalizing $\sum_Y \det(L_Y)$	$O(D^3)$
Marginalizing & Conditioning	$O(D^3 + D^2k^2)$
Sampling $\mathbf{Y} \sim \mathcal{P}_L$	$O(ND^2k)$

EXPONENTIAL N

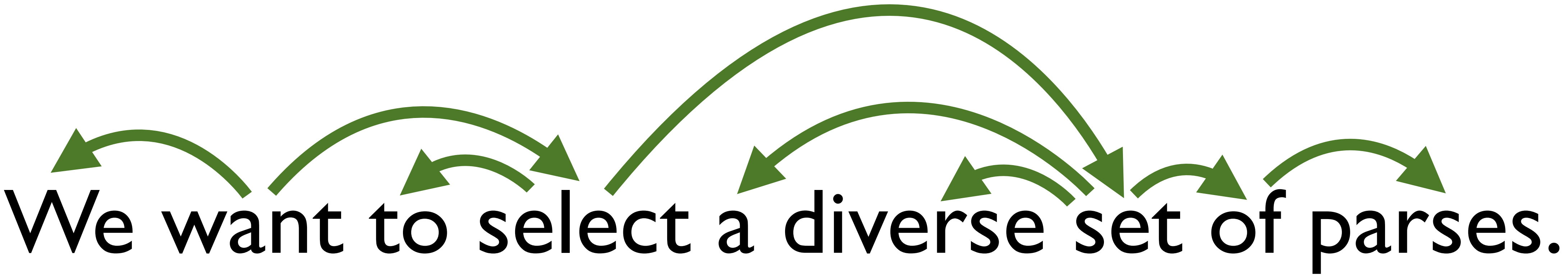
EXPONENTIAL N



We want to select a diverse set of parses.

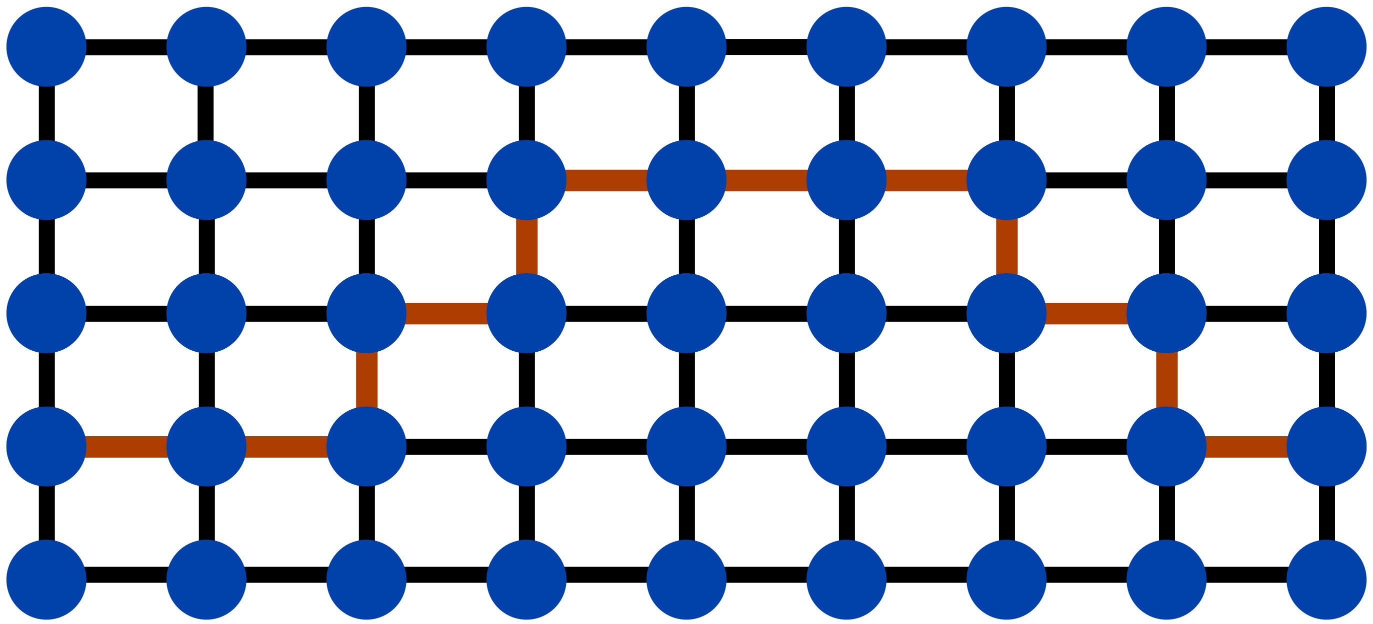
$$N = O(\{\text{sentence length}\}^{\{\text{sentence length}\}})$$

EXPONENTIAL N



We want to select a diverse set of parses.

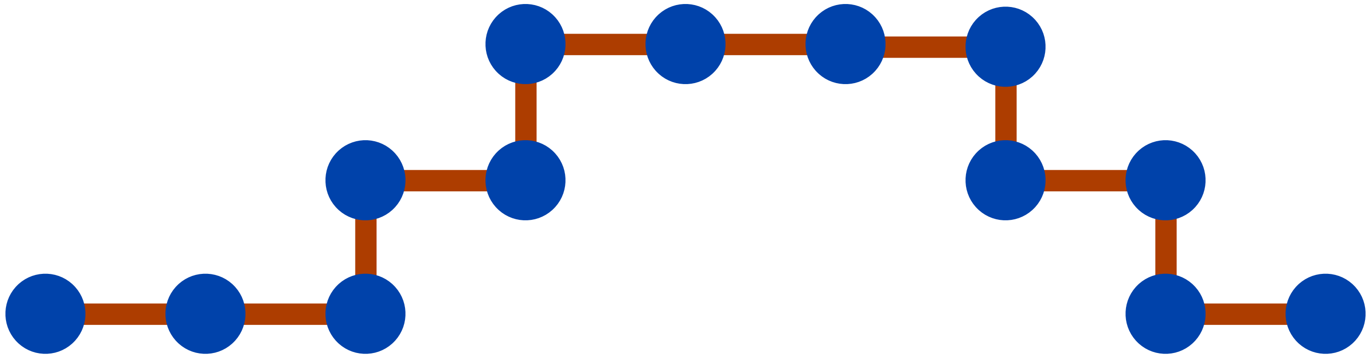
$$N = O(\{\text{sentence length}\}^{\{\text{sentence length}\}})$$



$$N = O(\{\text{node degree}\}^{\{\text{path length}\}})$$

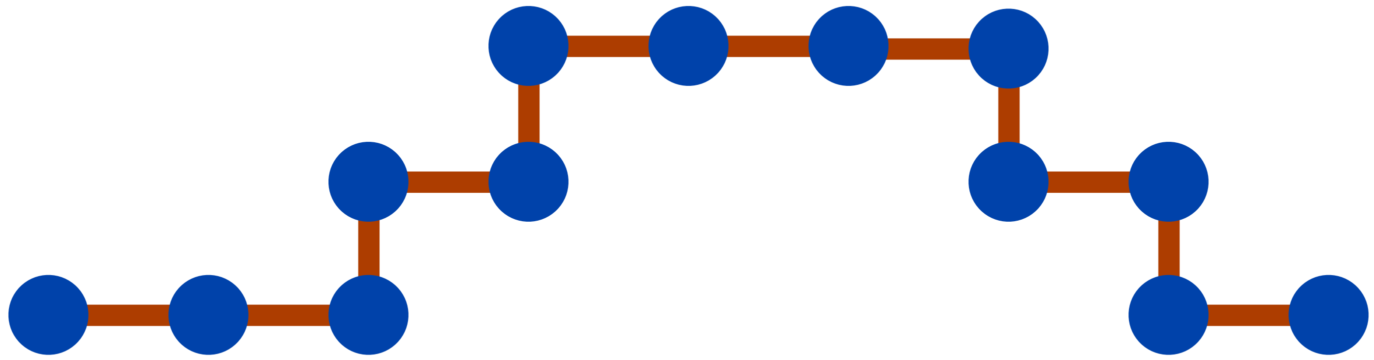
STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$\mathbf{i} =$ 

STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$\mathbf{i} =$ 

$$B_{\mathbf{i}} = q(\mathbf{i})\phi(\mathbf{i})$$

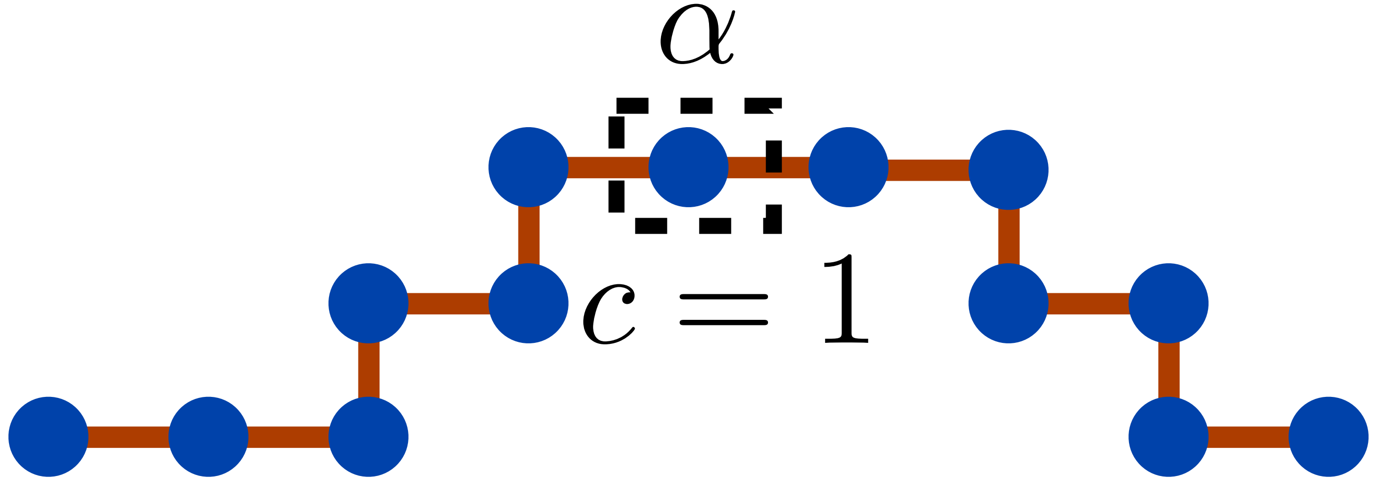
quality similarity

STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$$\mathbf{i} = \{i_\alpha\}_{\alpha \in F}$$

$$\mathbf{i} =$$



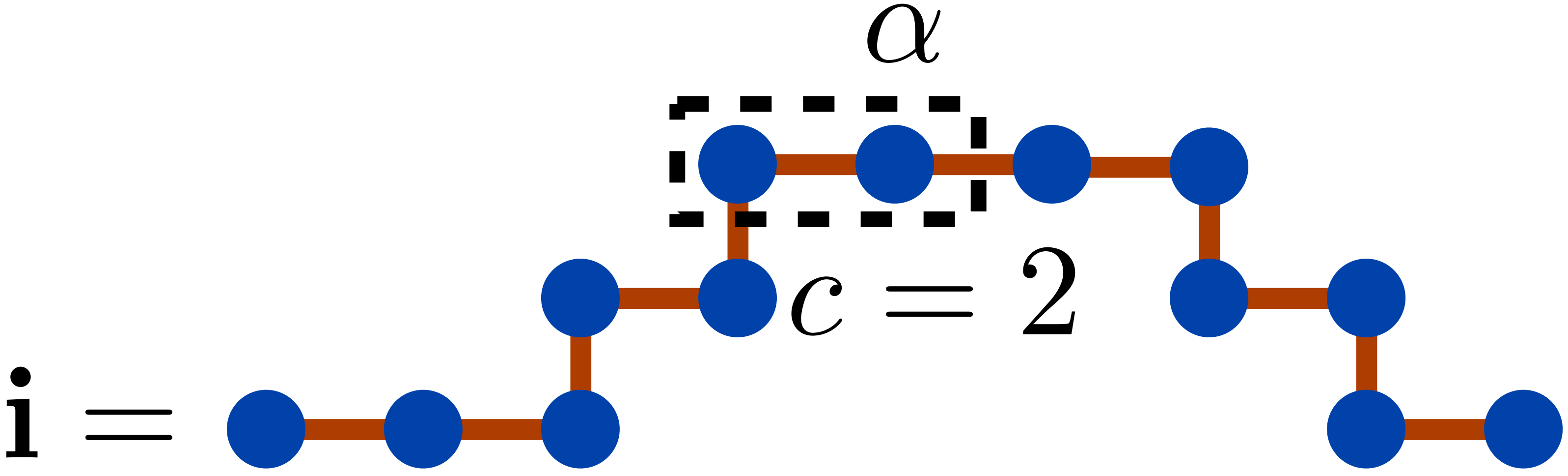
$$B_{\mathbf{i}} = q(\mathbf{i})\phi(\mathbf{i})$$

quality similarity

STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$$\mathbf{i} = \{i_\alpha\}_{\alpha \in F}$$



$$B_{\mathbf{i}} = q(\mathbf{i})\phi(\mathbf{i})$$

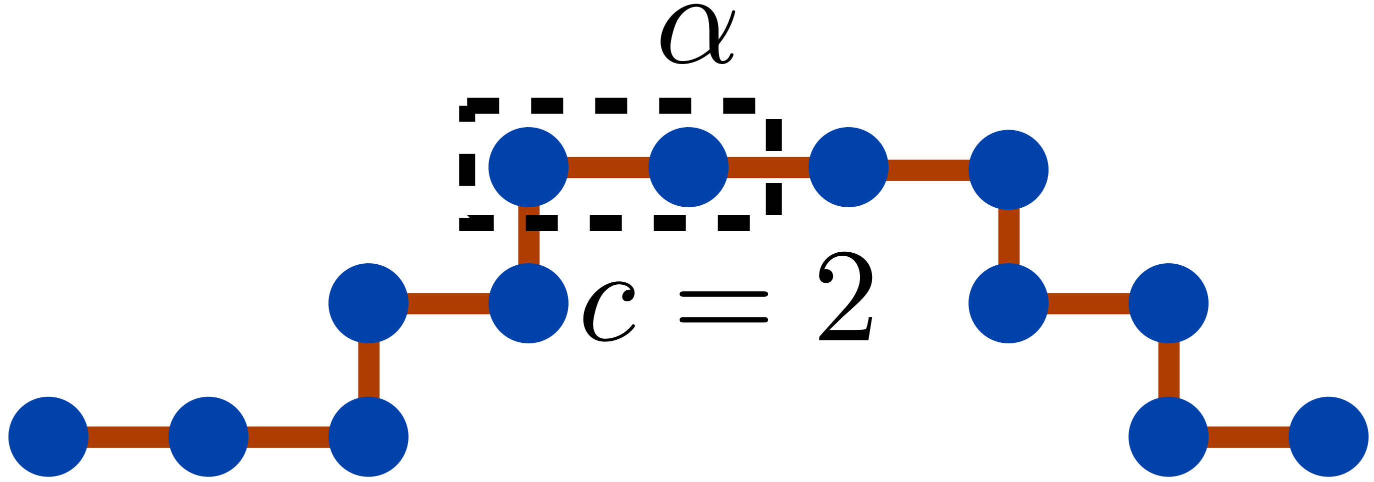
quality similarity

STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$$\mathbf{i} = \{i_\alpha\}_{\alpha \in F}$$

$\mathbf{i} =$



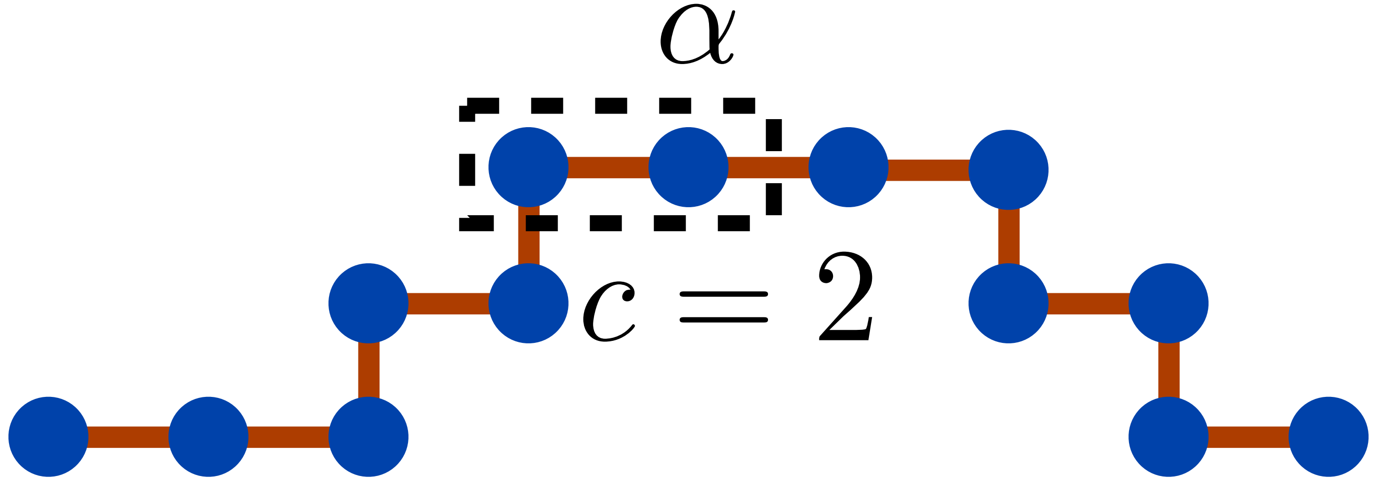
$$B_{\mathbf{i}} = \left[\prod_{\alpha \in F} q(i_\alpha)\right] \phi(\mathbf{i})$$

STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$$\mathbf{i} = \{i_\alpha\}_{\alpha \in F}$$

$\mathbf{i} =$

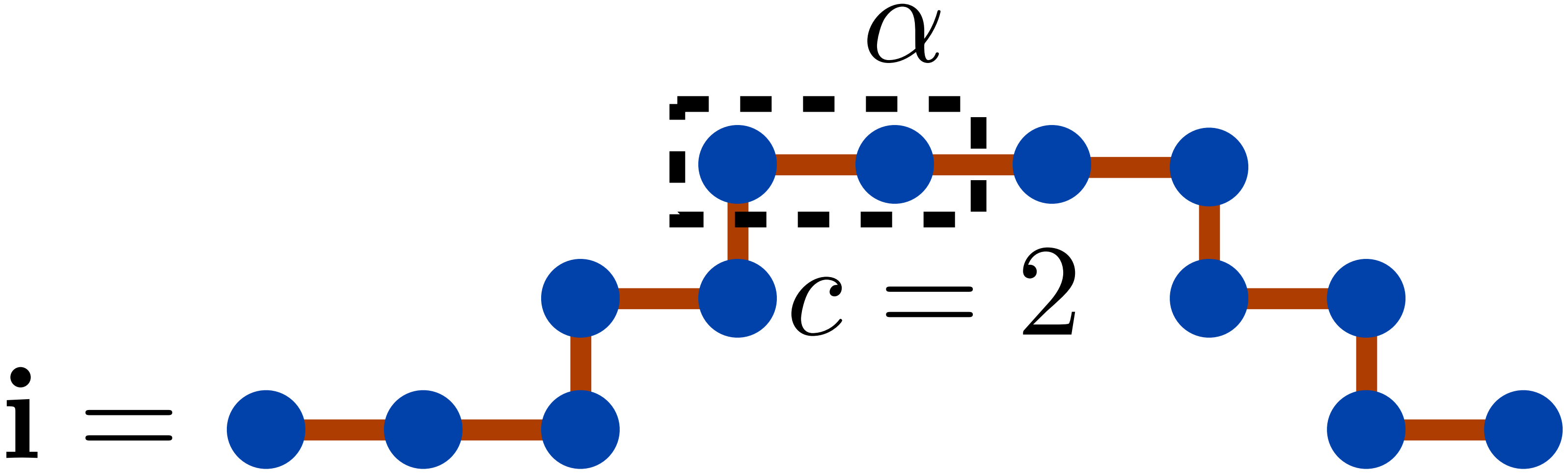


$$B_{\mathbf{i}} = \left[\prod_{\alpha \in F} q(i_\alpha)\right]\left[\sum_{\alpha \in F} \phi(i_\alpha)\right]$$

STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$$\mathbf{i} = \{i_\alpha\}_{\alpha \in F}$$



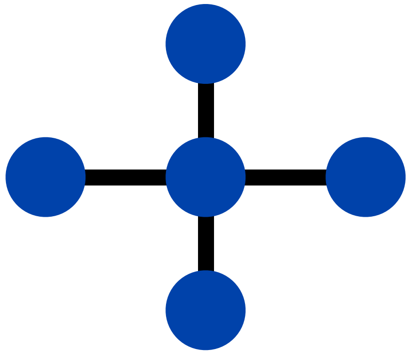
$$B_{\mathbf{i}} = \left[\prod_{\alpha \in F} q(i_\alpha)\right]\left[\sum_{\alpha \in F} \phi(i_\alpha)\right]$$

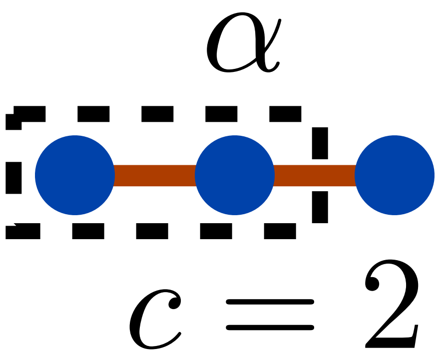
$$\mathbf{Y} \sim \mathcal{P}_L \qquad O(ND^2k)$$

STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$M =$





$R =$

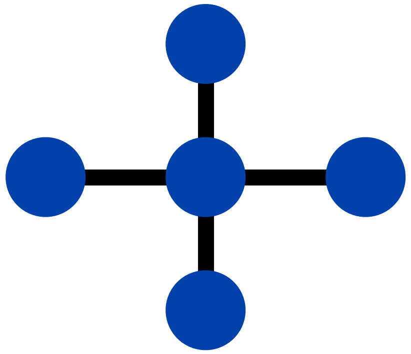
$$B_{\mathbf{i}} = \left[\prod_{\alpha\in F} q(i_\alpha)\right]\left[\sum_{\alpha\in F}\phi(i_\alpha)\right]$$

$$\mathbf{Y}\sim\mathcal{P}_L \quad O(D^2k^3 + Dk^2M^cR)$$

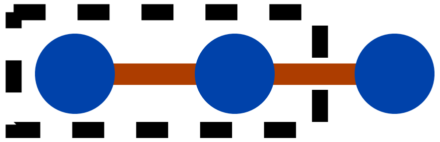
STRUCTURE FACTORIZATION

KULESZA AND TASKAR (NIPS 2010)

$M =$



α



$c = 2$

$R =$

$$B_{\mathbf{i}} = \left[\prod_{\alpha\in F} q(i_\alpha)\right]\left[\sum_{\alpha\in F} \phi(i_\alpha)\right]$$

$$\mathbf{Y} \sim \mathcal{P}_L \quad O(D^2k^3 + Dk^2M^cR)$$

$$M^cR = 4^2 * 12 = 192 \ll N = 4^{12} = 16{,}777{,}216$$

LARGE FEATURE SETS

LARGE FEATURE SETS

$N = \#\text{ of items}$

$D = \#\text{ of features}$

	Large	**Exponential**
Small	dual	dual + structure

LARGE FEATURE SETS

$N = \#$ of items

$D = \#$ of features

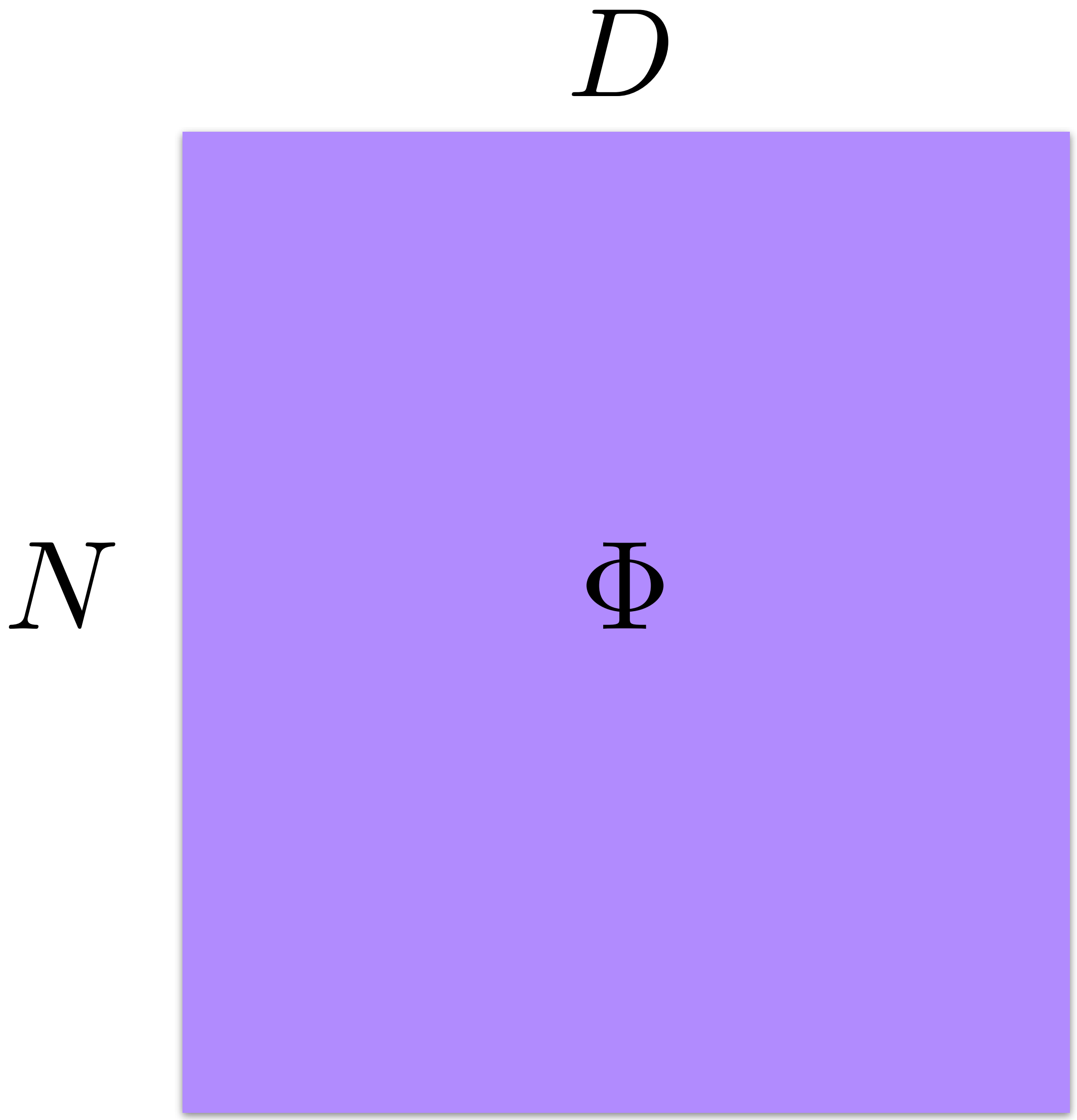
	Large	**Exponential**
Small	dual	dual + structure
Large	?	?

RANDOM PROJECTIONS

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

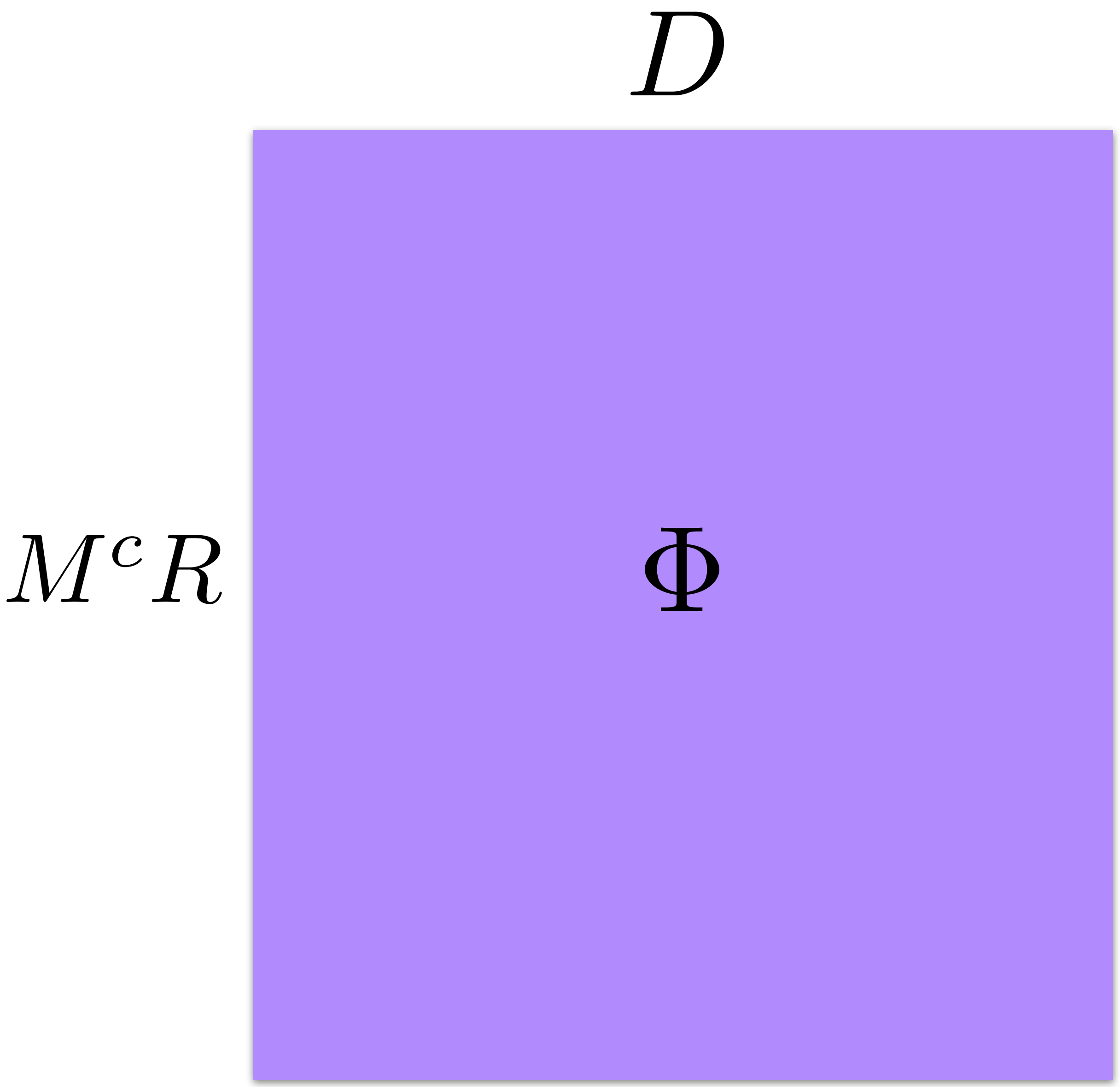
RANDOM PROJECTIONS

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)



RANDOM PROJECTIONS

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)



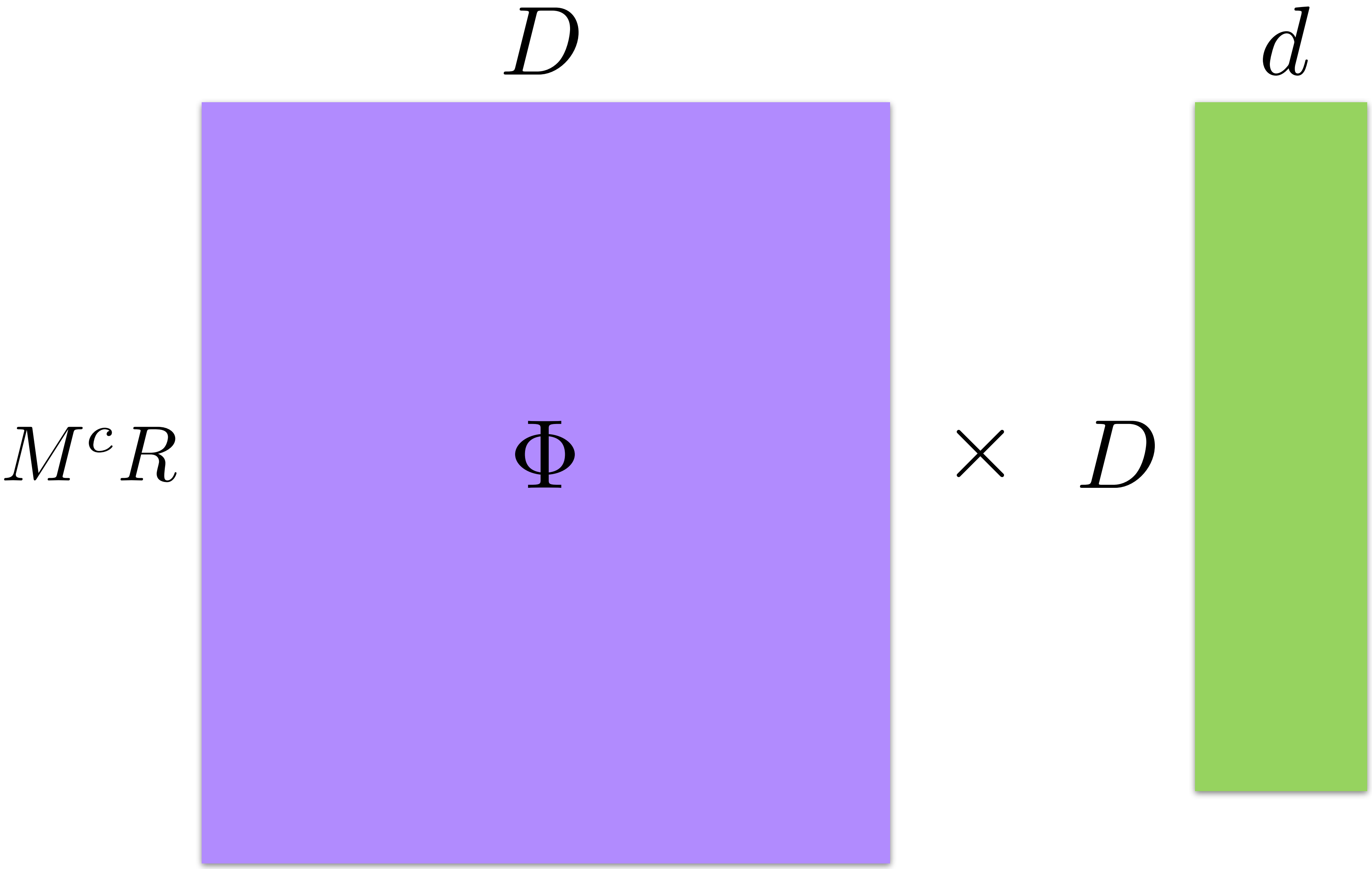
D

$M^c R$

Φ

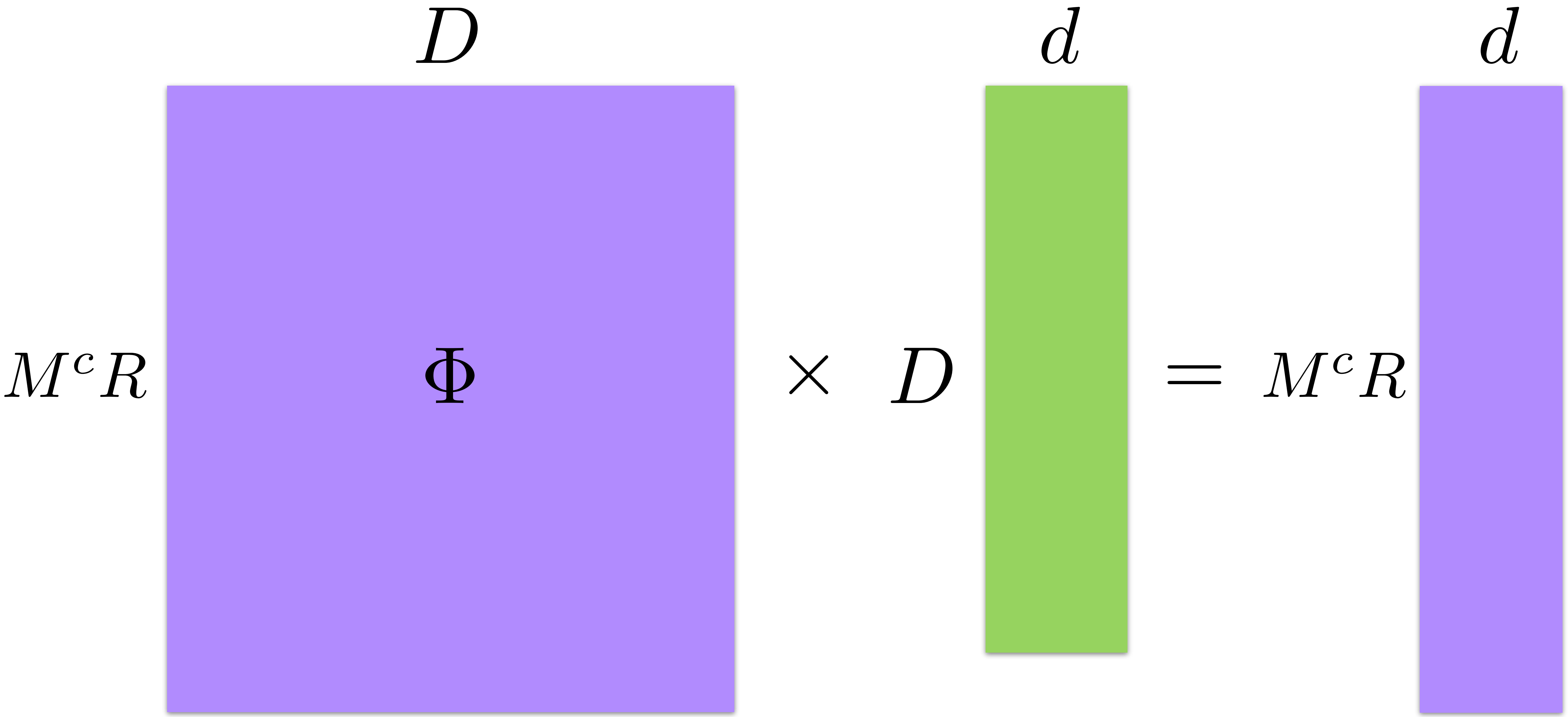
RANDOM PROJECTIONS

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)



