Geometric Inference in Bayesian Hierarchical Models with Applications to Topic Modeling

Mikhail Yurochkin

University of Michigan

April 16, 2018

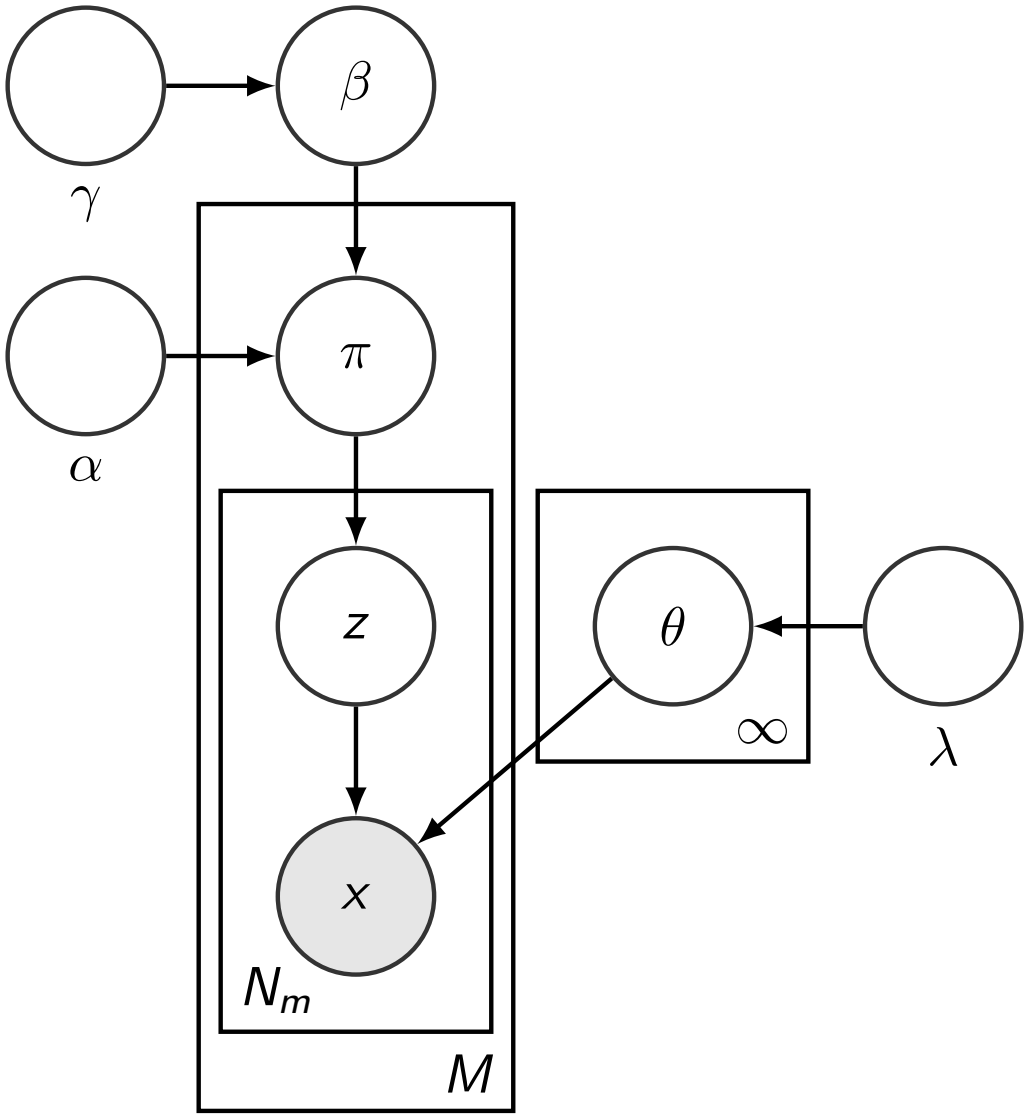
Research summary

- Unsupervised learning
  - ⋆ Geometric inference and modeling for topic models (Yurochkin & Nguyen, 2016; Yurochkin et al., 2017a, 2018a)
  - ⋆ Clustering grouped data using Optimal Transport (Ho et al., 2017)
- Supervised learning
  - ⋆ Regression with interaction selection of unbounded order (Yurochkin et al., 2017b)
  - ⋆ Neural networks based on convolution on graphs. Joint learning of the neural network parameters and graph adjacency matrix (Yurochkin et al., 2018b)

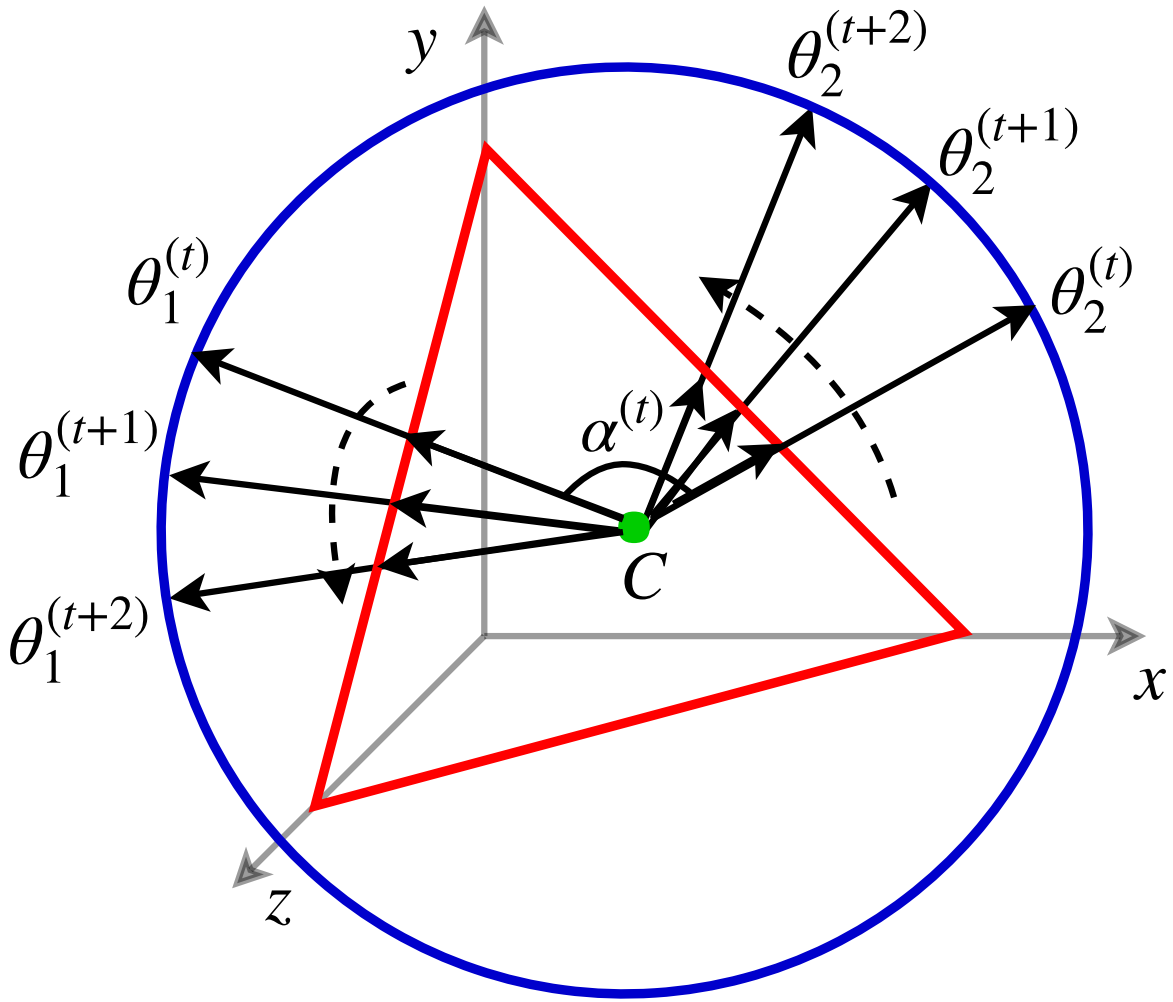
Unsupervised learning

Bayesian approach

Graphical models



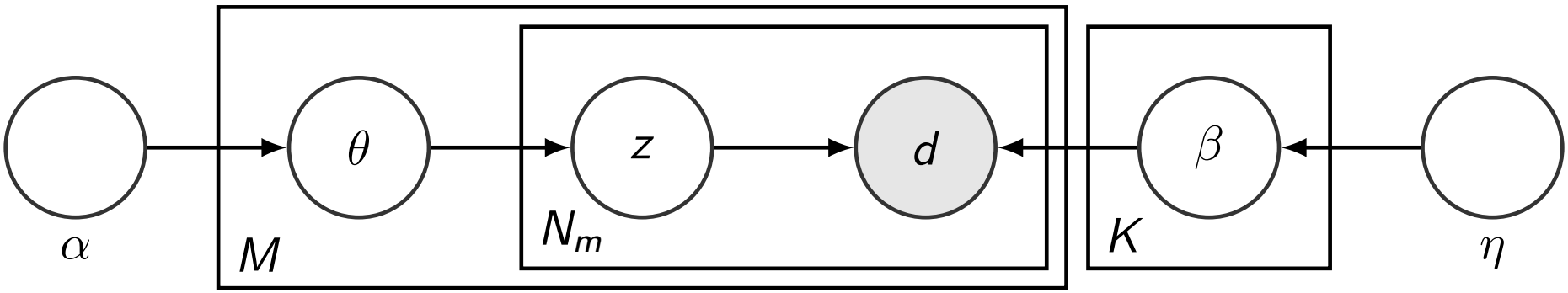
Geometry of Bayesian



Outline

Outline

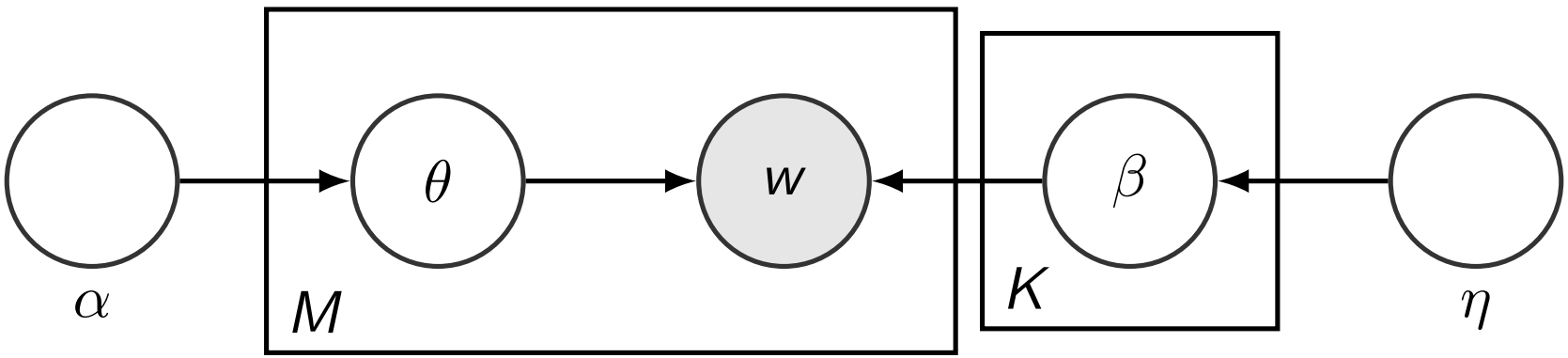
LDA: data generation



Latent Dirichlet Allocation (Blei et al., 2003) generative process:

- For $k = 1, \ldots, K$
  - ⋆ sample a topic $\beta_k \sim \text{Dir}_V(\eta)$ over V unique words
- For each document $m = 1, \ldots, M$
  - ⋆ sample topic proportions $\theta_m \sim \text{Dir}_K(\alpha)$
  - ⋆ for each word position $n_m = 1, \ldots, N_M$
    - ▸ pick a topic label $z_{n_m}|\theta_m \sim \text{Categorical}(\theta_m)$
    - ▸ sample a word $d_{n_m}|z_{n_m}, \beta_1, \ldots \beta_K \sim \text{Categorical}(\beta_{z_{n_m}})$

LDA without z

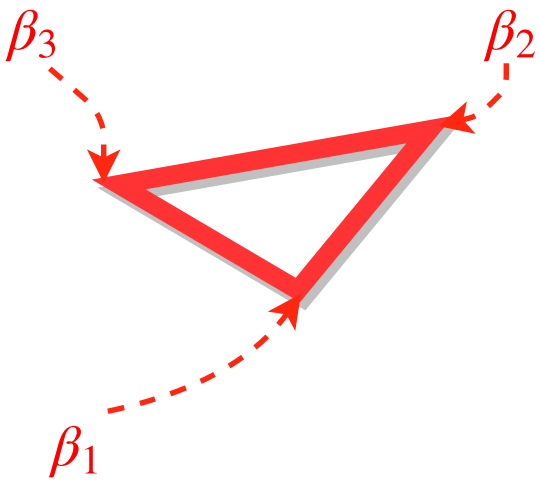


Integrating z out for each of the M documents:

- For each document $m = 1, \ldots, M$
  - ⋆ sample topic proportions $\theta_m \sim \text{Dir}_K(\alpha)$
  - ⋆ compute word probabilities for a document $p_m = \sum_{k=1}^{K} \theta_{mk}\beta_k \in \Delta^{V-1}$
  - ⋆ generate a document $w_m|p_m \sim \text{Multinomial}(p_m, N_m)$