RANDOM PROJECTIONS

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

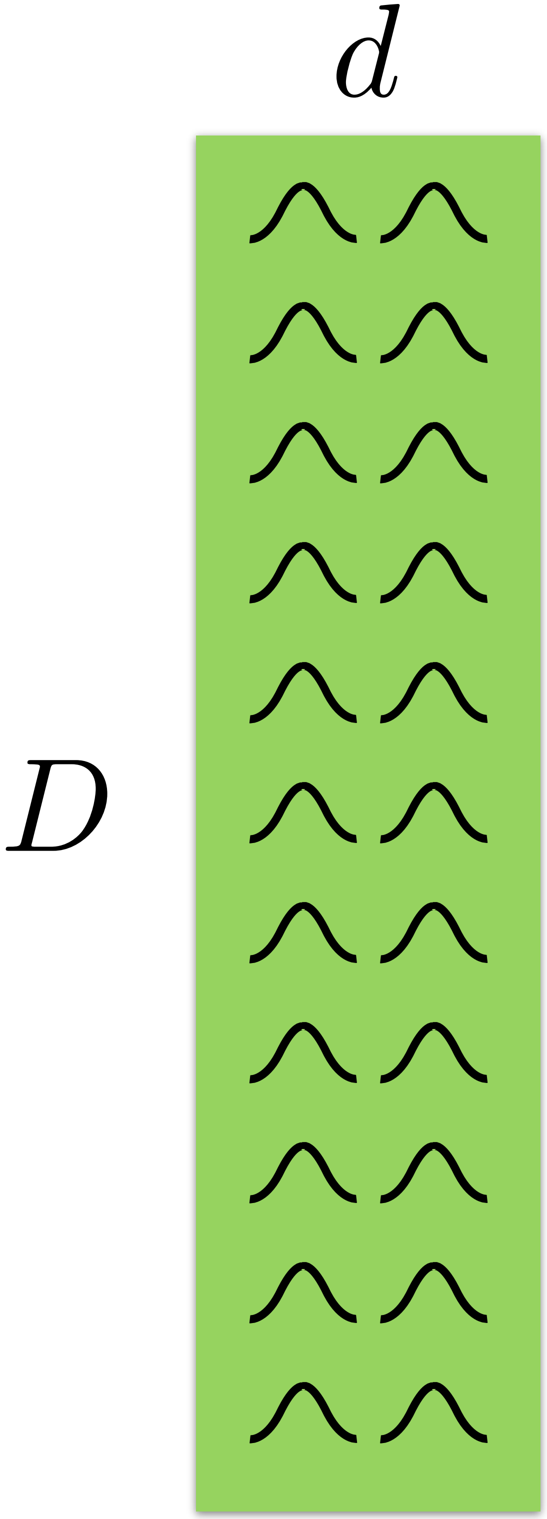


RANDOM PROJECTIONS

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

RANDOM PROJECTIONS

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)



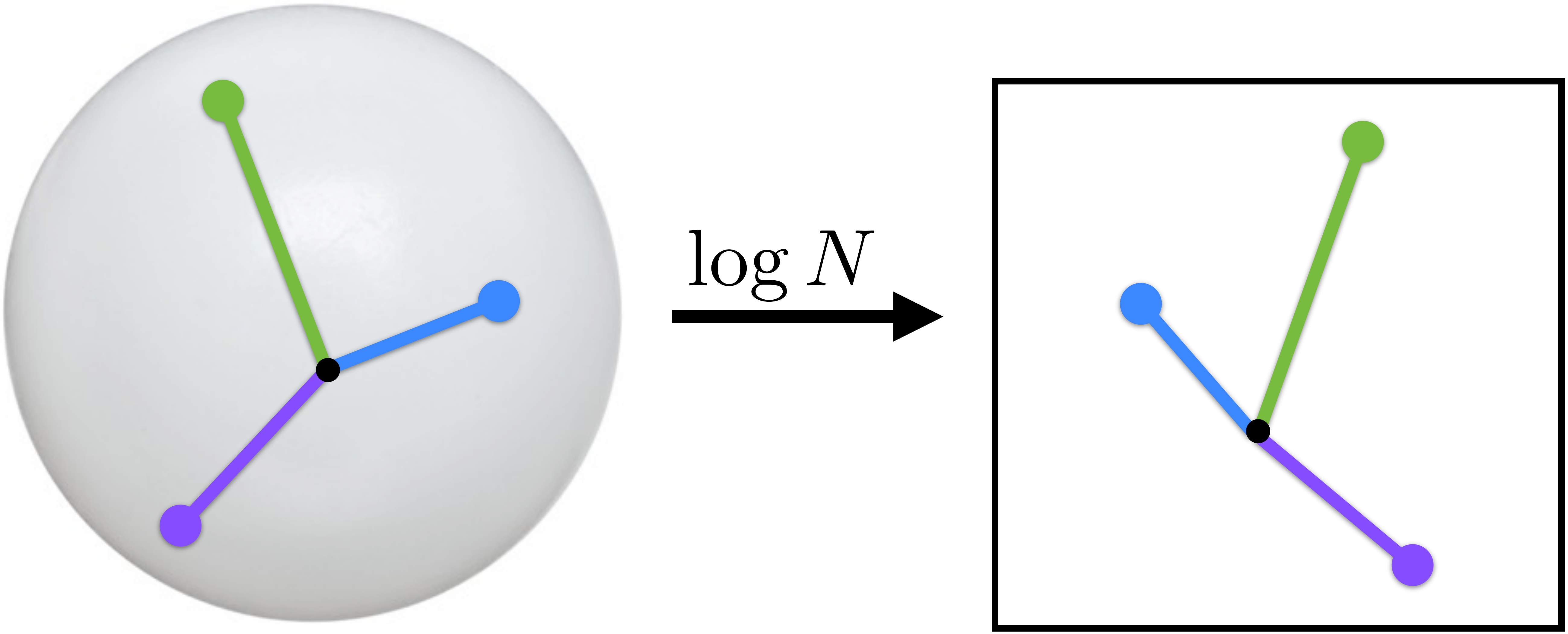
VOLUME PRESERVATION

VOLUME PRESERVATION

JOHNSON AND LINDENSTRAUSS (1984)

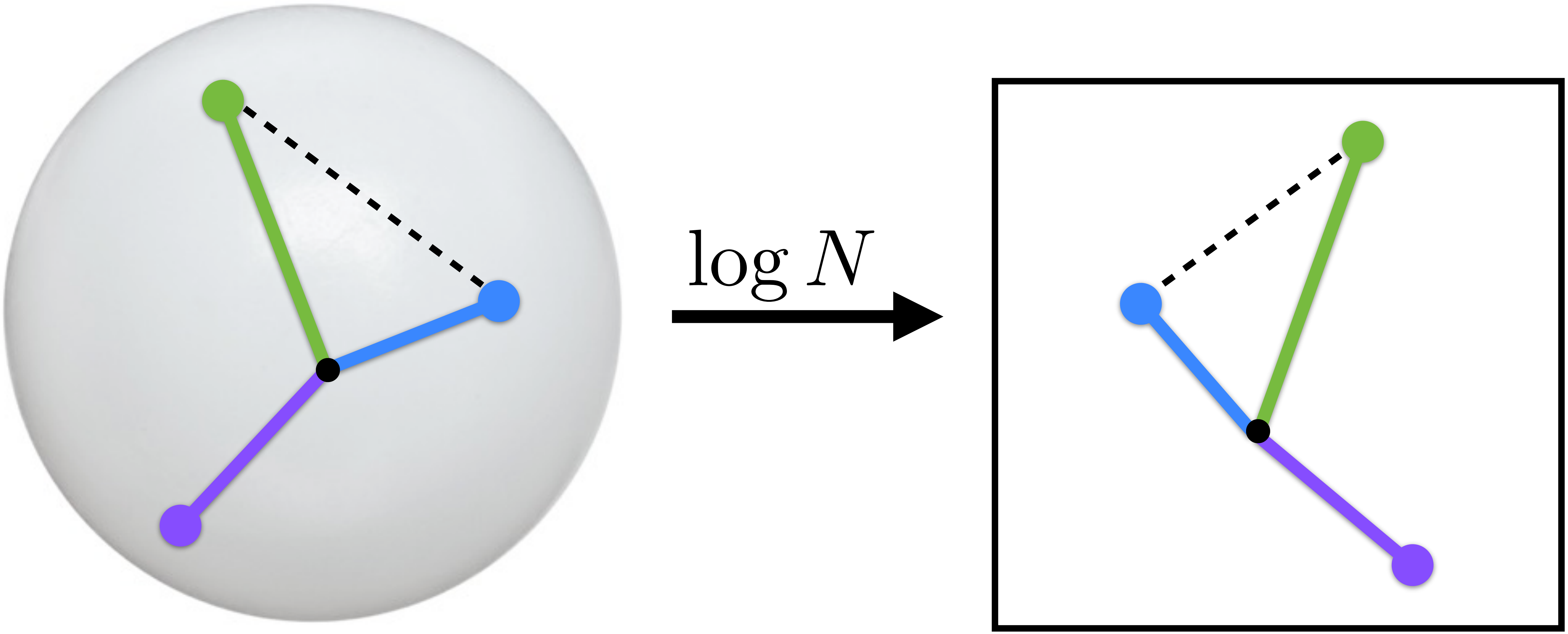
VOLUME PRESERVATION

JOHNSON AND LINDENSTRAUSS (1984)



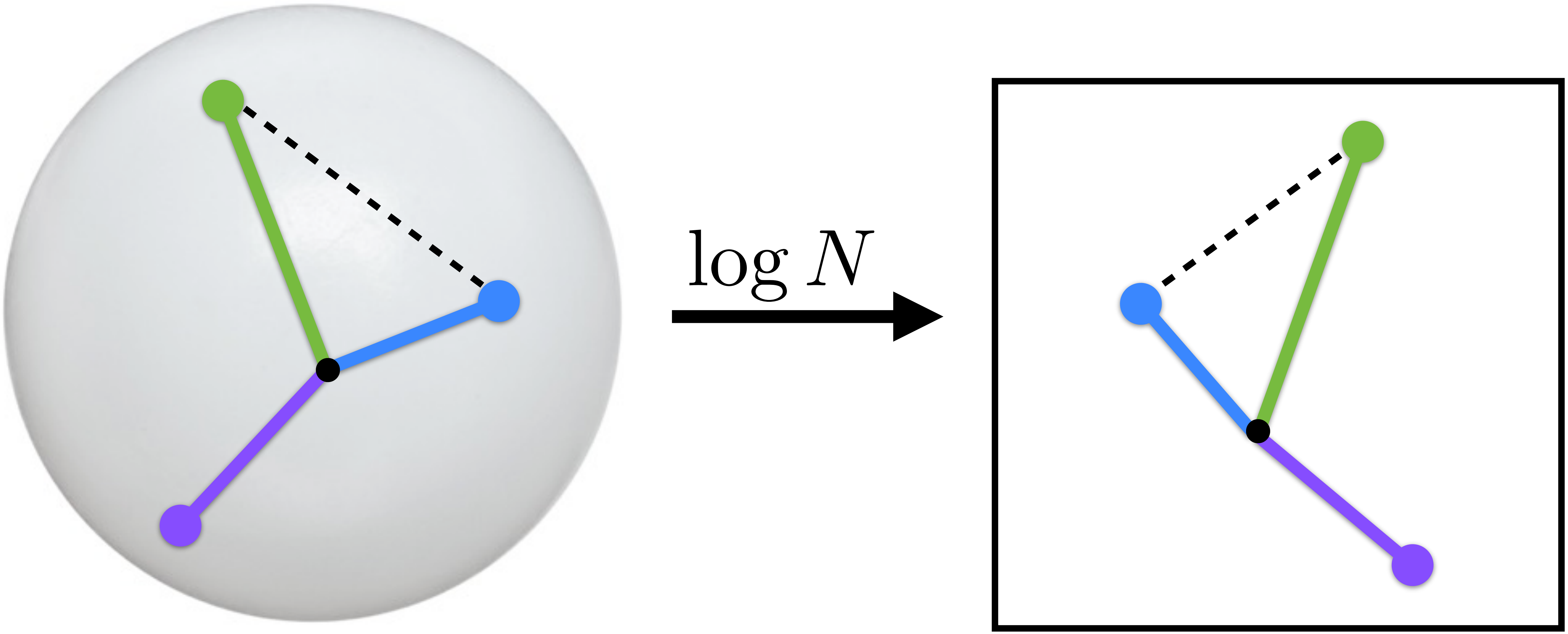
VOLUME PRESERVATION

JOHNSON AND LINDENSTRAUSS (1984)



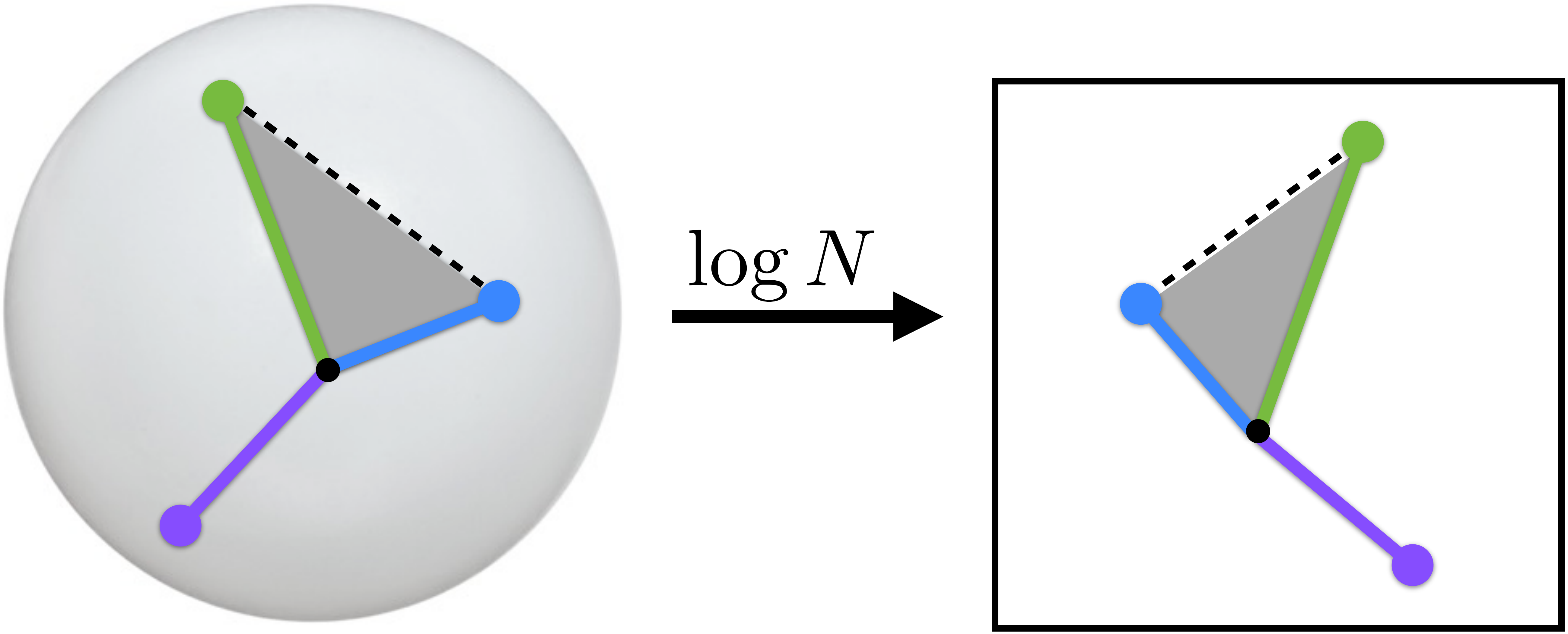
VOLUME PRESERVATION

MAGEN AND ZOUZIAS (2008)



VOLUME PRESERVATION

MAGEN AND ZOUZIAS (2008)



DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

$$\mathrm{vol}^2 = \det$$

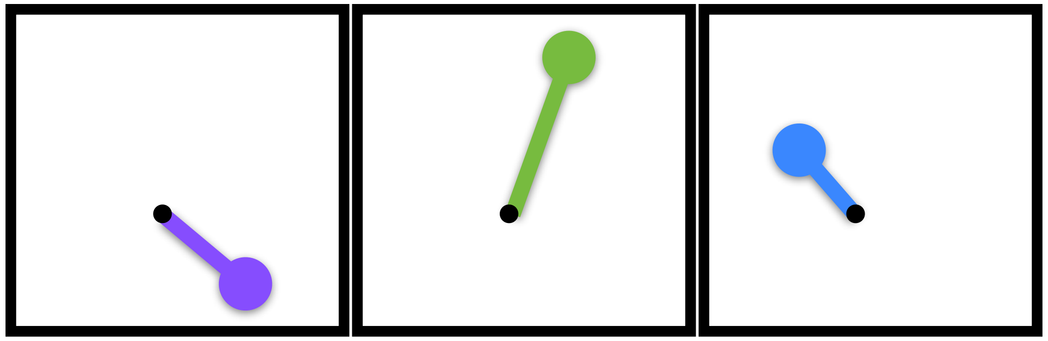
DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

$$\mathrm{vol}^2 = \det$$

$$k = 1$$

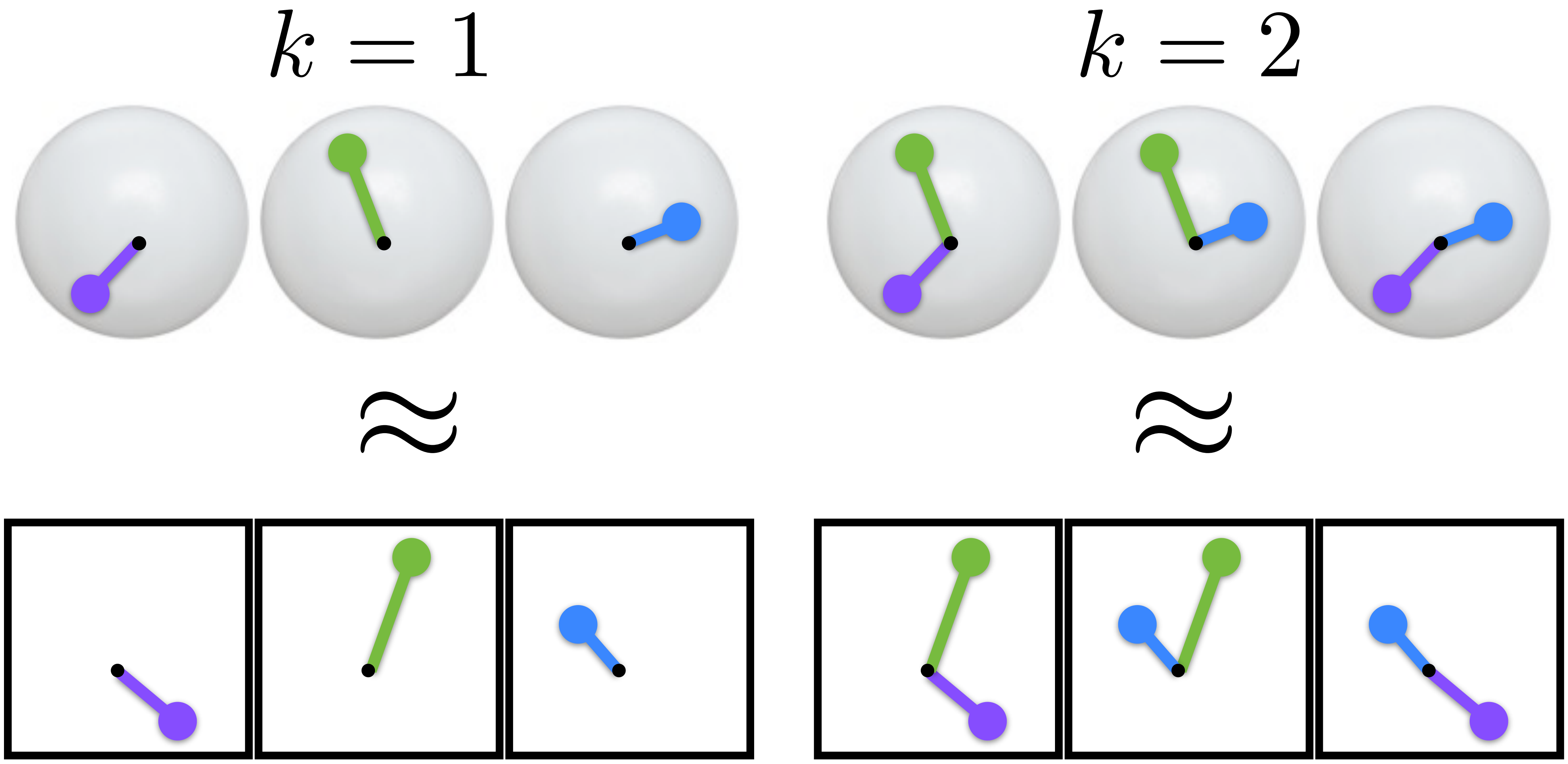
$$\approx$$



DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

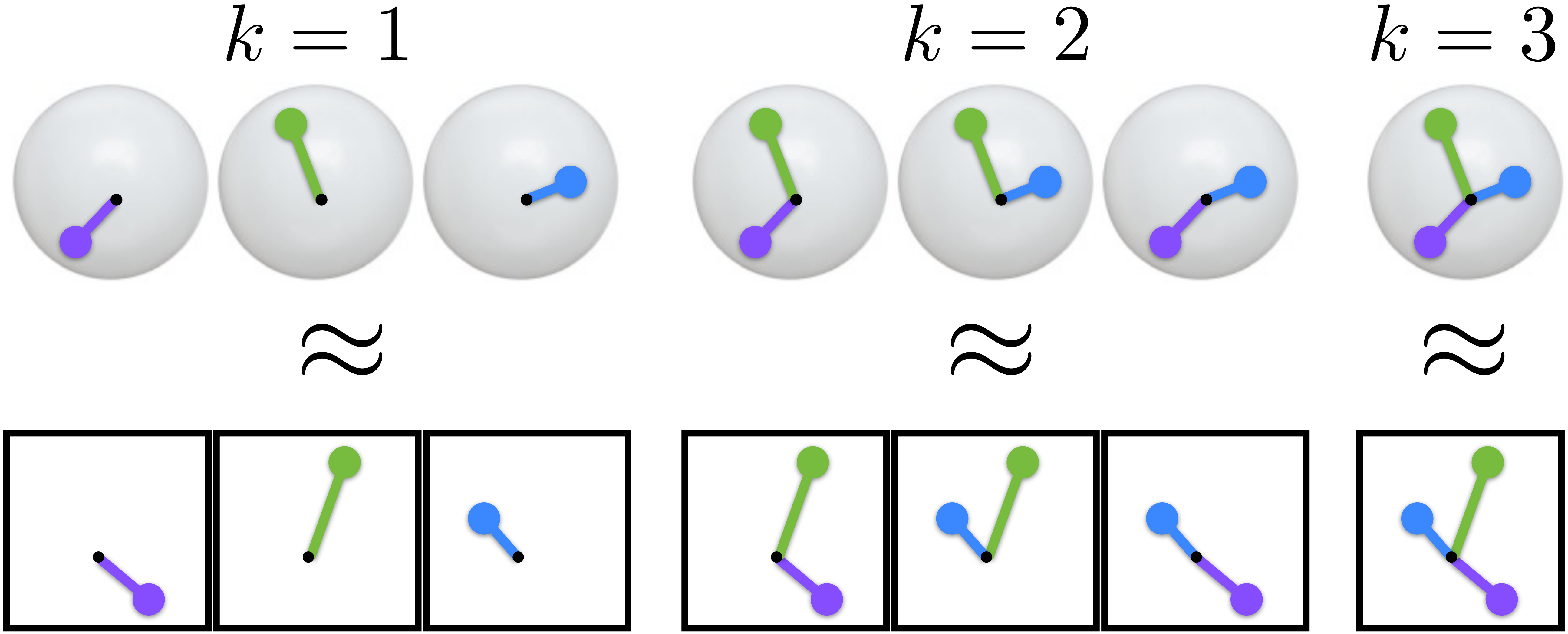
$$\mathrm{vol}^2 = \det$$



DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

$$\mathrm{vol}^2 = \det$$



DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

$$d = O\left(\max\left\{\frac{k}{\epsilon}, \frac{\log(1/\delta)+\log(N)}{\epsilon^2} + k\right\}\right)$$

DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

$$d = O\left(\max\left\{\frac{\overset{\text{subset size}}{\overset{\big\downarrow}{k}}}{\epsilon}, \frac{\log(1/\delta)+\log(\overset{\text{total \# of items}}{\overset{\big\downarrow}{N}})}{\epsilon^2} + k\right\}\right)$$

DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

$$d = O\left(\max\left\{\frac{\overset{\substack{\text{subset size}\\ \downarrow}}{k}}{\epsilon}, \frac{\log(1/\delta)+\log(\overset{\substack{\text{total \# of items}\\ \downarrow}}{N})}{\epsilon^2} + k\right\}\right)$$

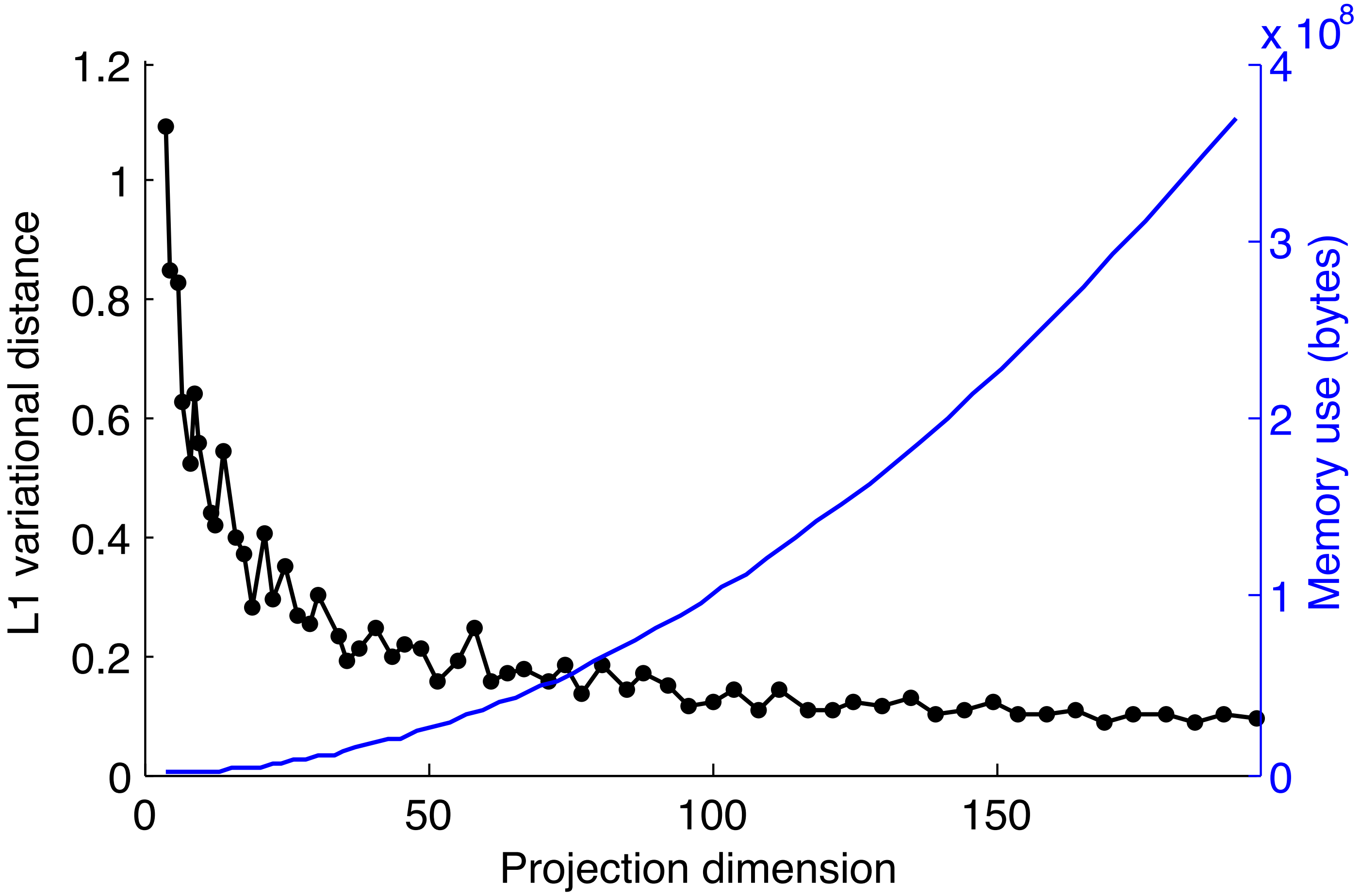
$$\text{w.p. } 1-\delta: \quad \|\mathcal{P}^k - \tilde{\mathcal{P}}^k\|_1 \le e^{6k\epsilon} - 1$$

DPP PRESERVATION

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

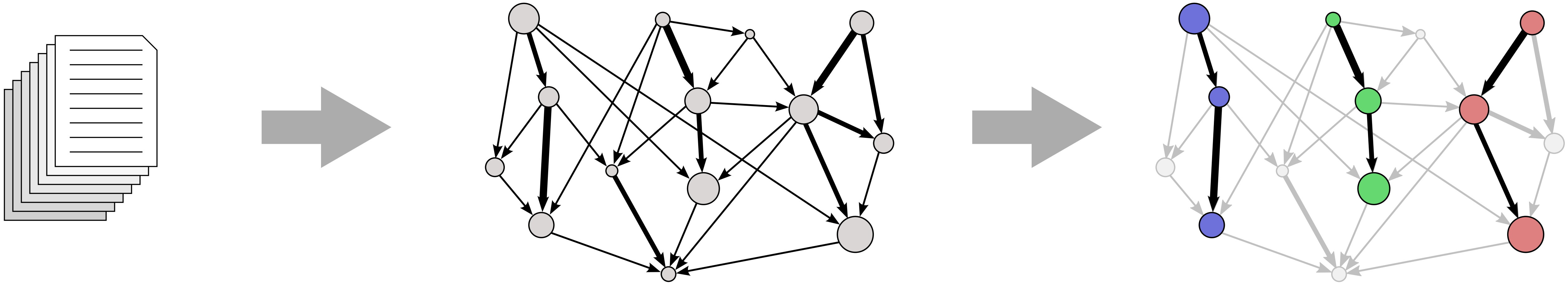
$$d = O\left(\max\left\{\overset{\substack{\text{subset size}\\ \downarrow}}{\frac{k}{\epsilon}}, \frac{\log(1/\delta)+\log(\overset{\substack{\text{total \# of items}\\ \downarrow}}{N})}{\epsilon^2} + k\right\}\right)$$

$$\text{w.p. } 1-\delta: \quad \|\mathcal{P}^k - \tilde{\mathcal{P}}^k\|_1 \leq e^{6k\epsilon} - 1$$

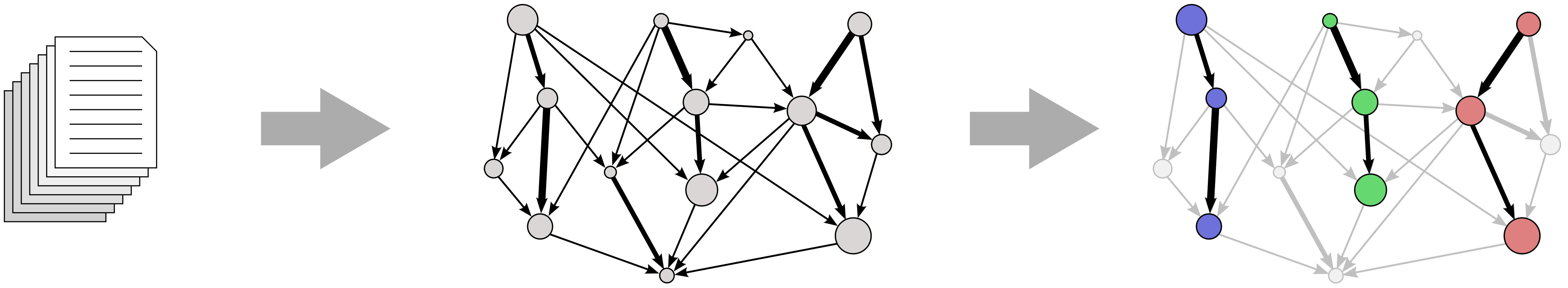


NEWS THREADING

NEWS THREADING

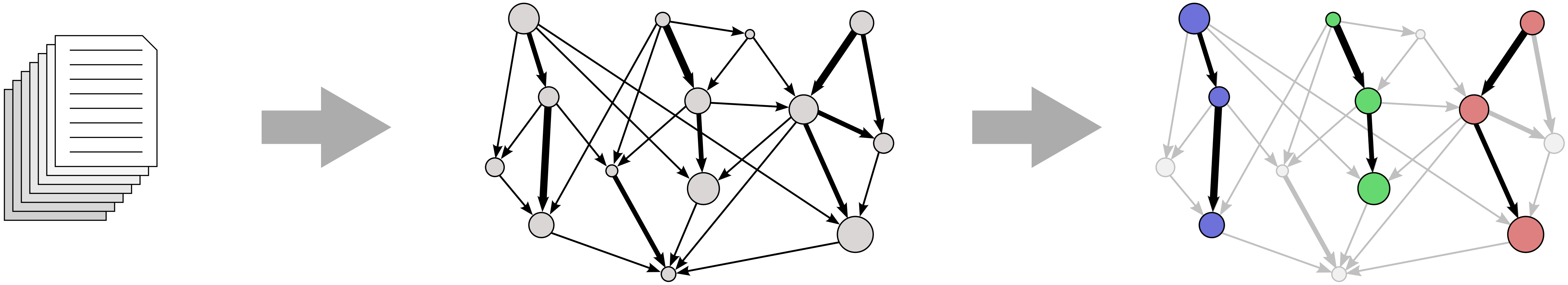


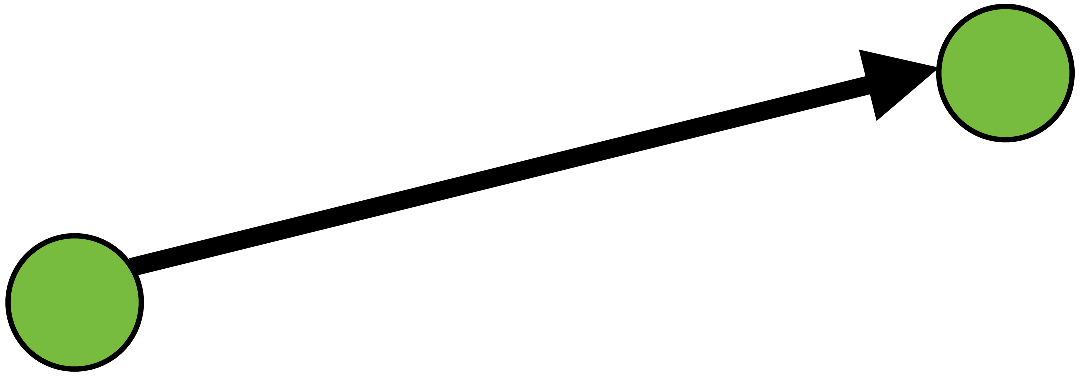
NEWS THREADING



March 28: Health officials confirm Ebola outbreak in Guinea's capital

NEWS THREADING

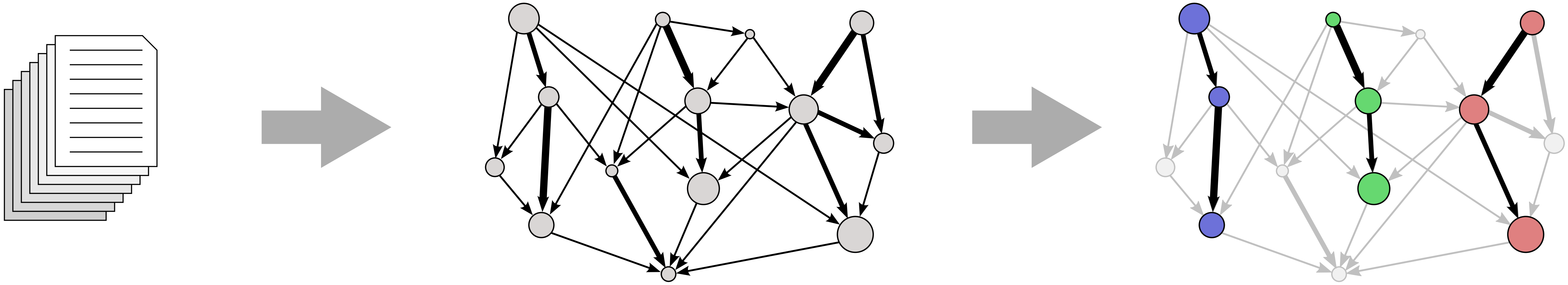


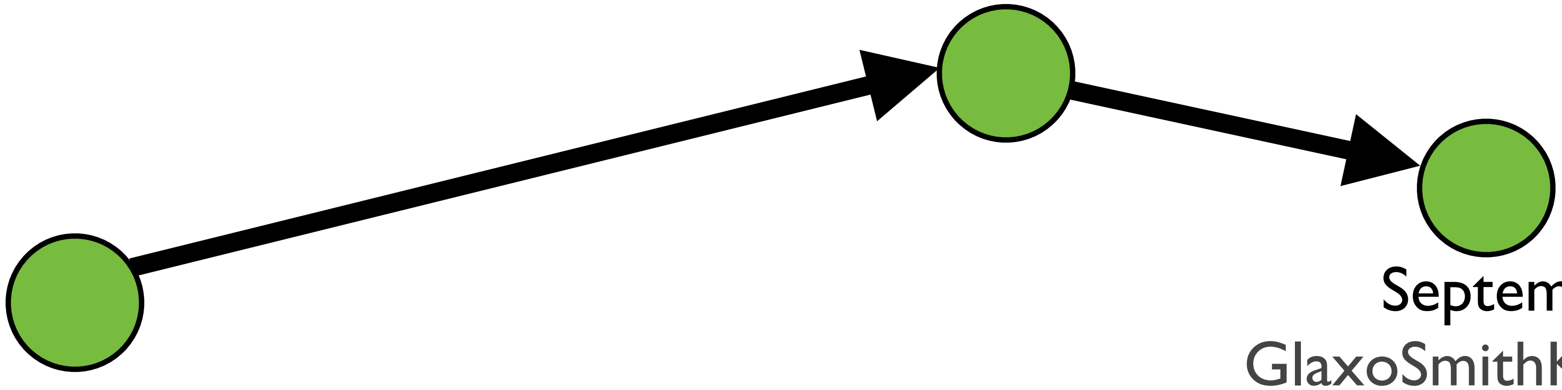


August 8: World Health Organization declares Ebola epidemic an international health emergency

March 28: Health officials confirm Ebola outbreak in Guinea's capital

NEWS THREADING



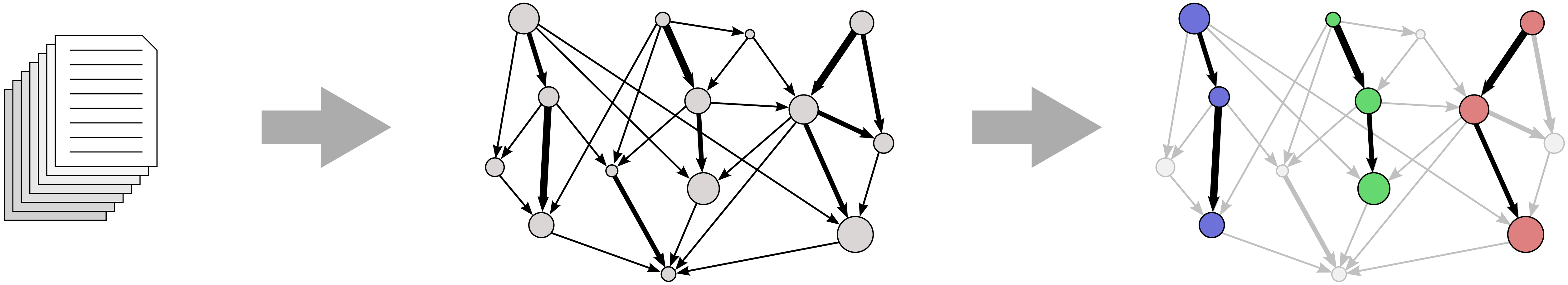


August 8: World Health Organization declares Ebola epidemic an international health emergency

September 2: GlaxoSmithKlein begins Ebola vaccine drug trial

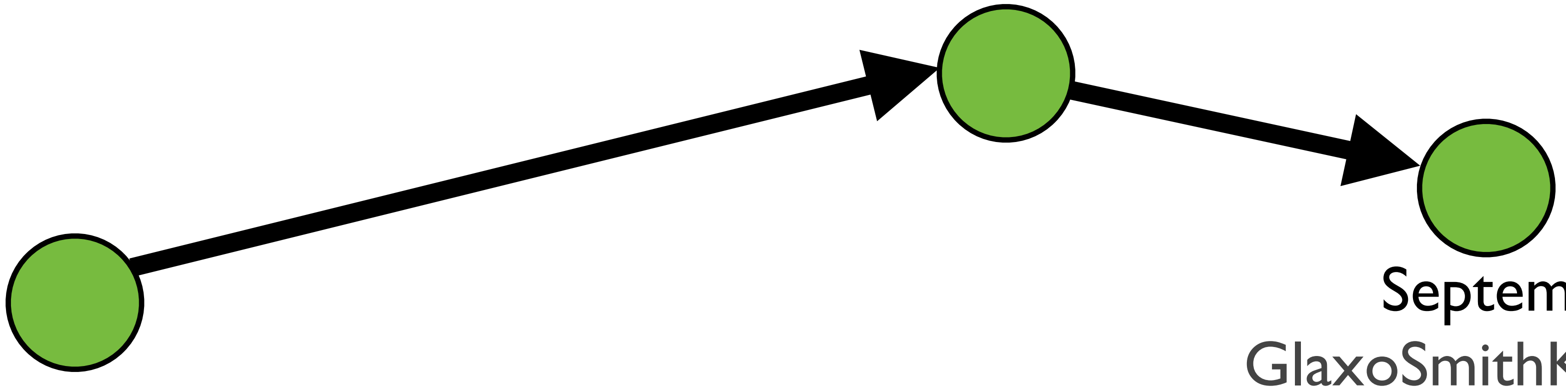
March 28: Health officials confirm Ebola outbreak in Guinea's capital

NEWS THREADING



$M \approx 35{,}000$

10^{360}



August 8: World Health Organization declares Ebola epidemic an international health emergency

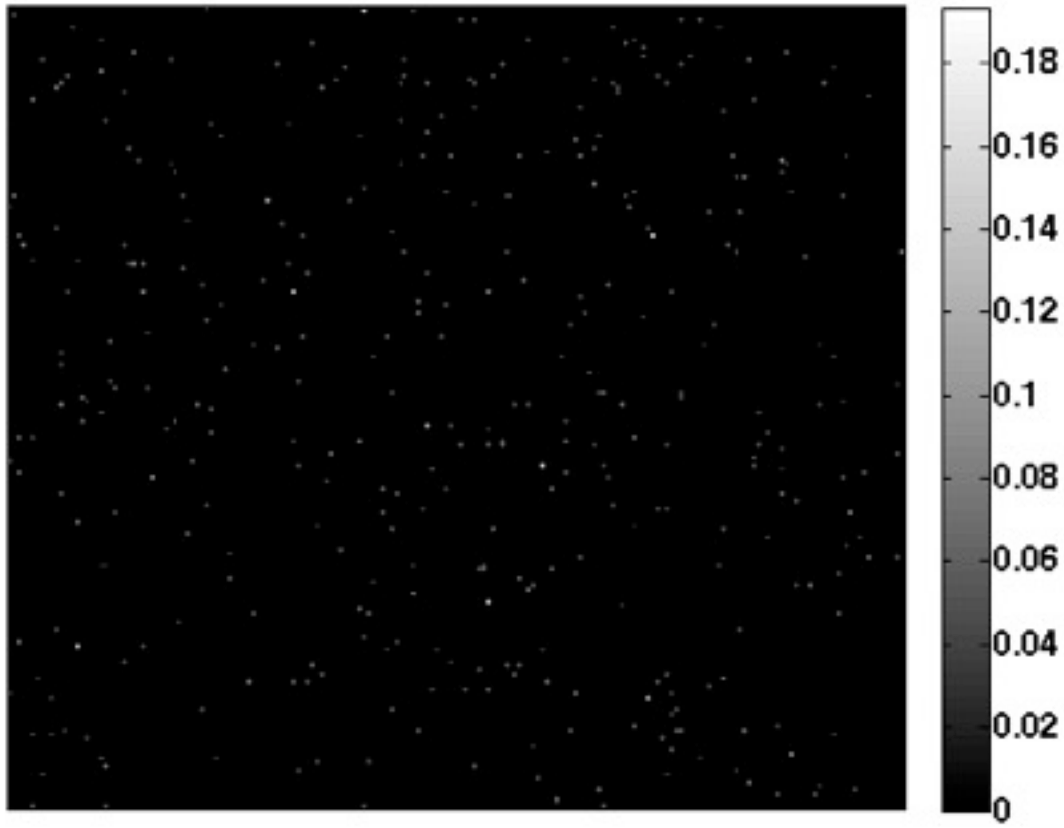
March 28: Health officials confirm Ebola outbreak in Guinea's capital

September 2: GlaxoSmithKlein begins Ebola vaccine drug trial

PROJECTING NEWS FEATURES

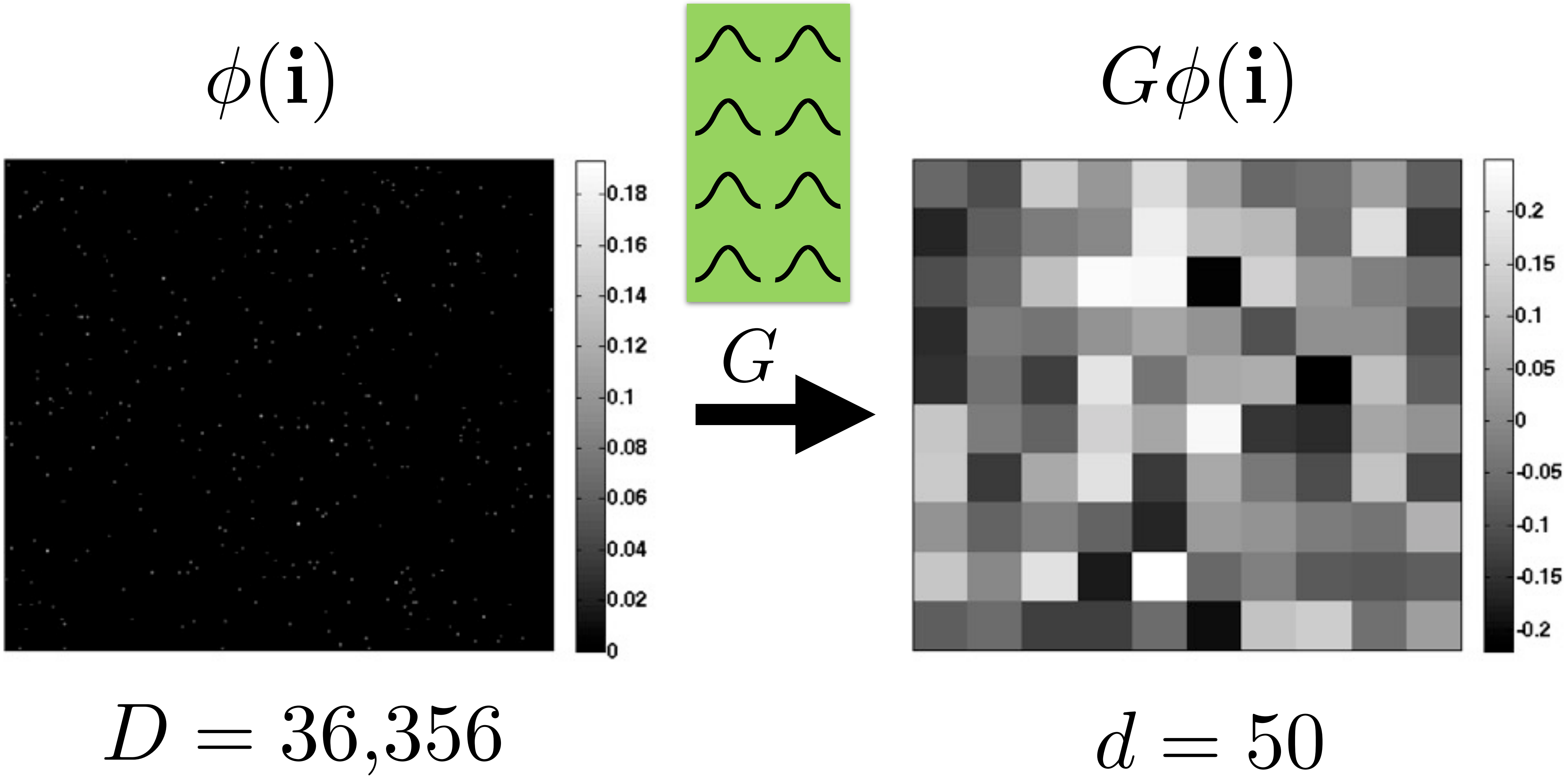
PROJECTING NEWS FEATURES

$\phi(\mathbf{i})$



$D = 36{,}356$

PROJECTING NEWS FEATURES



$\phi(\mathbf{i})$

$G\phi(\mathbf{i})$

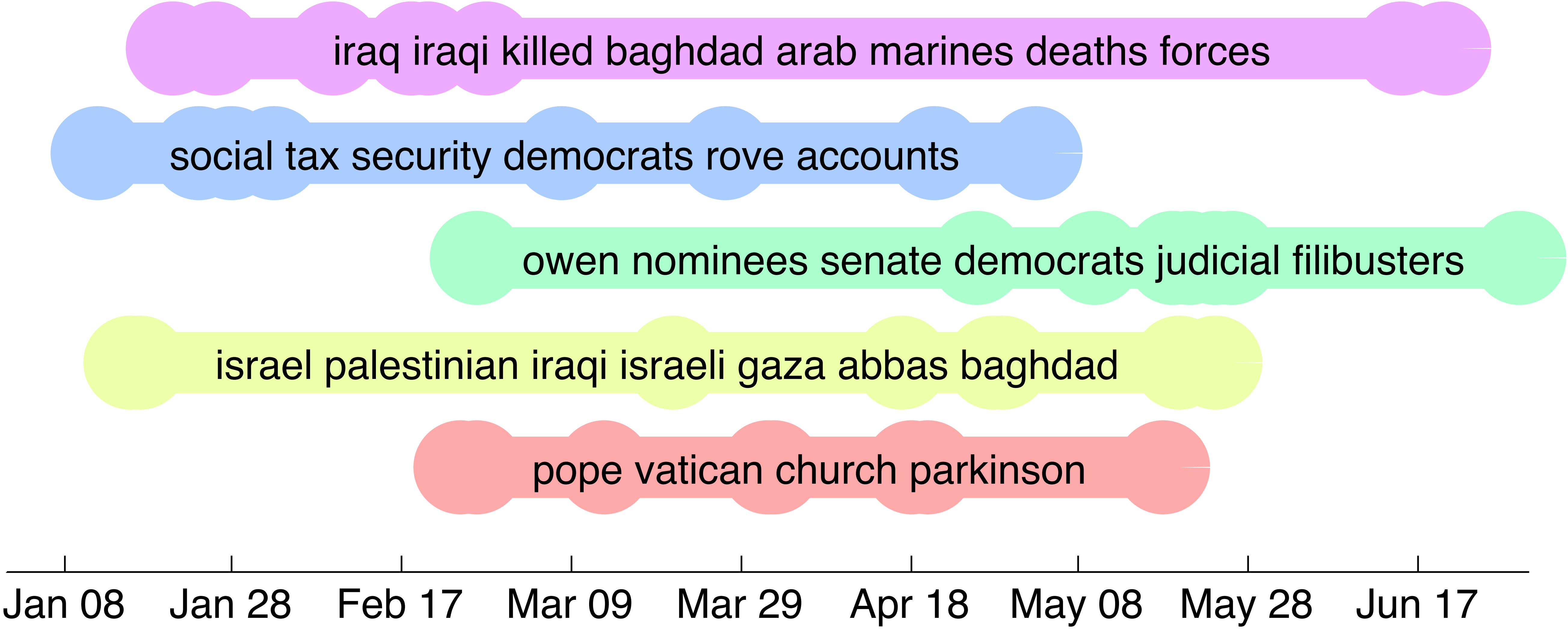
$\xrightarrow{G}$

$D = 36{,}356$

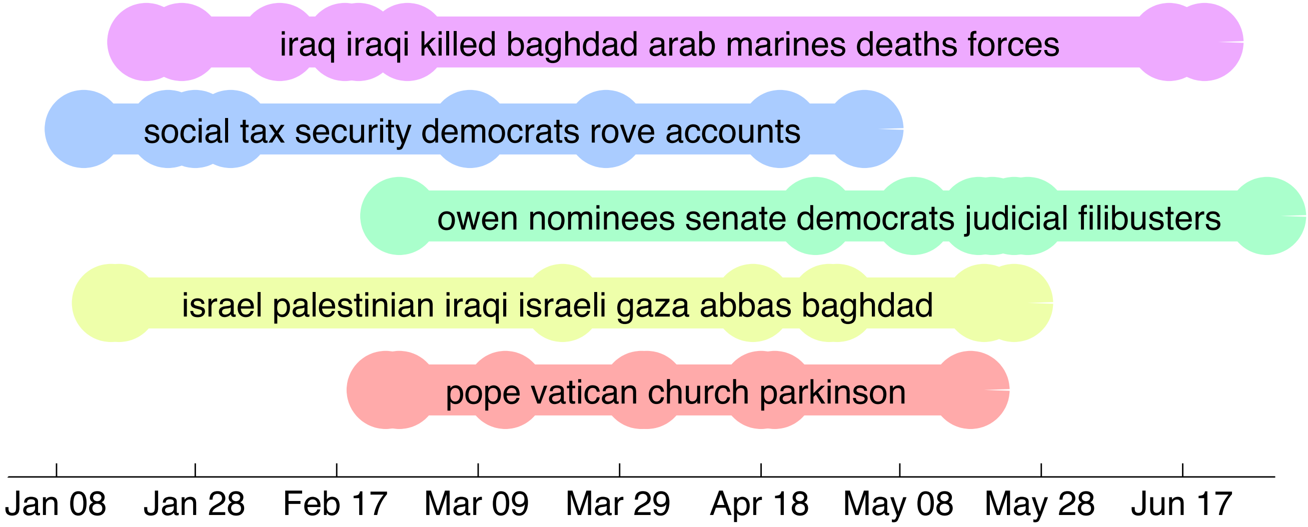
$d = 50$

DPP THREADS

DPP THREADS


iraq iraqi killed baghdad arab marines deaths forces
social tax security democrats rove accounts
owen nominees senate democrats judicial filibusters
israel palestinian iraqi israeli gaza abbas baghdad
pope vatican church parkinson
Jan 08
Jan 28
Feb 17
Mar 09
Mar 29
Apr 18
May 08
May 28
Jun 17

DPP THREADS



Feb 24: Parkinson's Disease Increases Risks to Pope
Feb 26: Pope's Health Raises Questions About His Ability to Lead
Mar 13: Pope Returns Home After 18 Days at Hospital
Apr 01: Pope's Condition Worsens as World Prepares for End of Papacy
Apr 02: Pope, Though Gravely Ill, Utters Thanks for Prayers
Apr 18: Europeans Fast Falling Away from Church
Apr 20: In Developing World, Choice [of Pope] Met with Skepticism
May 18: Pope Sends Message with Choice of Name

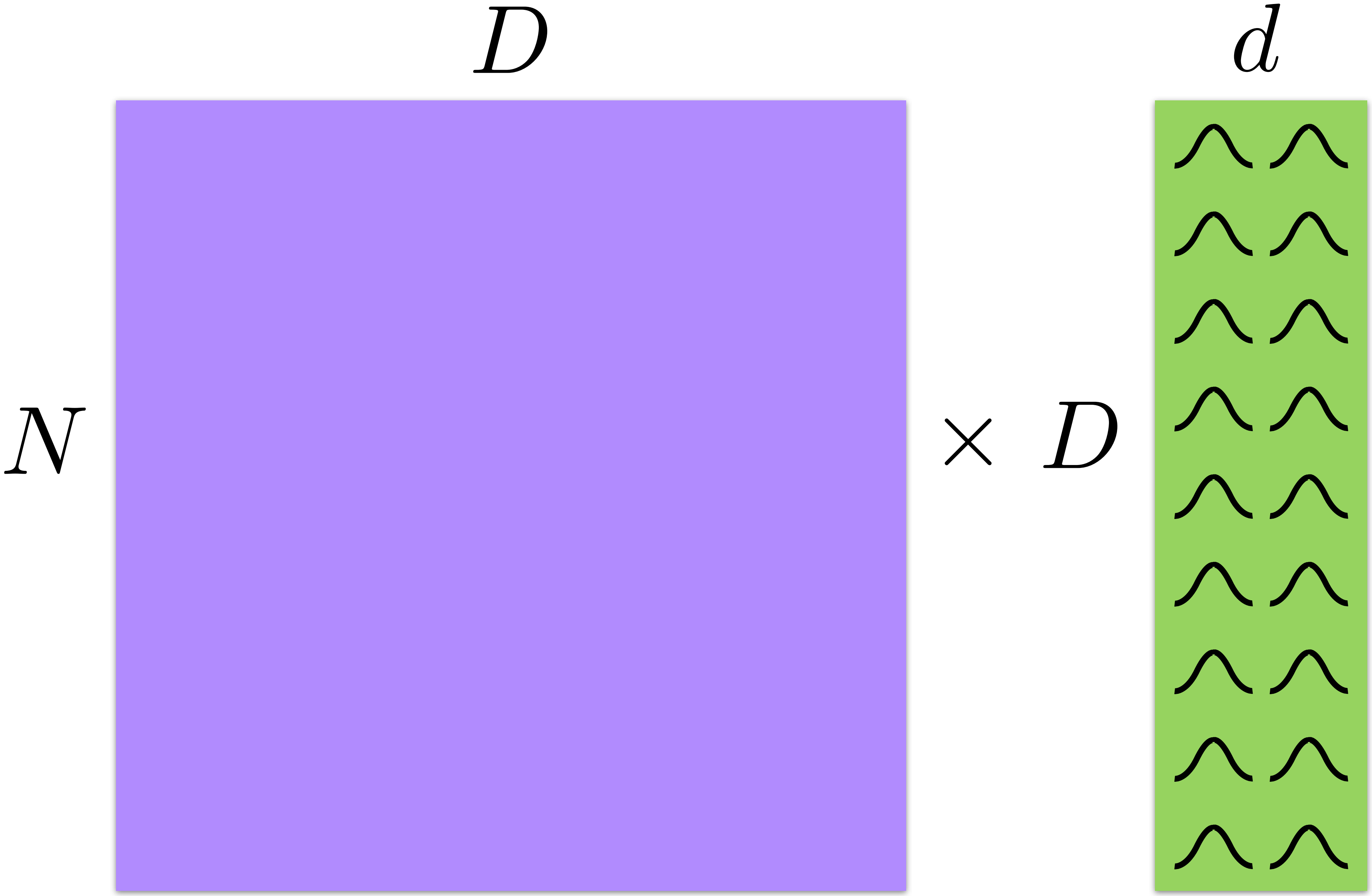
PROPOSED WORK

PROPOSED WORK

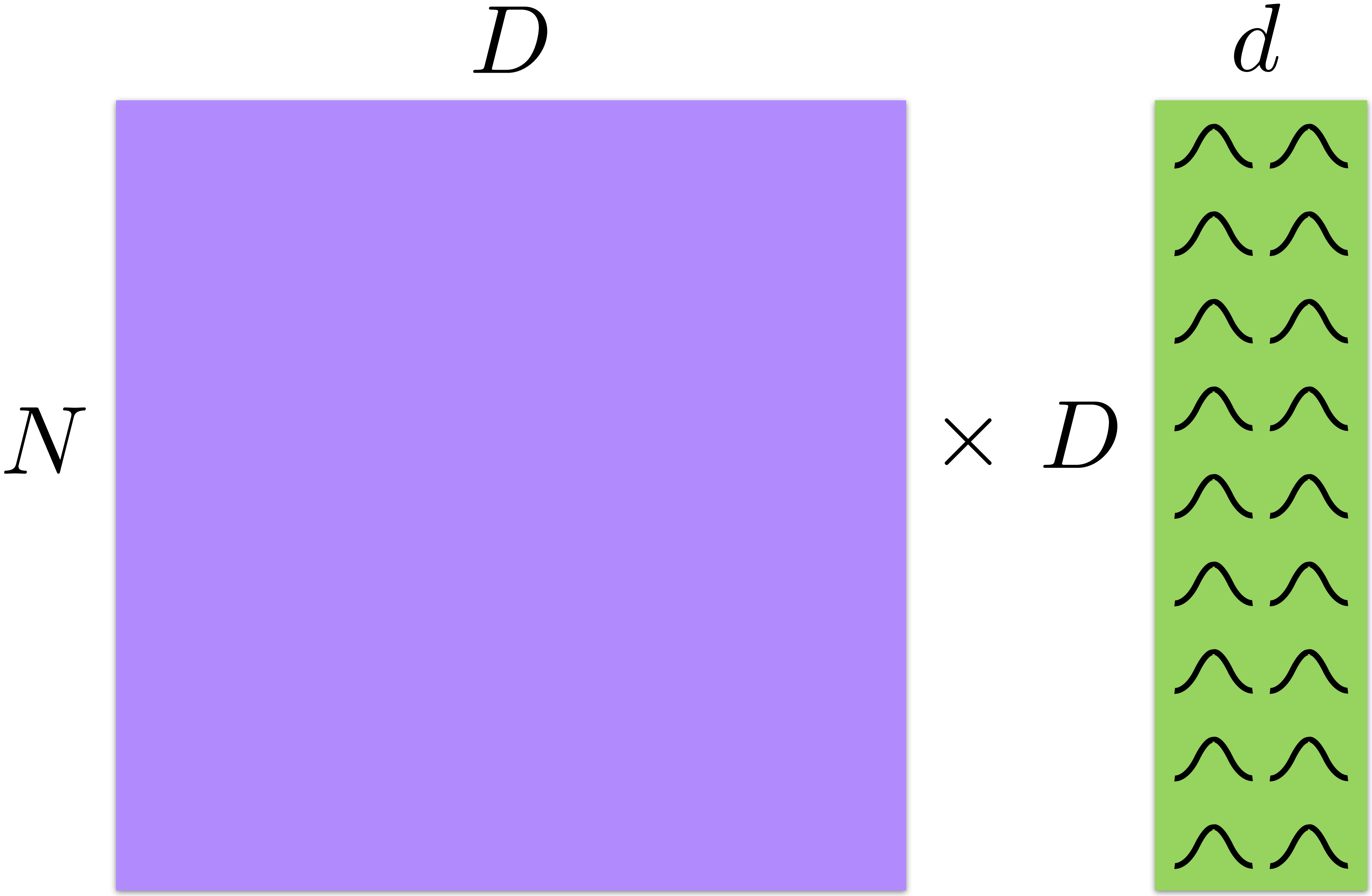
- Survey algorithms for large-scale eigendecomps

LARGE-SCALE EIGENDECOMP

LARGE-SCALE EIGENDECOMP



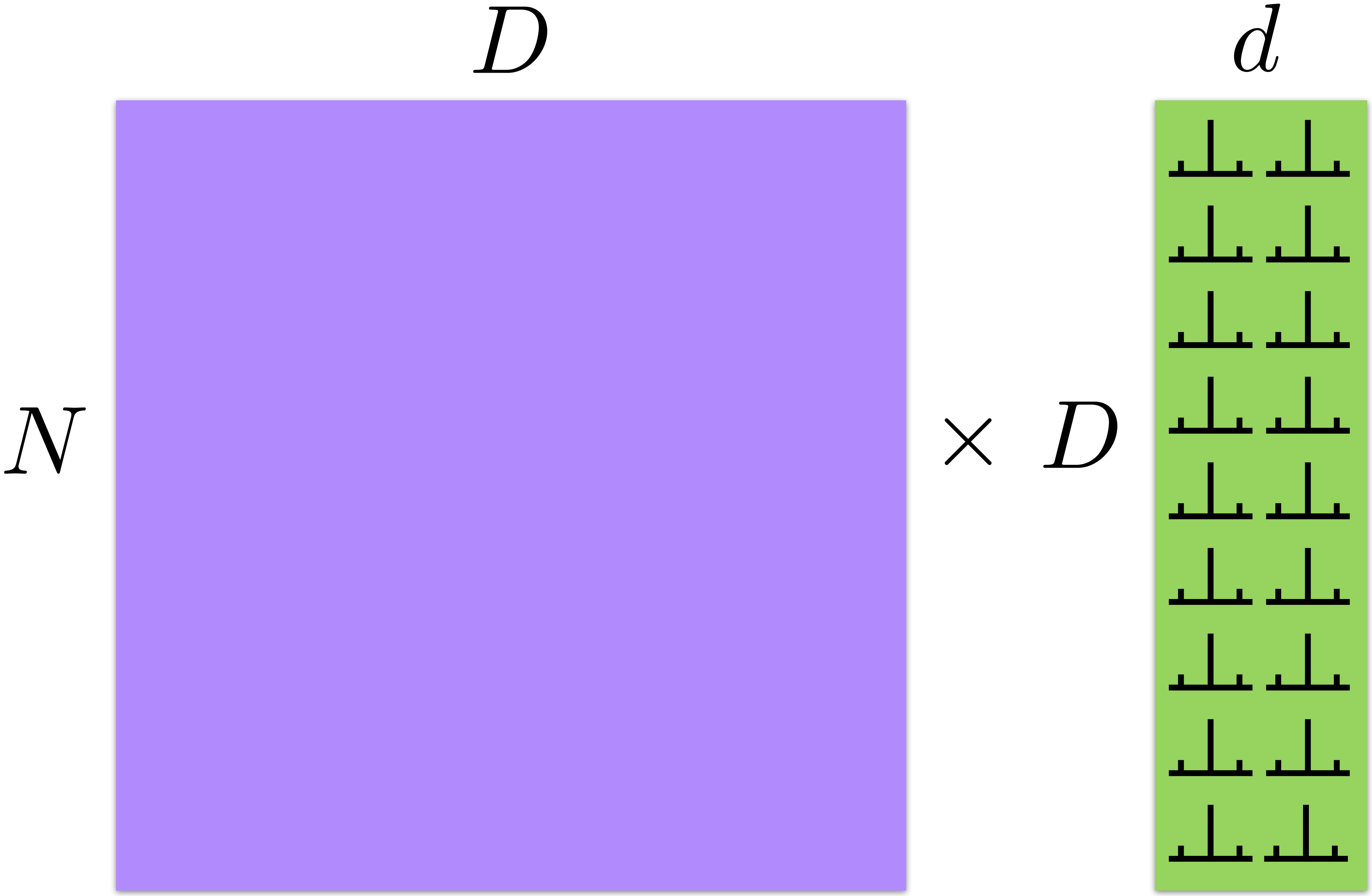
LARGE-SCALE EIGENDECOMP



$O(NDd)$

LARGE-SCALE EIGENDECOMP

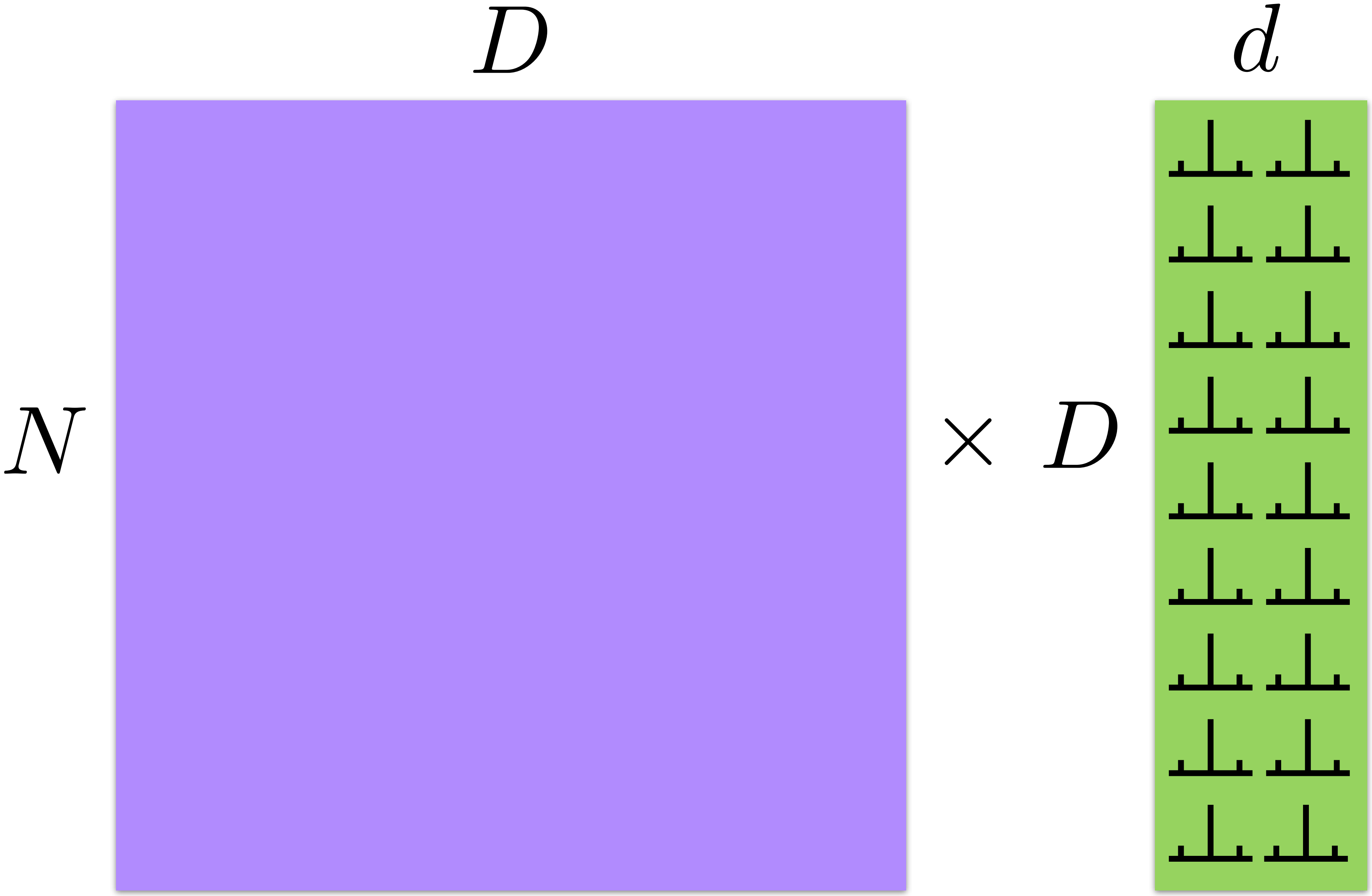
ACHLIOPTAS (2002)



$O(NDd)$

LARGE-SCALE EIGENDECOMP

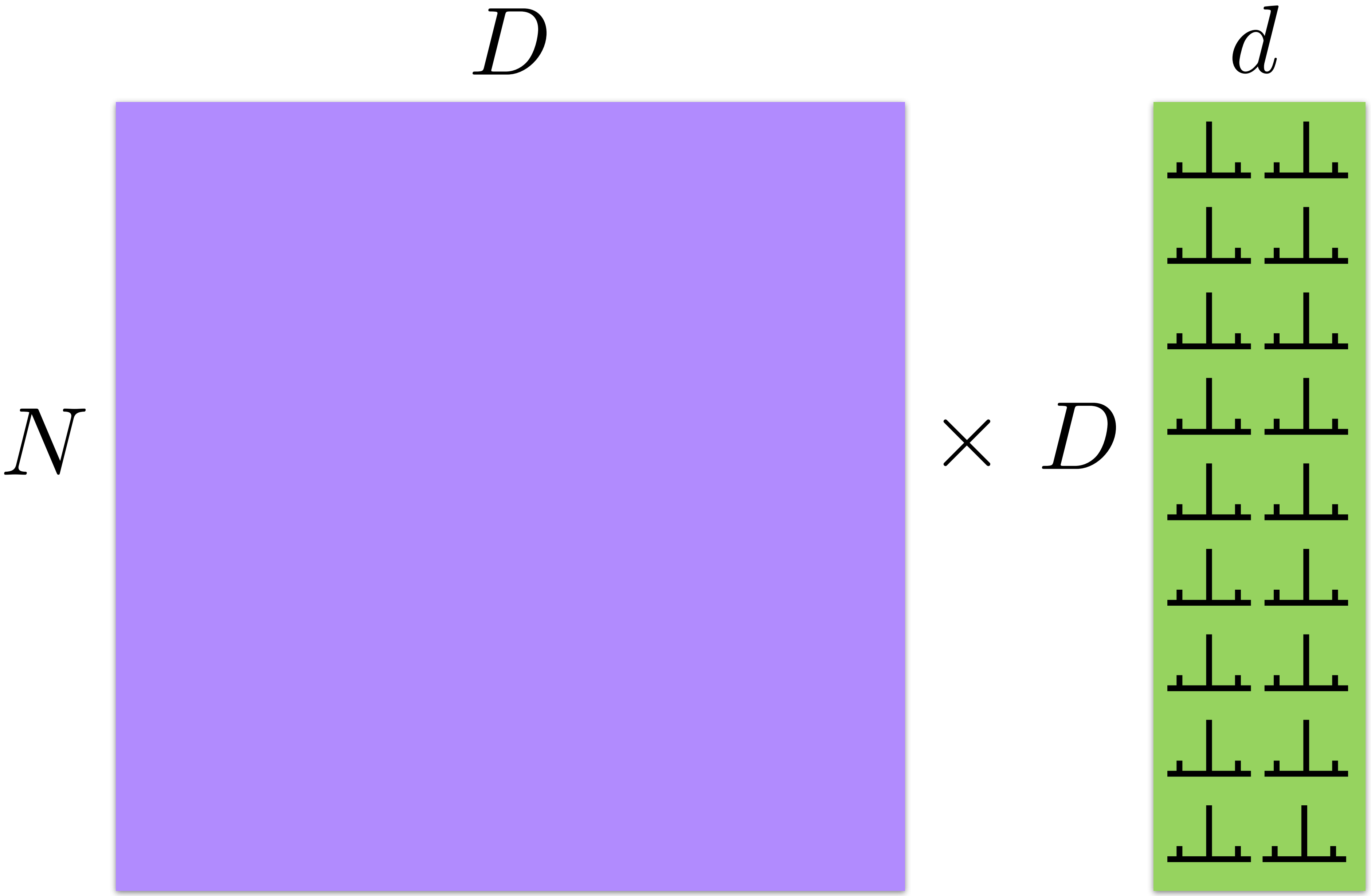
ACHLIOPTAS (2002)



$$\frac{1}{3}O(NDd)$$

LARGE-SCALE EIGENDECOMP

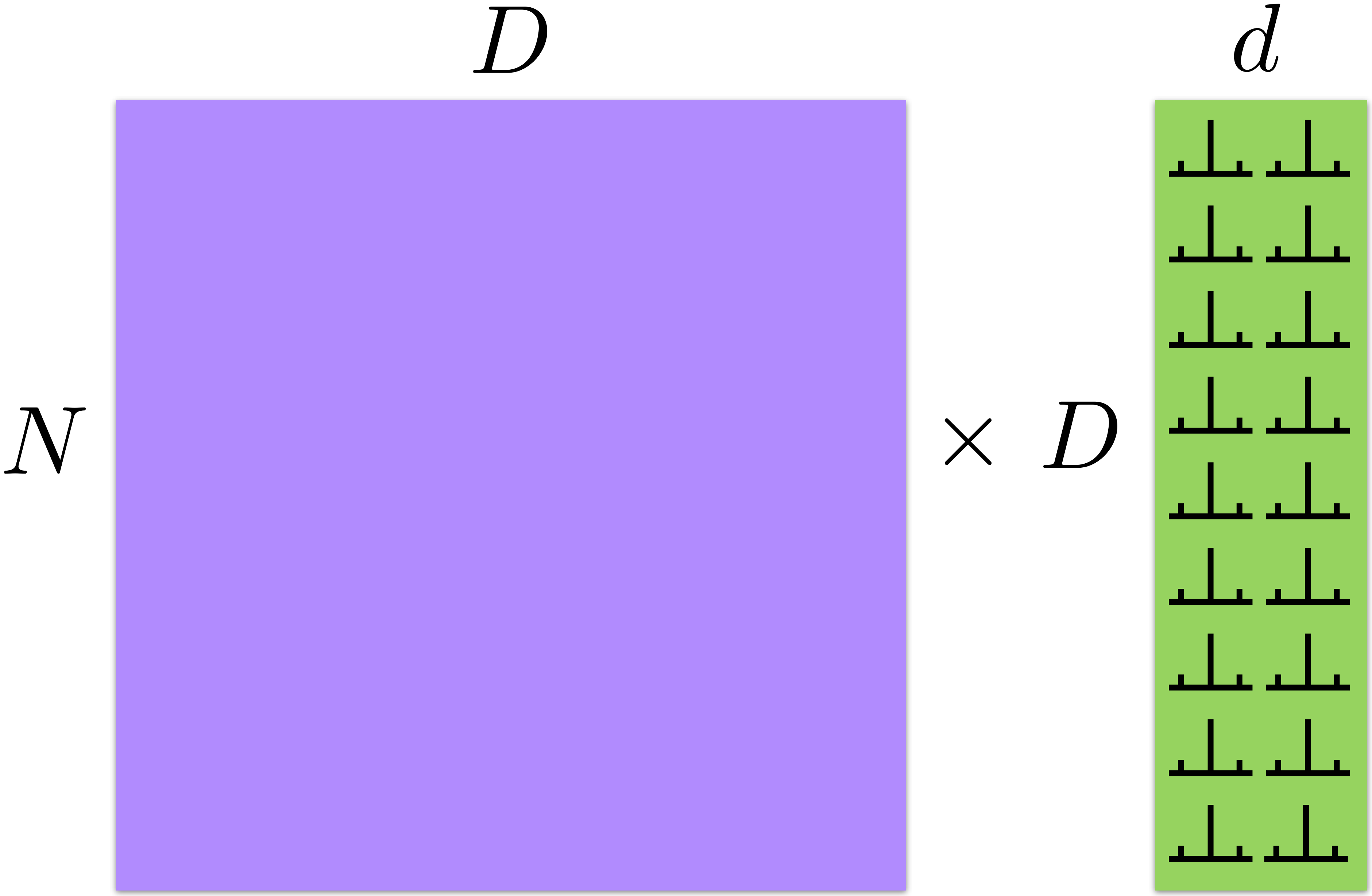
AILON AND CHAZELLE (2009)



$$\frac{1}{3}O(NDd)$$

LARGE-SCALE EIGENDECOMP

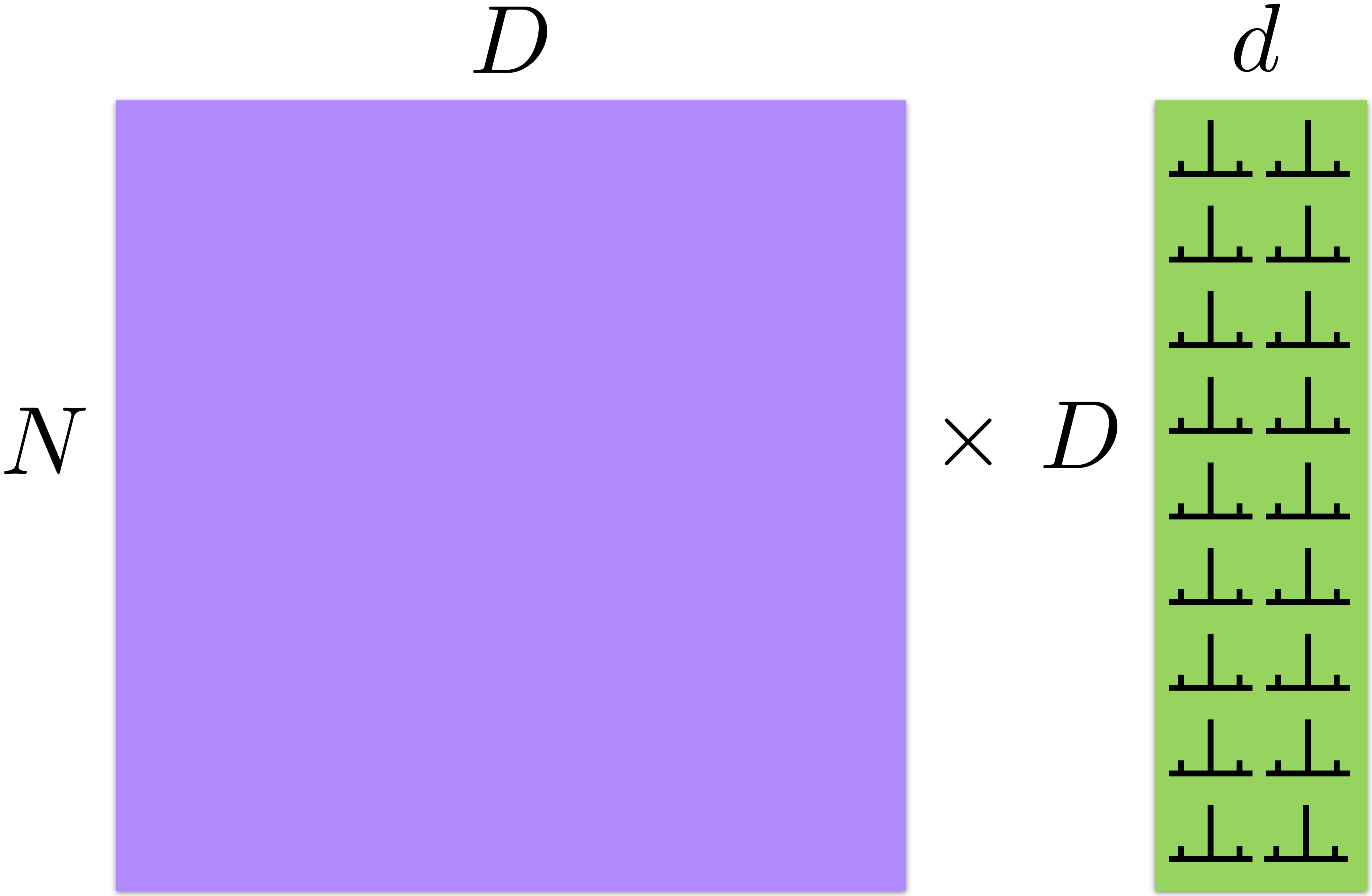
AILON AND CHAZELLE (2009)