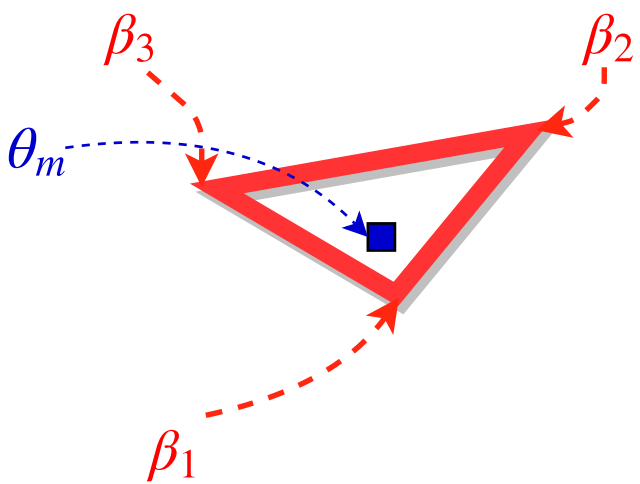
Outline

Geometric perspective of LDA



Topic poytope: $B = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$

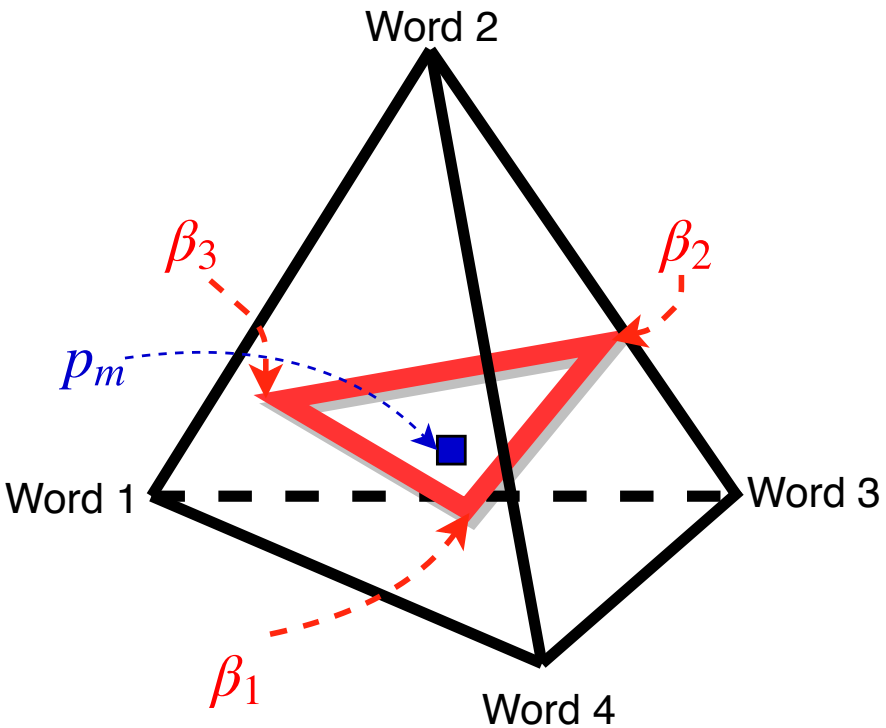
Geometric perspective of LDA



Topic poytope: $B = \text{Conv}(\beta_1, \ldots, \beta_K)$

Barycentric coordinates: $\theta_m \sim \text{Dir}_K(\alpha)$

Geometric perspective of LDA

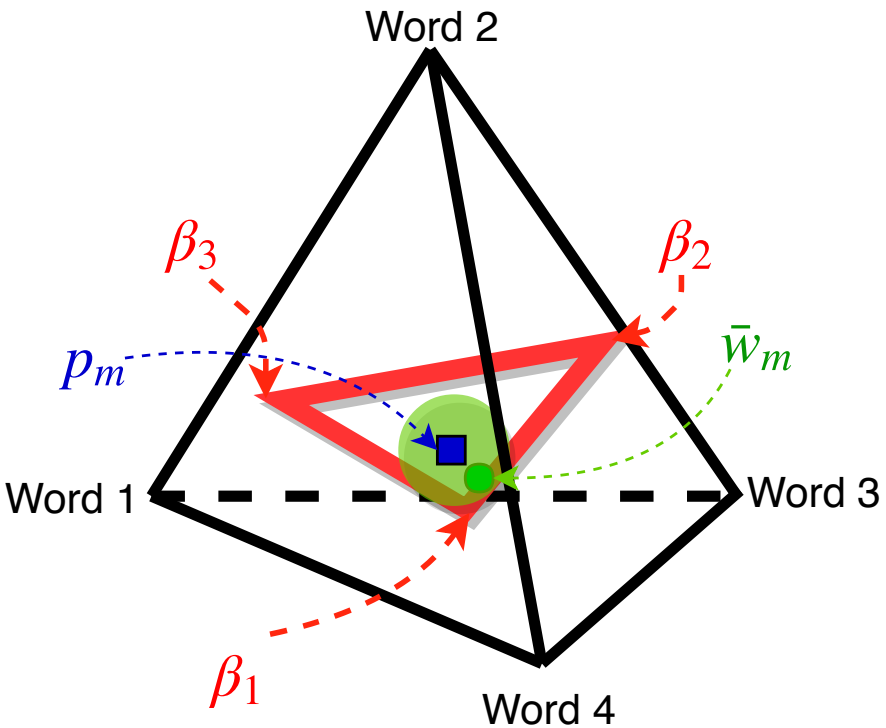


Topic poytope: $B = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$

Barycentric coordinates: $\theta_m \sim \mathrm{Dir}_K(\alpha)$

Cartesian coordinates: $p_m = \sum_{k=1}^{K} \theta_{mk} \beta_k$

Geometric perspective of LDA



Topic poytope: $B = \text{Conv}(\beta_1, \ldots, \beta_K)$

Barycentric coordinates: $\theta_m \sim \text{Dir}_K(\alpha)$

Cartesian coordinates: $p_m = \sum_{k=1}^{K} \theta_{mk} \beta_k$

Document: $w_m \sim \text{Multinomial}(p_m, N_m)$

Normalized document: $\bar{w}_m = w_m / N_m$

Geometric perspective of LDA

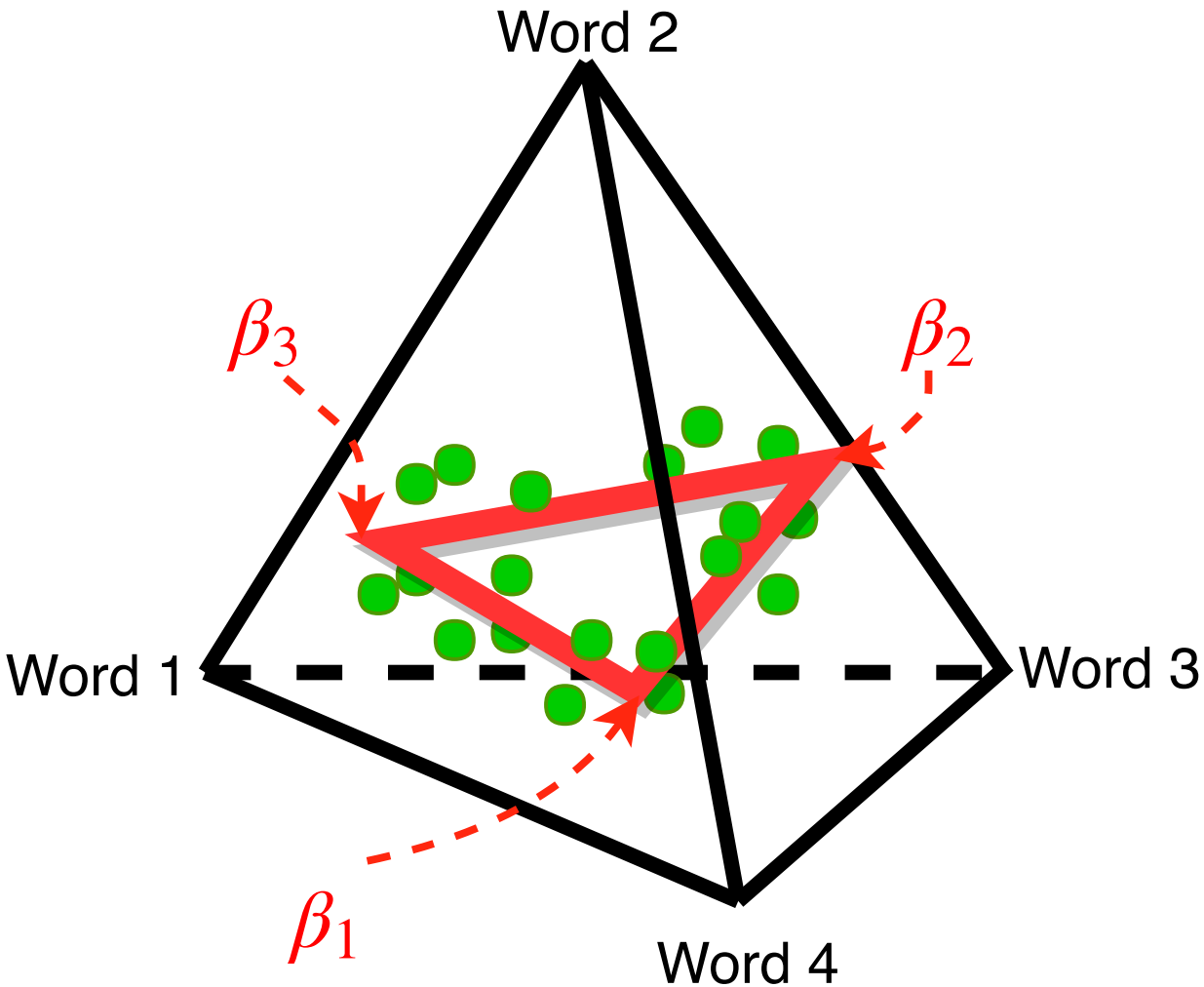


Figure: Observed normalized documents

Geometric Surrogate Loss to the Likelihood

Multinomial maximum log-likelihood: $\sum_{m=1}^{M} \sum_{i=1}^{V} w_{mi} \log \bar{w}_{mi}$

LDA log-likelihood: $\sum_{m=1}^{M} \sum_{i=1}^{V} w_{mi} \log \left(\sum_{k=1}^{K} \theta_{mk} \beta_{ki} \right)$

Goal: find $B = \text{Conv}(\beta_1, \ldots, \beta_K)$ containing all $\bar{w}_1, \ldots, \bar{w}_M$

Geometric objective: $\sum_{m=1}^{M} N_m \min_{x \in B} \|x - \bar{w}_m\|_2^2$

Geometric Surrogate Loss to the Likelihood

Multinomial maximum log-likelihood: $\sum_{m=1}^{M}\sum_{i=1}^{V} w_{mi} \log \bar{w}_{mi}$

LDA log-likelihood: $\sum_{m=1}^{M}\sum_{i=1}^{V} w_{mi} \log \left(\sum_{k=1}^{K} \theta_{mk}\beta_{ki}\right)$

Goal: find $B = \text{Conv}(\beta_1, \ldots, \beta_K)$ containing all $\bar{w}_1, \ldots, \bar{w}_M$

Geometric objective: $\sum_{m=1}^{M} N_m \min_{x \in B} \|x - \bar{w}_m\|_2^2$

Topic proportions: given B, θ_{M+1} can be obtained as the barycentric coordinates of the projection of $\bar{w}_{M+1}$ onto B for a new document.

Step 1: weighted k-means

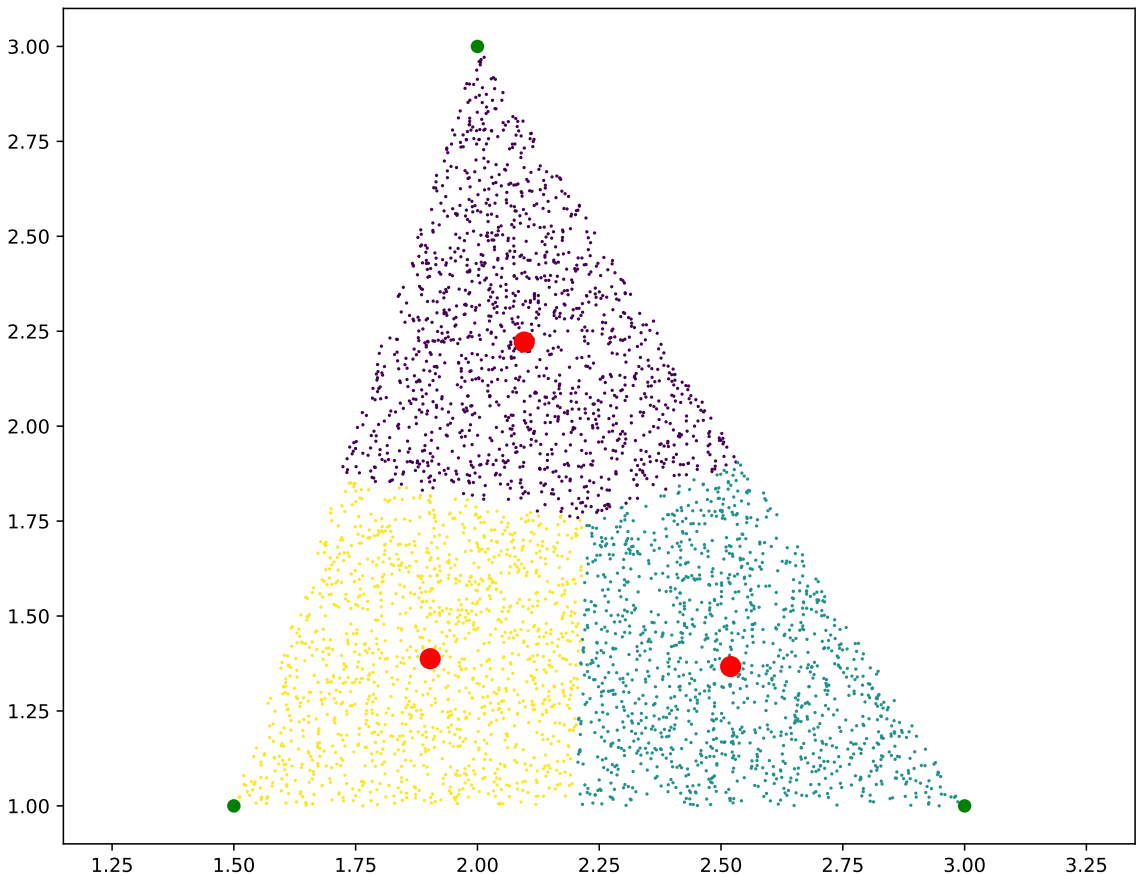
Proposition (Y. and Nguyen, 2016)

Subspace spanned by optimal weighted k-means centroids is equal to the subspace of weighted low rank matrix approximation under mild relaxations.

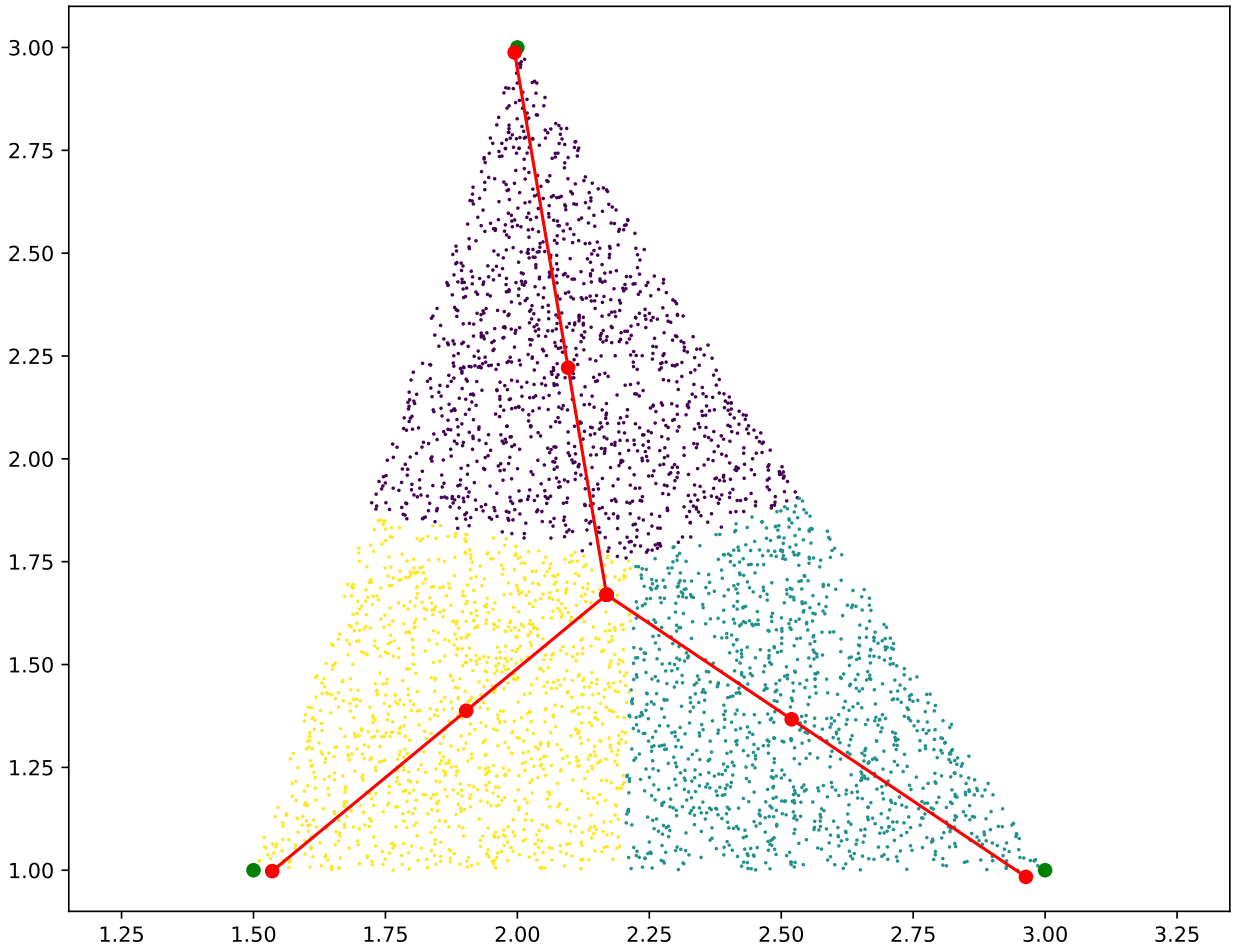
Step 1: weighted k-means

Proposition (Y. and Nguyen, 2016)

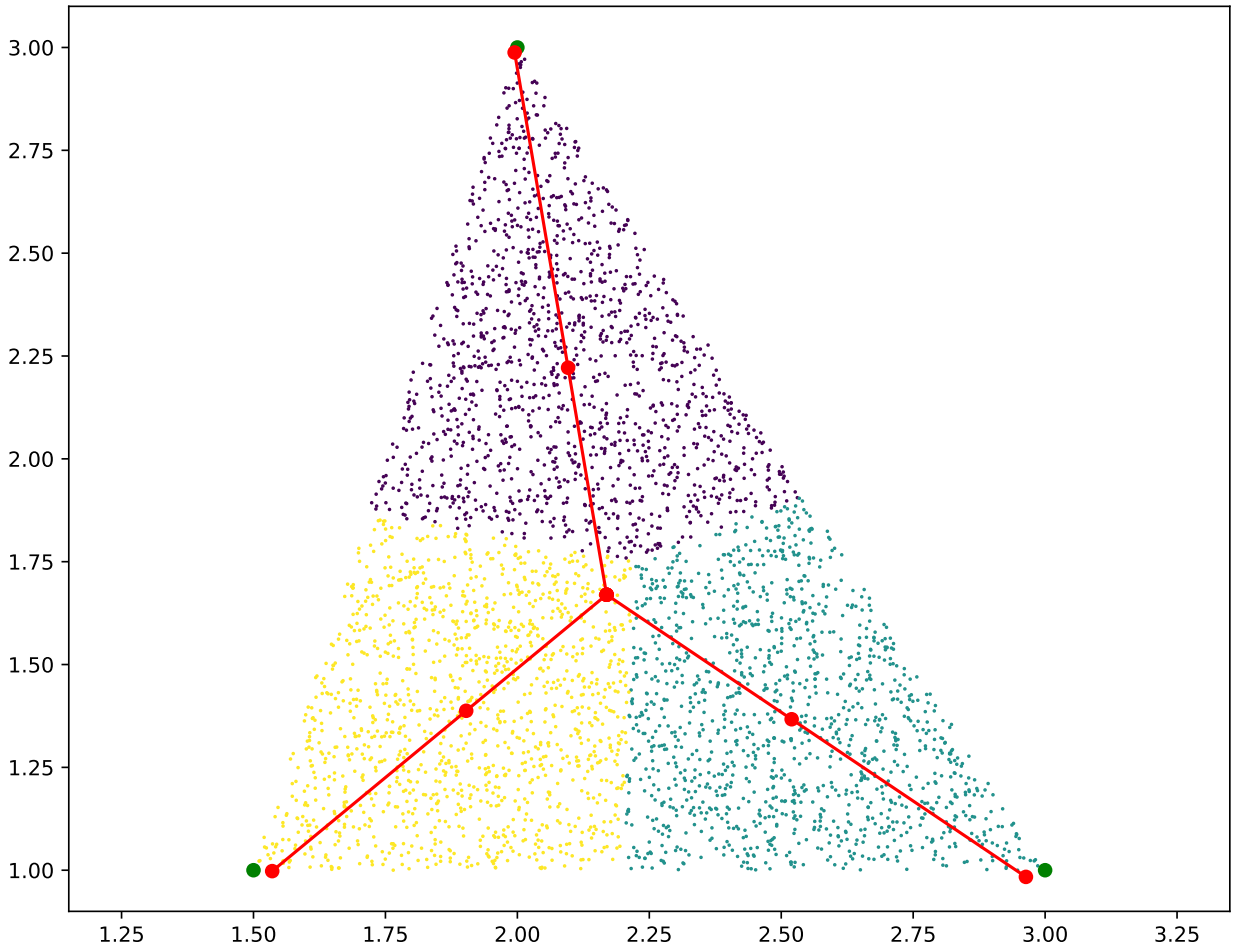
Subspace spanned by optimal weighted k-means centroids is equal to the subspace of weighted low rank matrix approximation under mild relaxations.



Step 2: geometric correction



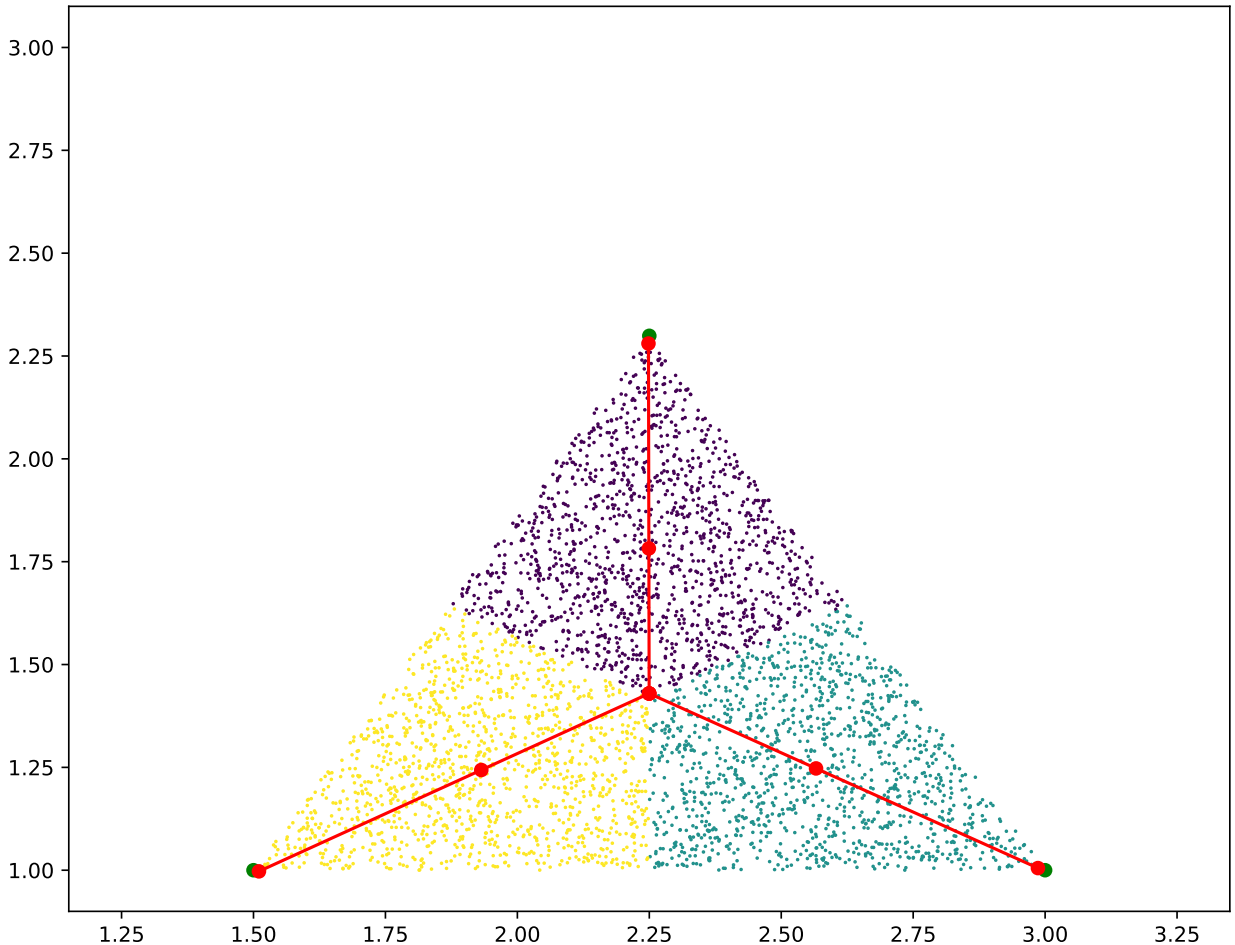
Step 2: geometric correction



Theorem (Y. and Nguyen, 2016)

If given samples from the true topic polytope, GDM is consistent if either true topic polytope is equilateral or $\alpha \to 0$.

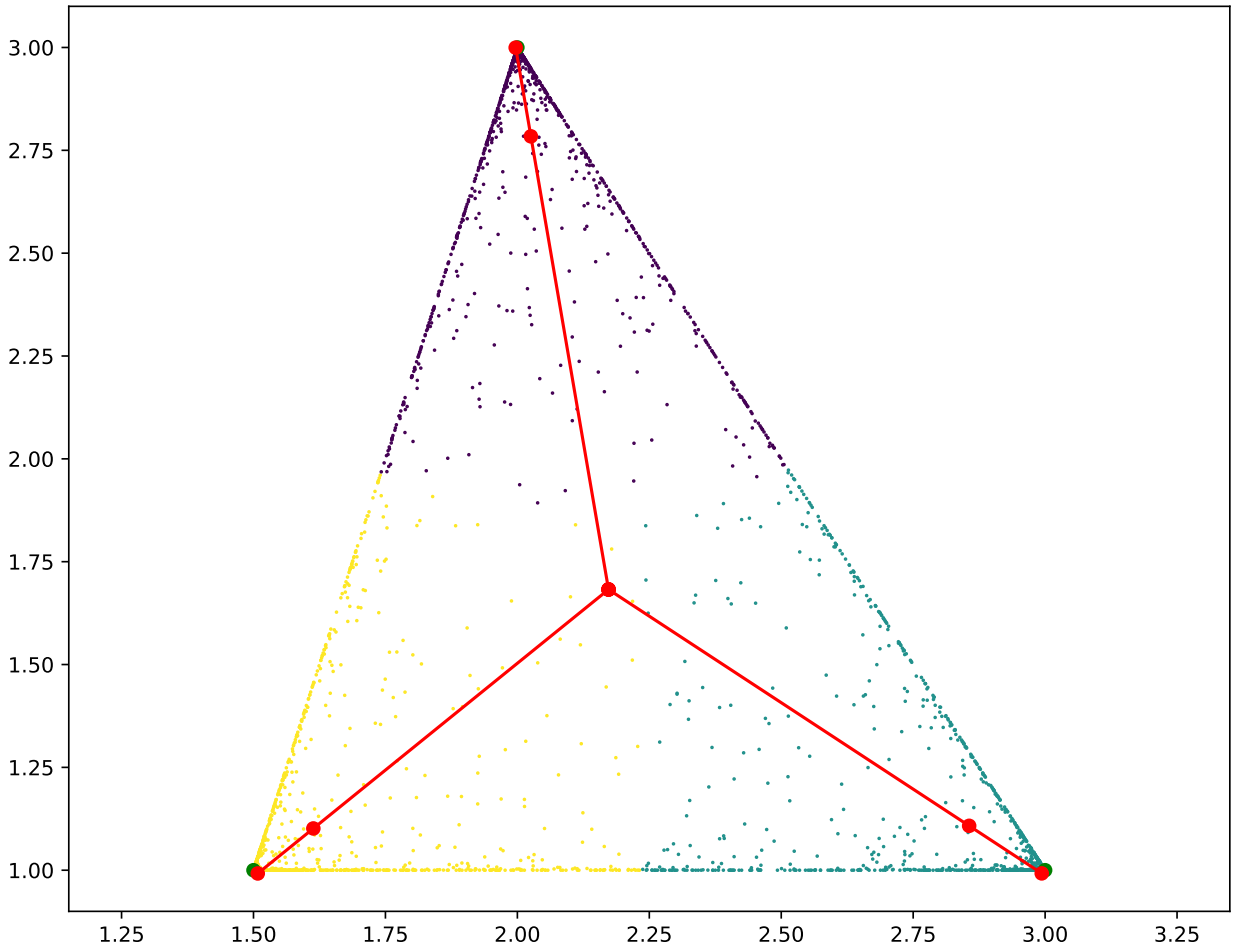
Step 2: geometric correction



Theorem (Y. and Nguyen, 2016)

If given samples from the true topic polytope, GDM is consistent if either true topic polytope is equilateral or $\alpha \to 0$.