$O(ND(< d))$

LARGE-SCALE EIGENDECOMP

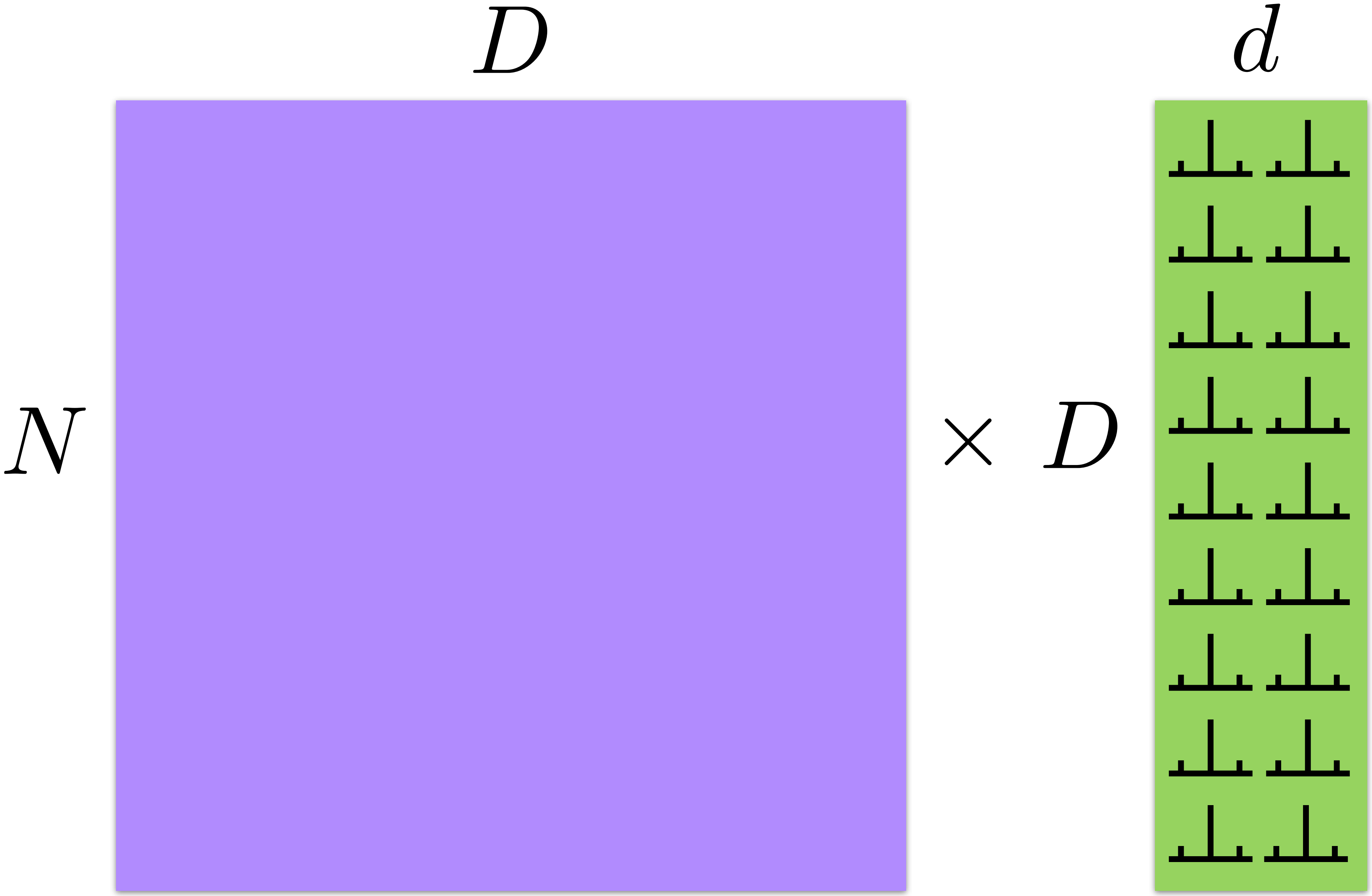
MAGEN AND ZOUZIAS (2008)



$O(ND(< d))$

LARGE-SCALE EIGENDECOMP

MAGEN AND ZOUZIAS (2008)



$O(ND(< d))$

Take-away message for these other random projection methods:
1) no volume-preservation guarantees, and 2) runtime will be dominated by DPP sampling anyway.

PROPOSED WORK

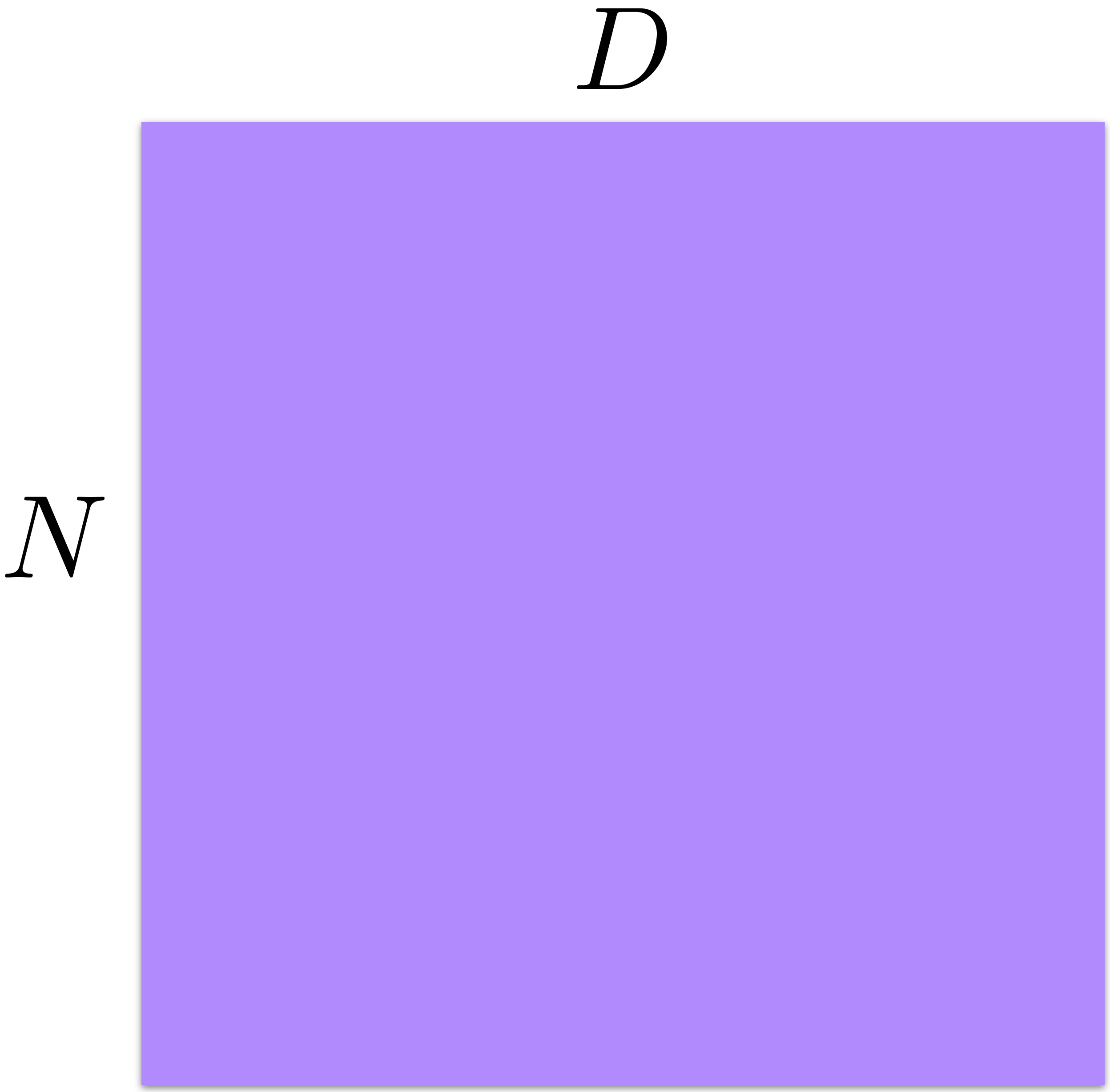
PROPOSED WORK

- Survey algorithms for large-scale eigendecomps

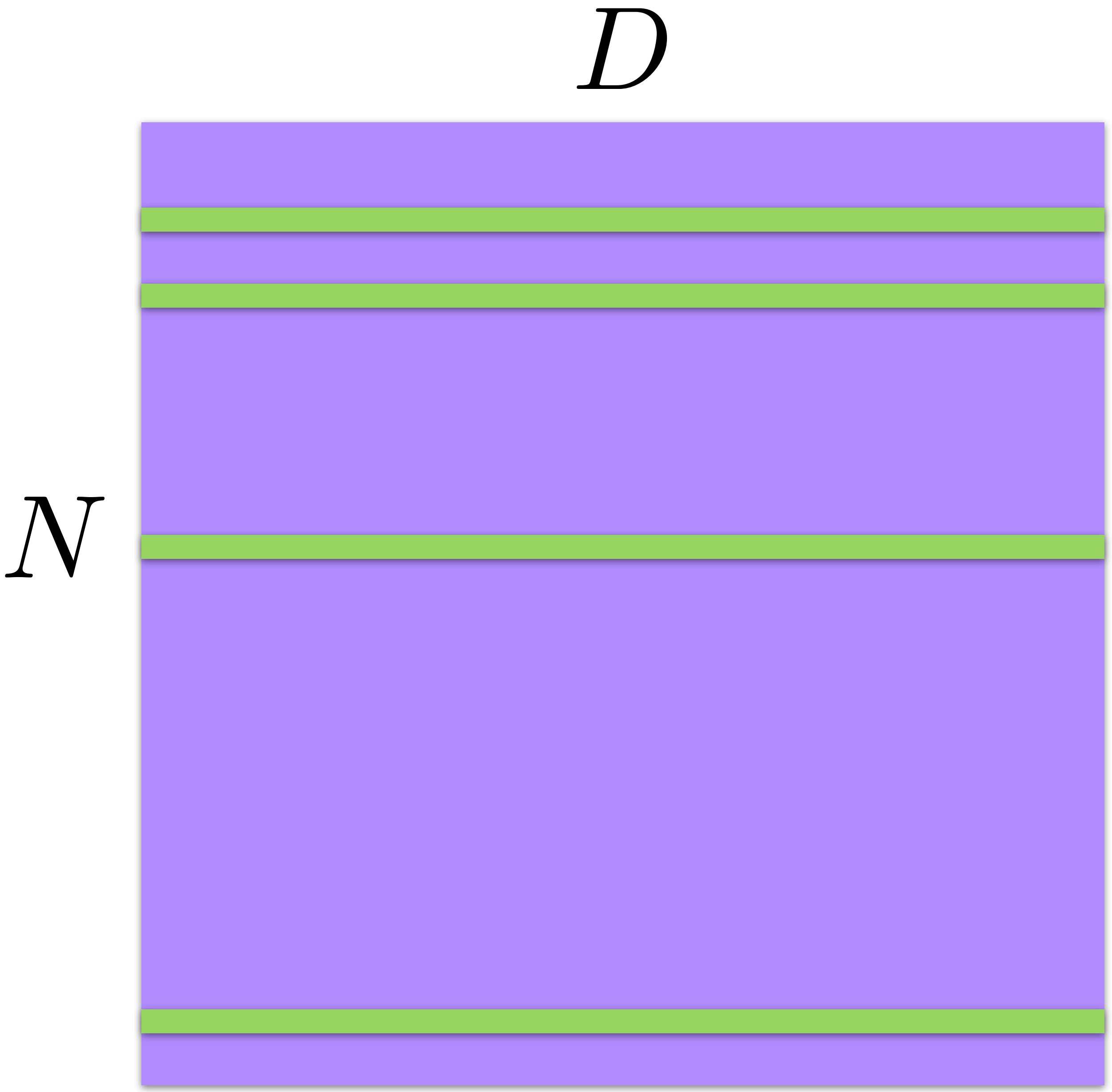
PROPOSED WORK

- Survey algorithms for large-scale eigendecomps
- Try to extend Nyström approximation from Affandi et al. (AIStats 2013) to structured DPPs

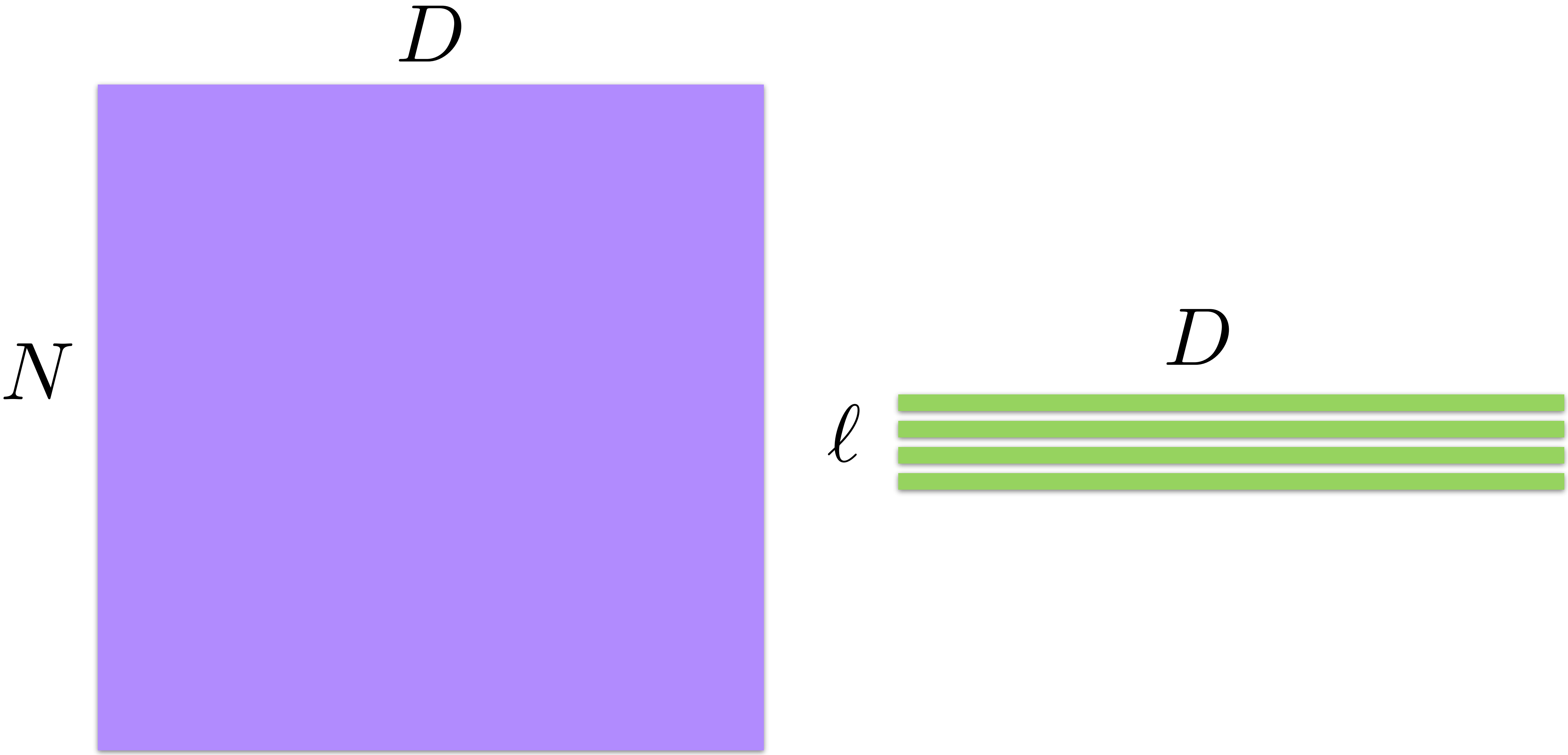
NYSTRÖM APPROX



NYSTRÖM APPROX

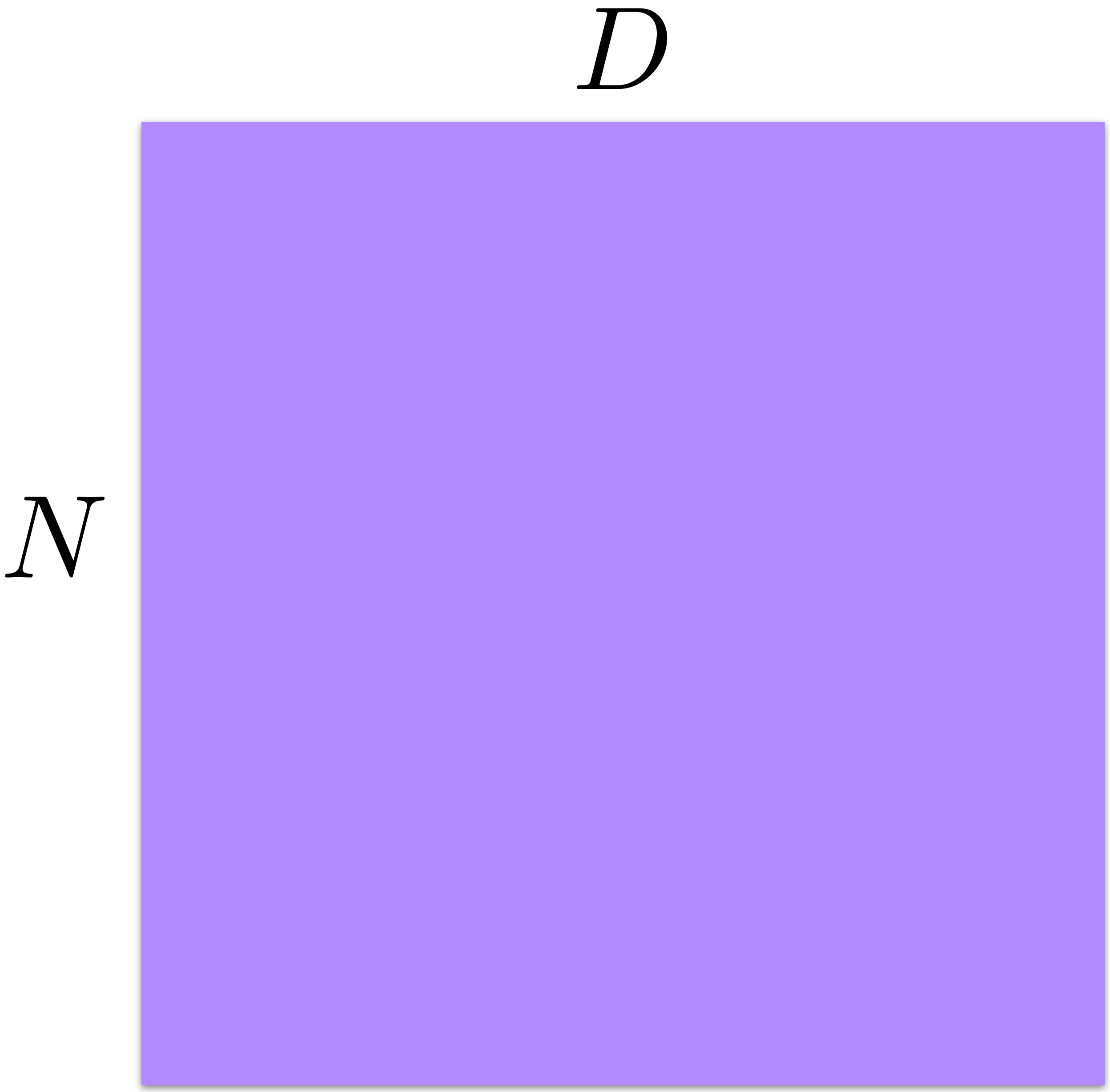


NYSTRÖM APPROX

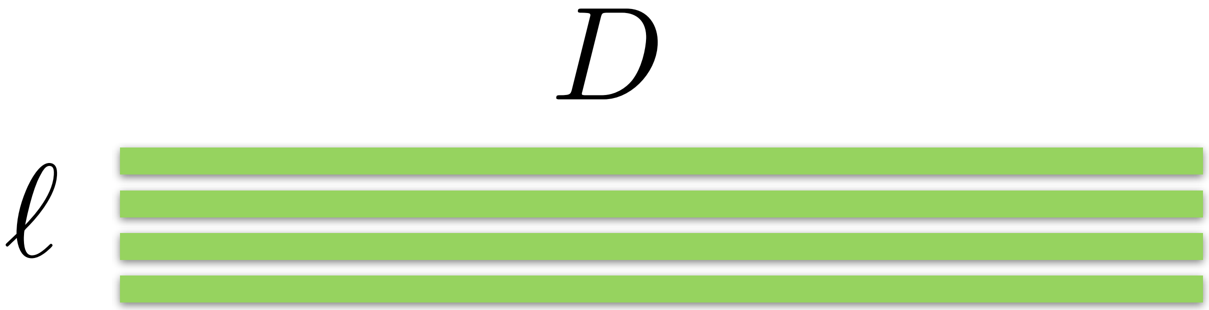


NYSTRÖM APPROX

AFFANDI, KULESZA, FOX, AND TASKAR (AISTATS 2013)

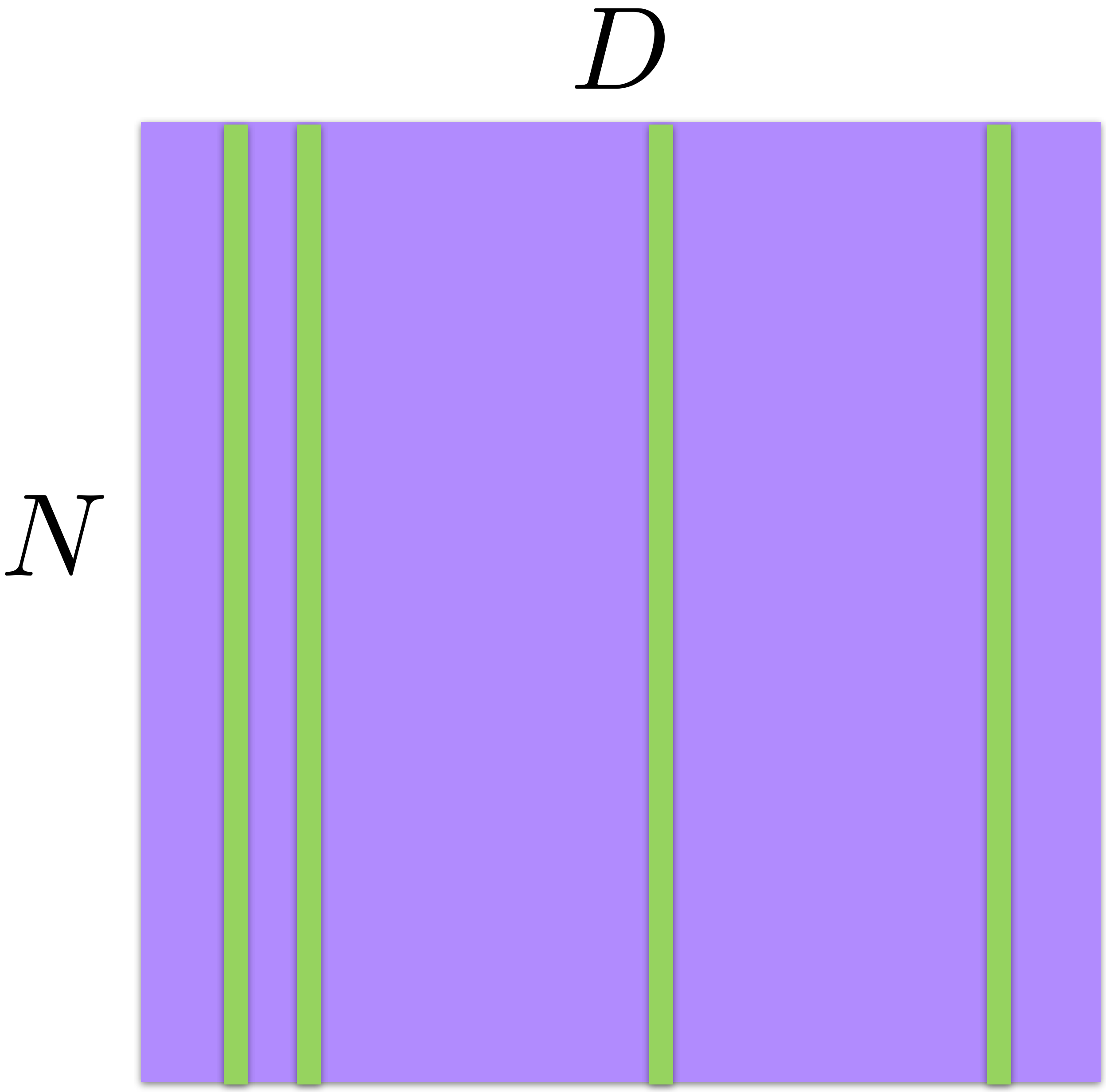


$|\mathcal{P}(Y) - \tilde{\mathcal{P}}(Y)|$ bounded



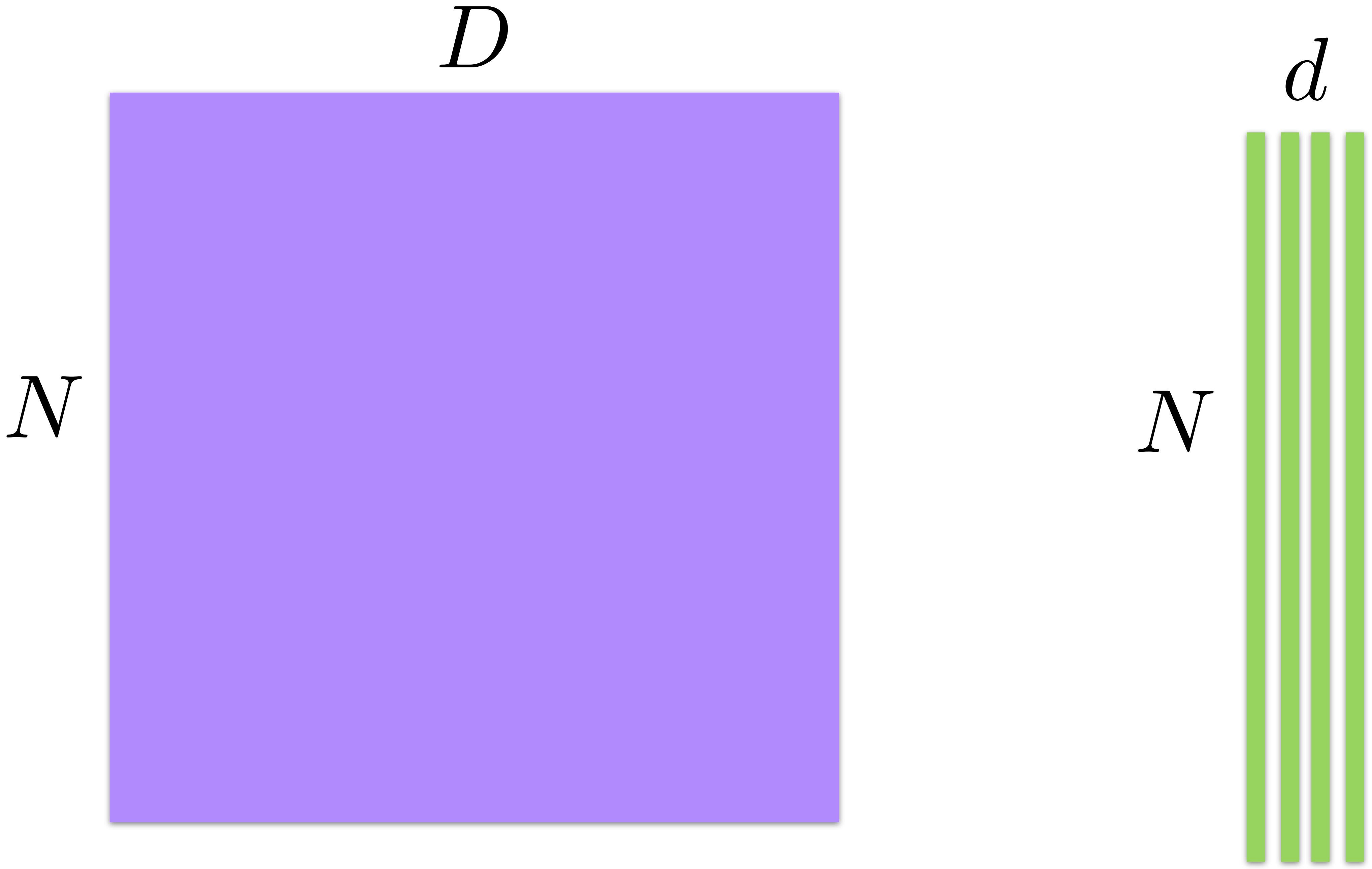
NYSTRÖM APPROX

AFFANDI, KULESZA, FOX, AND TASKAR (AISTATS 2013)



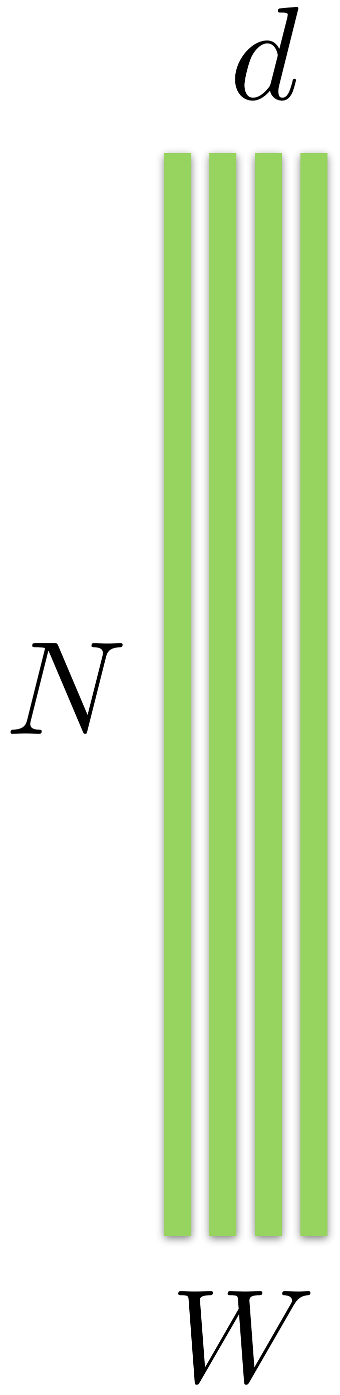
NYSTRÖM APPROX

AFFANDI, KULESZA, FOX, AND TASKAR (AISTATS 2013)



NYSTRÖM APPROX

NYSTRÖM APPROX



$$\tilde{C} = \begin{pmatrix} C_W & C_{W,\overline{W}} \\ C_{\overline{W},W} & C_{\overline{W},W} C_W^{+} C_{W,\overline{W}} \end{pmatrix}$$

NYSTRÖM APPROX

$$\tilde{C} = \begin{pmatrix} C_W & C_{W,\overline{W}} \\ C_{\overline{W},W} & C_{\overline{W},W} C_W^{+} C_{W,\overline{W}} \end{pmatrix}$$

$$B_{\mathbf{i}} = \left[\prod_{\alpha\in F} q(i_\alpha)\right]\left[\sum_{\alpha\in F}\phi(i_\alpha)\right]$$

NYSTRÖM APPROX

$$\tilde{C} = \begin{pmatrix} C_W & C_{W,\overline{W}} \\ C_{\overline{W},W} & C_{\overline{W},W} C_W^{+} C_{W,\overline{W}} \end{pmatrix}$$

$$B_{\mathbf{i}} = \left[\prod_{\alpha \in F} q(i_\alpha) \right] \left[\sum_{\alpha \in F} \phi(i_\alpha) \right]$$

$$\tilde{C} = \tilde{V} \tilde{\Lambda} \tilde{V}^\top$$

$$L \approx B^\top \tilde{V} \tilde{V}^\top B$$

PROPOSED WORK

PROPOSED WORK

- Survey algorithms for large-scale eigendecomps
- Try to extend Nyström approximation from Affandi et al. (AIStats 2013) to structured DPPs

PROPOSED WORK

- Survey algorithms for large-scale eigendecomps
- Try to extend Nyström approximation from Affandi et al. (AIStats 2013) to structured DPPs
- Try to establish mixing rates for MCMC sampling algorithms from Kang (NIPS 2013)

PROPOSED WORK

- Survey algorithms for large-scale eigendecomps

- Try to extend Nyström approximation from Affandi et al. (AIStats 2013) to structured DPPs

- Try to establish mixing rates for MCMC sampling algorithms from Kang (NIPS 2013)

$$O(k^3 \log(k/\epsilon))$$

PROPOSED WORK

- Survey algorithms for large-scale eigendecomps
- Try to extend Nyström approximation from Affandi et al. (AIStats 2013) to structured DPPs
- Try to establish mixing rates for MCMC sampling algorithms from Kang (NIPS 2013)

$$O(\xcancel{k^3 \log(k/\epsilon)})$$

$$\text{mixing time} \to \infty \text{ as eigenvalues of } L \to 0$$

PROPOSED WORK

- Survey algorithms for large-scale eigendecomps

- Try to extend Nyström approximation from Affandi et al. (AIStats 2013) to structured DPPs