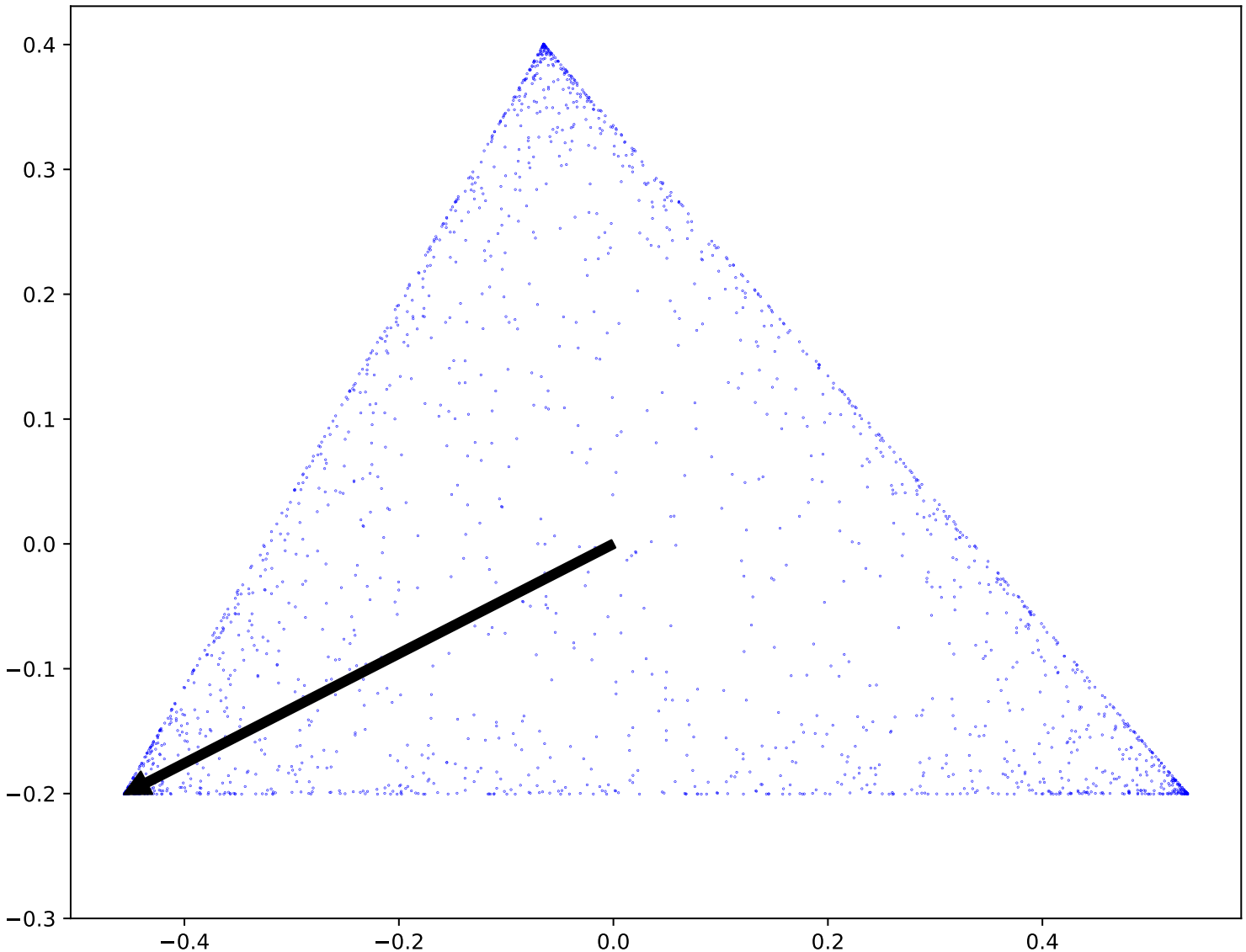
Step 2: geometric correction



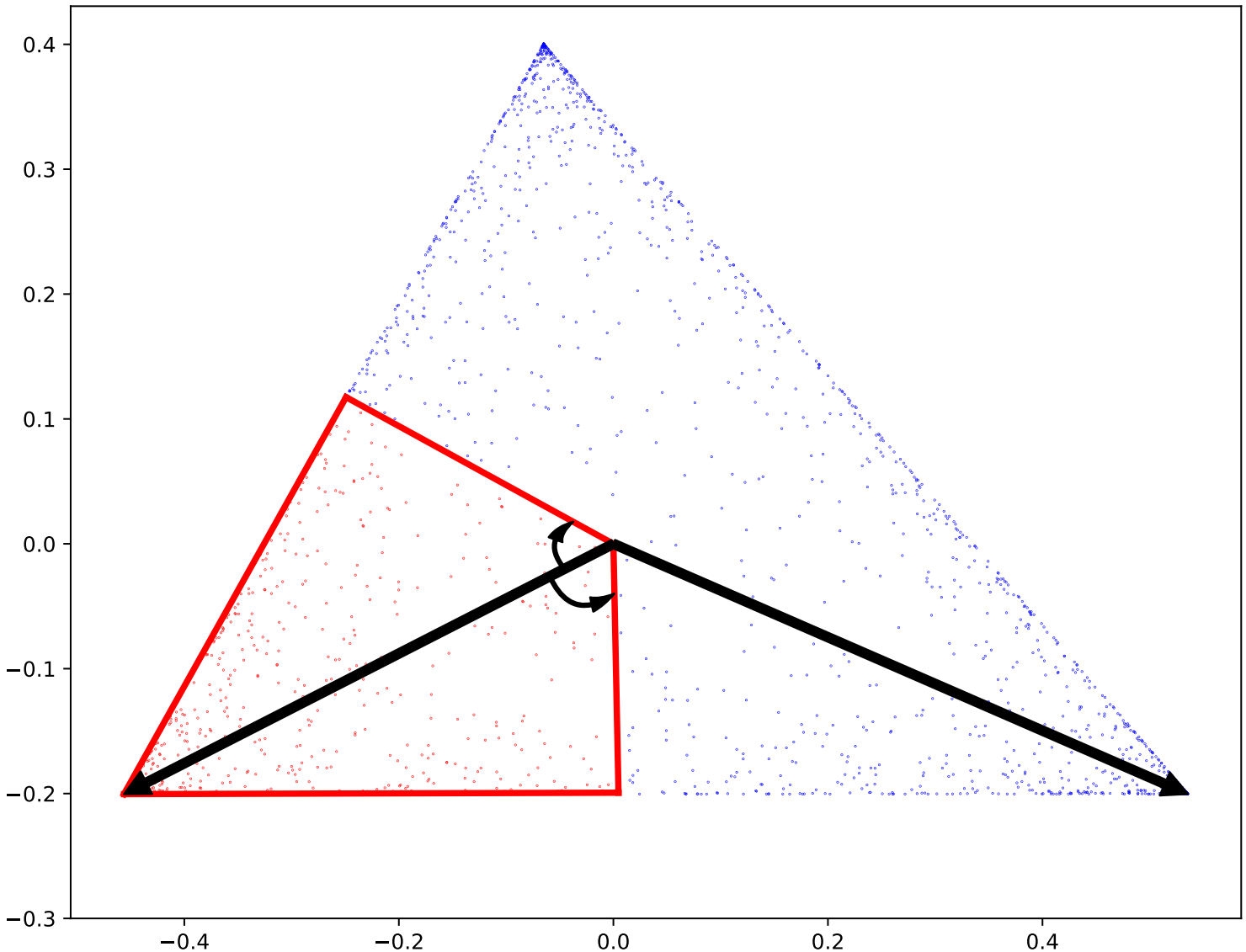
Theorem (Y. and Nguyen, 2016)

If given samples from the true topic polytope, GDM is consistent if either true topic polytope is equilateral or $\alpha \to 0$.

A Cone of light



A Cone of light



Conic Scan-and-Cover (CoSAC): Cones

$\hat{C}_p = \frac{1}{M}\sum_m p_m$
$\tilde{p}_m := p_m - \hat{C}_p,\ m = 1, \ldots, M$
$A_1 = \{1, \ldots, M\};\ k = 1$
while $A_k \neq \emptyset$ **do**
  $v_k = \underset{\tilde{p}_m : m \in A_k}{\operatorname{argmax}} \|\tilde{p}_m\|_2$
  $\mathcal{S}_\omega(v_k) = \{m : d_{\cos}(\tilde{p}_m, v_k) < \omega\}$
  $A_k = A_k \setminus \mathcal{S}_\omega(v_k)$
  $k = k + 1$
end while

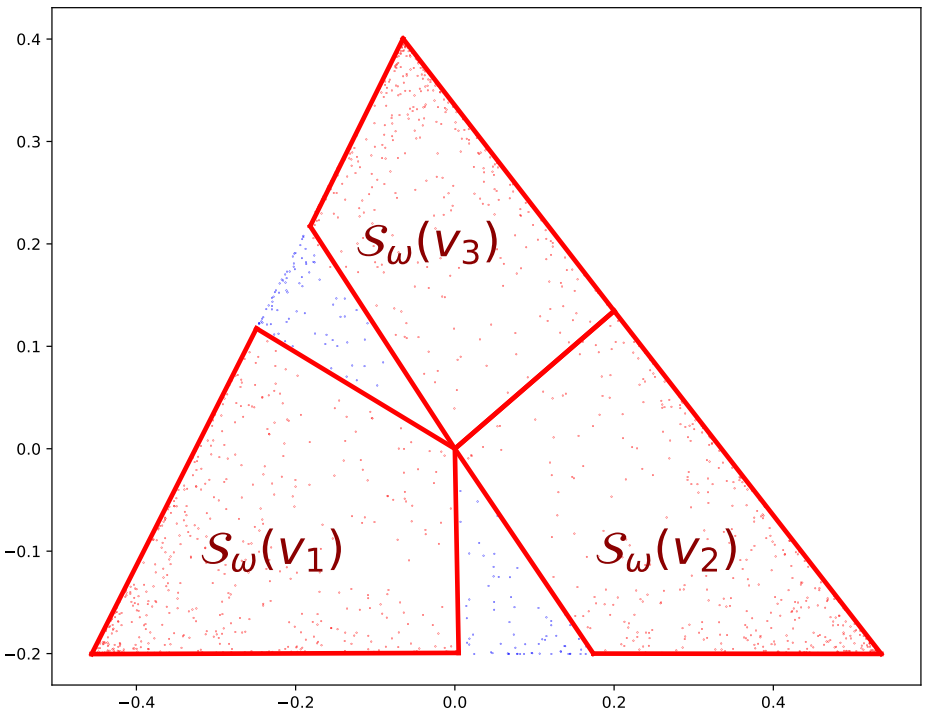


Figure: Incomplete coverage using cones

Coverage with Cones

Proposition (Y., Guha and Nguyen, 2017)

For choice of ω in some range (ω_1, ω_2), a complete coverage of the topic polytope is achievable with cones only, such that each cone contains exactly one topic vertex.

Coverage with Cones

Proposition (Y., Guha and Nguyen, 2017)

For choice of ω in some range (ω_1, ω_2), a complete coverage of the topic polytope is achievable with cones only, such that each cone contains exactly one topic vertex.

- $\omega_1 = 1 - r/R_{max}$
  - ⋆ r — inradius
  - ⋆ R_{max} — maximum distance from incenter to a topic

Coverage with Cones

Proposition (Y., Guha and Nguyen, 2017)

For choice of ω in some range (ω_1, ω_2), a complete coverage of the topic polytope is achievable with cones only, such that each cone contains exactly one topic vertex.

- $\omega_1 = 1 - r/R_{max}$
  - ⋆ r — inradius
  - ⋆ R_{max} — maximum distance from incenter to a topic
- $\omega_2 = \max\{(a_{min}^2)/(2R_{max}^2), \max_{i,k=1,\dots,K}(1 - \cos(b_i, b_k)\}$
  - ⋆ a_{min} — minimum distance between topics
  - ⋆ $\cos(b_i, b_k)$ — cosine of an angle between topics i and k

Coverage with Cones

Proposition (Y., Guha and Nguyen, 2017)

For choice of ω in some range (ω_1, ω_2), a complete coverage of the topic polytope is achievable with cones only, such that each cone contains exactly one topic vertex.

- $\omega_1 = 1 - r/R_{max}$
  - ⋆ r — inradius
  - ⋆ R_{max} — maximum distance from incenter to a topic
- $\omega_2 = \max\{(a_{min}^2)/(2R_{max}^2), \max_{i,k=1,\dots,K}(1 - \cos(b_i, b_k)\}$
  - ⋆ a_{min} — minimum distance between topics
  - ⋆ $\cos(b_i, b_k)$ — cosine of an angle between topics i and k
- Angular separation — $\cos(b_i, b_k) \leq 0$ for any $i, k = 1, \dots, K$
  - ⋆ $\omega \in \left(1 - \frac{r}{R_{max}}, 1\right) \neq \emptyset$

Conic Scan-and-Cover (CoSAC): Cones and Sphere

Stop when $\|\tilde{p}_m\|_2 < \mathcal{R}\ \forall m \in A_k$

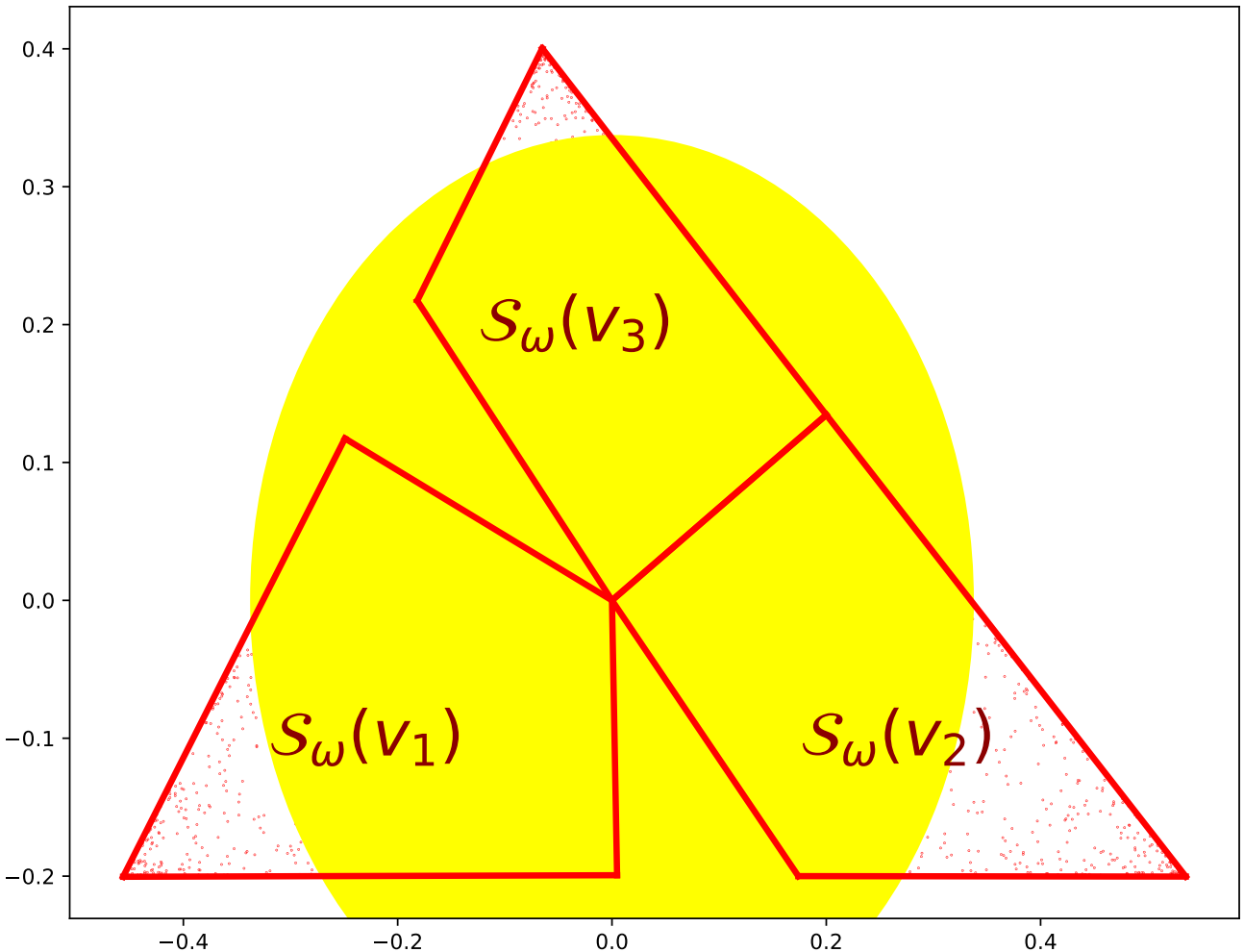


Figure: Complete coverage using cones and sphere

Coverage with Cones and Sphere

Proposition (Y., Guha and Nguyen, 2017)

For choice of ω in some range (ω_3, ω_2), we can choose a sphere of radius $\mathcal{R}$ along with cones, with each cone containing exactly one topic vertex, such that $\omega_3 < \omega_1$.

- If $R_{min} \approx R_{max}$
  - ⋆ $\omega_3 \to 0$ as $\mathcal{R} \to R_{min}$
- Recommended parameters
  - ⋆ for cones choose $\omega = 0.6$
  - ⋆ for sphere choose $\mathcal{R} =$ median of $\{\|\tilde{p}_1\|_2, \ldots, \|\tilde{p}_M\|_2\}$

Coverage with Cones and Sphere

Proposition (Y., Guha and Nguyen, 2017)

For choice of ω in some range (ω_3, ω_2), we can choose a sphere of radius $\mathcal{R}$ along with cones, with each cone containing exactly one topic vertex, such that $\omega_3 < \omega_1$.

- If $R_{min} \approx R_{max}$
  - $\star$ $\omega_3 \to 0$ as $\mathcal{R} \to R_{min}$
- Recommended parameters
  - $\star$ for cones choose $\omega = 0.6$
  - $\star$ for sphere choose $\mathcal{R} =$ median of $\{\|\tilde{p}_1\|_2, \ldots, \|\tilde{p}_M\|_2\}$

Theorem (Y., Guha and Nguyen, 2017)

As $M \to \infty$ the minimum matching distance between estimated and true topics $\to 0$ almost surely.The estimated number of topics also equals the true number of topics almost surely.

Conic Scan-and-Cover (CoSAC) at work

Data is simulated according to the LDA generative process

$\alpha = 0.1,\ \eta = 0.1,\ V = 2000,\ K = 30,\ M = 15000.$

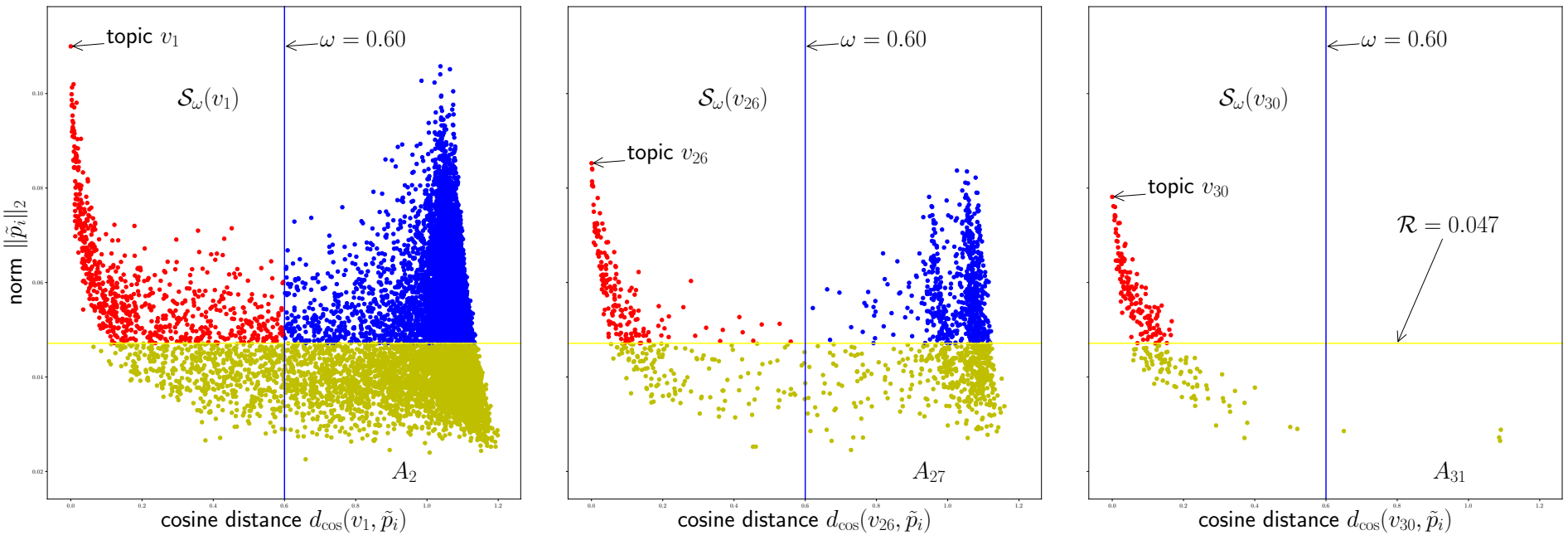
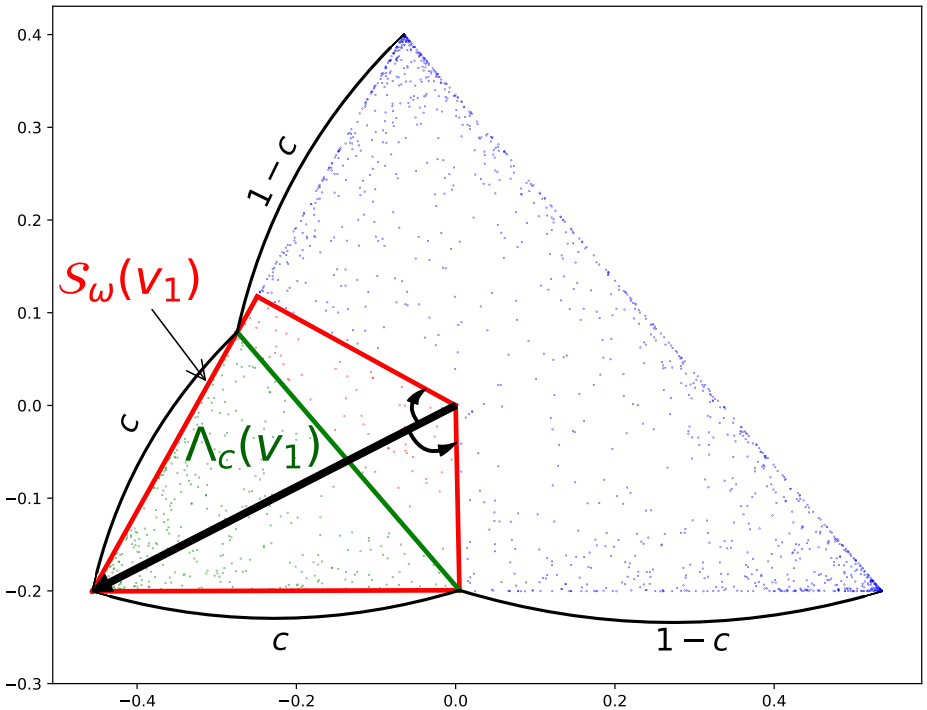


Figure: CoSAC iterations 1, 26, 30. Red: documents in the cone $\mathcal{S}_\omega(v_k)$; Blue: documents in the active set A_{k+1} for next iteration. Yellow: documents $\|\tilde{p}_m\|_2 < \mathcal{R}$.

Conic Scan-and-Cover (CoSAC): Outliers

- If cone contains less than λ portion of data points, ignore cone
- Idea: quantify $\mathbb{P}(\Lambda_c(v_1))$ using Beta distribution
- Recommended $\lambda = 0.001$



Proposition (Y., Guha and Nguyen, 2017)

For $\omega \in (\omega_\lambda, \omega_2)$, each cone around a topic contains at least λ proportion of documents.

Conic Scan-and-Cover (CoSAC): Mean-Shifting

Farthest document as topic estimate

$$v_k = \underset{\tilde{p}_m : m \in A_k}{\operatorname{argmax}} \|\tilde{p}_m\|_2 \text{ — high variance!}$$

1: **while** v_k not converged **do** {mean-shifting}
2: Find cone of near documents

$$\mathcal{S}_\omega(v_k) = \{m : d_{\cos}(\tilde{w}_m, v_k) < \omega\}$$

3: Update direction

$$v_k = \sum_{m \in \mathcal{S}_\omega(v_k)} \tilde{w}_m / \operatorname{card}(\mathcal{S}_\omega(v_k))$$

4: **end while**

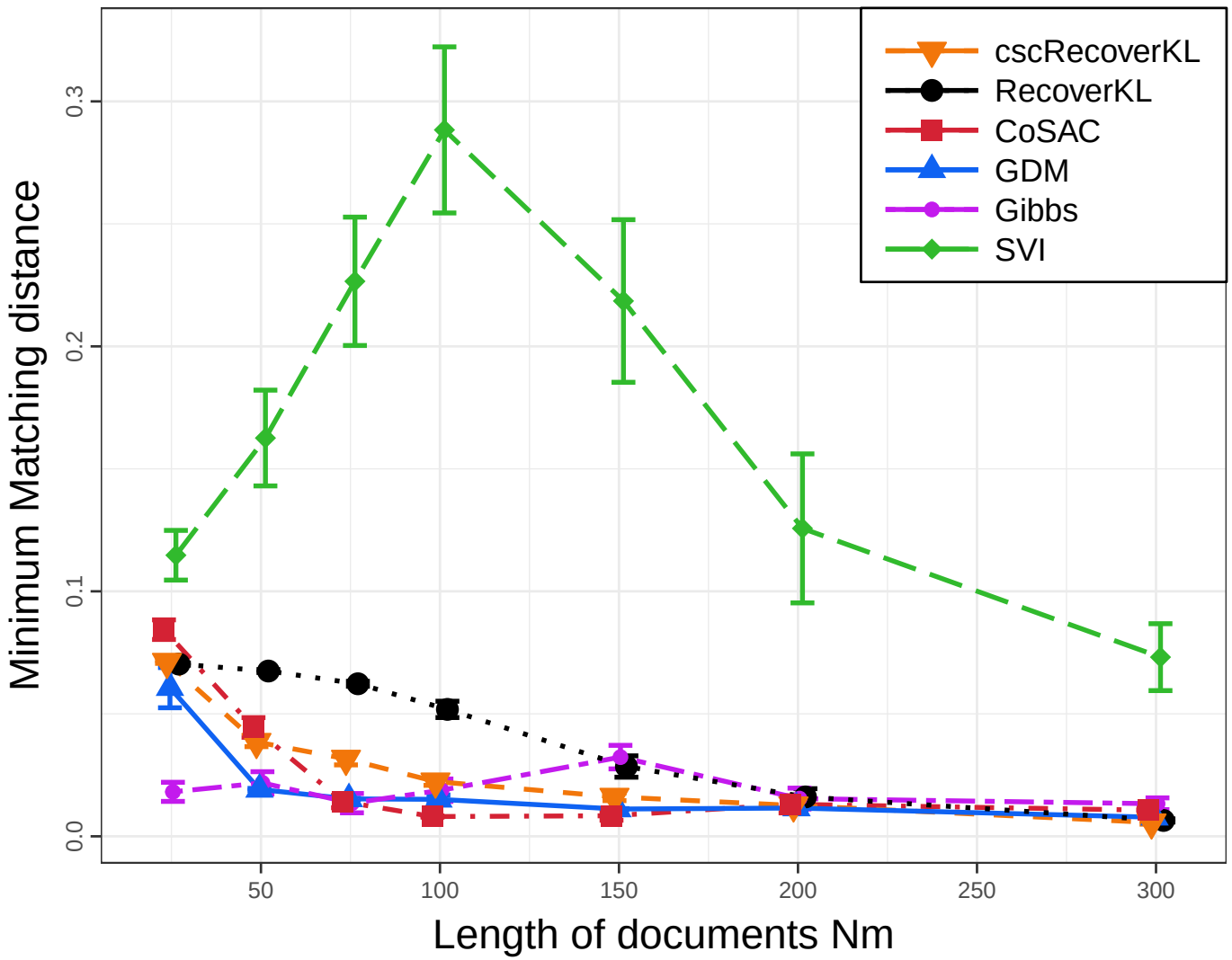
Mean-shifting leads to reduced variance of topic direction estimate.

CoSAC for Anchor words

- Anchor word has non-zero probability in a single topic
- Each topic is assumed to have an anchor word
- Anchor rows of word co-occurrence matrix form a simplex containing rest of the rows
- Use CoSAC to find anchor words
- Learn topics with RecoverKL (Arora et al., 2013)

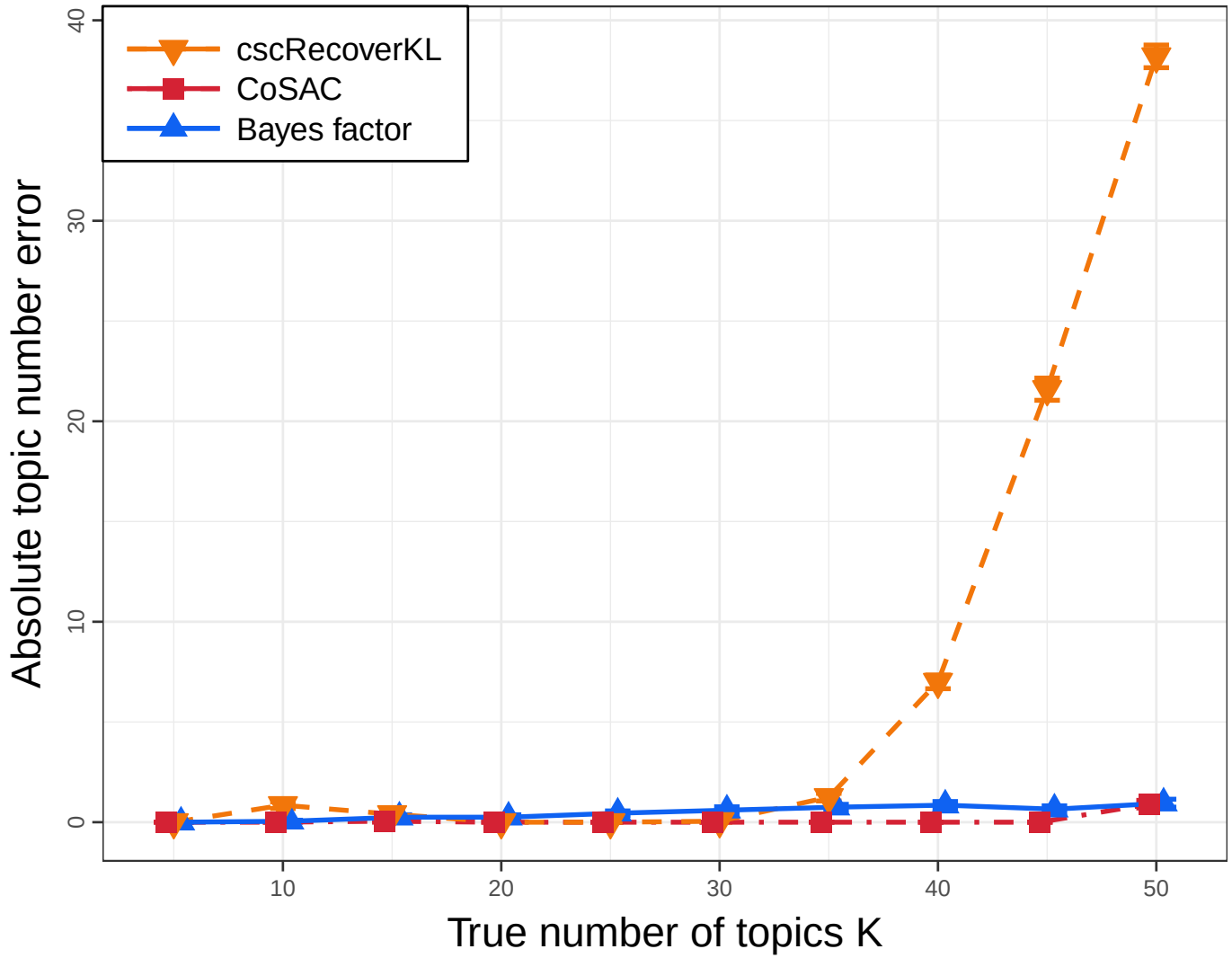
Simulations: document length

$\alpha = 0.1,\ \eta = 0.1,\ V = 2000,\ K = 15,\ M = 30000.$



Simulations: number of topics

$\alpha = 0.1,\ \eta = 0.1,\ V = 2000,\ M = 5000,\ N_m = 500.$



New York Times results

We analyzed 130000 NYT articles with vocabulary $V = 5320$.