- Try to establish mixing rates for MCMC sampling algorithms from Kang (NIPS 2013)

$$O(\cancel{k^3 \log(k/\epsilon)})$$

$$\text{mixing time} \to \infty \text{ as eigenvalues of } L \to 0$$

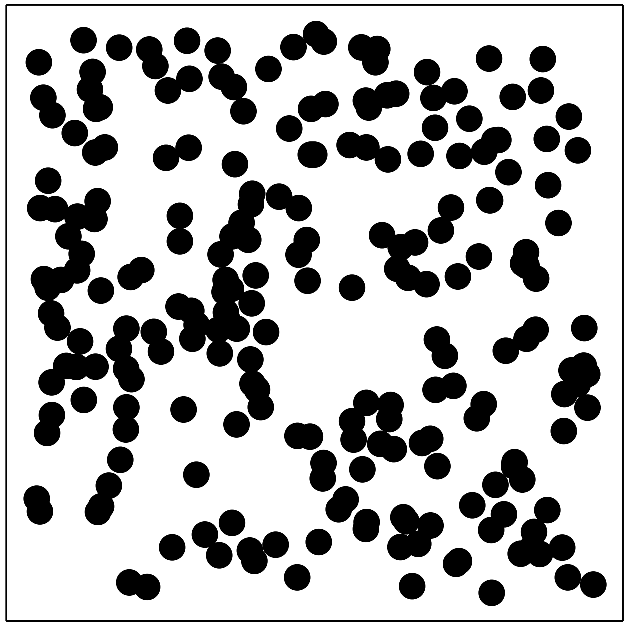
$$\text{condition number bounded: } \frac{\lambda_{\max}}{\lambda_{\min}} \leq c$$

$$\implies \text{mixing time bounded?}$$

2. MAP ESTIMATION

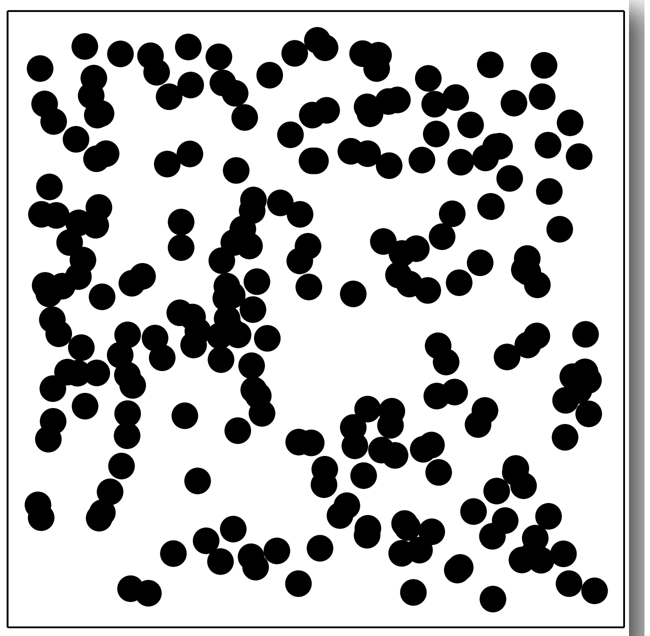
HARDNESS OF MAP

HARDNESS OF MAP

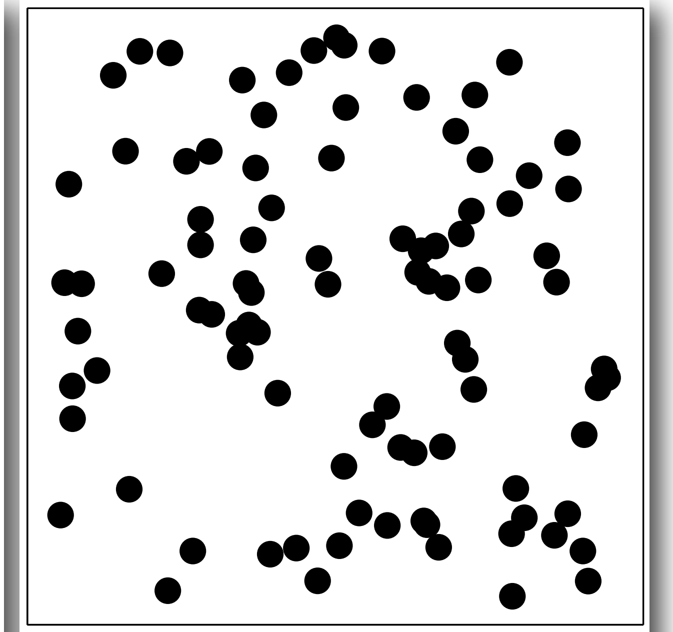


All points

HARDNESS OF MAP

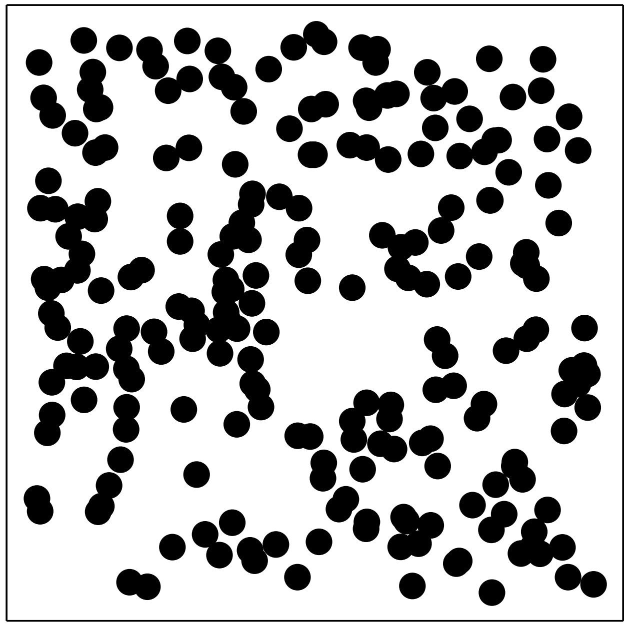


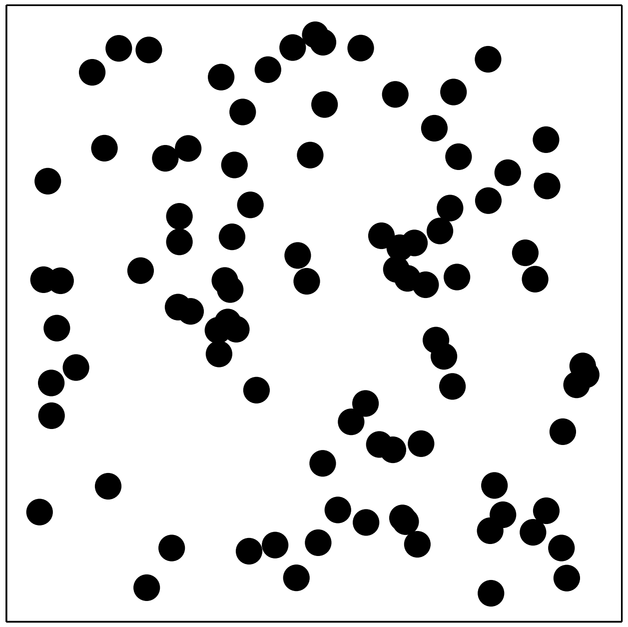
All points

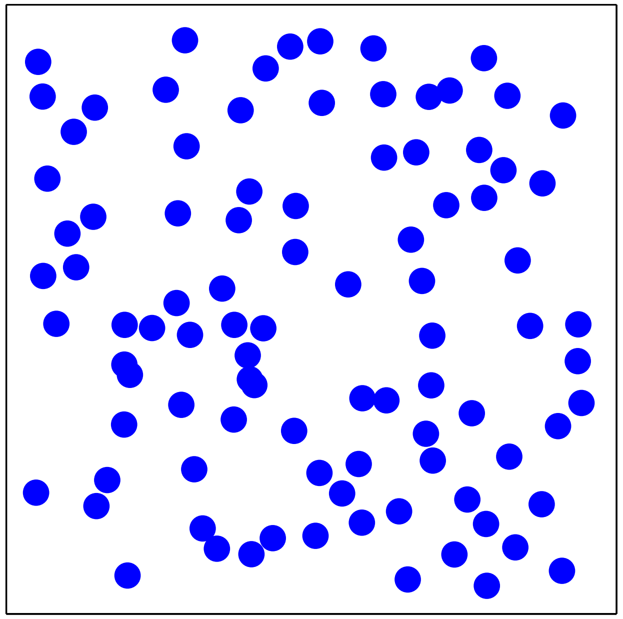


Independent sample

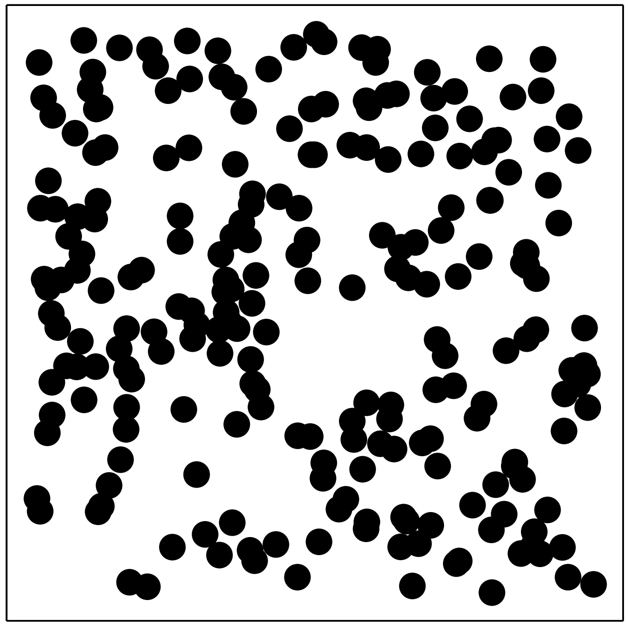
HARDNESS OF MAP


All points

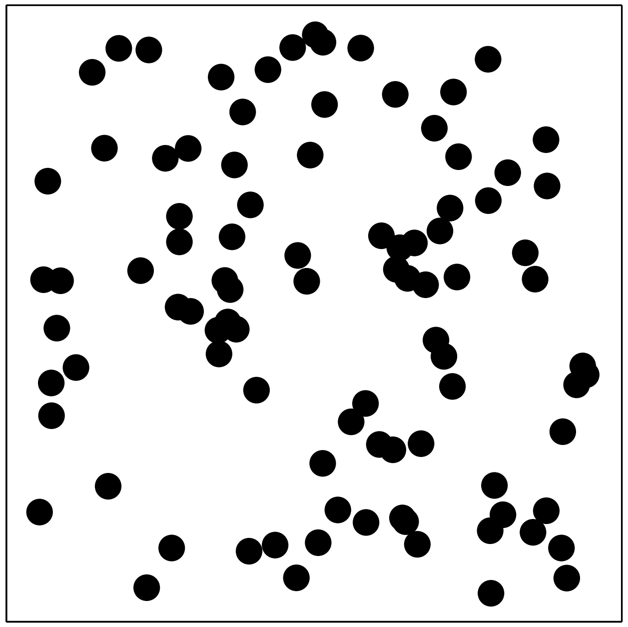

Independent sample


DPP sample

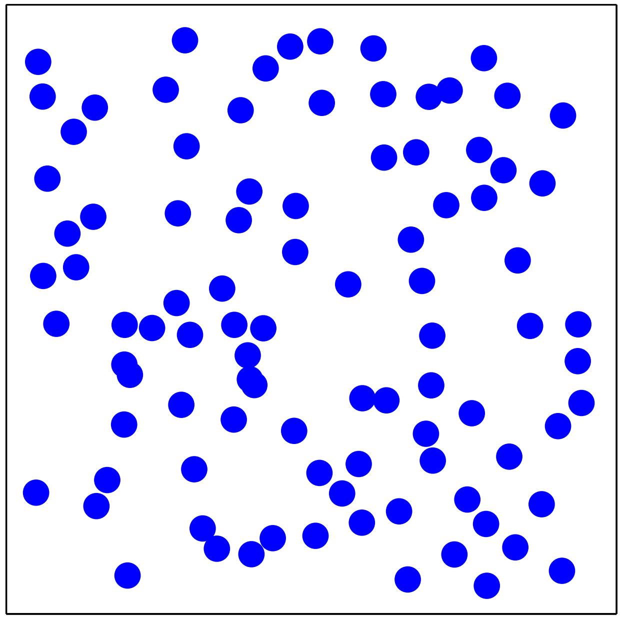
HARDNESS OF MAP



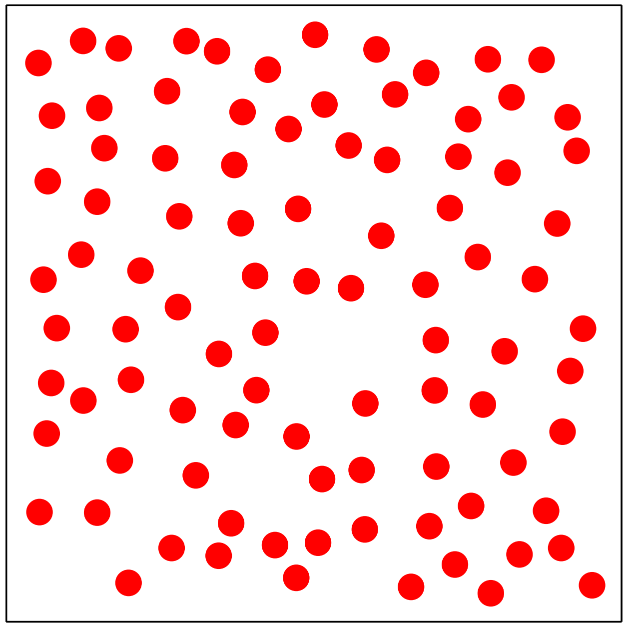
All points



Independent sample

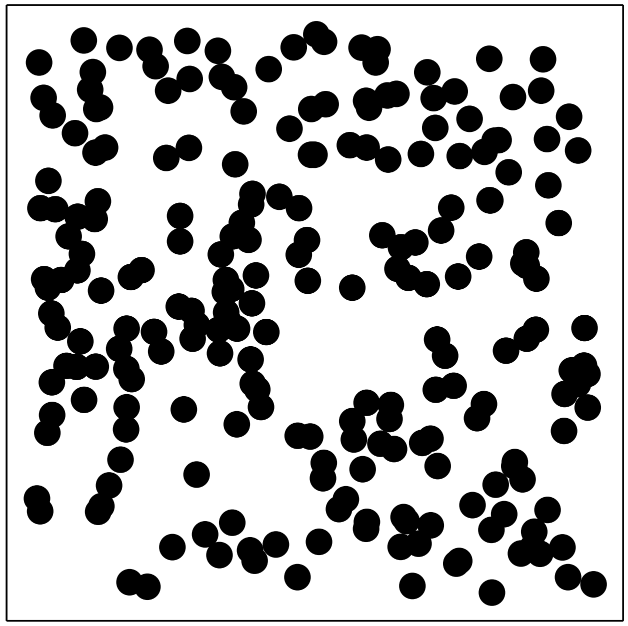


DPP sample

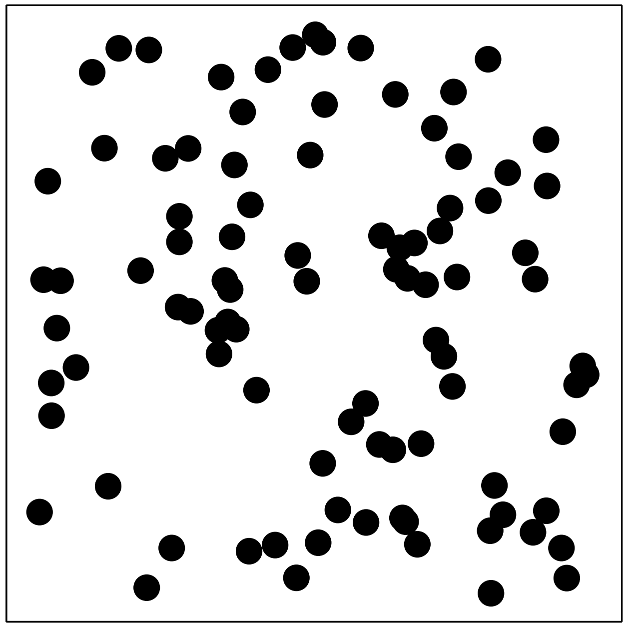


DPP (approx) MAP

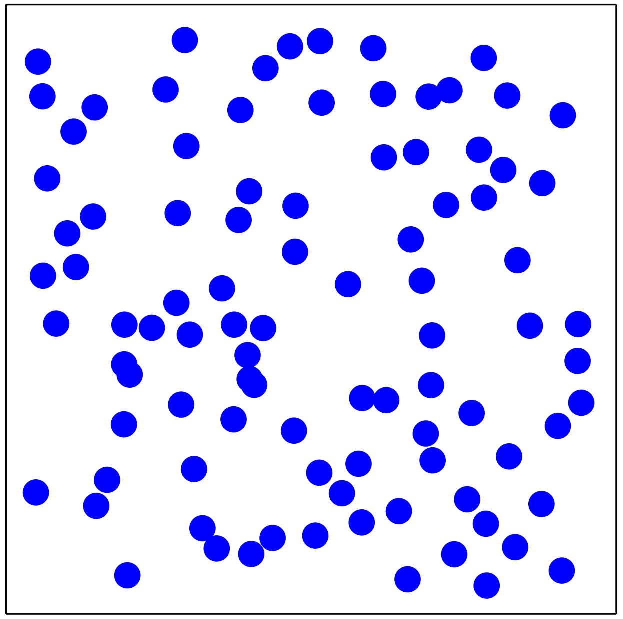
HARDNESS OF MAP



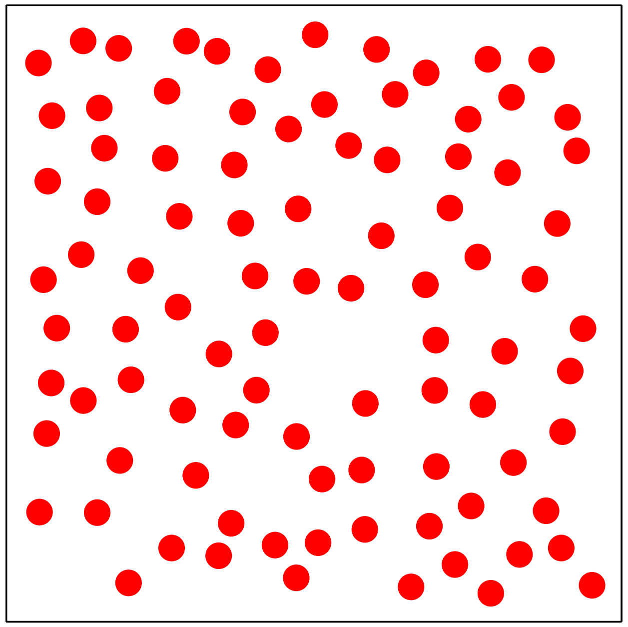
All points



Independent sample



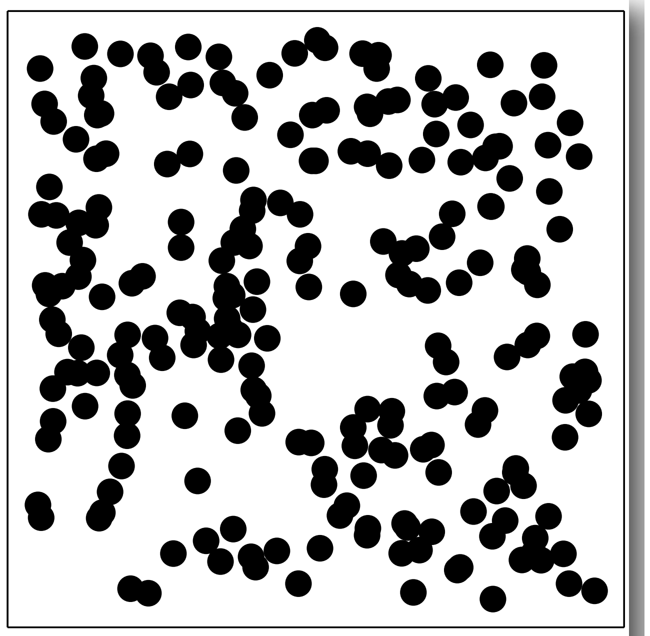
DPP sample



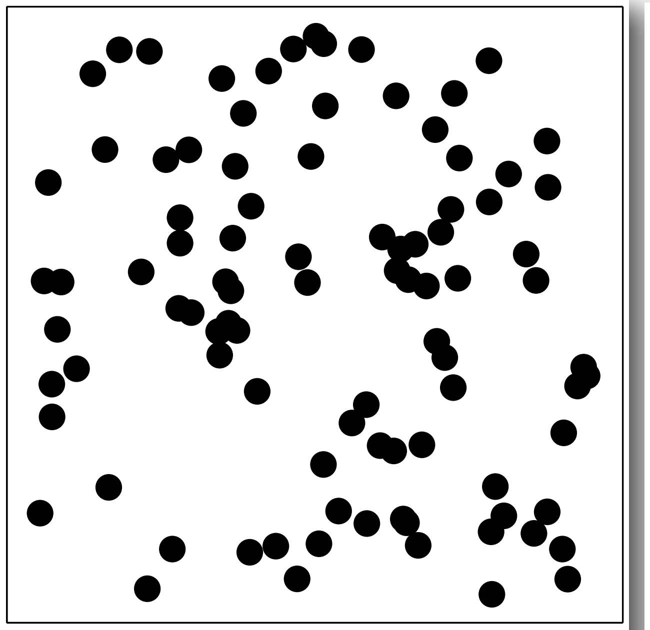
DPP (approx) MAP

$$Y = \underset{Y':Y'\subseteq\mathcal{Y}}{\arg\max}\ \det(L_{Y'})$$

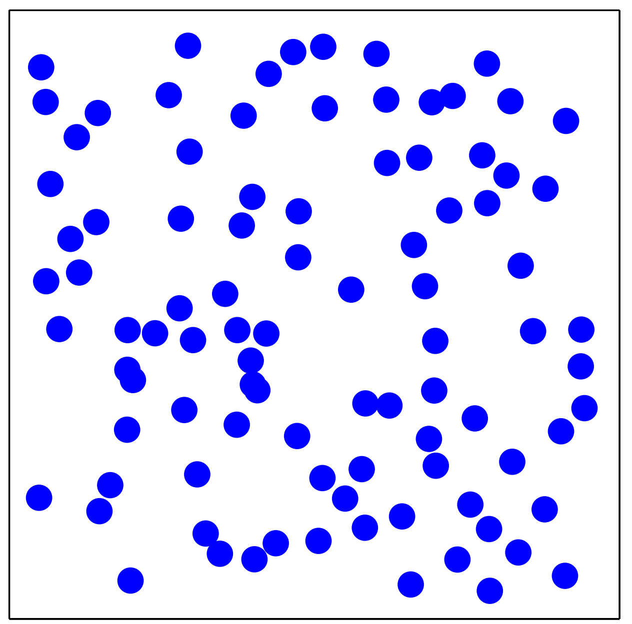
HARDNESS OF MAP



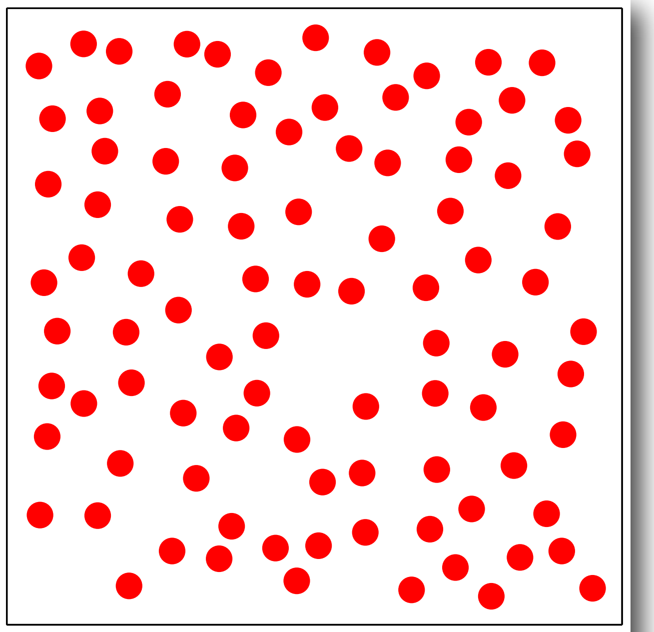
All points



Independent sample



DPP sample



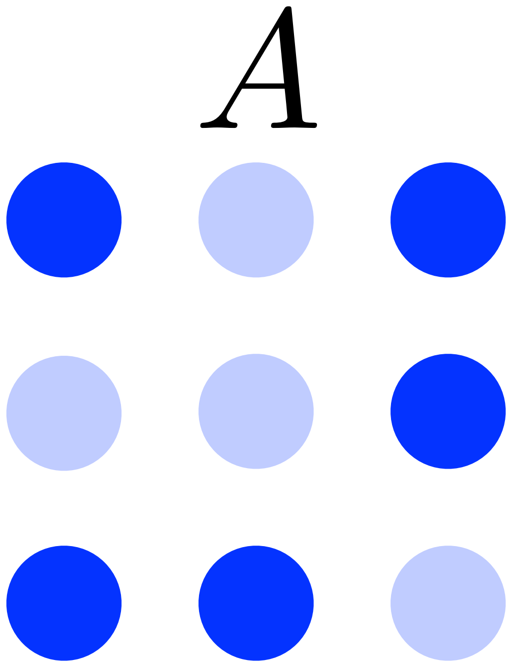
DPP (approx) MAP

$$Y = \underset{Y' : Y' \subseteq \mathcal{Y}}{\arg\max} \; \det(L_{Y'})$$

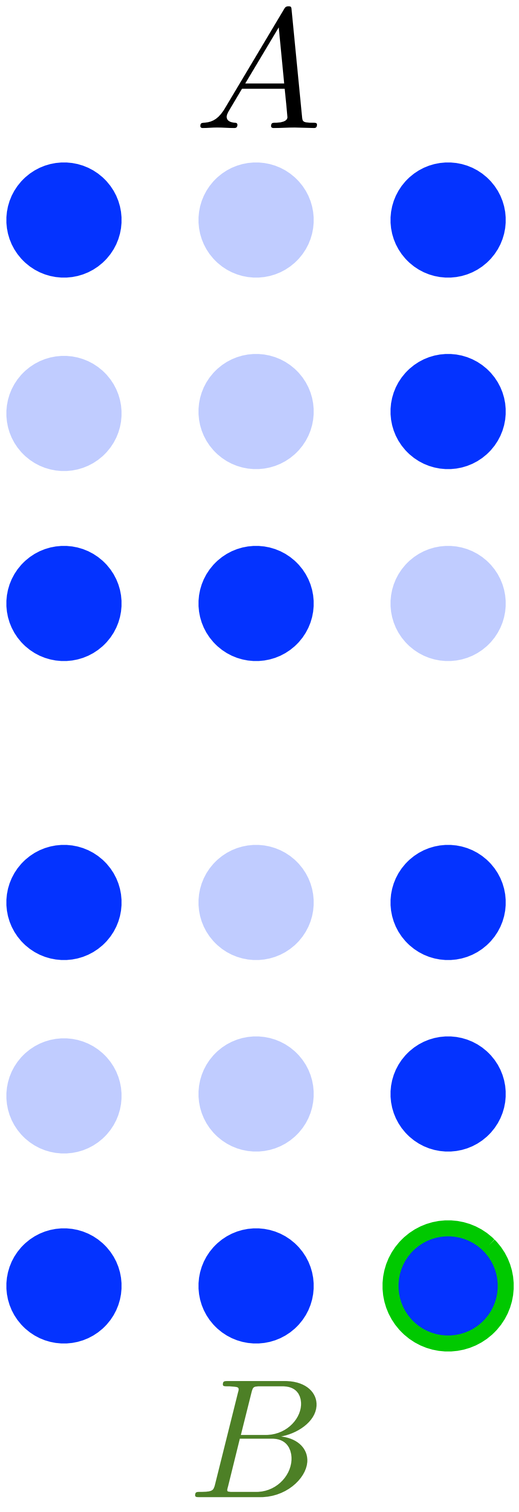
NP-hard, no PTAS: $\det(L_{\hat{Y}}) \geq \left(\frac{8}{9} + \epsilon\right) \det(L_{Y^*})$

SUBMODULARITY

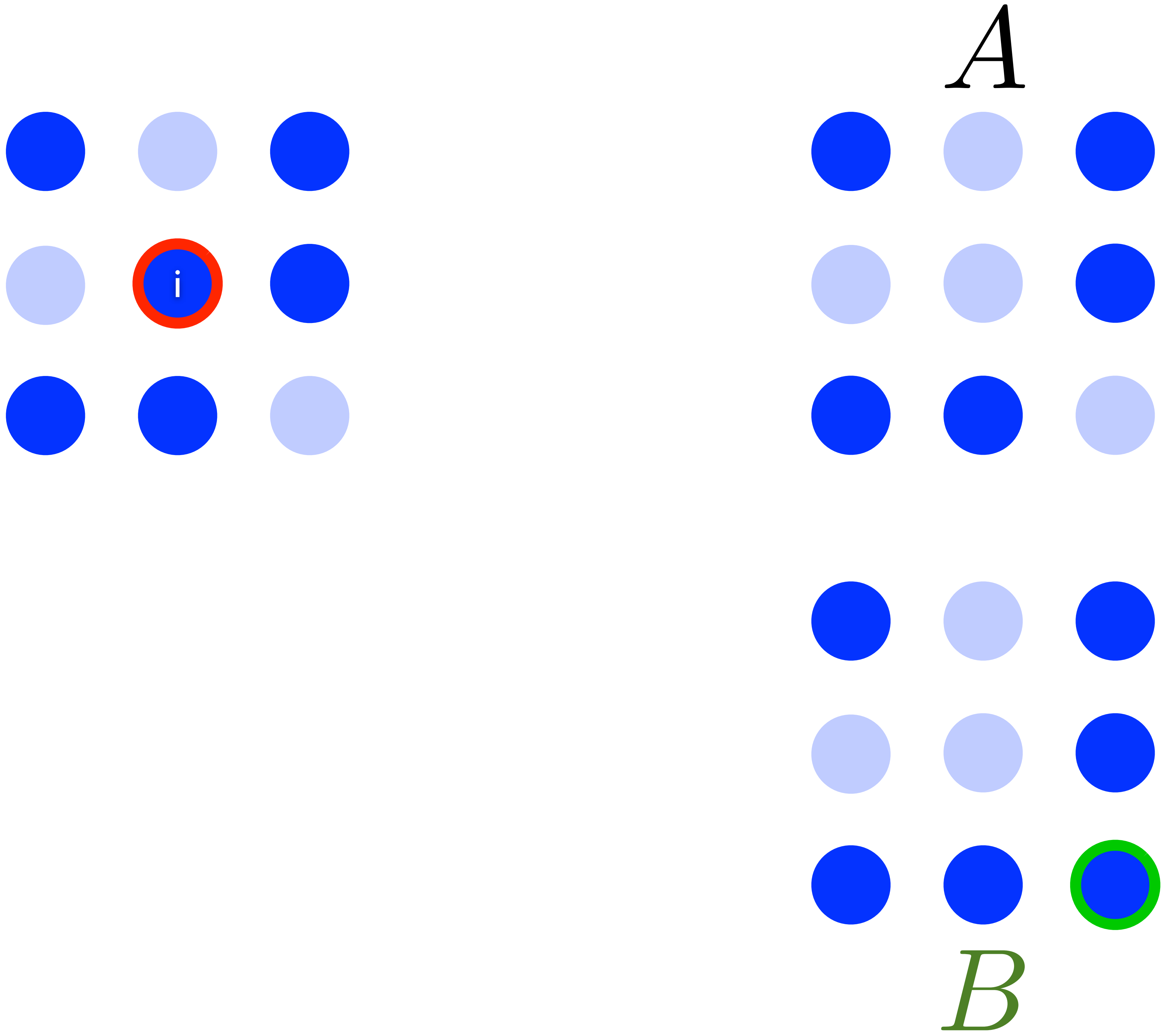
SUBMODULARITY



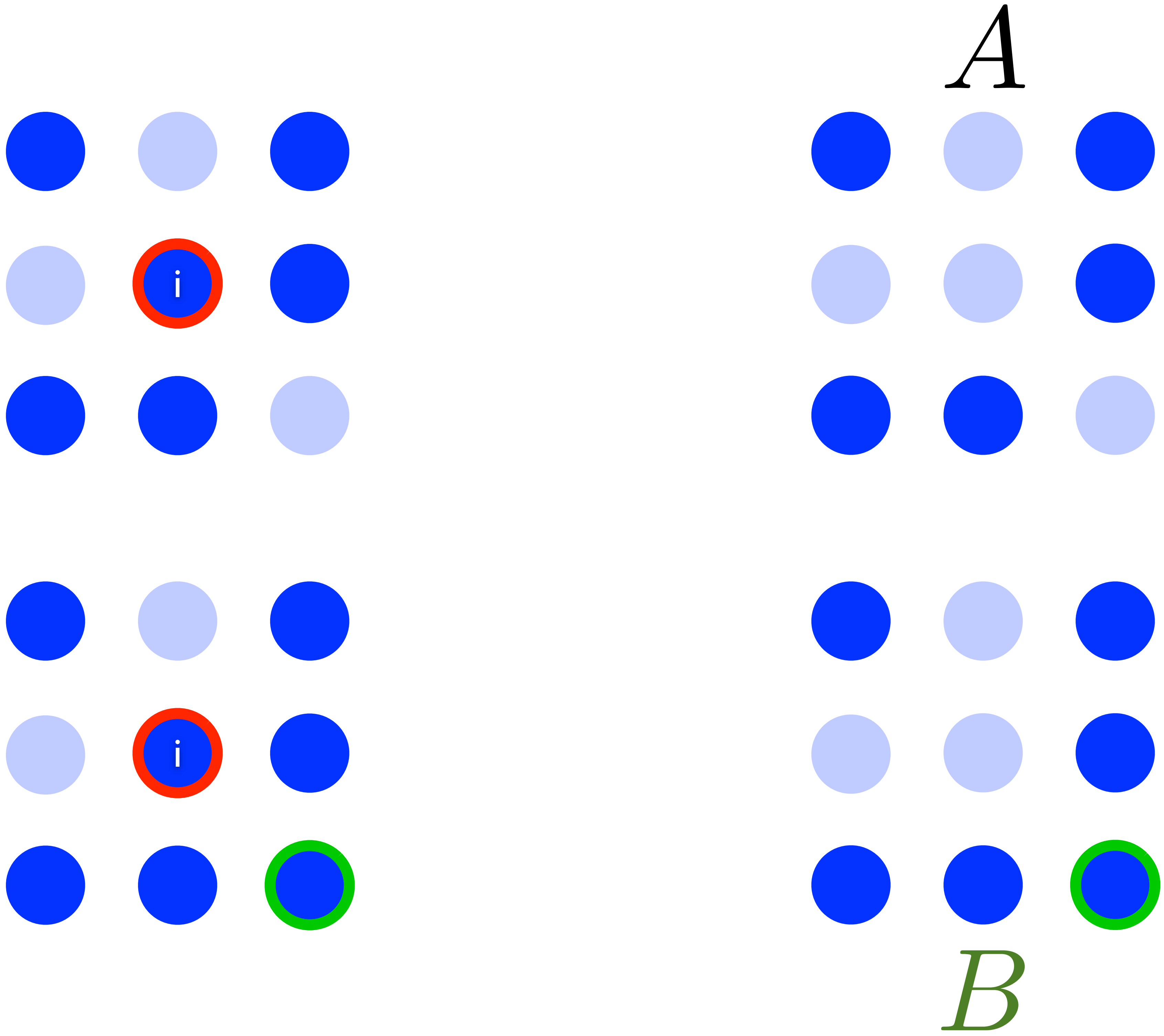
SUBMODULARITY



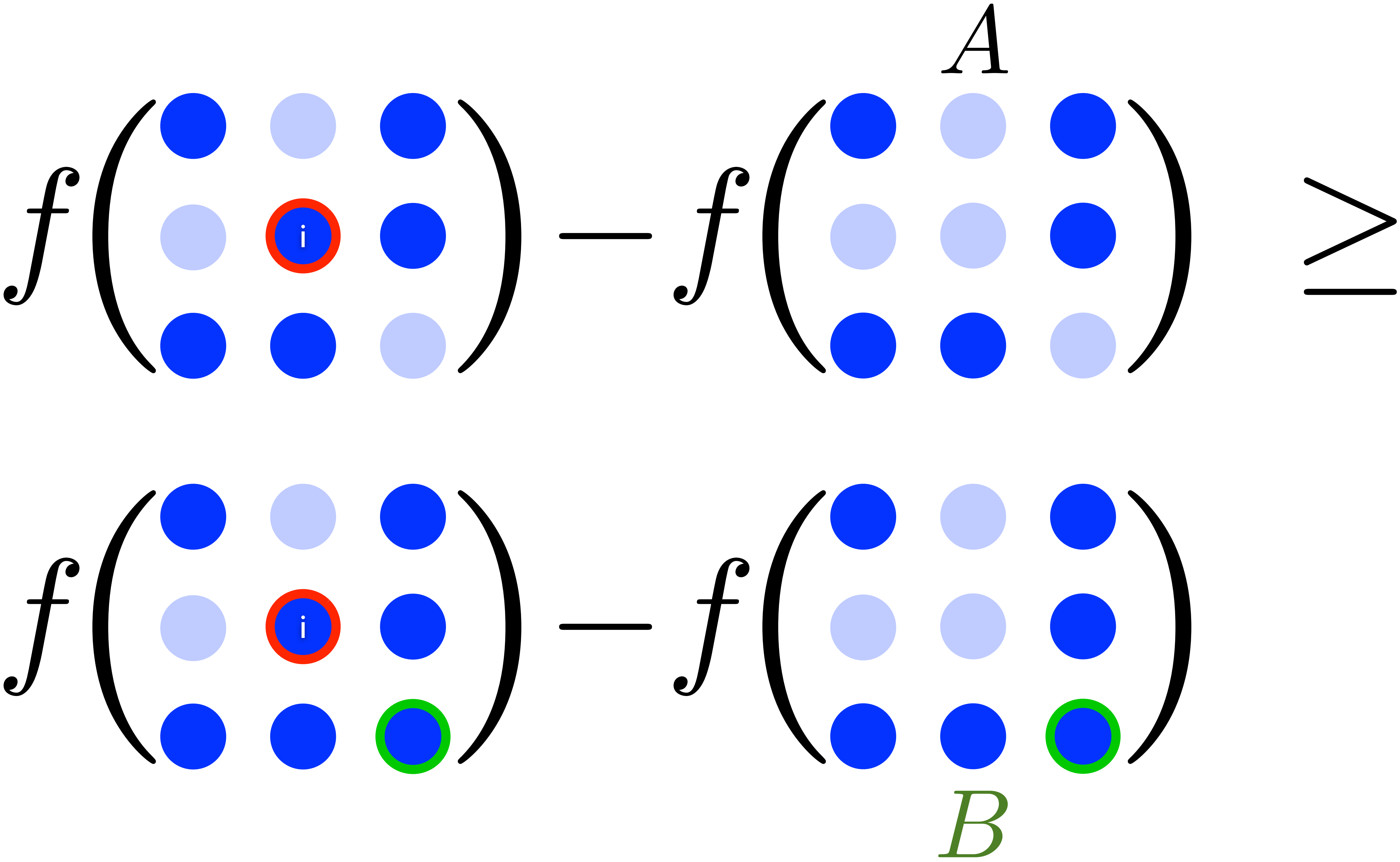
SUBMODULARITY



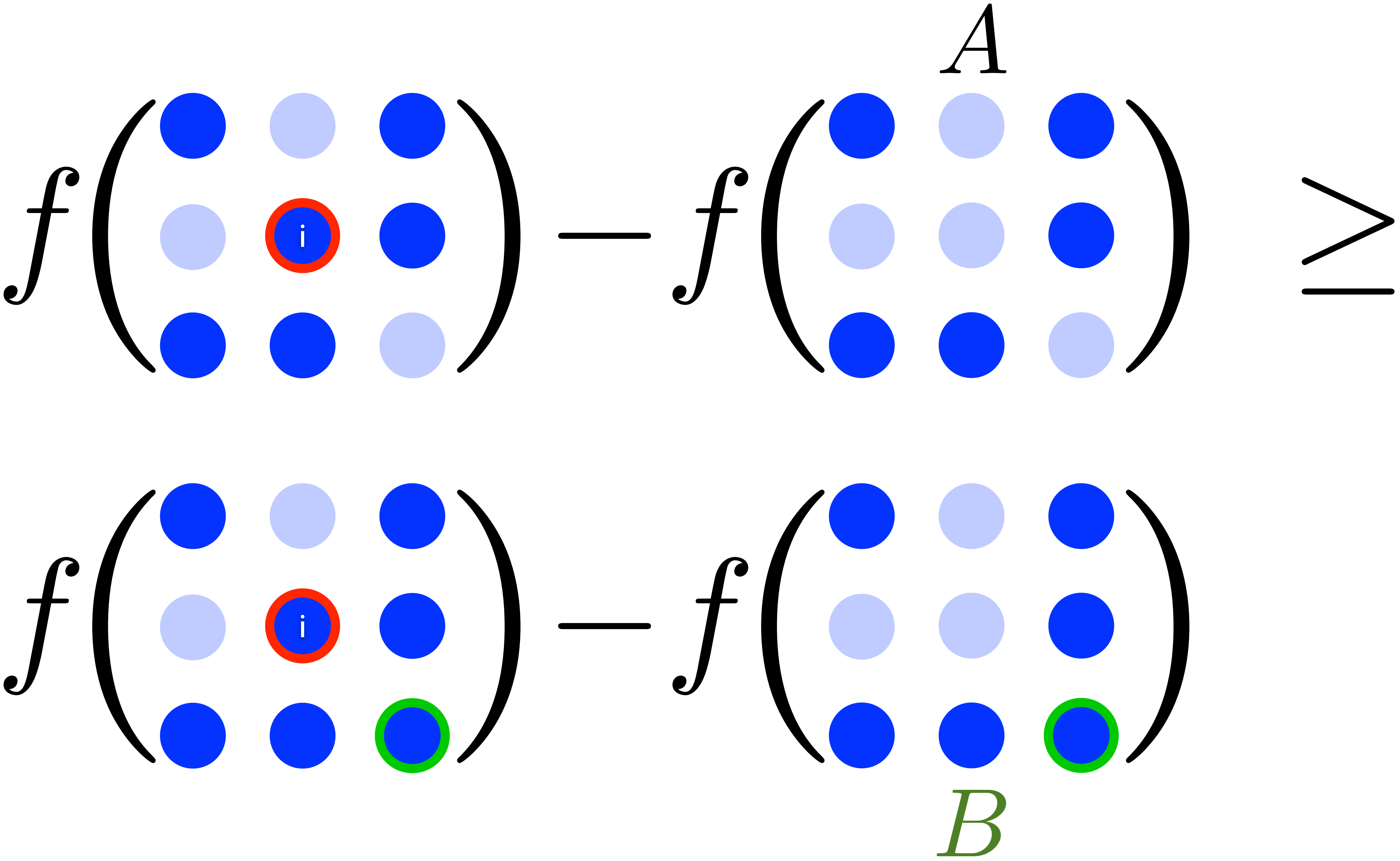
SUBMODULARITY



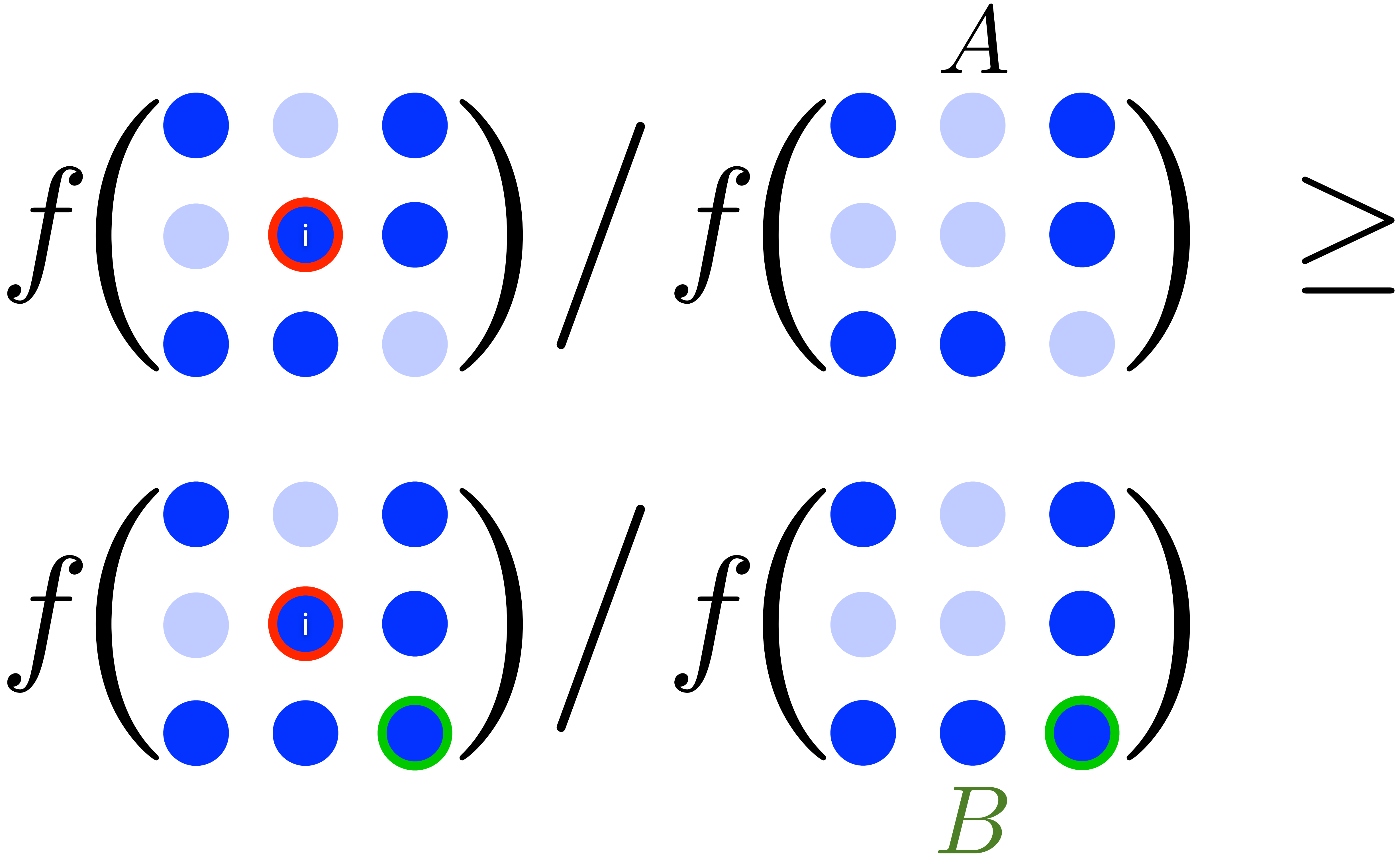
SUBMODULARITY



LOG-SUBMODULARITY



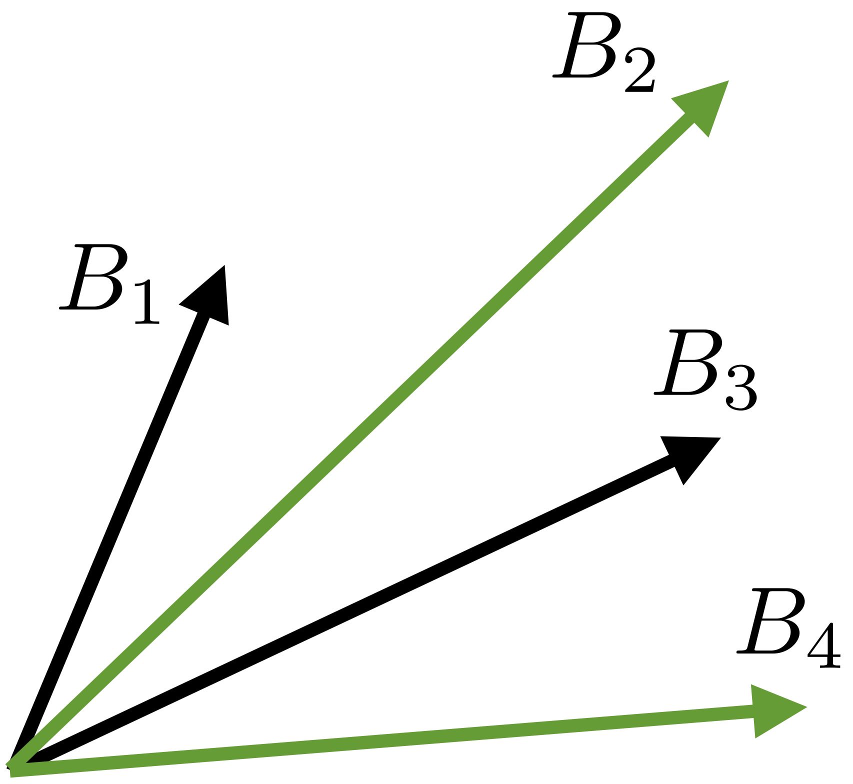
LOG-SUBMODULARITY



CHEKURI ET AL. 2011

CHEKURI ET AL. 2011

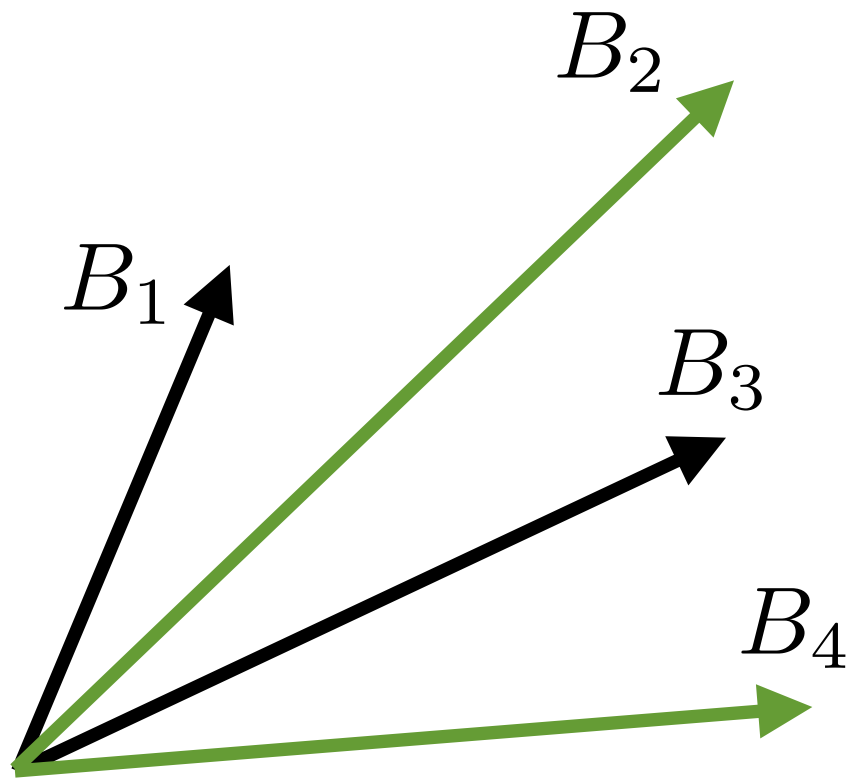
Step 1: Relax inclusion-exclusion



$Y = \{2, 4\}$

CHEKURI ET AL. 2011

Step 1: Relax inclusion-exclusion

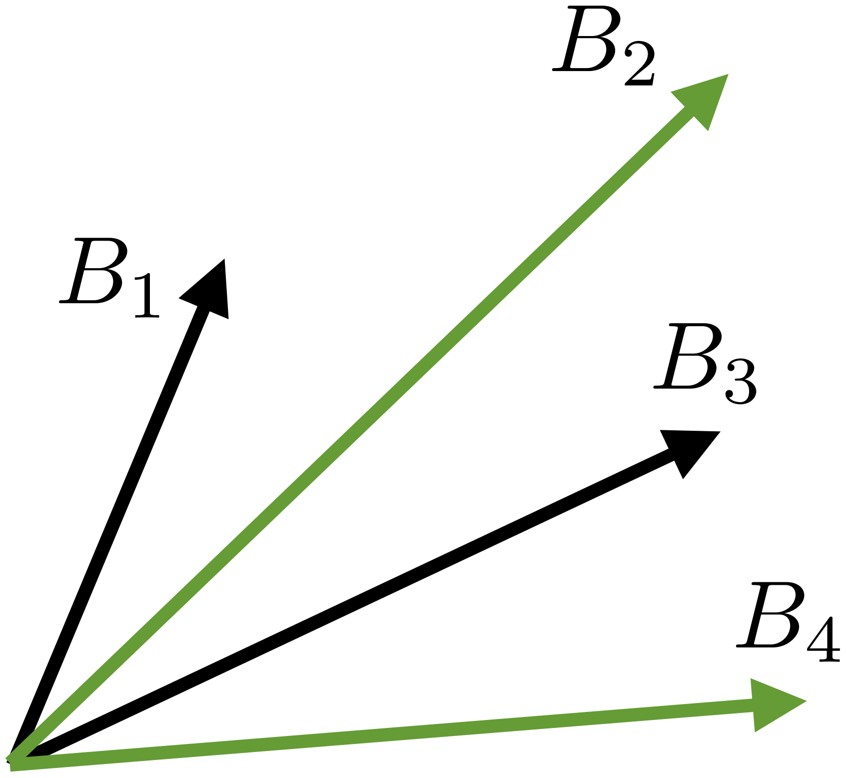


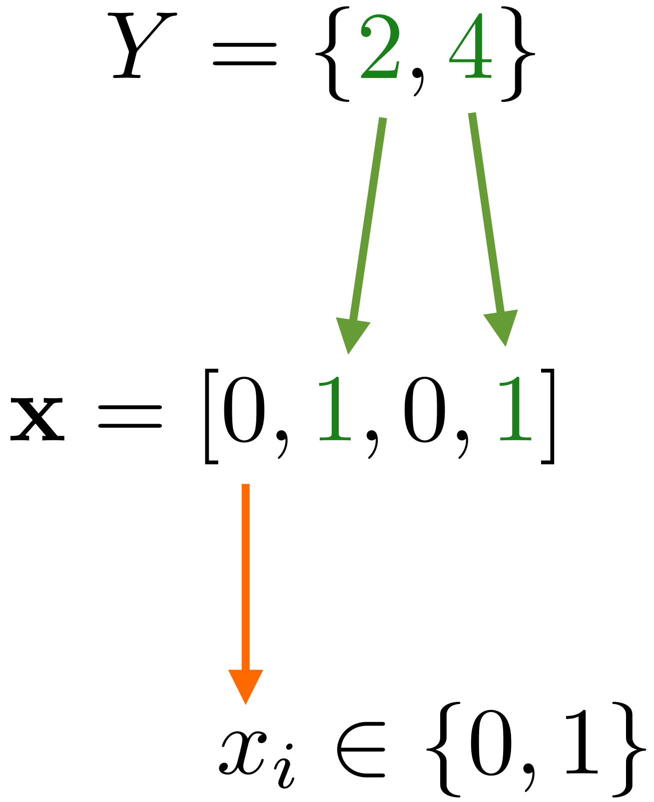
$$Y = \{2, 4\}$$

$$\mathbf{x} = [0, 1, 0, 1]$$

CHEKURI ET AL. 2011

Step 1: Relax inclusion-exclusion





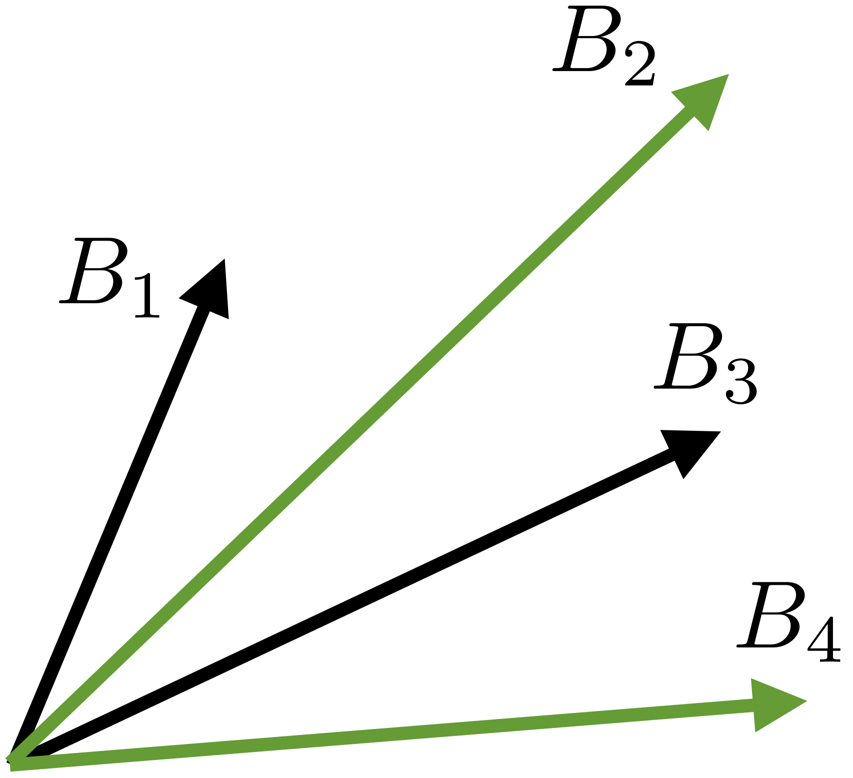
$$Y = \{2, 4\}$$

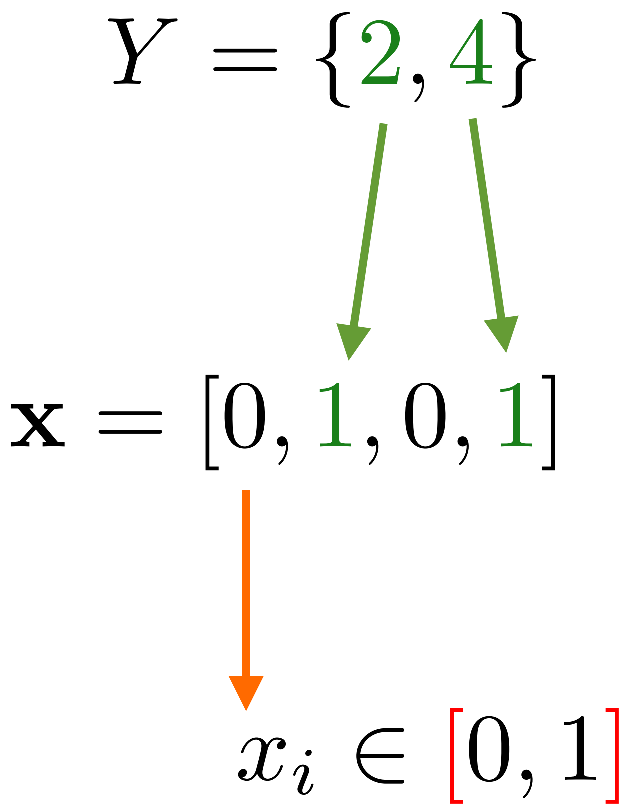
$$\mathbf{x} = [0, 1, 0, 1]$$

$$x_i \in \{0, 1\}$$

CHEKURI ET AL. 2011

Step 1: Relax inclusion-exclusion





$$Y = \{2, 4\}$$

$$\mathbf{x} = [0, 1, 0, 1]$$

$$x_i \in [0, 1]$$

CHEKURI ET AL. 2011

Step 2: Extend objective

$$F(\mathbf{x}) = E_{\mathbf{x}}[\log \boxed{f(Y)}]$$ multilinear extension

$\downarrow$

log-submodular, like $\det(L_Y)$

CHEKURI ET AL. 2011

Step 2: Extend objective

$$F(\mathbf{x}) = E_{\mathbf{x}}[\log f(Y)]$$ multilinear extension

$$= \sum_Y \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) \log f(Y)$$

CHEKURI ET AL. 2011

Step 2: Extend objective

$$F(\mathbf{x}) = E_{\mathbf{x}}[\log f(Y)] \qquad \text{multilinear extension}$$

$$= \sum_Y \underbrace{\boxed{\prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i)}}_{p_Y} \log f(Y)$$

CHEKURI ET AL. 2011

Step 2: Extend objective

$$F(\mathbf{x}) = E_{\mathbf{x}}[\log f(Y)]$$ multilinear extension

$$= \boxed{\textstyle\sum_Y} \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) \log f(Y)$$

2^N subsets

CHEKURI ET AL. 2011

Step 2: Extend objective

$$F(\mathbf{x}) = E_{\mathbf{x}}[\log f(Y)]$$ multilinear extension

$$= \underbrace{\boxed{\sum_{Y}}}_{2^N \ \text{subsets}} \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) \log f(Y)$$

Step 3: Optimize using gradient-based methods $\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}$

CHEKURI ET AL. 2011

Step 2: Extend objective

$$F(\mathbf{x}) = E_{\mathbf{x}}[\log f(Y)]$$ multilinear extension

$$= \boxed{\sum_Y} \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) \log f(Y)$$

2^N subsets

Step 3: Optimize using gradient-based methods $\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}$
Step 4: If unconstrained, solution will already be integer;
else, round solution: $x_i \in [0, 1] \rightarrow x_i \in \{0, 1\}$

CHEKURI ET AL. 2011

Step 2: Extend objective

$$F(\mathbf{x}) = E_{\mathbf{x}}[\log f(Y)]$$ multilinear extension

$$= \boxed{\textstyle\sum_Y} \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) \log f(Y)$$

$$2^N \text{ subsets } \implies \text{Monte Carlo required}$$

Step 3: Optimize using gradient-based methods $\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}$

Step 4: If unconstrained, solution will already be integer; else, round solution: $x_i \in [0, 1] \rightarrow x_i \in \{0, 1\}$

SOFTMAX EXTENSION

GILLENWATER, KULESZA, AND TASKAR (NIPS 2012)

Multilinear: $F(\mathbf{x}) = \sum_Y \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) \log f(Y)$

SOFTMAX EXTENSION

GILLENWATER, KULESZA, AND TASKAR (NIPS 2012)

Multilinear: $F(\mathbf{x}) = \sum_Y \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) \log f(Y)$

Softmax: $\tilde{F}(\mathbf{x}) = \log \sum_Y \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) f(Y)$

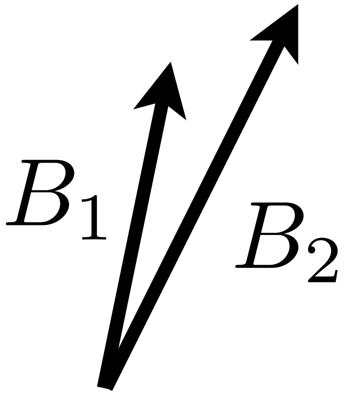
SOFTMAX EXTENSION

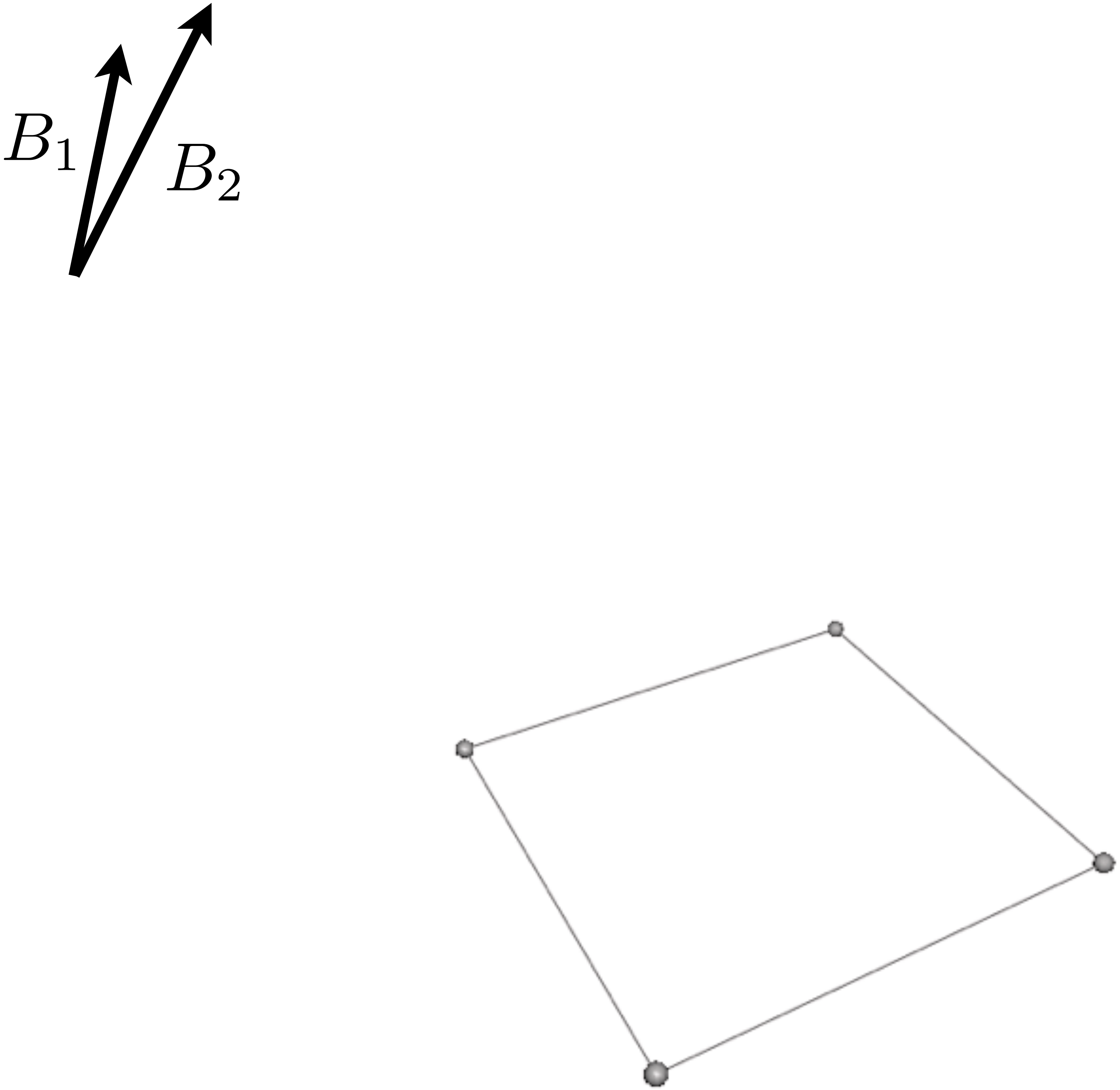
GILLENWATER, KULESZA, AND TASKAR (NIPS 2012)

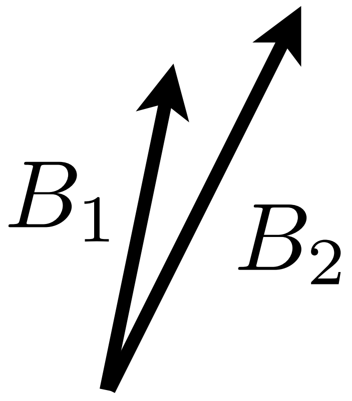
Multilinear: $F(\mathbf{x}) = \sum_Y \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) \log f(Y)$

Softmax: $\tilde{F}(\mathbf{x}) = \log \sum_Y \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) f(Y)$

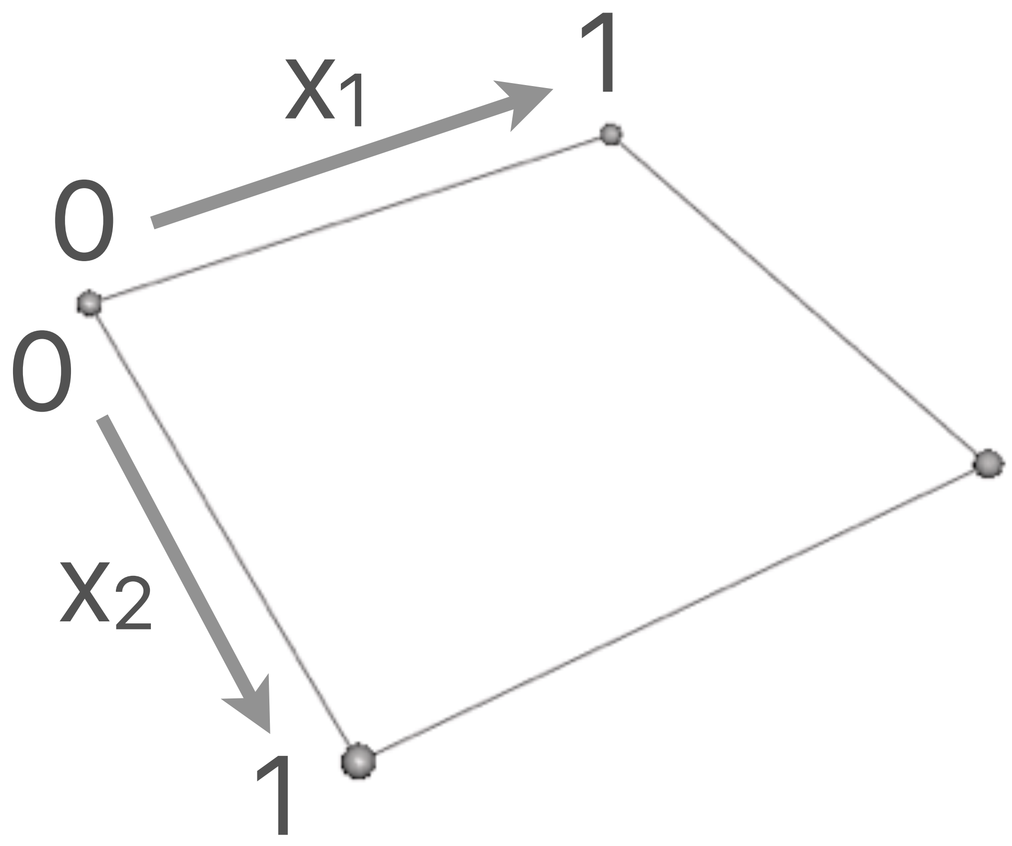
$N = 2$

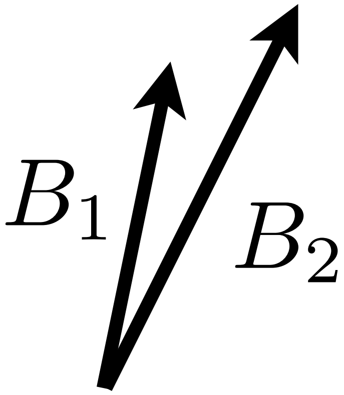



B_1
B_2

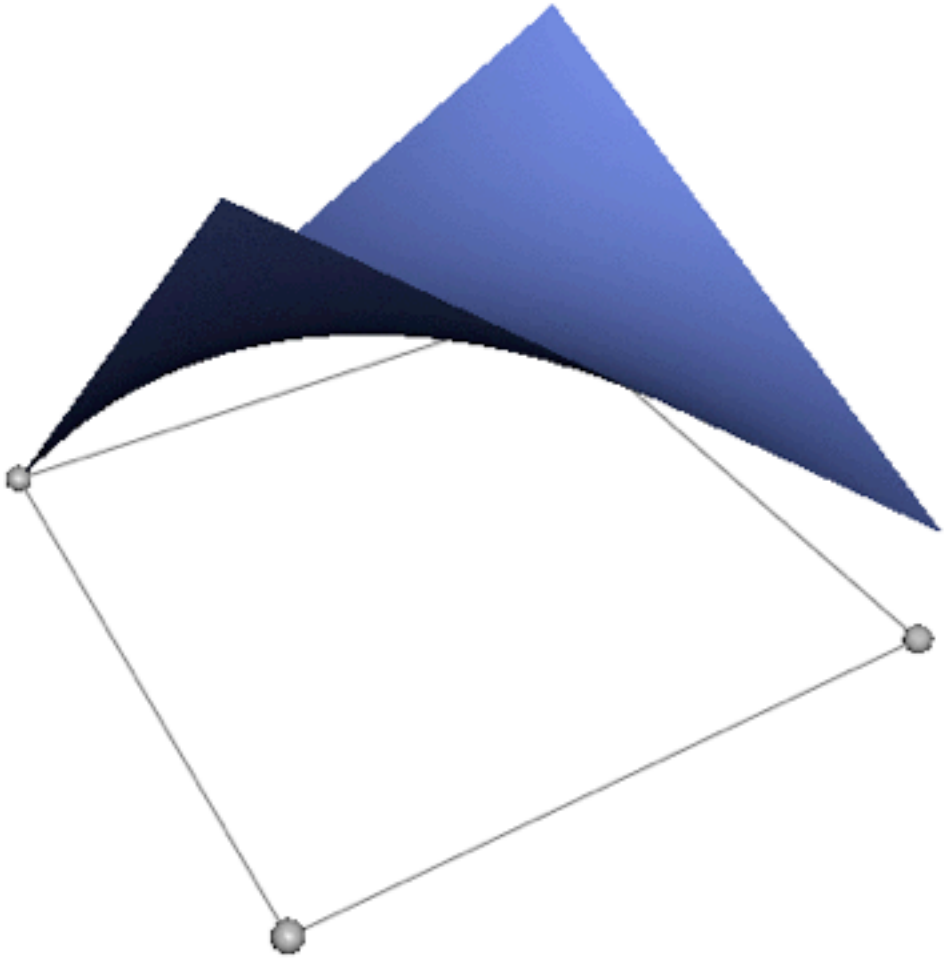


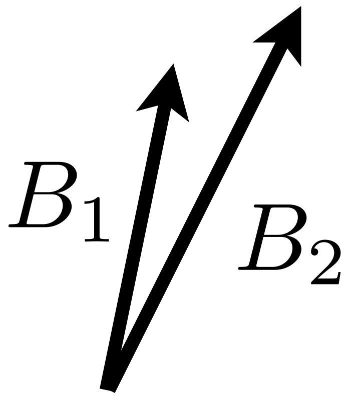
Relaxed domain





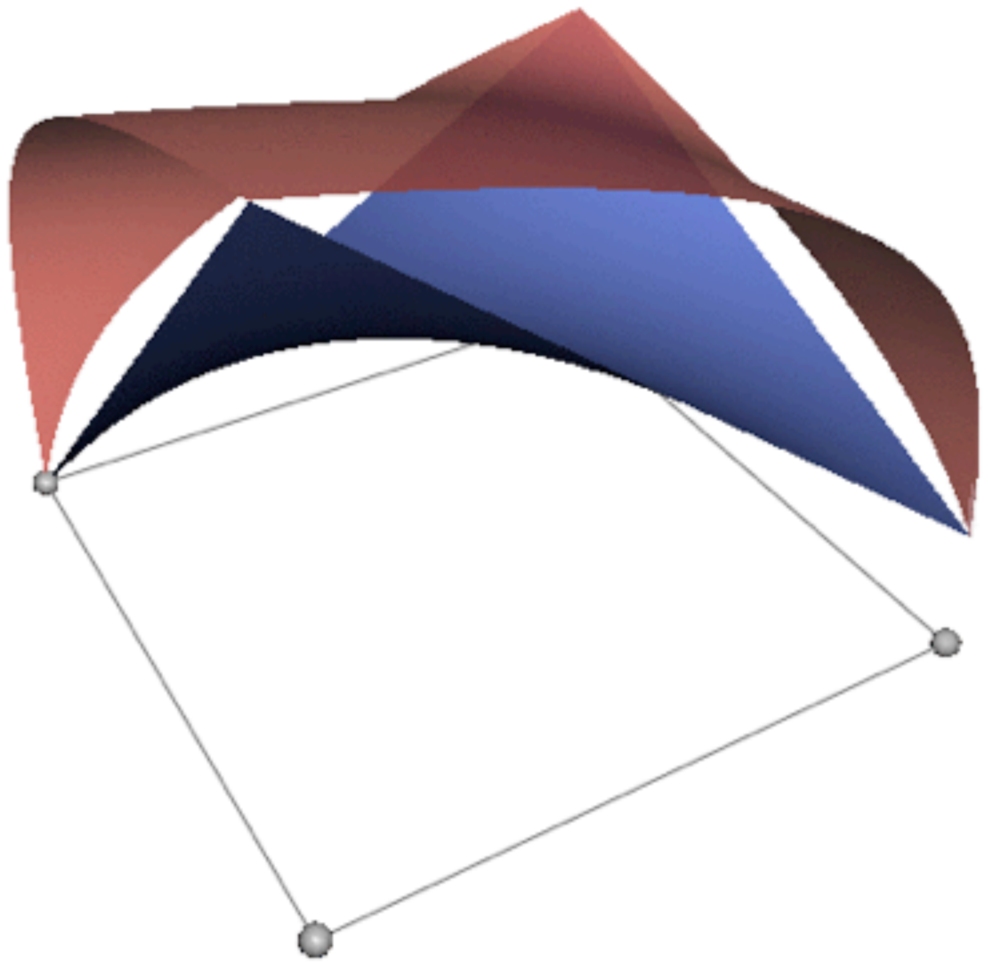
Multilinear extension

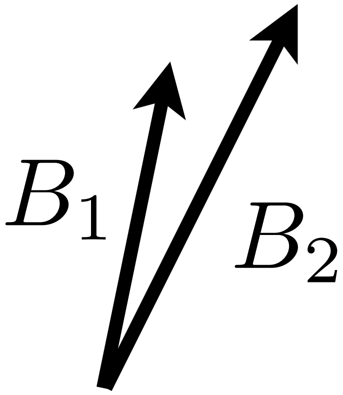




Softmax extension

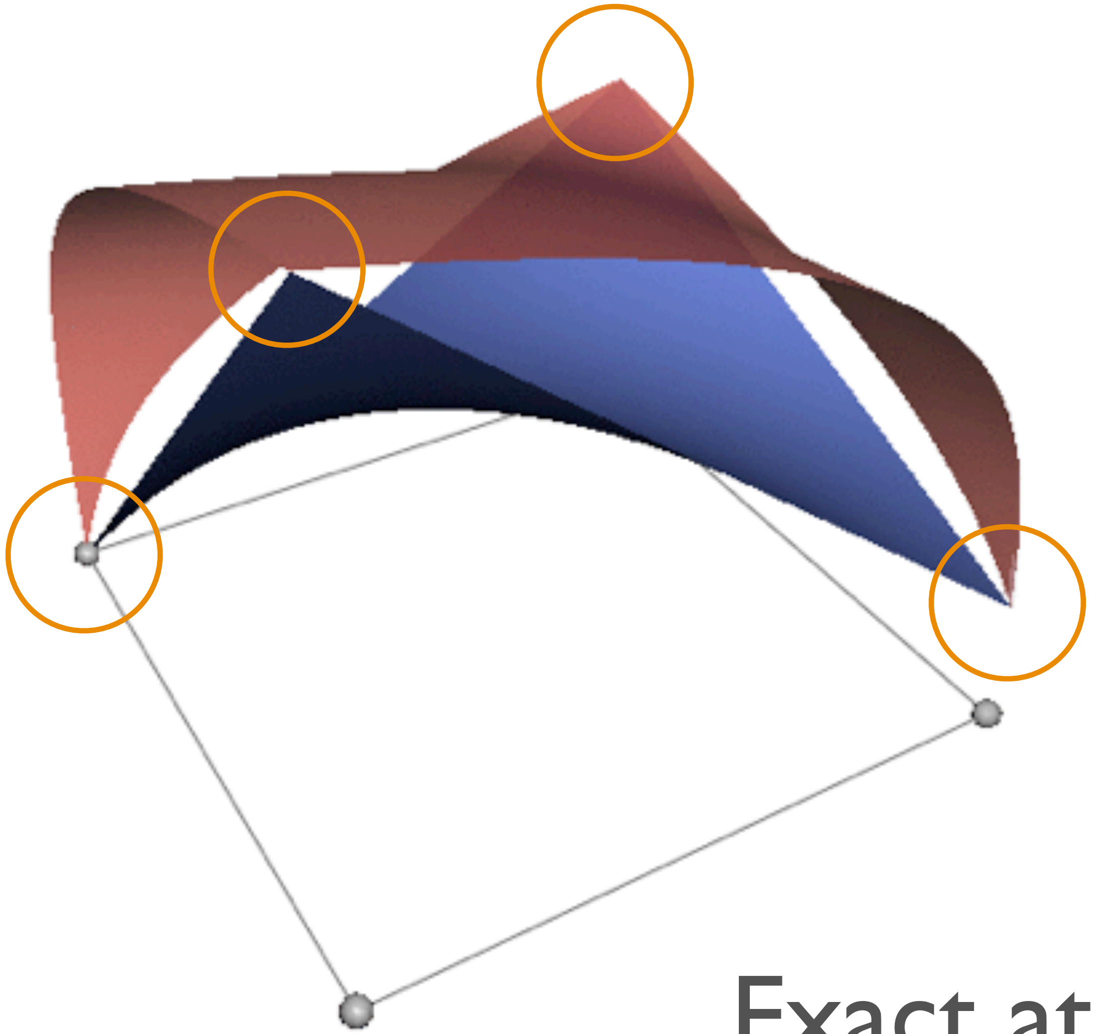
Multilinear extension





Softmax extension

Multilinear extension



Exact at integral points

SOFTMAX EXTENSION

GILLENWATER, KULESZA, AND TASKAR (NIPS 2012)

$$\tilde{F}(\mathbf{x}) = \log \sum_Y \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) f(Y)$$

SOFTMAX EXTENSION

GILLENWATER, KULESZA, AND TASKAR (NIPS 2012)

$$\tilde{F}(\mathbf{x}) = \log \sum_Y \prod_{i \in Y} x_i \prod_{i \notin Y} (1 - x_i) f(Y)$$

Theorem:

Efficiently computable for $f(Y) = \det(L_Y)$

$$O(N^3)$$

$$\tilde{F}(\mathbf{x}) = \log \det(\operatorname{diag}(\mathbf{x})(L - I) + I)$$

EFFICIENCY PROOF

$$\exp(\tilde{F}(\boldsymbol{x})) = \sum_{Y : Y \subseteq \mathcal{Y}} \det(L_Y) \prod_{i : i \in Y} x_i \prod_{i : i \notin Y} (1 - x_i)$$

EFFICIENCY PROOF

$$\begin{aligned}\exp(\tilde{F}(\boldsymbol{x})) &= \sum_{Y:Y\subseteq\mathcal{Y}} \det(L_Y) \prod_{i:i\in Y} x_i \prod_{i:i\notin Y} (1-x_i) \\ &= \prod_{i=1}^{N}(1-x_i) \sum_{Y:Y\subseteq\mathcal{Y}} \det(L_Y) \prod_{i:i\in Y} \frac{x_i}{1-x_i}\end{aligned}$$