	K	Perplexity	Coherence	Time
cscRecoverKL	27	2603	-238	37 min
HDP Gibbs	221 ± 5	1477 ± 1.6	-442 ± 1.7	35 hours
LDA Gibbs	80	1520 ± 1.5	-300 ± 0.7	5.3 hours
CoSAC	159	1568	-322	**19 min**

New York Times topics

Cooking	*Stem Cells*	*Antitrust*	*LGBT*	*Elections*
cup	cell	Microsoft	gay	ballot
minutes	stem	window	lesbian	Al Gore
tablespoon	research	company	right	election
add	human	software	sex	votes
teaspoon	scientist	case	marriage	recount
pepper	cloning	system	group	Florida
oil	patient	operating	couples	court
sugar	disease	computer	sexual	vote
butter	phones	antitrust	partner	voter
pan	researcher	court	issue	count

Topics evolve

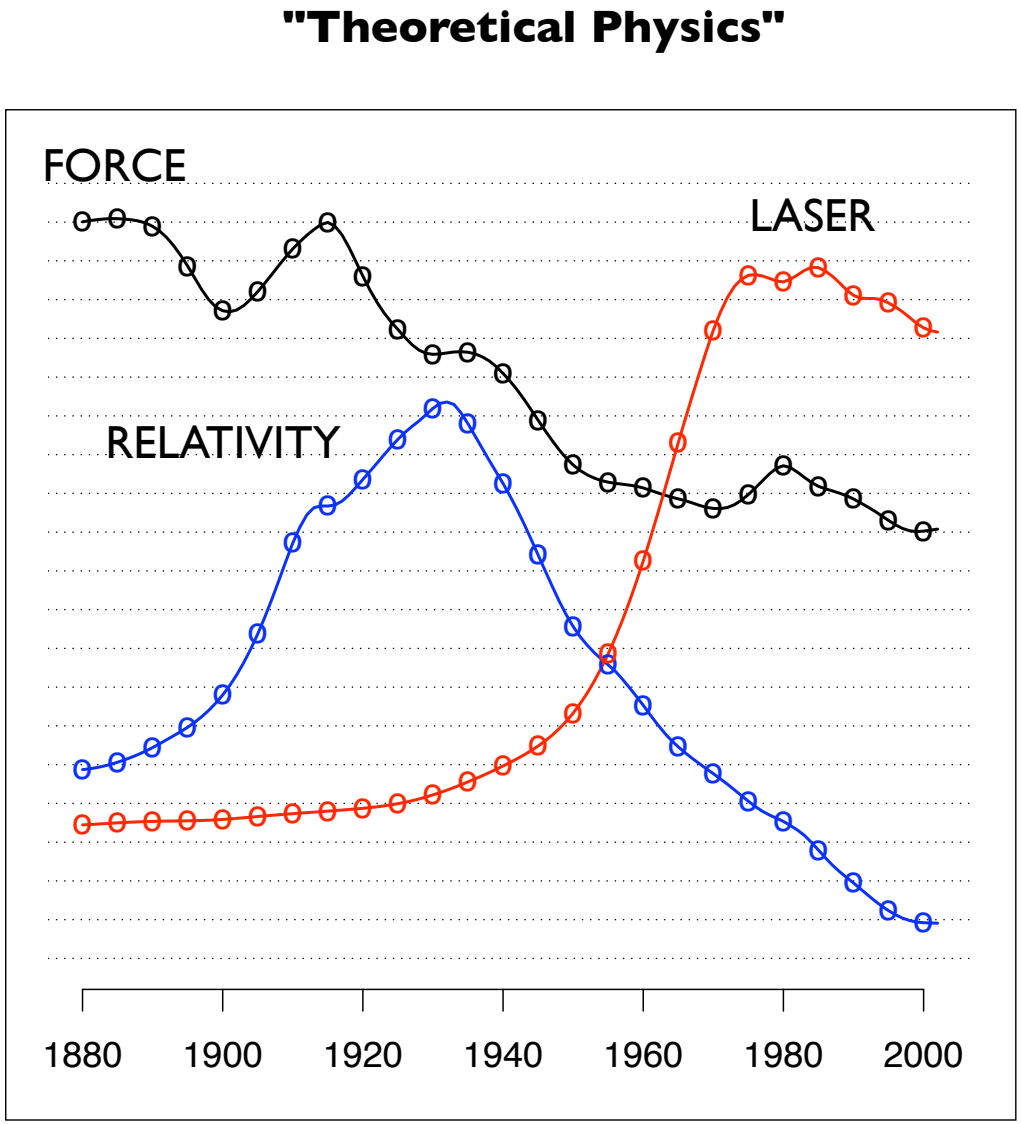


Figure: Dynamic Topic Models (Blei & Lafferty, 2006) on *Science*

Online learning

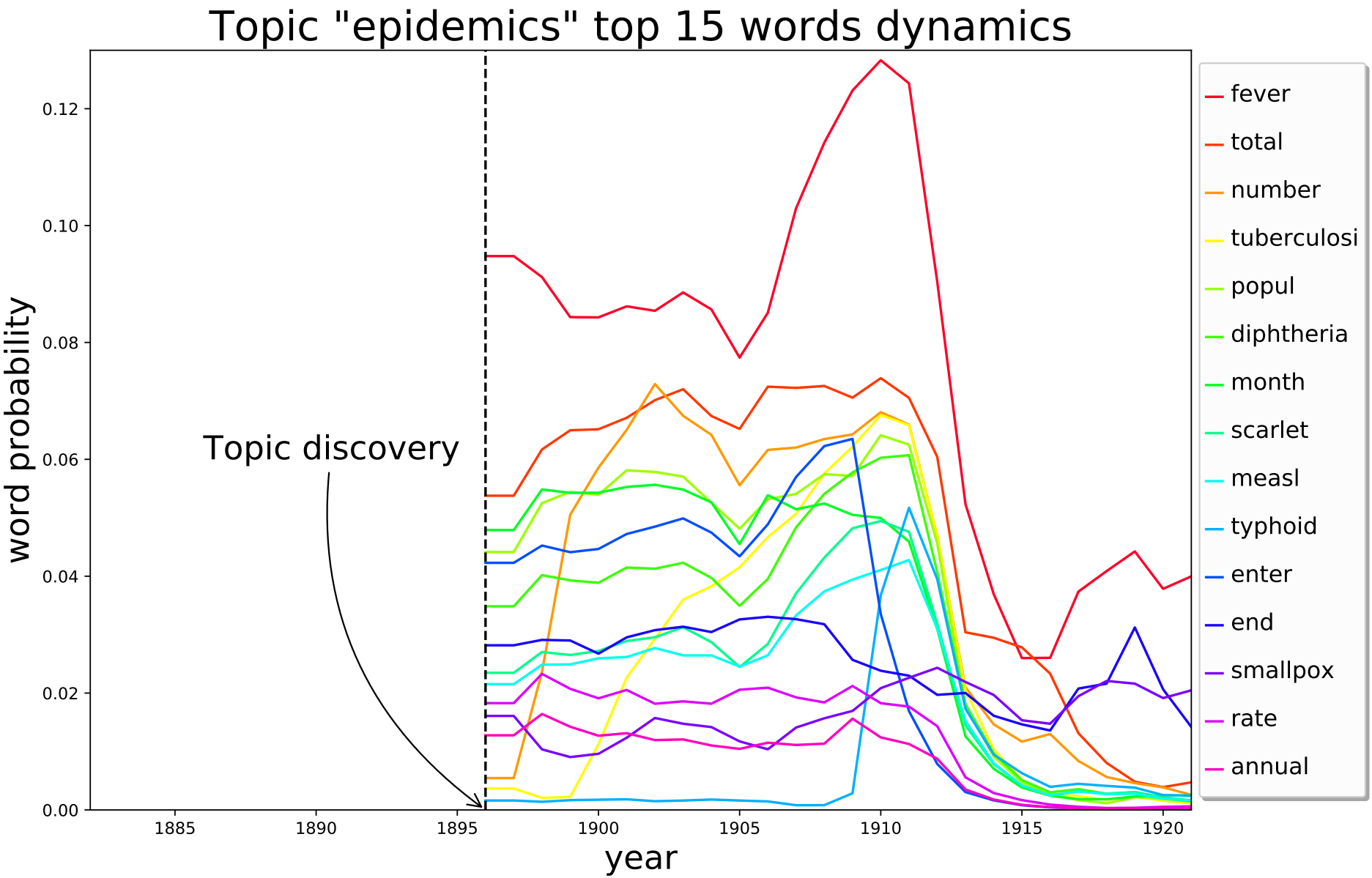
$\approx$ 240k tweets every 30 seconds

Streaming Dynamic Matching

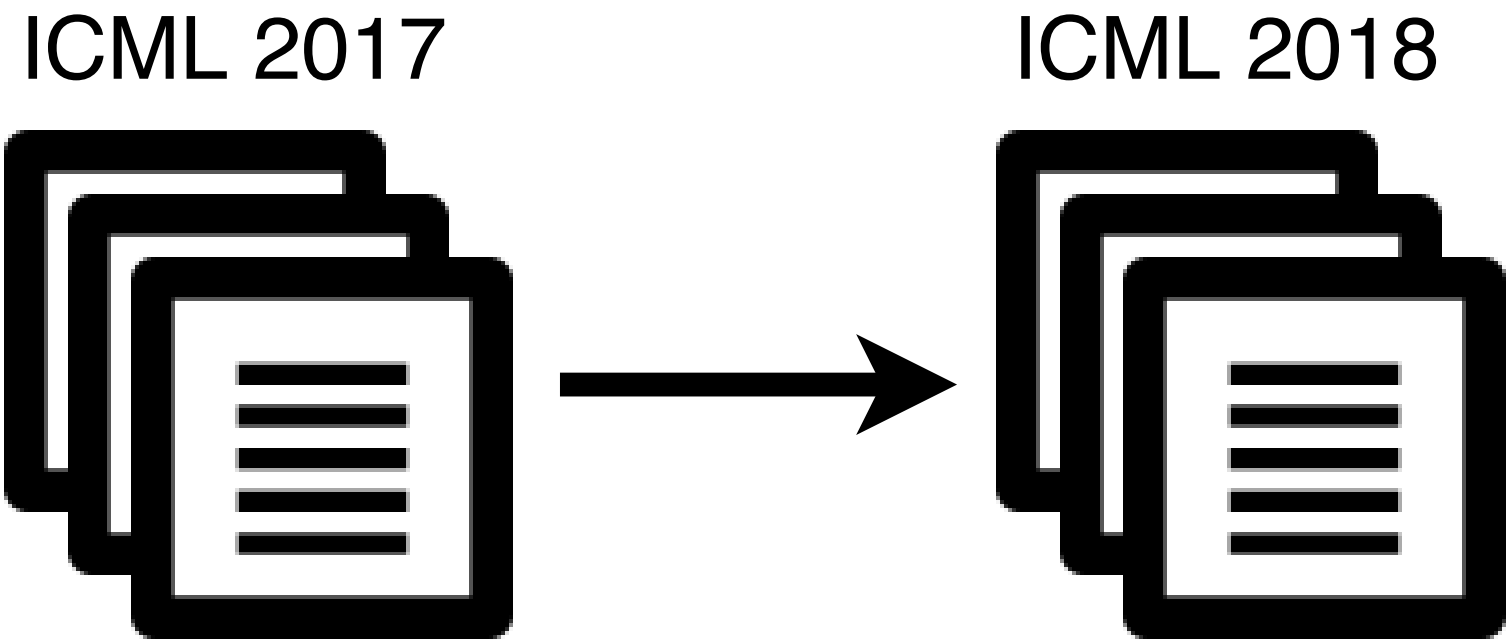
Early Journal Content: 400k articles, 4500 unique words, 40 years.