EFFICIENCY PROOF

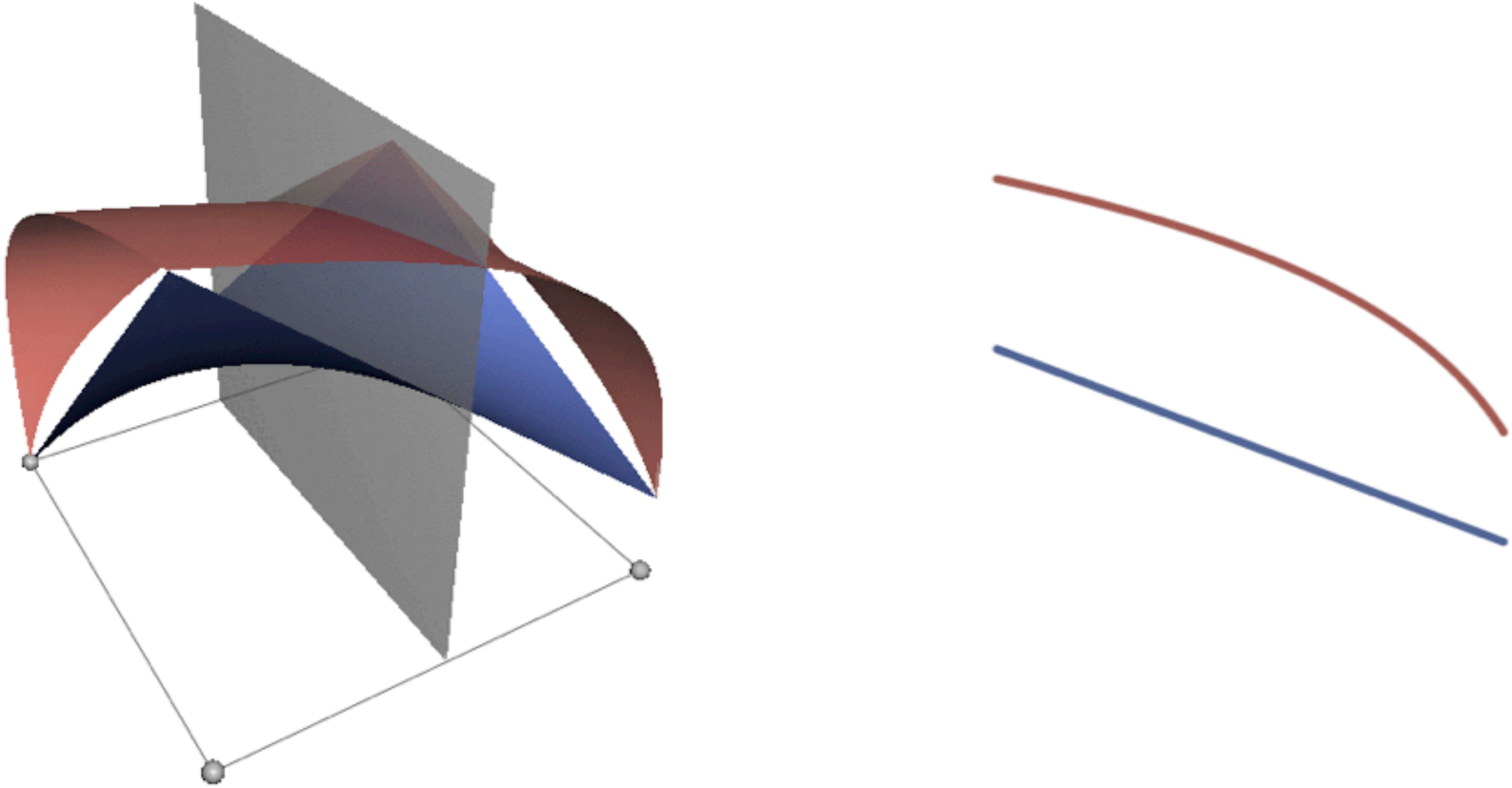
$$\begin{aligned}\exp(\tilde{F}(\boldsymbol{x})) &= \sum_{Y:Y\subseteq\mathcal{Y}} \det(L_Y) \prod_{i:i\in Y} x_i \prod_{i:i\notin Y} (1-x_i) \\ &= \prod_{i=1}^{N}(1-x_i) \sum_{Y:Y\subseteq\mathcal{Y}} \det(L_Y) \prod_{i:i\in Y} \frac{x_i}{1-x_i} \\ &= \prod_{i=1}^{N}(1-x_i) \sum_{Y:Y\subseteq\mathcal{Y}} \det\left(\left[\mathrm{diag}(\boldsymbol{x})\mathrm{diag}(1-\boldsymbol{x})^{-1}L\right]_Y\right)\end{aligned}$$

EFFICIENCY PROOF

$$
\begin{aligned}
\exp(\tilde{F}(\boldsymbol{x})) &= \sum_{Y:Y\subseteq\mathcal{Y}} \det(L_Y) \prod_{i:i\in Y} x_i \prod_{i:i\notin Y} (1-x_i) \\
&= \prod_{i=1}^{N}(1-x_i) \sum_{Y:Y\subseteq\mathcal{Y}} \det(L_Y) \prod_{i:i\in Y} \frac{x_i}{1-x_i} \\
&= \prod_{i=1}^{N}(1-x_i) \sum_{Y:Y\subseteq\mathcal{Y}} \det\left(\left[\mathrm{diag}(\boldsymbol{x})\mathrm{diag}(1-\boldsymbol{x})^{-1}L\right]_Y\right) \\
&= \prod_{i=1}^{N}(1-x_i)\det(\mathrm{diag}(\boldsymbol{x})\mathrm{diag}(1-\boldsymbol{x})^{-1}L+I)
\end{aligned}
$$

Concave in all-positive/all-negative directions

Concave in all-positive/all-negative directions



CONCAVITY PROOF

For $\boldsymbol{u} \geq \mathbf{0}$ and $\mathbf{0} < \boldsymbol{x} + s\boldsymbol{u} < \mathbf{1}$: $\frac{\partial^2}{\partial s^2} \tilde{F}(\boldsymbol{x} + s\boldsymbol{u}) \leq 0.$

CONCAVITY PROOF

For $\boldsymbol{u} \geq \mathbf{0}$ and $\mathbf{0} < \boldsymbol{x} + s\boldsymbol{u} < \mathbf{1}$: $\frac{\partial^2}{\partial s^2}\tilde{F}(\boldsymbol{x} + s\boldsymbol{u}) \leq 0.$

$$\frac{\partial^2}{\partial s^2}\tilde{F}(\boldsymbol{x} + s\boldsymbol{u}) = \mathrm{tr}(-D^{-2}\mathrm{diag}(\boldsymbol{u}^2) \\ + 2M^{-1}D^{-3}\mathrm{diag}(\boldsymbol{u}^2) \\ - (M^{-1}D^{-2}\mathrm{diag}(\boldsymbol{u}))^2)$$

CONCAVITY PROOF

For $\boldsymbol{u} \geq \mathbf{0}$ and $\mathbf{0} < \boldsymbol{x} + s\boldsymbol{u} < \mathbf{1}$: $\frac{\partial^2}{\partial s^2}\tilde{F}(\boldsymbol{x} + s\boldsymbol{u}) \leq 0.$

$$\begin{aligned}\frac{\partial^2}{\partial s^2}\tilde{F}(\boldsymbol{x} + s\boldsymbol{u}) &= \mathrm{tr}(-D^{-2}\mathrm{diag}(\boldsymbol{u}^2) \\ &\qquad + 2M^{-1}D^{-3}\mathrm{diag}(\boldsymbol{u}^2) \\ &\qquad - (M^{-1}D^{-2}\mathrm{diag}(\boldsymbol{u}))^2) \\ &= -\mathrm{tr}(SS^\top) \\ &\quad \text{where } S = D^{-1}\mathrm{diag}(\boldsymbol{u}) \\ &\qquad\qquad - M^{-1}D^{-2}\mathrm{diag}(\boldsymbol{u})\end{aligned}$$

CONCAVITY PROOF

For $\boldsymbol{u} \geq \mathbf{0}$ and $\mathbf{0} < \boldsymbol{x} + s\boldsymbol{u} < \mathbf{1}$: $\frac{\partial^2}{\partial s^2}\tilde{F}(\boldsymbol{x} + s\boldsymbol{u}) \leq 0.$

$$\begin{aligned}
\frac{\partial^2}{\partial s^2}\tilde{F}(\boldsymbol{x} + s\boldsymbol{u}) = \operatorname{tr}(&-D^{-2}\operatorname{diag}(\boldsymbol{u}^2) \\
&+ 2M^{-1}D^{-3}\operatorname{diag}(\boldsymbol{u}^2) \\
&- (M^{-1}D^{-2}\operatorname{diag}(\boldsymbol{u}))^2) \\
= \; &-\operatorname{tr}(SS^\top) \\
&\text{where } S = D^{-1}\operatorname{diag}(\boldsymbol{u}) \\
&\qquad\qquad - M^{-1}D^{-2}\operatorname{diag}(\boldsymbol{u}) \\
\leq 0 &
\end{aligned}$$

APPROXIMATION GUARANTEE

Theorem: Concavity in all-positive directions

APPROXIMATION GUARANTEE

Theorem: Concavity in all-positive directions

+ Submodularity $\Longrightarrow$

$\text{LOCAL OPT of } \tilde{F} \geq \frac{1}{4} \max_{\mathbf{x}} \tilde{F}(\mathbf{x})$

APPROXIMATION GUARANTEE

Theorem: Concavity in all-positive directions

+ Submodularity $\Longrightarrow$

$$\text{LOCAL OPT of } \tilde{F} \geq \tfrac{1}{4}\max_{\mathbf{x}} \tilde{F}(\mathbf{x})$$

Theorem: In the unconstrained case, LOCAL OPT will be integer (no rounding necessary).

APPROXIMATION GUARANTEE

Theorem: Concavity in all-positive directions

+ Submodularity $\Longrightarrow$

$$\text{LOCAL OPT of } \tilde{F} \geq \tfrac{1}{4} \max_{\mathbf{x}} \tilde{F}(\mathbf{x}) \geq \tfrac{1}{4} \max_{Y} \log\det(L_Y)$$

Theorem: In the unconstrained case, LOCAL OPT will be integer (no rounding necessary).

APPROXIMATION GUARANTEE

Theorem: Concavity in all-positive directions

+ Submodularity $\Longrightarrow$

$$\text{LOCAL OPT of } \tilde{F} \geq \tfrac{1}{4} \max_{\mathbf{x}} \tilde{F}(\mathbf{x}) \geq \tfrac{1}{4} \max_{Y} \log\det(L_Y)$$

Theorem: In the unconstrained case, LOCAL OPT will be integer (no rounding necessary).

Constrained: No guarantees, but in practice pipage rounding and thresholding work well.

$\max_{Y \in S}$

MATCHED SUMMARIZATION

MATCHED SUMMARIZATION

20 Republican primary debates

MATCHED SUMMARIZATION

20 Republican primary debates

MATCHED SUMMARIZATION

20 Republican primary debates

Average of 179 quotes per candidate

MATCHED SUMMARIZATION

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**
- **R2** (law): **We're not going to have Sharia law applied in U.S. courts.**

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**
- **R2** (law): **We're not going to have Sharia law applied in U.S. courts.**
- **R3** (**healthcare**): **I will … grant a waiver from Obamacare to all 50 states.**

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**
- **R2** (law): **We're not going to have Sharia law applied in U.S. courts.**
- **R3** (healthcare): **I will … grant a waiver from Obamacare to all 50 states.**
- **R4** (aid): **We're spending more on foreign aid than we ought to.**

- **R1 (taxes): No tax on interest, dividends, or capital gains.**
- **R2 (law): We're not going to have Sharia law applied in U.S. courts.**
- **R3 (healthcare): I will ... grant a waiver from Obamacare to all 50 states.**
- **R4 (aid): We're spending more on foreign aid than we ought to.**
- **R5(healthcare): If you think what we did in Massachusetts and what President Obama did are the same, boy, take a closer look.**

- **R1 (taxes): No tax on interest, dividends, or capital gains.**
- **R2 (law): We're not going to have Sharia law applied in U.S. courts.**
- **R3 (healthcare): I will ... grant a waiver from Obamacare to all 50 states.**
- **R4 (aid): We're spending more on foreign aid than we ought to.**
- **R5(healthcare): If you think what we did in Massachusetts and what President Obama did are the same, boy, take a closer look.**

- **R1 (taxes): No tax on interest, dividends, or capital gains.**
- **R2 (law): We're not going to have Sharia law applied in U.S. courts.**
- **R3 (healthcare): I will ... grant a waiver from Obamacare to all 50 states.**
- **R4 (aid): We're spending more on foreign aid than we ought to.**
- **R5(healthcare): If you think what we did in Massachusetts and what President Obama did are the same, boy, take a closer look.**

- **S1 (taxes): I don't believe in a zero capital gains tax rate.**

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**
- **R2** (law): **We're not going to have Sharia law applied in U.S. courts.**
- **R3** (healthcare): **I will … grant a waiver from Obamacare to all 50 states.**
- **R4** (aid): **We're spending more on foreign aid than we ought to.**
- **R5**(healthcare): **If you think what we did in Massachusetts and what President Obama did are the same, boy, take a closer look.**

- **S1** (taxes): **I don't believe in a zero capital gains tax rate.**
- **S2** (taxes): **Manufacture in America, you aren't going to pay any taxes.**

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**
- **R2** (law): **We're not going to have Sharia law applied in U.S. courts.**
- **R3** (healthcare): **I will ... grant a waiver from Obamacare to all 50 states.**
- **R4** (aid): **We're spending more on foreign aid than we ought to.**
- **R5**(healthcare): **If you think what we did in Massachusetts and what President Obama did are the same, boy, take a closer look.**

- **S1** (taxes): **I don't believe in a zero capital gains tax rate.**
- **S2** (taxes): **Manufacture in America, you aren't going to pay any taxes.**
- **S3** (aid): **Zeroing out foreign aid ... that's absolutely the wrong course.**

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**
- **R2** (law): **We're not going to have Sharia law applied in U.S. courts.**
- **R3** (**healthcare**): **I will ... grant a waiver from Obamacare to all 50 states.**
- **R4** (aid): **We're spending more on foreign aid than we ought to.**
- **R5**(**healthcare**): **If you think what we did in Massachusetts and what President Obama did are the same, boy, take a closer look.**

- **S1** (taxes): **I don't believe in a zero capital gains tax rate.**
- **S2** (**taxes**): **Manufacture in America, you aren't going to pay any taxes.**
- **S3** (aid): **Zeroing out foreign aid ... that's absolutely the wrong course.**
- **S4** (ethanol): **I voted against ethanol subsidies my entire time in Congress.**

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**
- **R2** (law): **We're not going to have Sharia law applied in U.S. courts.**
- **R3** (healthcare): **I will ... grant a waiver from Obamacare to all 50 states.**
- **R4** (aid): **We're spending more on foreign aid than we ought to.**
- **R5**(healthcare): **If you think what we did in Massachusetts and what President Obama did are the same, boy, take a closer look.**

- **S1** (taxes): **I don't believe in a zero capital gains tax rate.**
- **S2** (taxes): **Manufacture in America, you aren't going to pay any taxes.**
- **S3** (aid): **Zeroing out foreign aid ... that's absolutely the wrong course.**
- **S4** (ethanol): **I voted against ethanol subsidies my entire time in Congress.**
- **S5** (healthcare): **Obamacare ... is going to blow a hole in the budget.**

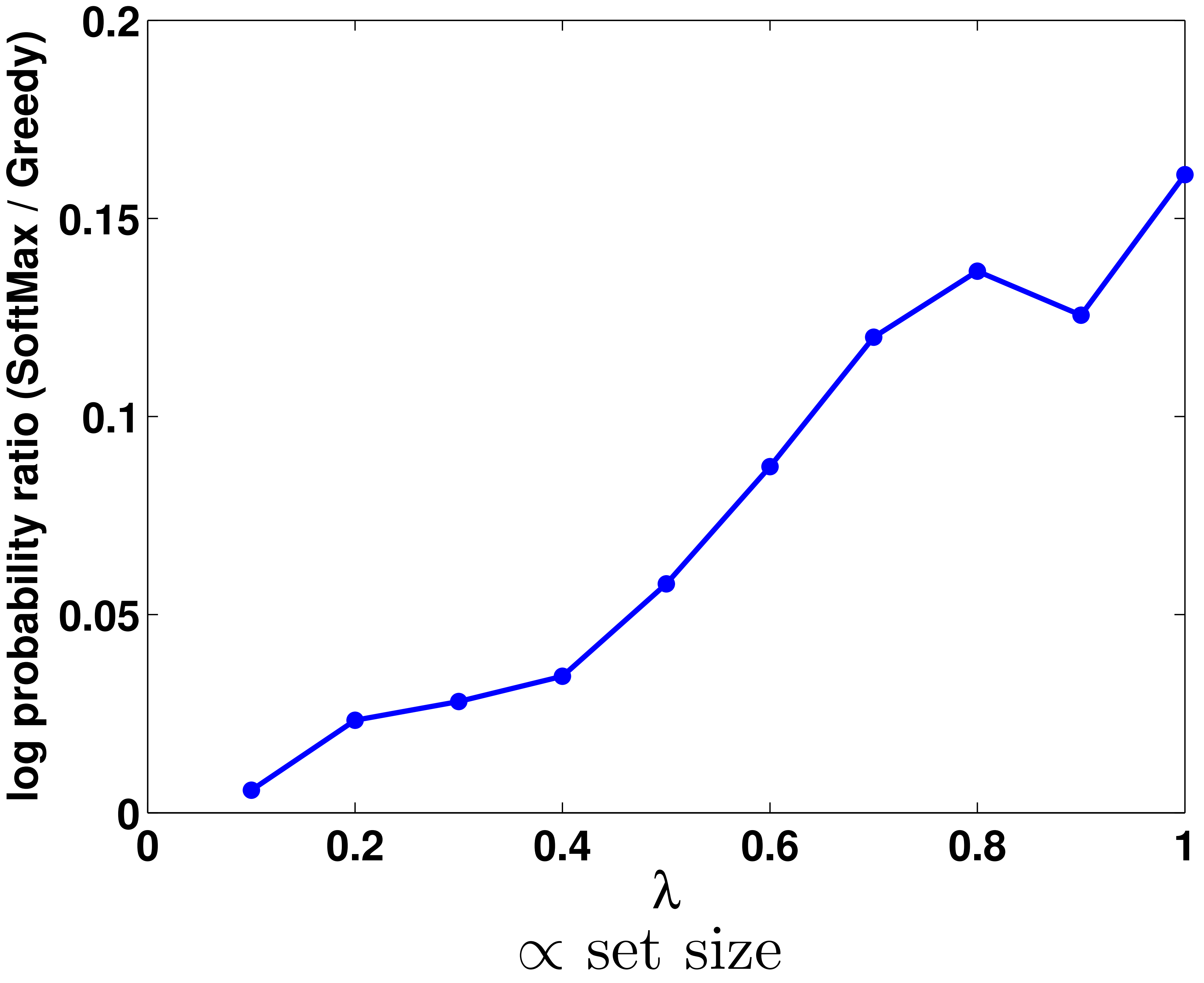
Matched summary

- **R1** (taxes): **No tax on interest, dividends, or capital gains.**
- **S1** (taxes): **I don’t believe in a zero capital gains tax rate.**

- **R3** (healthcare): **I will … grant a waiver from Obamacare to all 50 states.**
- **S5** (healthcare): **Obamacare … is going to blow a hole in the budget.**

- **R4** (aid): **We’re spending more on foreign aid than we ought to.**
- **S3** (aid): **Zeroing out foreign aid … that’s absolutely the wrong course.**

PERFORMANCE



PROPOSED WORK

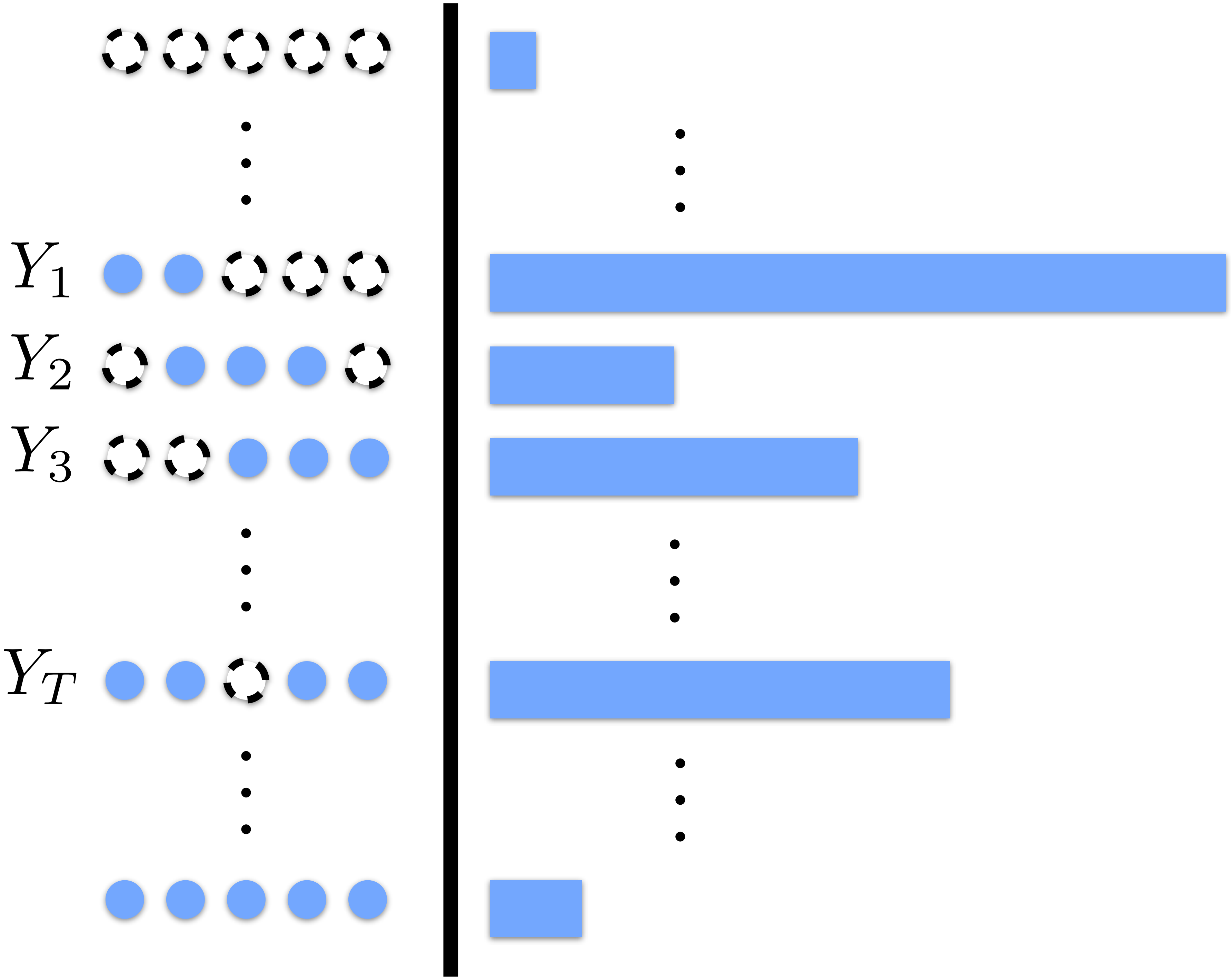
- Update experiments with faster algorithms and testing with random restarts
- k-DPP MAP estimation algorithm based on marginals, plus extension to structured k-DPPs

FUTURE WORK

- Update experiments with faster algorithms and testing with random restarts
- k-DPP MAP estimation algorithm based on marginals, plus extension to structured k-DPPs

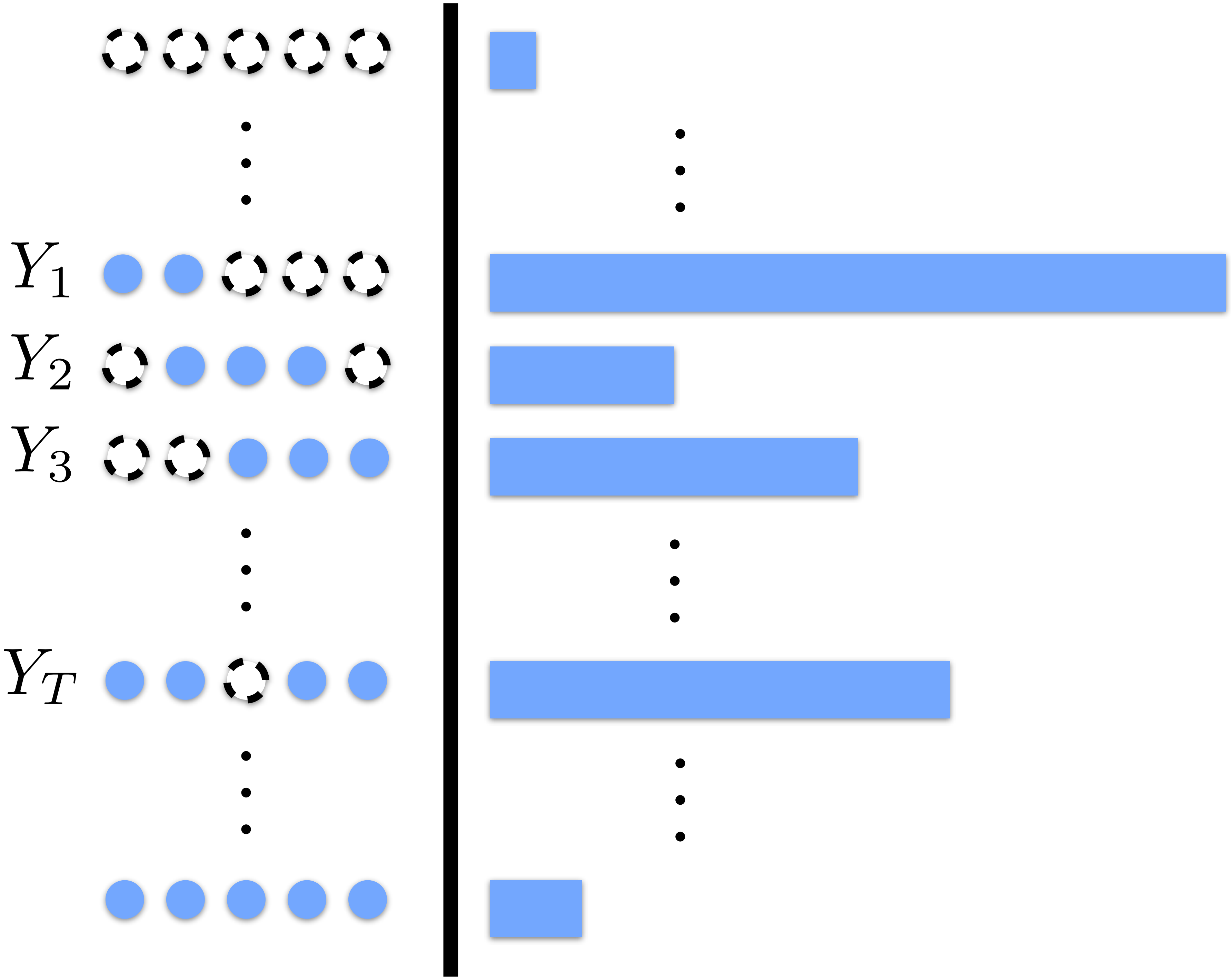
3. LIKELIHOOD MAXIMIZATION

LIKELIHOOD



LIKELIHOOD

$\mathcal{P}$ is a DPP



$$\mathcal{L}(L) = \sum_{t=1}^{T} \log\left(\frac{\det(L_{Y_t})}{\det(L+I)}\right)$$

OR

$$\mathcal{L}(K) = \sum_{t=1}^{T} \log\left(|\det(K - I_{\overline{Y}_t})|\right)$$

NON-CONCAVITY

NON-CONCAVITY

$$\mathcal{L}(L) = \sum_{t=1}^{T} \left[\log\det(L_{Y_t}) - \log\det(L + I)\right]$$

NON-CONCAVITY

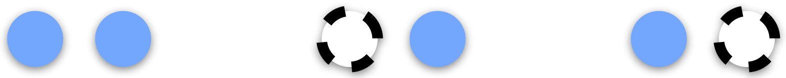
$$\mathcal{L}(L) = \sum_{t=1}^{T} \left[\log\det(L_{Y_t}) - \log\det(L + I) \right]$$

$$\max_{L: L \succeq 0} \mathcal{L}(L)$$

NON-CONCAVITY

$$\mathcal{L}(L) = \sum_{t=1}^{T} \left[\log\det(L_{Y_t}) - \log\det(L+I)\right]$$

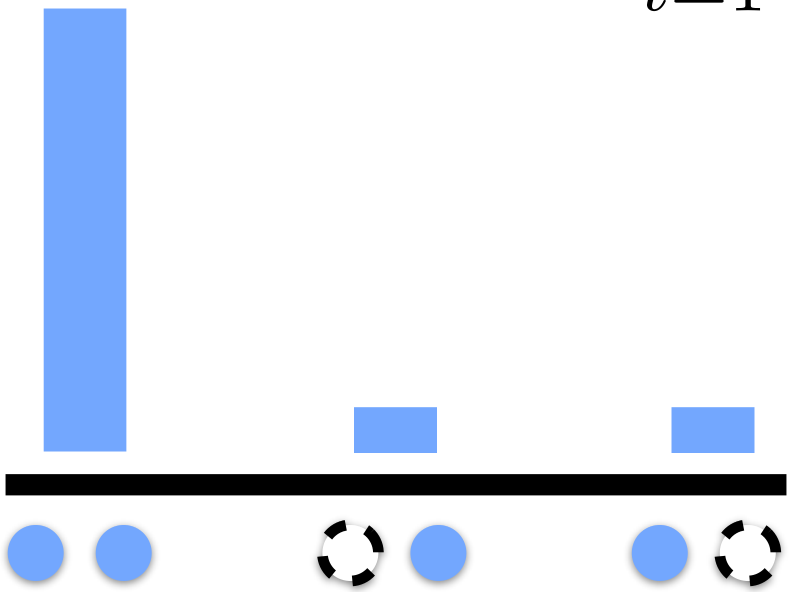
$$\max_{L:L\succeq 0} \mathcal{L}(L)$$



NON-CONCAVITY

$$\mathcal{L}(L) = \sum_{t=1}^{T} \left[\log\det(L_{Y_t}) - \log\det(L + I)\right]$$

$$\max_{L:L\succeq 0} \mathcal{L}(L)$$

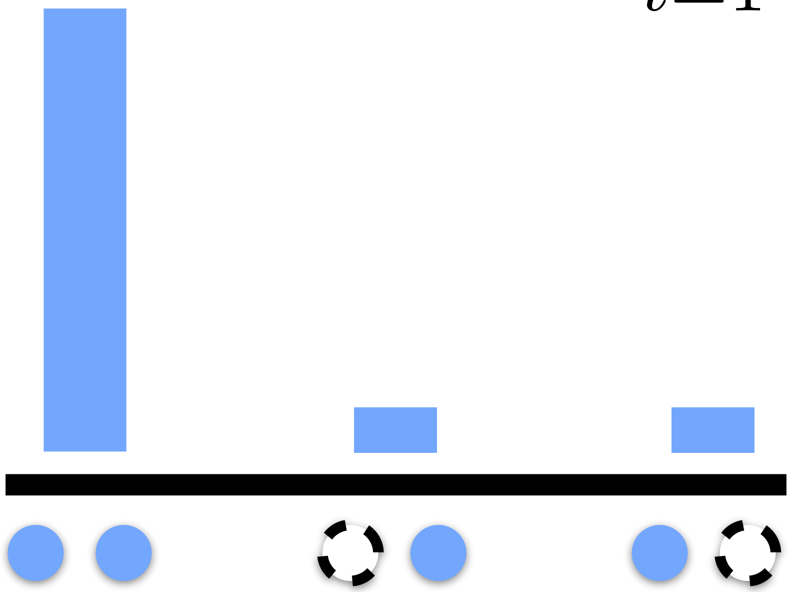


$$L = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

NON-CONCAVITY

$$\mathcal{L}(L) = \sum_{t=1}^{T} \left[\log\det(L_{Y_t}) - \log\det(L+I)\right]$$

$$\max_{L : L \succeq 0} \mathcal{L}(L)$$



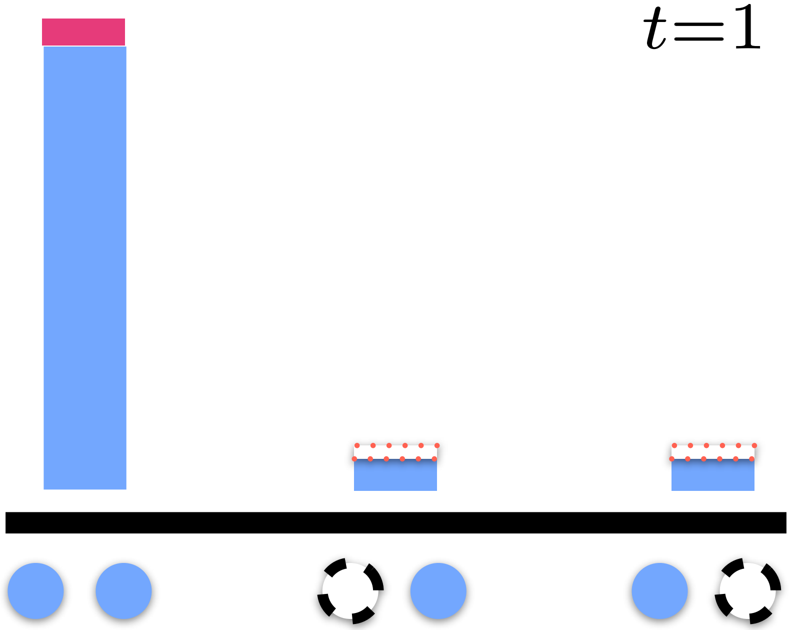
$$L = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

$$+ \eta \begin{bmatrix} 0.1 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}$$

NON-CONCAVITY

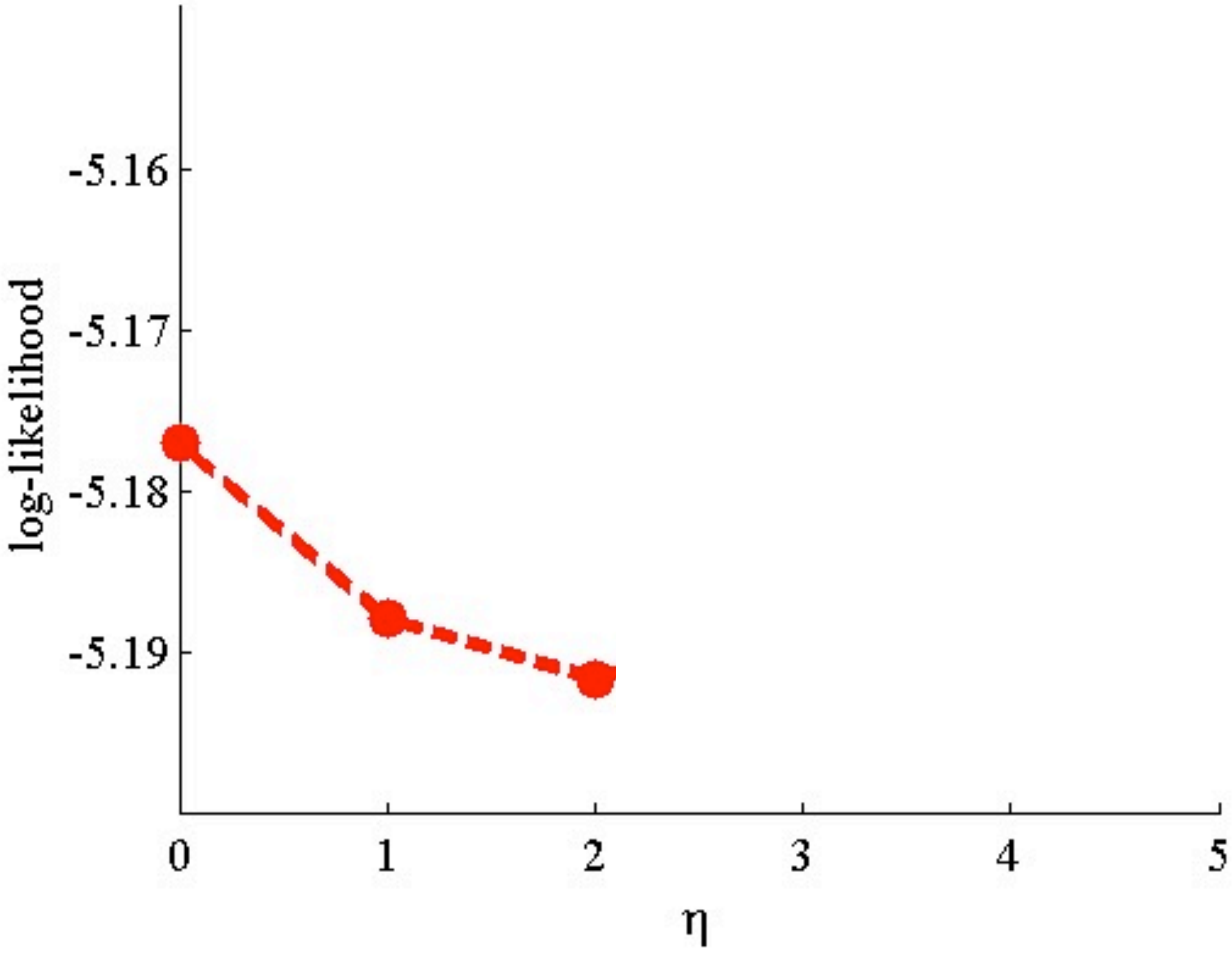
$$\mathcal{L}(L) = \sum_{t=1}^{T} \left[\log\det(L_{Y_t}) - \log\det(L+I)\right]$$

$$\max_{L:L\succeq 0} \mathcal{L}(L)$$



$$L = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

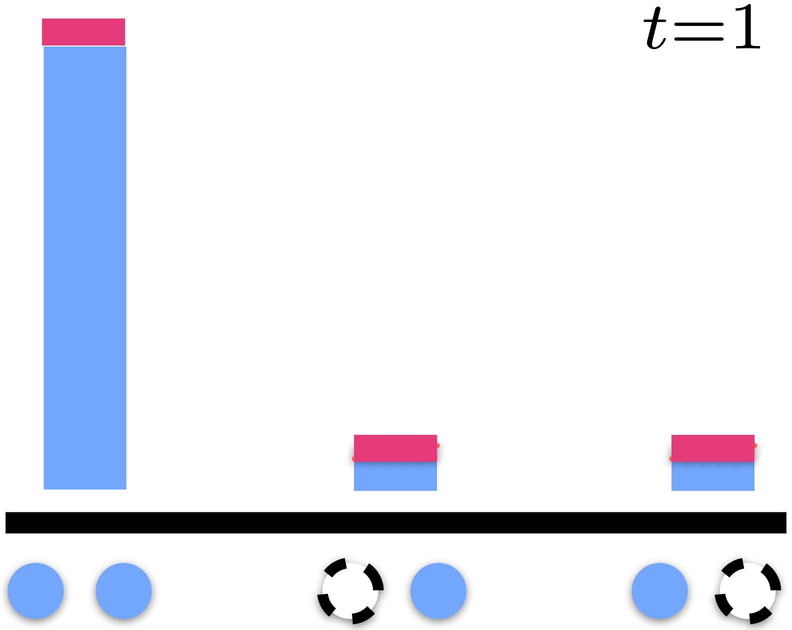
$$+\,\eta \begin{bmatrix} 0.1 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}$$



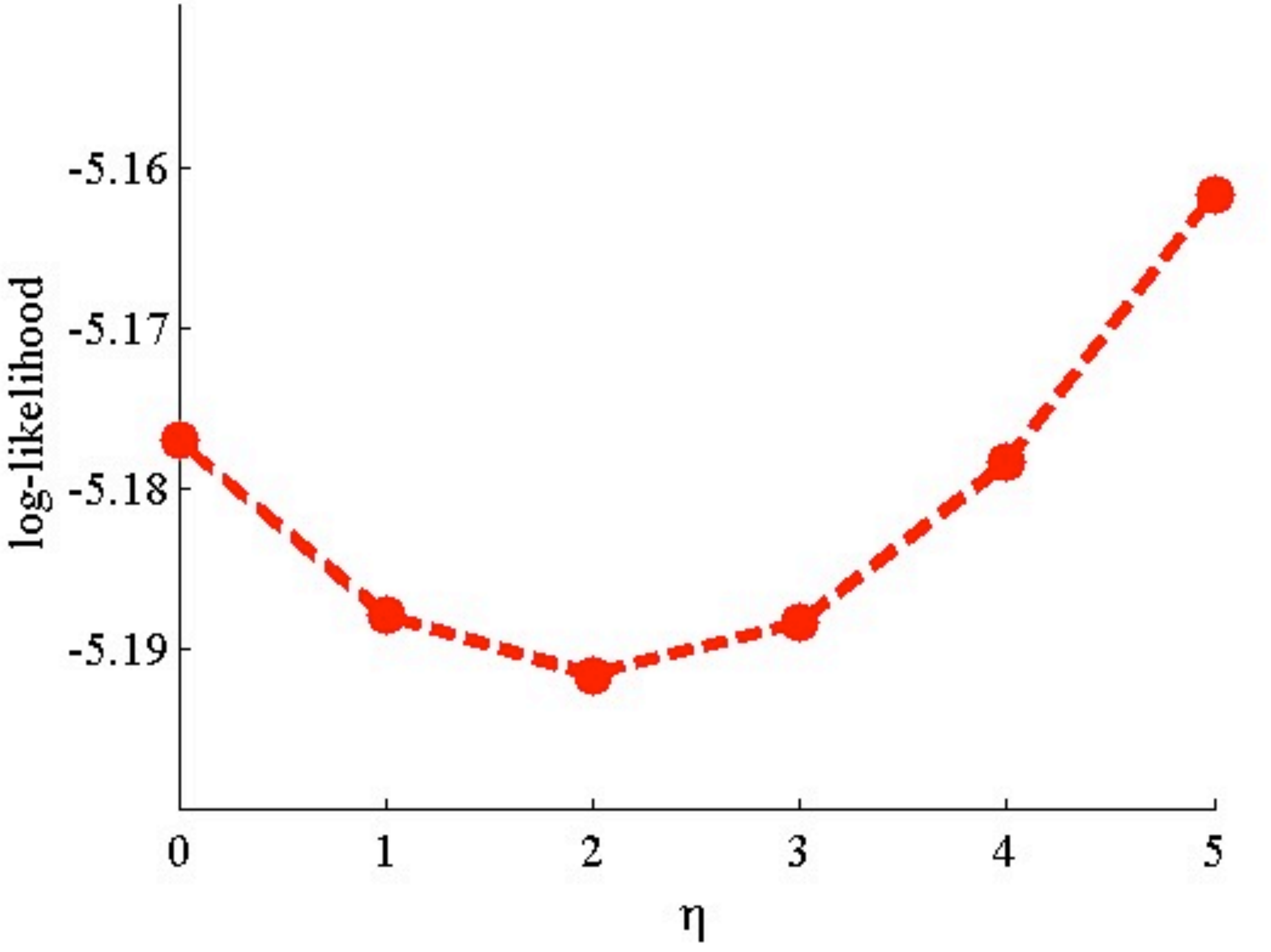
NON-CONCAVITY

$$\mathcal{L}(L) = \sum_{t=1}^{T} \left[\log\det(L_{Y_t}) - \log\det(L+I)\right]$$

$$\max_{L:L\succeq 0} \mathcal{L}(L)$$



$$L = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} + \eta \begin{bmatrix} 0.1 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}$$



K PARAMETERIZATION

K PARAMETERIZATION

Change of variables for developing EM algorithm:

$$L = V \operatorname{diag}(\boldsymbol{\lambda}) V^\top$$

$$K = V \operatorname{diag}(\tfrac{\boldsymbol{\lambda}}{1+\boldsymbol{\lambda}}) V^\top$$

K PARAMETERIZATION

Change of variables for developing EM algorithm:

$$L = V \text{diag}(\boldsymbol{\lambda}) V^\top$$

$$K = V \text{diag}\left(\frac{\boldsymbol{\lambda}}{1+\boldsymbol{\lambda}}\right) V^\top$$

$$\mathcal{L}(K) = \sum_{t=1}^{T} \log |\det(K - I_{\overline{Y}_t})|$$

K PARAMETERIZATION

Change of variables for developing EM algorithm:

$$L = V \text{diag}(\boldsymbol{\lambda}) V^\top$$

$$K = V \text{diag}(\tfrac{\boldsymbol{\lambda}}{1+\boldsymbol{\lambda}}) V^\top$$

$$\mathcal{L}(K) = \sum_{t=1}^{T} \log |\det(K - I_{\overline{Y}_t})|$$

$$\max_{\substack{K:K \succeq 0, \\ I-K \succeq 0}} \mathcal{L}(K)$$

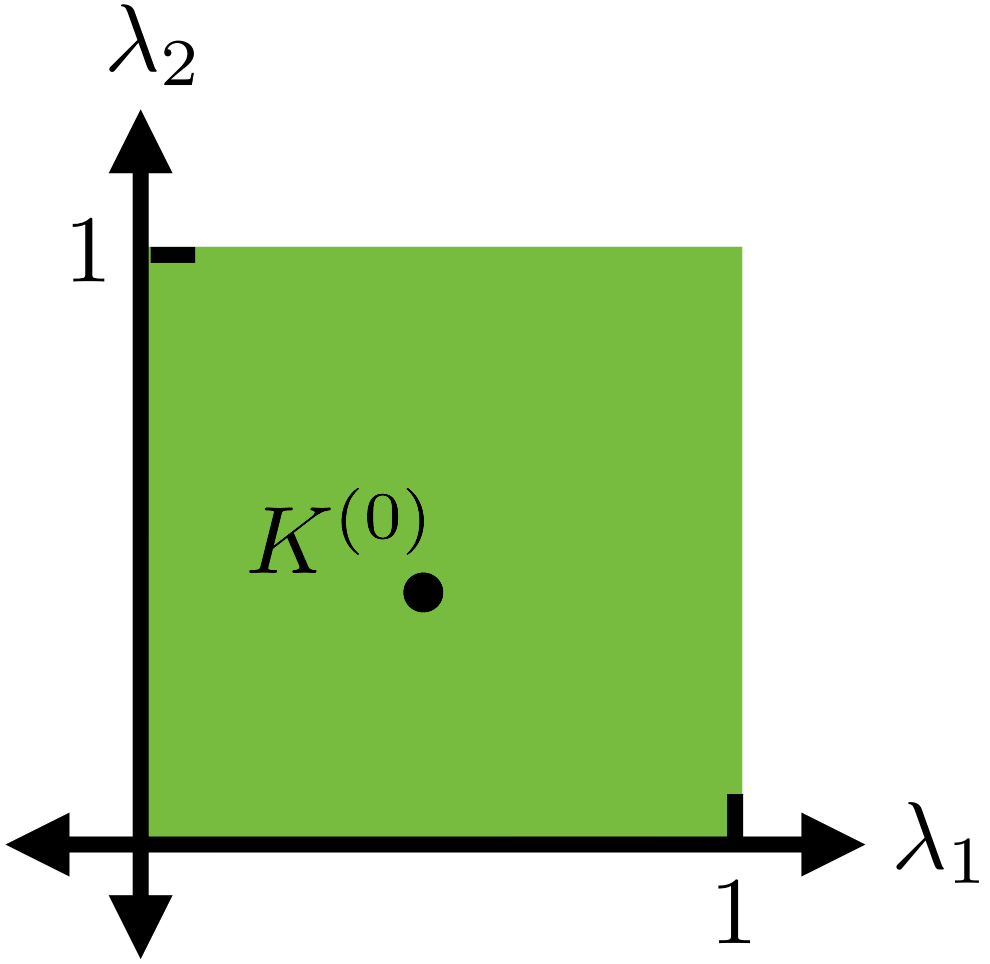
GRADIENT ASCENT

GRADIENT ASCENT

$$\frac{\partial \mathcal{L}(K)}{\partial K} = \sum_{t=1}^{T} (K - I_{\overline{Y}_t})^{-1}$$

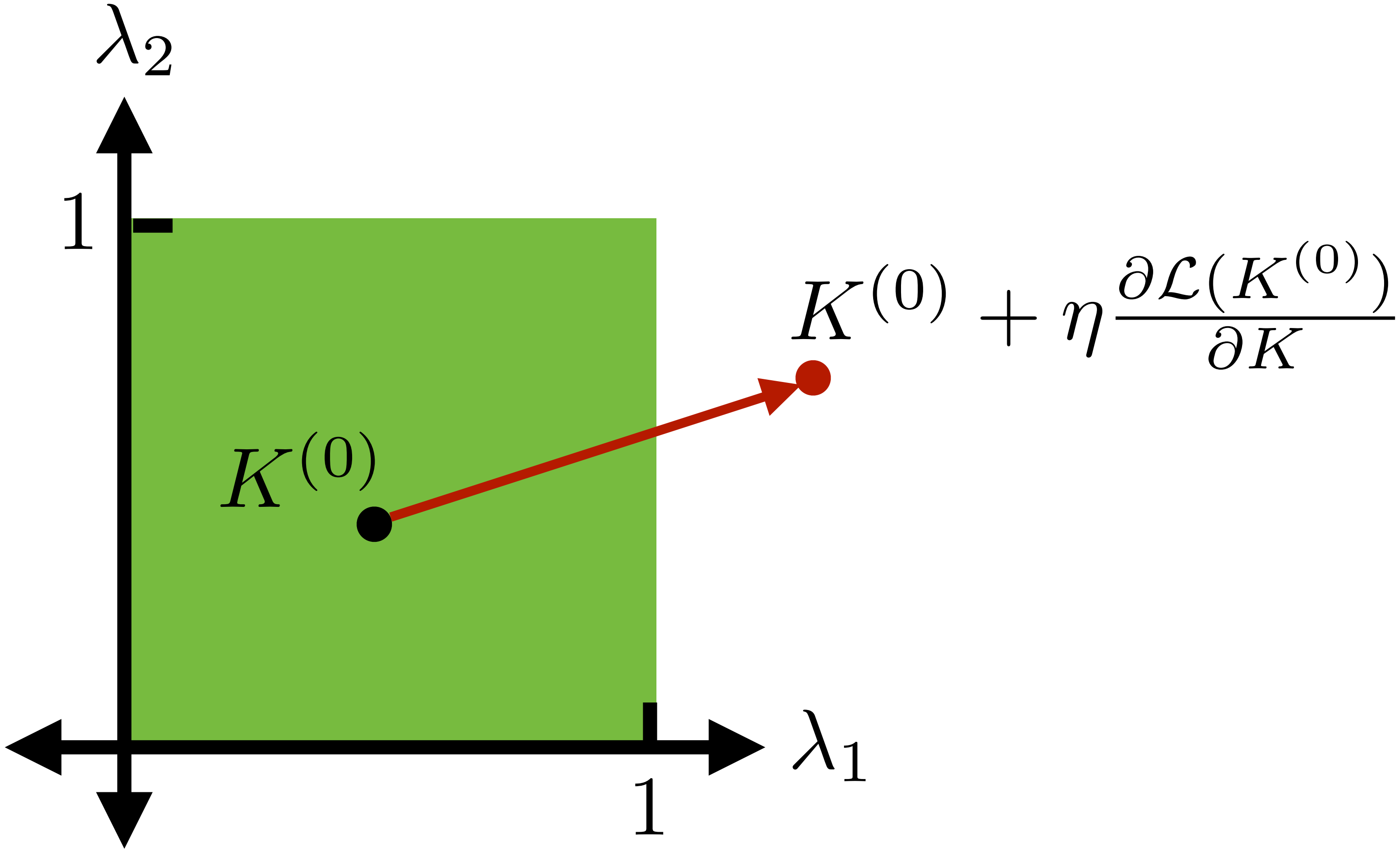
GRADIENT ASCENT

$$\frac{\partial \mathcal{L}(K)}{\partial K} = \sum_{t=1}^{T} (K - I_{\overline{Y}_t})^{-1}$$



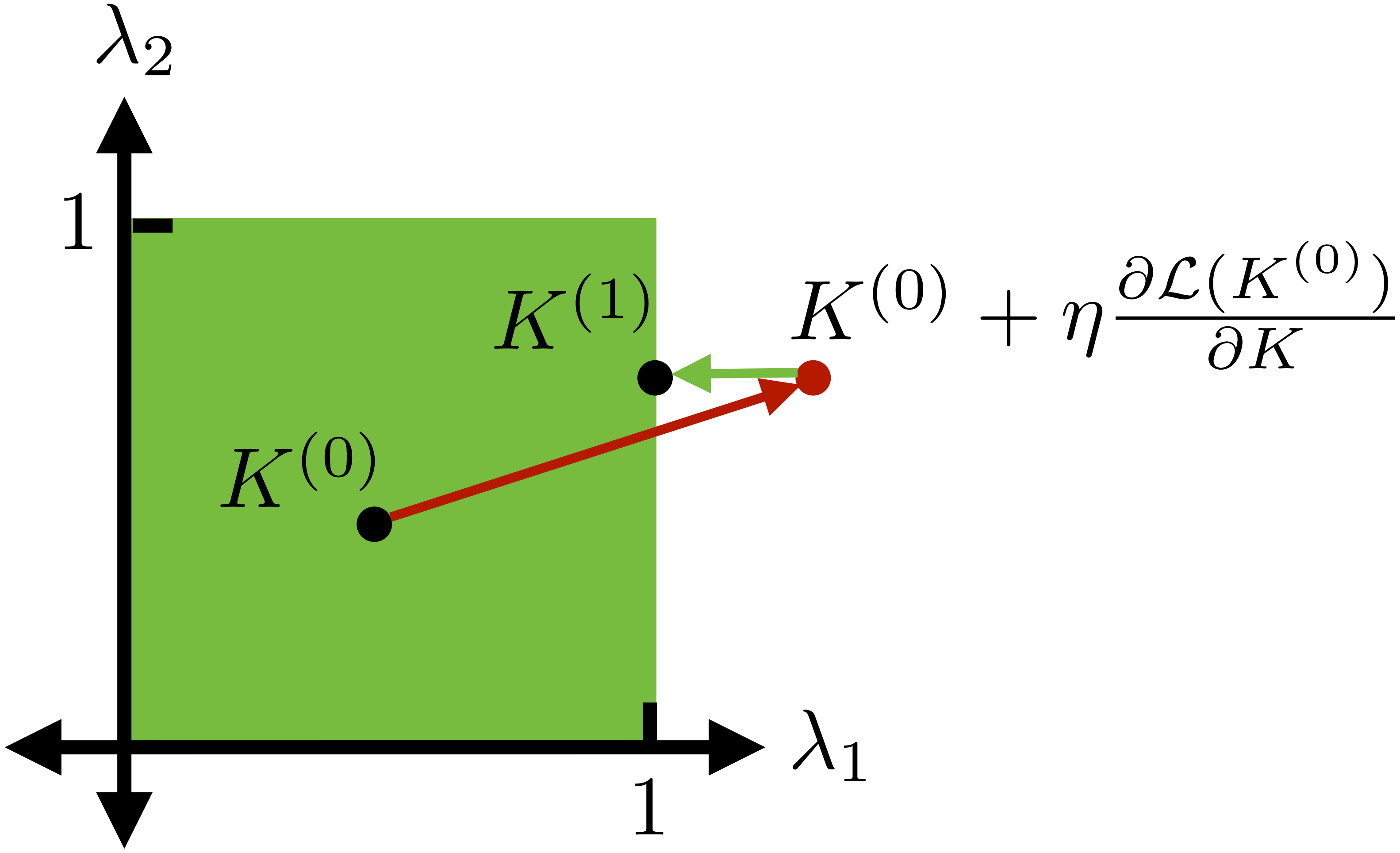
GRADIENT ASCENT

$$\frac{\partial \mathcal{L}(K)}{\partial K} = \sum_{t=1}^{T} (K - I_{\overline{Y}_t})^{-1}$$



GRADIENT ASCENT

$$\frac{\partial \mathcal{L}(K)}{\partial K} = \sum_{t=1}^{T} (K - I_{\overline{Y}_t})^{-1}$$



EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$K = V \Lambda V^\top$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$K = V\Lambda V^\top$$

$$K^J = V^J \mathrm{diag}(\mathbf{1})(V^J)^\top$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$K = V \Lambda V^\top$$

$$K^J = V^J \mathrm{diag}(\mathbf{1})(V^J)^\top$$

$$\mathcal{P}(Y) = \frac{\det(L_Y)}{\det(L+I)} = |\det(K - I_{\overline{Y}})|$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$K = V \Lambda V^\top$$

$$K^J = V^J \mathrm{diag}(\mathbf{1})(V^J)^\top$$

$$\mathcal{P}(Y) = \frac{\det(L_Y)}{\det(L+I)} = |\det(K - I_{\overline{Y}})|$$

$$= \sum_{J : J \subseteq \{1,\ldots,N\}} \mathcal{P}_{K^J}(Y) \prod_{j : j \in J} \lambda_j \prod_{j : j \notin J} (1 - \lambda_j)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$K = V \Lambda V^\top$$

$$K^J = V^J \mathrm{diag}(\mathbf{1})(V^J)^\top$$

$$\mathcal{P}(Y) = \frac{\det(L_Y)}{\det(L+I)} = |\det(K - I_{\overline{Y}})|$$

$$= \sum_{J: J \subseteq \{1,\dots,N\}} \mathcal{P}_{K^J}(Y) \prod_{j: j \in J} \lambda_j \prod_{j: j \notin J} (1 - \lambda_j)$$

$$= \sum_{J: J \subseteq \{1,\dots,N\}} \Pr_K(Y \mid J) \Pr_K(J)$$

Hidden variable J

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\mathcal{L}(K) = \mathcal{L}(V, \Lambda) = \sum_{t=1}^{T} \log p_K(Y_t) = \sum_{t=1}^{T} \log \left(\sum_{J} p_K(J, Y_t) \right)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\mathcal{L}(K) = \mathcal{L}(\overset{\substack{\text{eigenvectors}\\ \searrow}}{V}, \overset{\substack{\text{eigenvalues}\\ \swarrow}}{\Lambda}) = \sum_{t=1}^{T} \log p_K(Y_t) = \sum_{t=1}^{T} \log \left(\sum_{J} p_K(J, Y_t) \right)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\mathcal{L}(K) = \mathcal{L}(\overset{\substack{\text{eigenvectors}\\ \searrow}}{V}, \overset{\substack{\text{eigenvalues}\\ \swarrow}}{\Lambda}) = \sum_{t=1}^{T} \log p_K(Y_t) = \sum_{t=1}^{T} \log \left(\overset{\substack{\text{choice of eigenvectors}\\ \downarrow}}{\sum_{J}} p_K(J, Y_t) \right)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\mathcal{L}(K) = \mathcal{L}(\overset{\substack{\text{eigenvectors}\\ \searrow}}{V}, \overset{\substack{\text{eigenvalues}\\ \swarrow}}{\Lambda}) = \sum_{t=1}^{T} \log p_K(Y_t) = \sum_{t=1}^{T} \log \left(\overset{\substack{\text{choice of eigenvectors}\\ \downarrow}}{\sum_{J}} p_K(J, Y_t) \right)$$

$$= \sum_{t=1}^{T} \log \left(\sum_{J} q(J \mid Y_t) \frac{p_K(J, Y_t)}{q(J \mid Y_t)} \right)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$
\begin{aligned}
\mathcal{L}(K) = \mathcal{L}(\overset{\substack{\text{eigenvectors}\\ \searrow}}{V}, \overset{\substack{\text{eigenvalues}\\ \swarrow}}{\Lambda}) &= \sum_{t=1}^{T} \log p_K(Y_t) = \sum_{t=1}^{T} \log \left(\overset{\substack{\text{choice of eigenvectors}\\ \downarrow}}{\sum_{J}} p_K(J, Y_t) \right) \\
&= \sum_{t=1}^{T} \log \left(\sum_{J} q(J \mid Y_t) \frac{p_K(J, Y_t)}{q(J \mid Y_t)} \right) \\
&\geq \sum_{t=1}^{T} \sum_{J} q(J \mid Y_t) \log \left(\frac{p_K(J, Y_t)}{q(J \mid Y_t)} \right) \equiv F(q, V, \Lambda)
\end{aligned}
$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\begin{aligned}
\mathcal{L}(K) = \mathcal{L}(\overset{\substack{\text{eigenvectors}\\ \searrow}}{V}, \overset{\substack{\text{eigenvalues}\\ \swarrow}}{\Lambda}) &= \sum_{t=1}^{T} \log p_K(Y_t) = \sum_{t=1}^{T} \log \left(\overset{\substack{\text{choice of eigenvectors}\\ \downarrow}}{\sum_{J}} p_K(J, Y_t) \right) \\
&= \sum_{t=1}^{T} \log \left(\sum_{J} q(J \mid Y_t) \frac{p_K(J, Y_t)}{q(J \mid Y_t)} \right) \\
&\geq \sum_{t=1}^{T} \sum_{J} q(J \mid Y_t) \log \left(\frac{p_K(J, Y_t)}{q(J \mid Y_t)} \right) \equiv F(q, V, \Lambda)
\end{aligned}$$

$$\mathcal{P}(Y) = \sum_{J : J \subseteq \{1, \dots, N\}} \mathcal{P}_{K^J}(Y) \prod_{j : j \in J} \lambda_j \prod_{j : j \notin J} (1 - \lambda_j)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$
\begin{aligned}
\mathcal{L}(K) = \mathcal{L}(\overset{\substack{\text{eigenvectors}\\ \searrow}}{V}, \overset{\substack{\text{eigenvalues}\\ \swarrow}}{\Lambda}) &= \sum_{t=1}^{T} \log p_K(Y_t) = \sum_{t=1}^{T} \log \left(\overset{\substack{\text{choice of eigenvectors}\\ \downarrow}}{\sum_{J}} p_K(J, Y_t) \right) \\
&= \sum_{t=1}^{T} \log \left(\sum_{J} q(J \mid Y_t) \frac{p_K(J, Y_t)}{q(J \mid Y_t)} \right) \\
&\geq \sum_{t=1}^{T} \sum_{J} q(J \mid Y_t) \log \left(\frac{p_K(J, Y_t)}{q(J \mid Y_t)} \right) \equiv F(q, V, \Lambda)
\end{aligned}
$$

$$
\mathcal{P}(Y) = \sum_{J : J \subseteq \{1, \ldots, N\}} \mathcal{P}_{K^J}(Y) \prod_{j : j \in J} \lambda_j \prod_{j : j \notin J} (1 - \lambda_j)
$$

$$
\log(p_K(J, Y)) = \log \mathcal{P}_{K^J}(Y) + \sum_{j : j \in J} \log \lambda_j + \sum_{j : j \notin J} \log(1 - \lambda_j)
$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$F(q, V, \Lambda) = \sum_{t=1}^{T} -\mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t)) + \mathcal{L}(V, \Lambda)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$F(q, V, \Lambda) = \sum_{t=1}^{T} -\mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t)) + \mathcal{L}(V, \Lambda)$$

$$\text{E-step:} \quad \min_q \sum_{t=1}^{T} \mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t))$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$F(q, V, \Lambda) = \sum_{t=1}^{T} -\mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t)) + \mathcal{L}(V, \Lambda)$$

$$= \sum_{t=1}^{T} \mathbb{E}_q[\log p_K(J, Y_t)] + H(q)$$

$$\text{E-step:} \quad \min_q \sum_{t=1}^{T} \mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t))$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$F(q, V, \Lambda) = \sum_{t=1}^{T} -\mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t)) + \mathcal{L}(V, \Lambda)$$

$$= \sum_{t=1}^{T} \mathbb{E}_q[\log p_K(J, Y_t)] + H(q)$$

$$\text{E-step:} \quad \min_q \sum_{t=1}^{T} \mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t))$$

$$\text{M-step:} \quad \max_{V,\Lambda} \sum_{t=1}^{T} \mathbb{E}_q[\log p_K(J, Y_t)] \;\; \text{s.t.} \;\; \mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{1}, V^\top V = I$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\text{E-step:} \quad \min_{q} \sum_{t=1}^{T} \mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t))$$

$$q(J \mid Y_t) = p_K(J \mid Y_t)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\text{E-step:} \quad \min_{q} \sum_{t=1}^{T} \mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t))$$

$$q(J \mid Y_t) = p_K(J \mid Y_t)$$

$$P(Y_t \mid J) \text{ is a DPP} \implies P(J \mid Y_t) \text{ is a DPP}$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\text{E-step:} \quad \min_q \sum_{t=1}^{T} \mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t))$$

$$q(J \mid Y_t) = p_K(J \mid Y_t)$$

$$P(Y_t \mid J) \text{ is a DPP} \implies P(J \mid Y_t) \text{ is a DPP}$$

$$\text{M-step:} \;\; \lambda_j = \frac{1}{T} \sum_{t=1}^{T} \sum_{J : j \in J} q(J \mid Y_t)$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\text{E-step:}\quad \min_q \sum_{t=1}^{T} \mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t))$$

$$q(J \mid Y_t) = p_K(J \mid Y_t)$$

$$P(Y_t \mid J) \text{ is a DPP} \implies P(J \mid Y_t) \text{ is a DPP}$$

$$\text{M-step:}\ \ \lambda_j = \frac{1}{T} \sum_{t=1}^{T} \sum_{J:j \in J} q(J \mid Y_t)$$

$$O(TNk^2), \text{ for } k = \max_t |Y_t|$$

EXPECTATION-MAXIMIZATION

GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

$$\text{E-step:} \quad \min_q \sum_{t=1}^{T} \mathbf{KL}(q(J \mid Y_t) \parallel p_K(J \mid Y_t))$$

$$q(J \mid Y_t) = p_K(J \mid Y_t)$$

$$P(Y_t \mid J) \text{ is a DPP} \implies P(J \mid Y_t) \text{ is a DPP}$$

$$\text{M-step:} \;\; \lambda_j = \frac{1}{T} \sum_{t=1}^{T} \sum_{J : j \in J} q(J \mid Y_t)$$

$O(TNk^2)$, for $k = \max_t |Y_t|$

V update is more complicated, but still efficient

PRODUCT RECOMMENDATION

furniture

carseats

toys

PRODUCT RECOMMENDATION

$\sim$ 30,000 Amazon baby registries

furniture

carseats

toys

PRODUCT RECOMMENDATION

$\sim$ 30,000 Amazon baby registries

13 categories

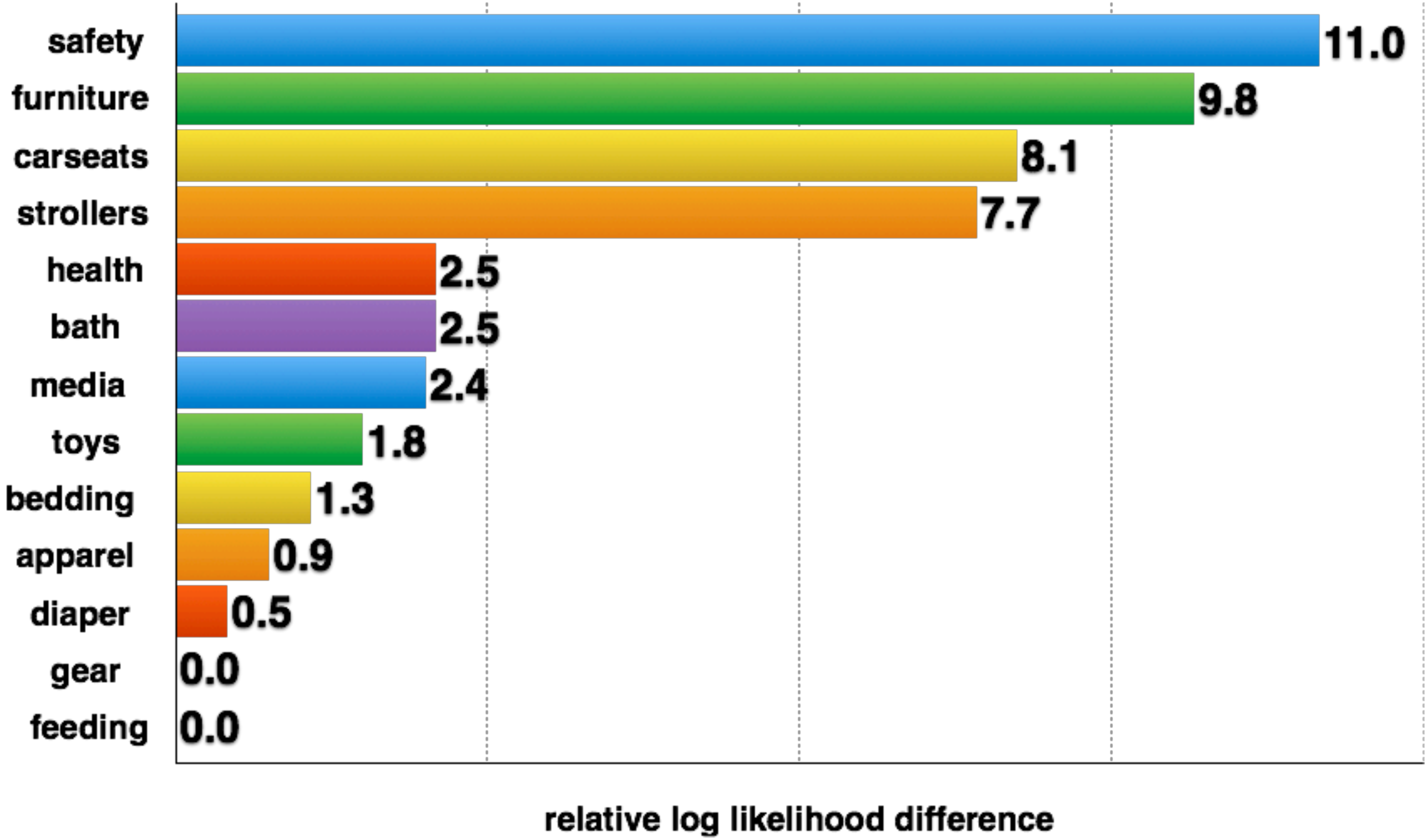
furniture

carseats

toys

EM VS PROJECTED GRADIENT

EM VS PROJECTED GRADIENT



“SAFETY” SELECTION

“SAFETY” SELECTION

Graco Sweet Slumber
Sound Machine

Boppy Noggin Nest
Head Support

Cloud b Twilight
Night Light

Braun ThermoScan
Lens Filters

Aquatopia Bath
Thermometer Alarm

Britax EZ-Cling
Sun Shades

TL Care Organic
Cotton Mittens

Regalo Easy Step
Walk Thru Gate

VTech Comm.
Audio Monitor

Infant Optics
Video Monitor

"SAFETY" SELECTION

Graco Sweet Slumber
Sound Machine

Motorola
Video Monitor

Cloud b Twilight
Night Light

Summer Infant
Video Monitor

Aquatopia Bath
Thermometer Alarm

Britax EZ-Cling
Sun Shades

TL Care Organic
Cotton Mittens

Regalo Easy Step
Walk Thru Gate

VTech Comm.
Audio Monitor

Infant Optics
Video Monitor

PROPOSED WORK

PROPOSED WORK

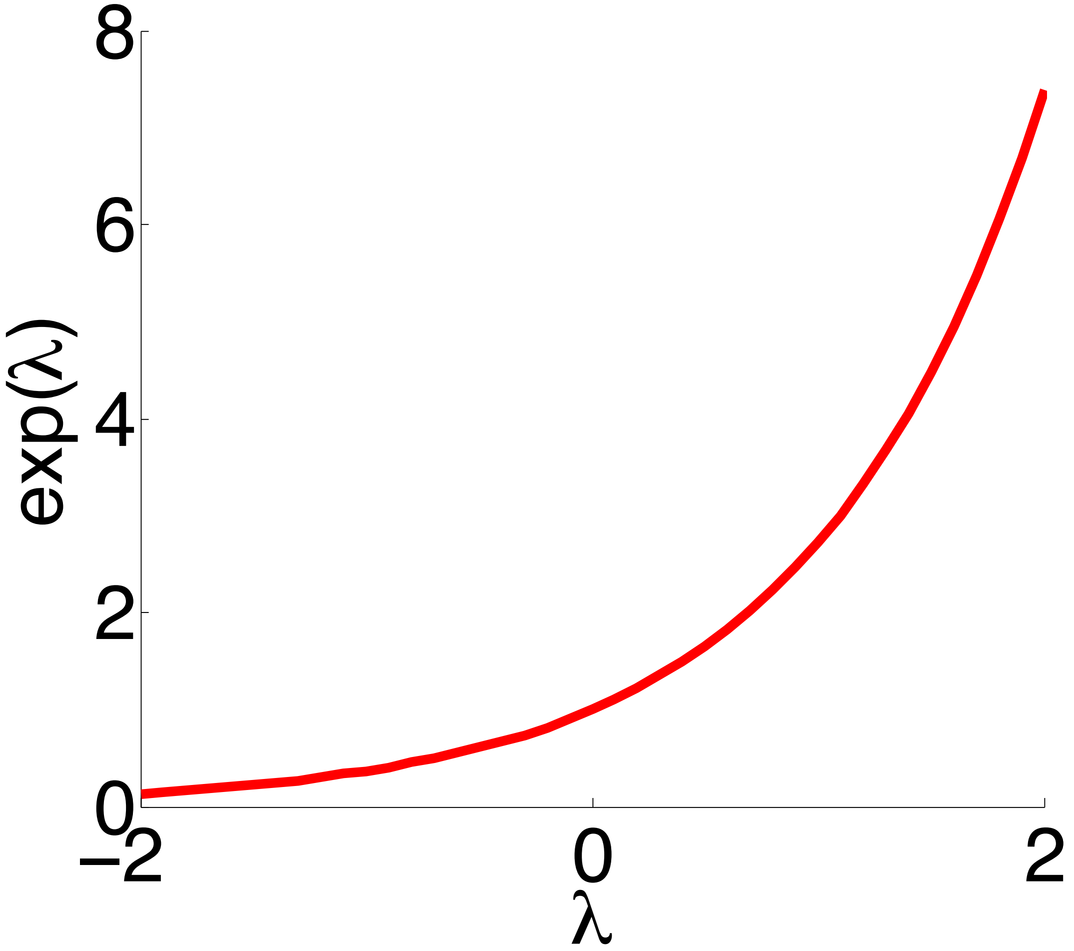
Applying matrix *exponentiated* gradient, proposed by Tsuda et al. (JMLR 2005)

$$K^{(t+1)} \leftarrow \exp\left(\log(K^{(t)}) + \eta \nabla \mathcal{L}(K^{(t)})\right)$$

PROPOSED WORK

Applying matrix *exponentiated* gradient, proposed by Tsuda et al. (JMLR 2005)

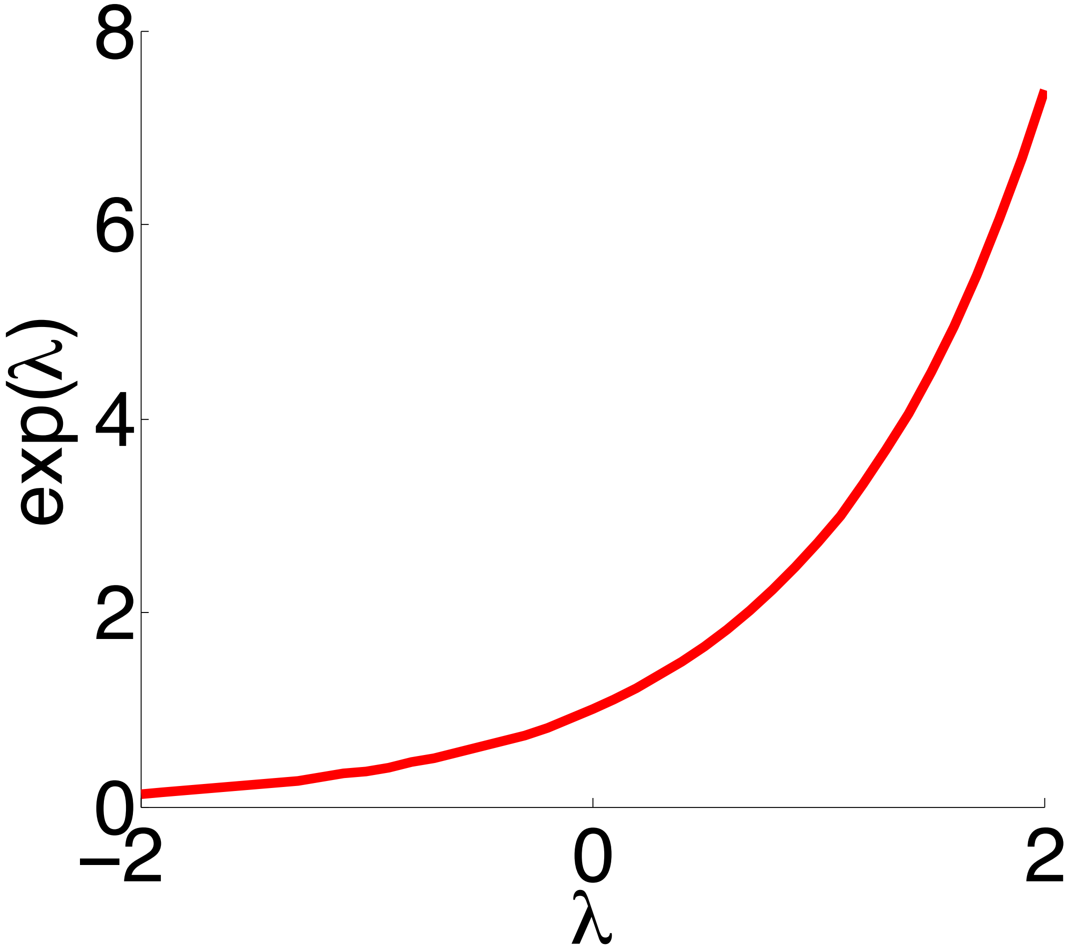
$$K^{(t+1)} \leftarrow \exp\left(\log(K^{(t)}) + \eta \nabla \mathcal{L}(K^{(t)})\right)$$



PROPOSED WORK

Applying matrix *exponentiated* gradient, proposed by Tsuda et al. (JMLR 2005)

$$K^{(t+1)} \leftarrow \exp\left(\log(K^{(t)}) + \eta\nabla\mathcal{L}(K^{(t)})\right)$$



Relative log-likelihood

$\frac{|EM-EG|}{|EG|}$: 0.57%

APPROXIMATE INFERENCE
FOR
DETERMINANTAL POINT PROCESSES

APPROXIMATE INFERENCE
FOR
DETERMINANTAL POINT PROCESSES

APPROXIMATE INFERENCE
FOR
DETERMINANTAL POINT PROCESSES

1. Dimensionality Reduction

GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

APPROXIMATE INFERENCE
FOR
DETERMINANTAL POINT PROCESSES

1. **Dimensionality Reduction**

   GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)

2. **MAP Estimation**

   GILLENWATER, KULESZA, AND TASKAR (NIPS 2012)

APPROXIMATE INFERENCE
FOR
DETERMINANTAL POINT PROCESSES

1. **Dimensionality Reduction**

   GILLENWATER, KULESZA, AND TASKAR (EMNLP 2012)
2. **MAP Estimation**

   GILLENWATER, KULESZA, AND TASKAR (NIPS 2012)
3. **Likelihood Maximization**

   GILLENWATER, KULESZA, FOX, AND TASKAR (NIPS 2014)

APPROXIMATE INFERENCE
FOR
DETERMINANTAL POINT PROCESSES

Thanks to the committee:

Michael Kearns
Sasha Rakhlin
Ali Jadbabaie
Jeff Bilmes