	Perplexity	Time	Topics	Cores used
SDM	**1181**	**24min**	124	1
DTM	1194	56hours	100	1

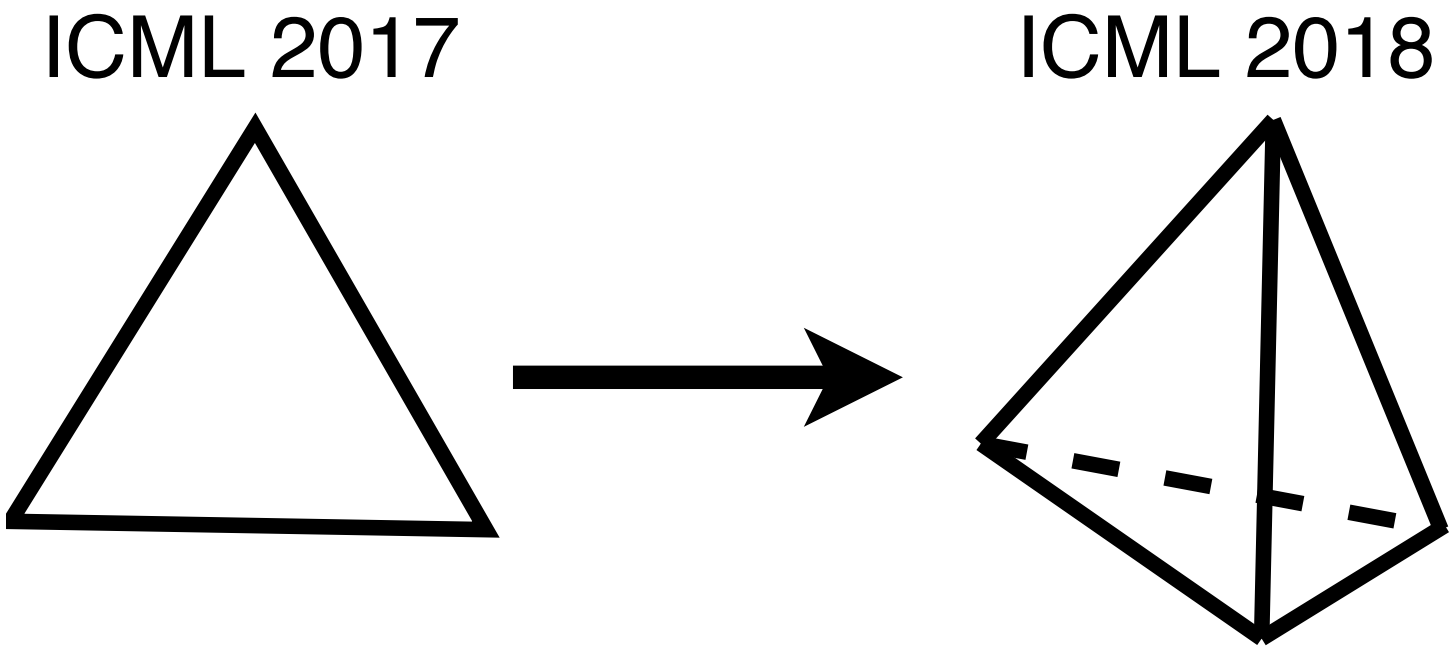
Streaming Dynamic Matching



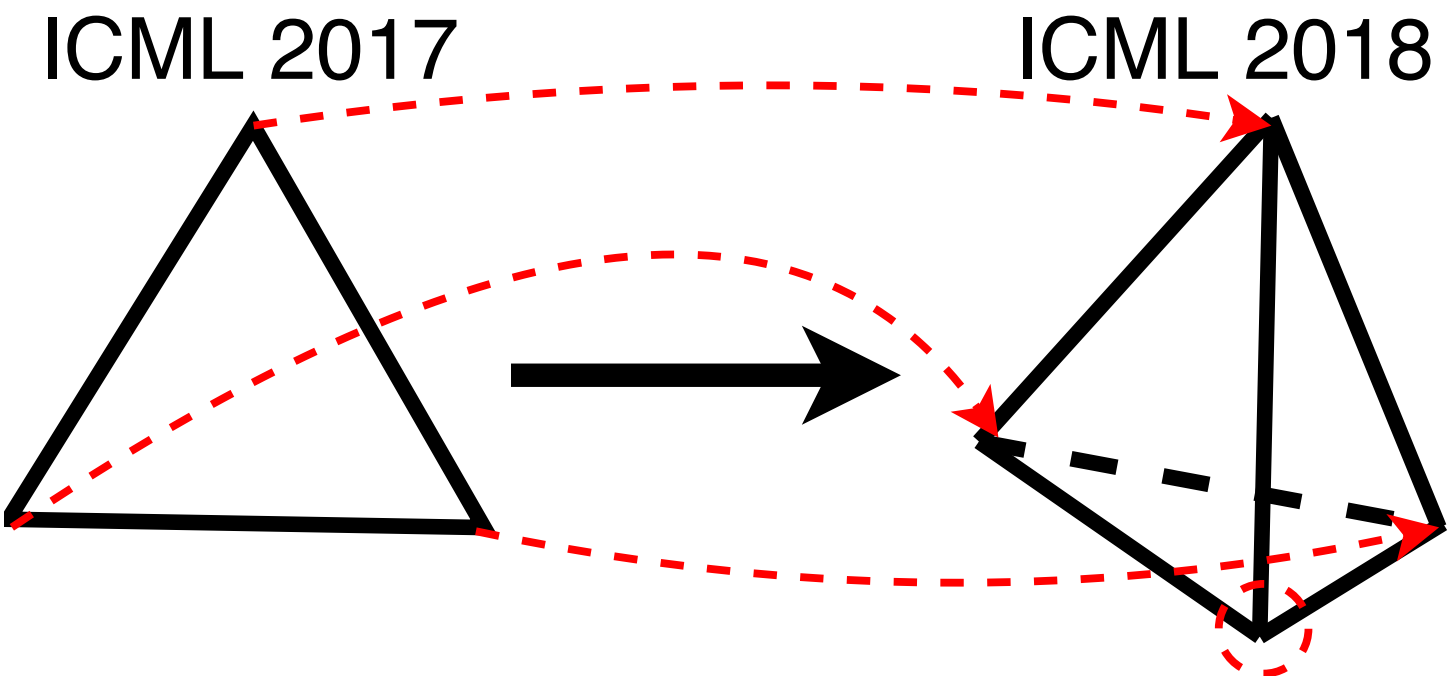
Geometry evolves



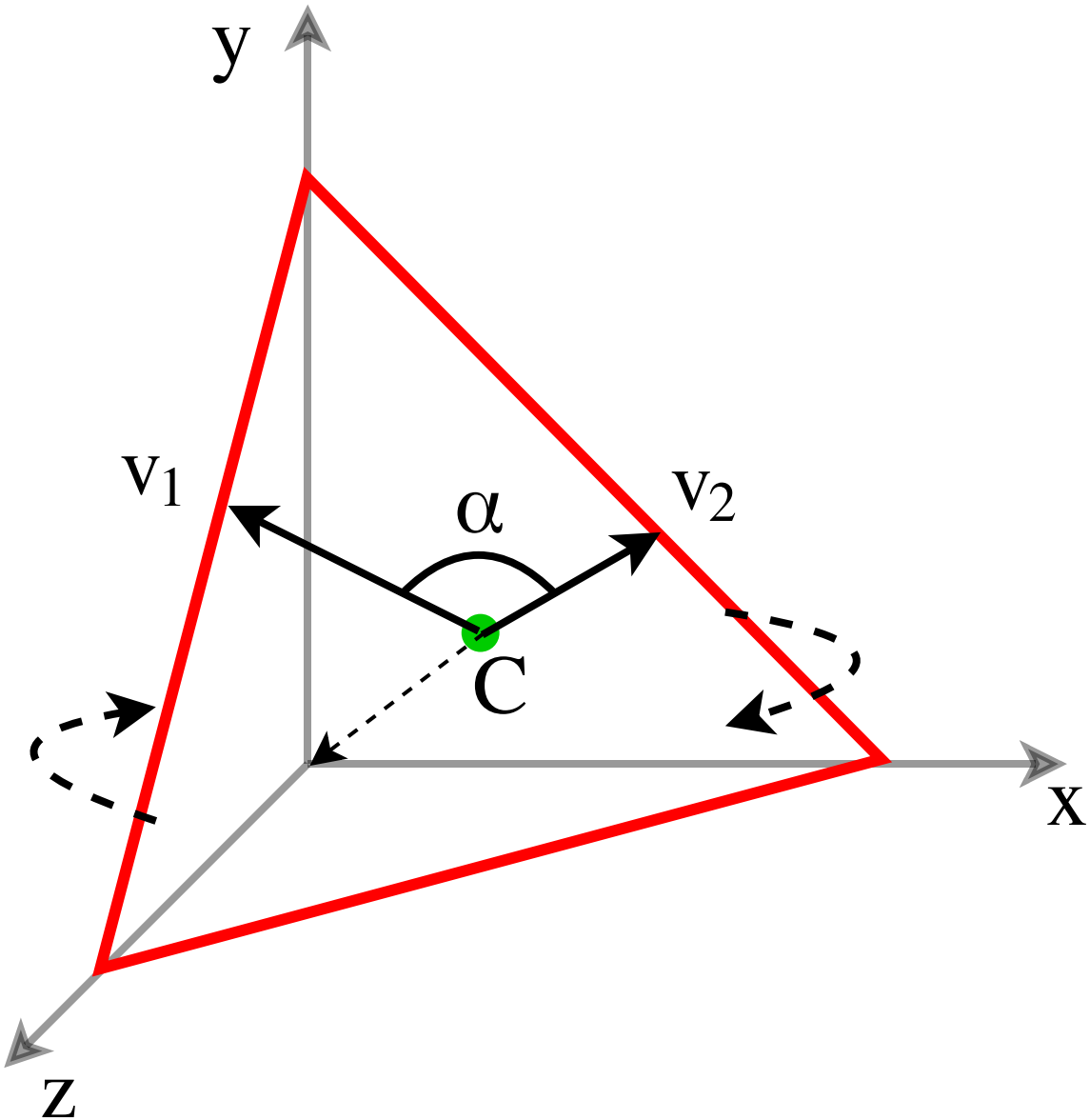
Geometry evolves



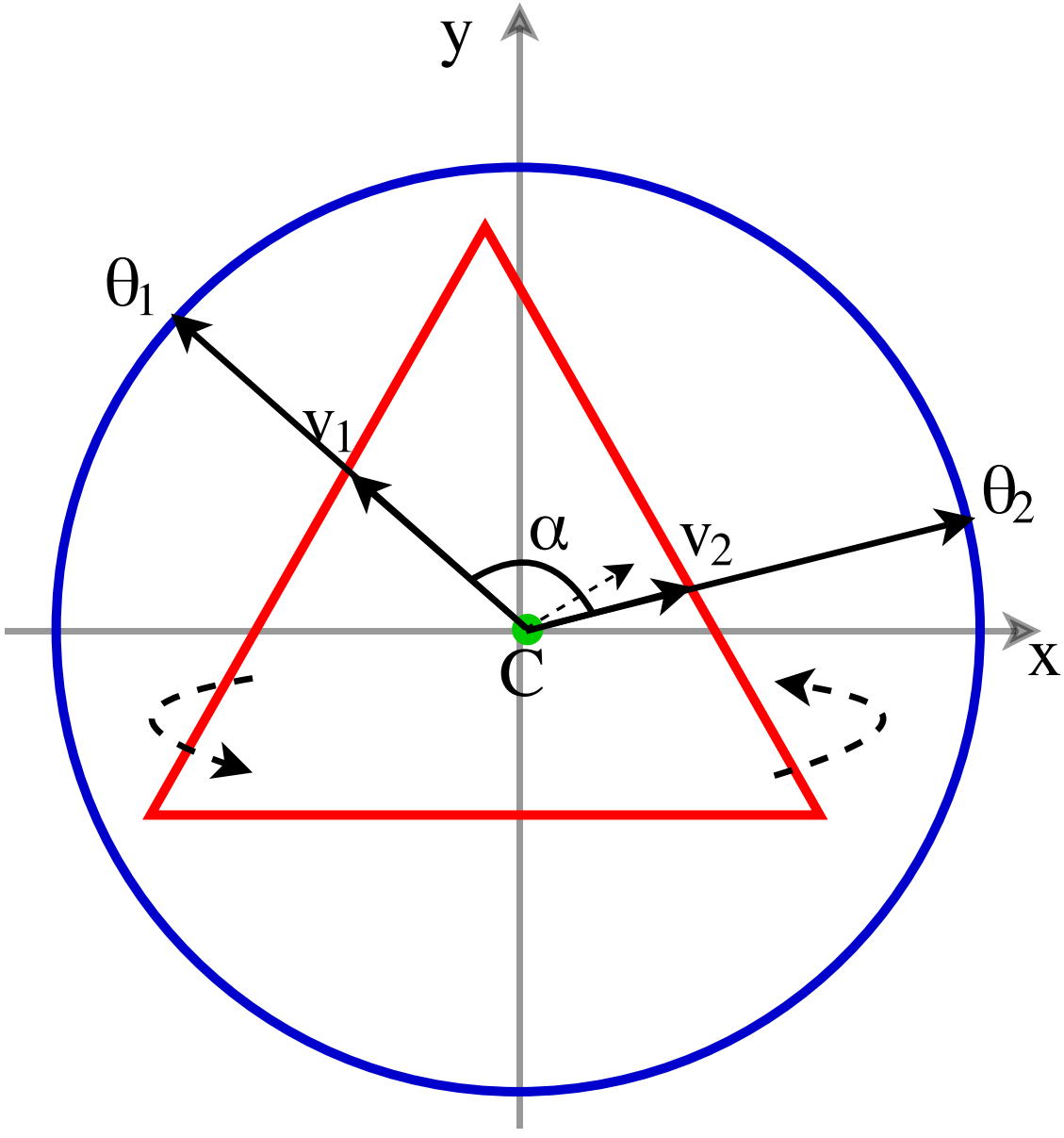
Geometry evolves



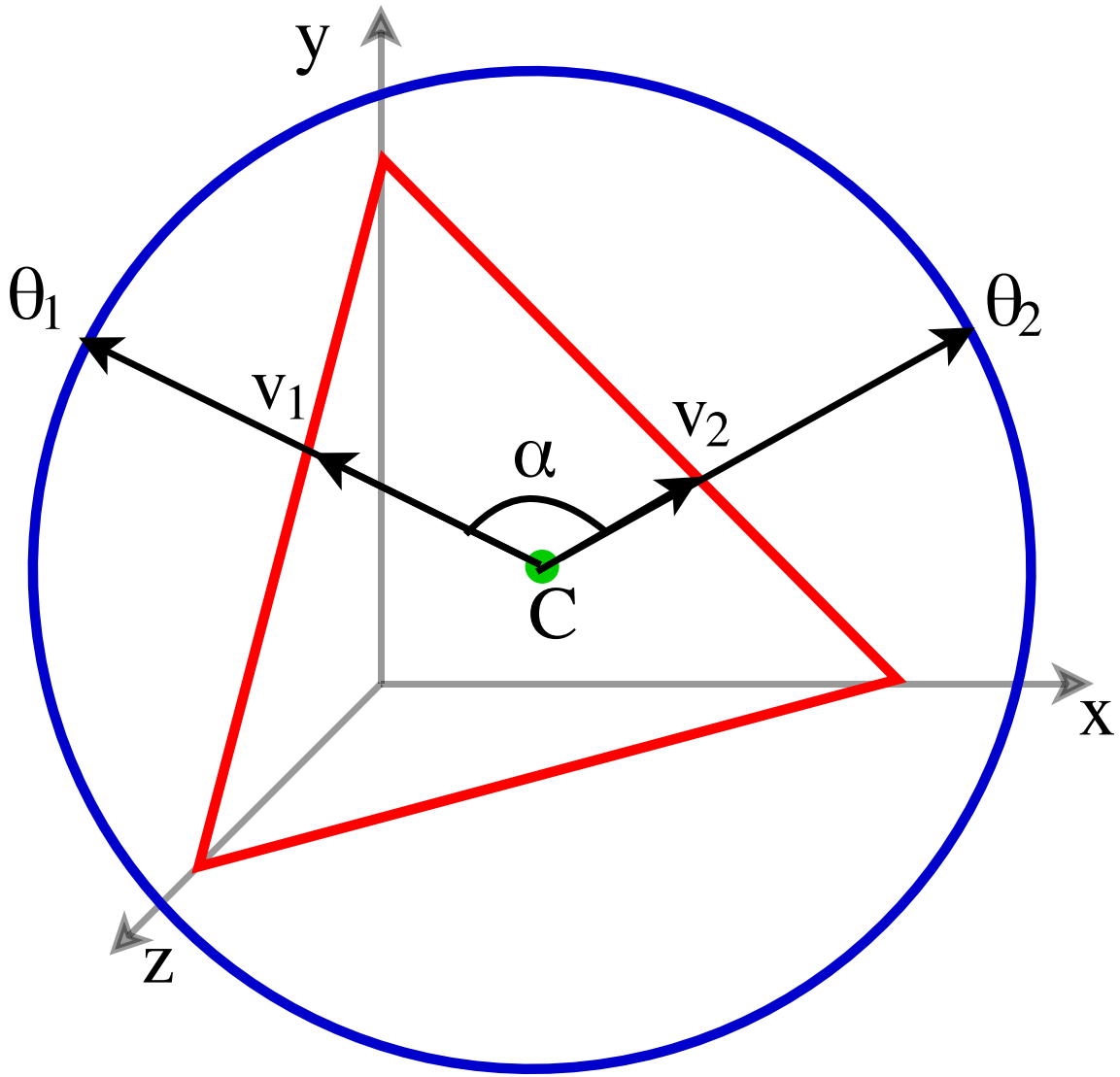
Topic polytope dynamics on sphere



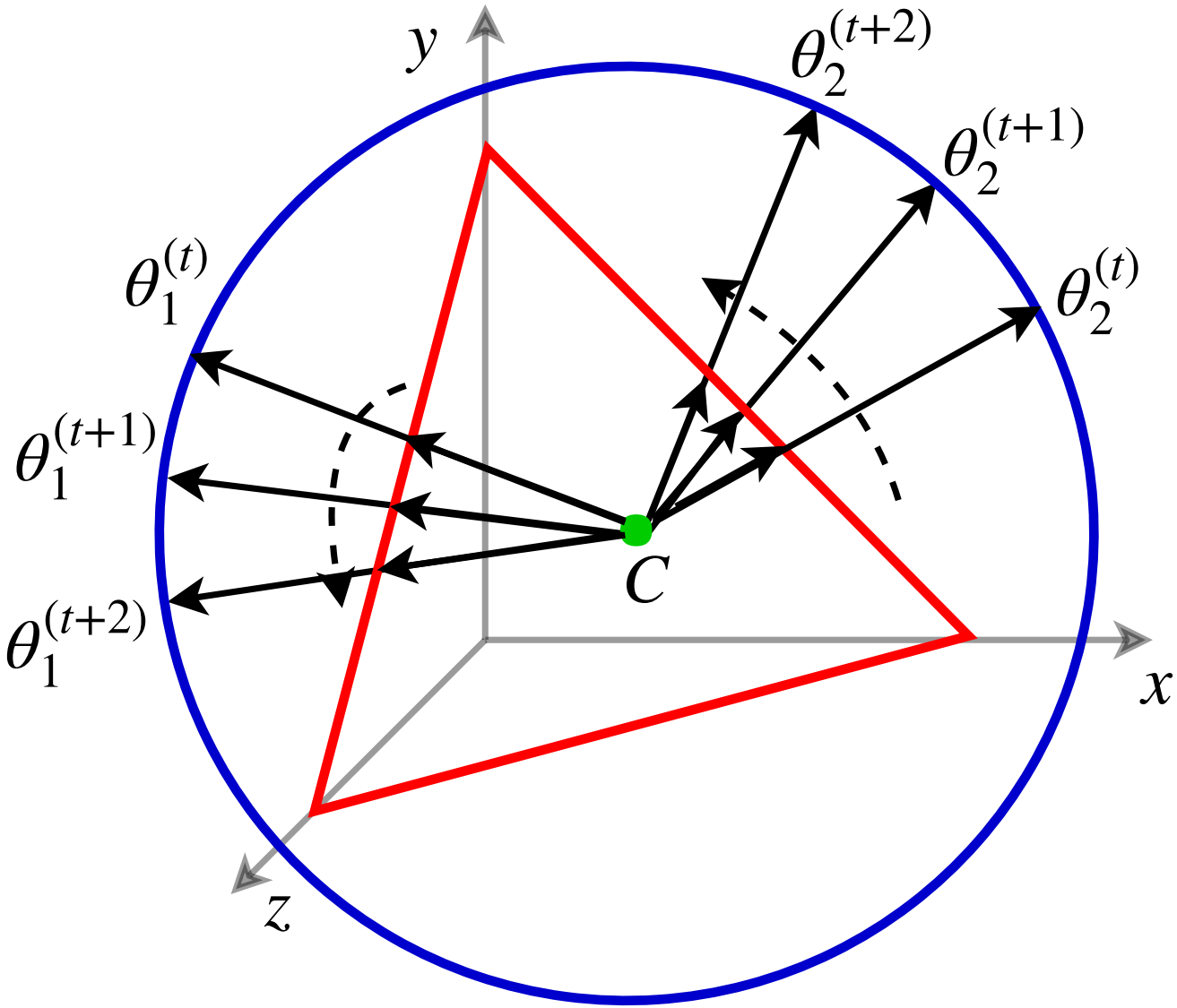
Topic polytope dynamics on sphere



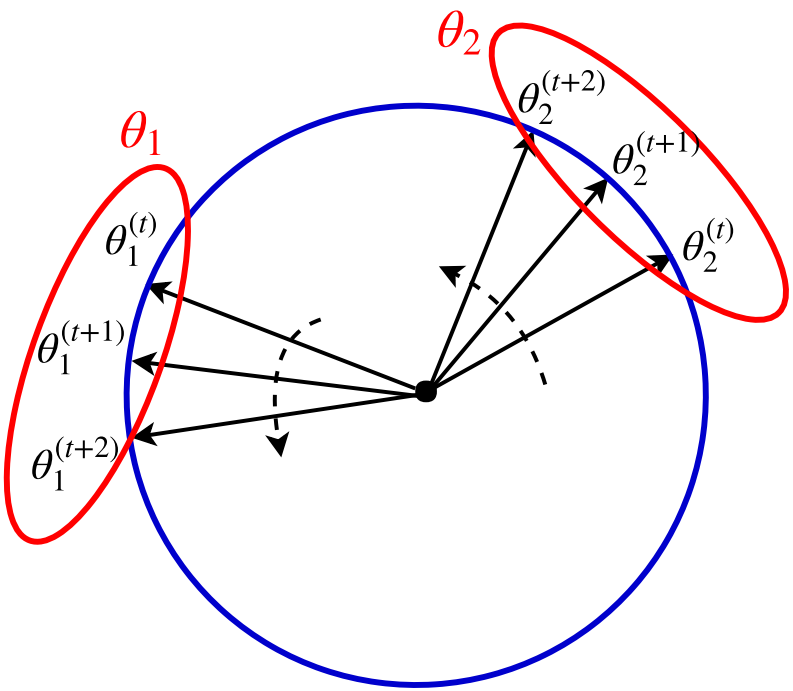
Topic polytope dynamics on sphere



Topic polytope dynamics on sphere



Dynamic Beta process: Global topics

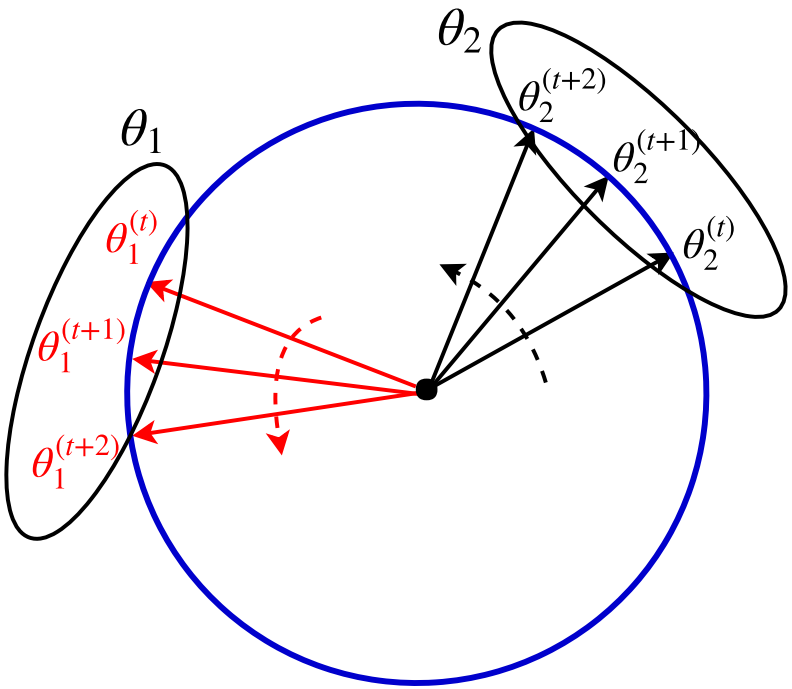


Global topics: $Q|\gamma_0, H \sim \mathrm{BP}(\gamma_0, H)$.

$$Q = \sum_i q_i \delta_{\theta_i}.$$

Topic i: $\theta_i := \{\theta_i^{(t)}\}_{t=1}^{T} \sim H, q_i \in [0,1]$.

Dynamic Beta process: Global topics



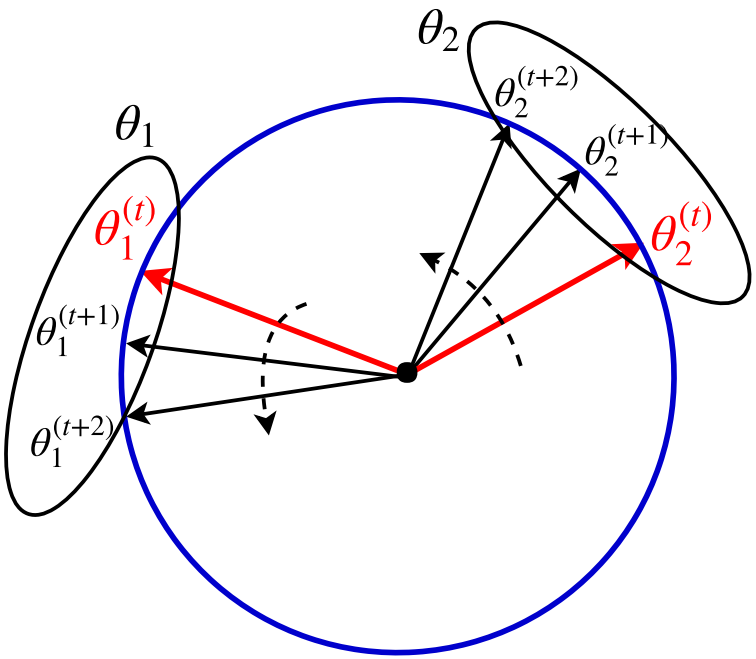
Global topics: $Q|\gamma_0, H \sim \mathrm{BP}(\gamma_0, H)$.

Base measure H :

$\theta_i^{(t)}|\theta_i^{(t-1)} \sim \mathrm{vMF}(\theta_i^{(t-1)}, \tau_0)$, $t = 1, \ldots, T$,

$\theta_i^{(0)} \sim \mathrm{vMF}(\cdot, 0)$ – uniform on sphere.

Dynamic Beta process: Global topics



Global topics: $Q|\gamma_0, H \sim \mathrm{BP}(\gamma_0, H)$.

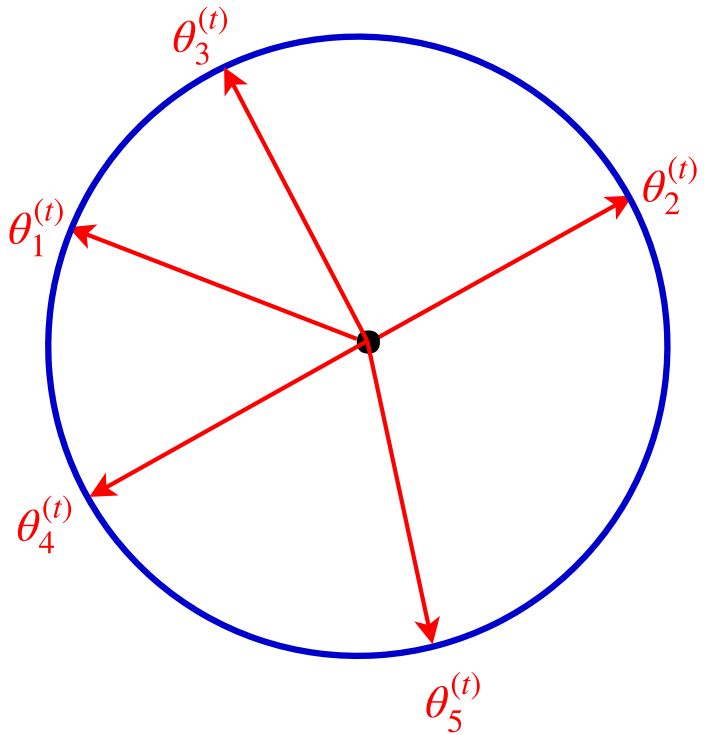
Base measure H :

$\theta_i^{(t)}|\theta_i^{(t-1)} \sim \mathrm{vMF}(\theta_i^{(t-1)}, \tau_0),\ t = 1, \ldots, T,$

$\theta_i^{(0)} \sim \mathrm{vMF}(\cdot, 0)$ – uniform on sphere.

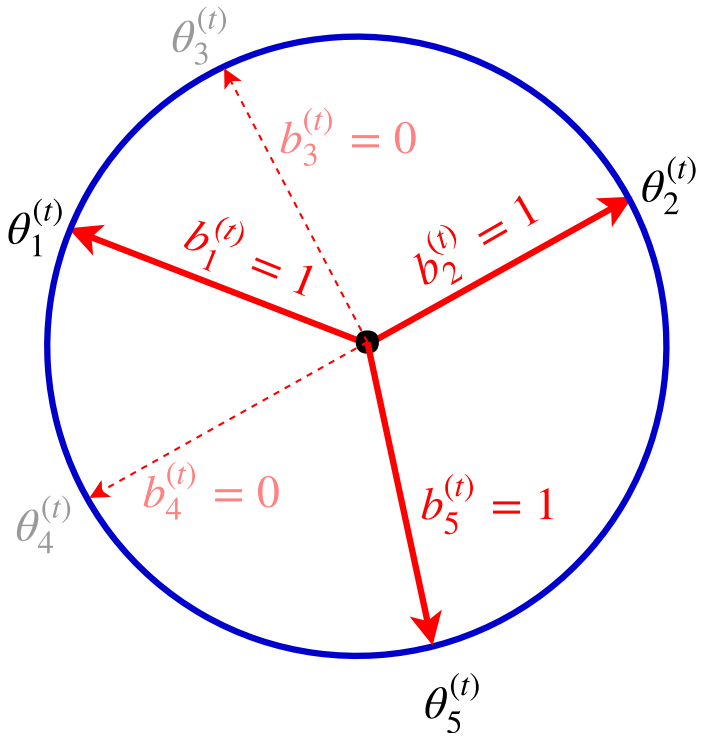
Global topics at time t: $Q_t = \sum_i q_i \delta_{\theta_i^{(t)}}$.

Dynamic Beta process: Local topics



Global topics at time t: $Q_t = \sum_i q_i \delta_{\theta_i^{(t)}}$.

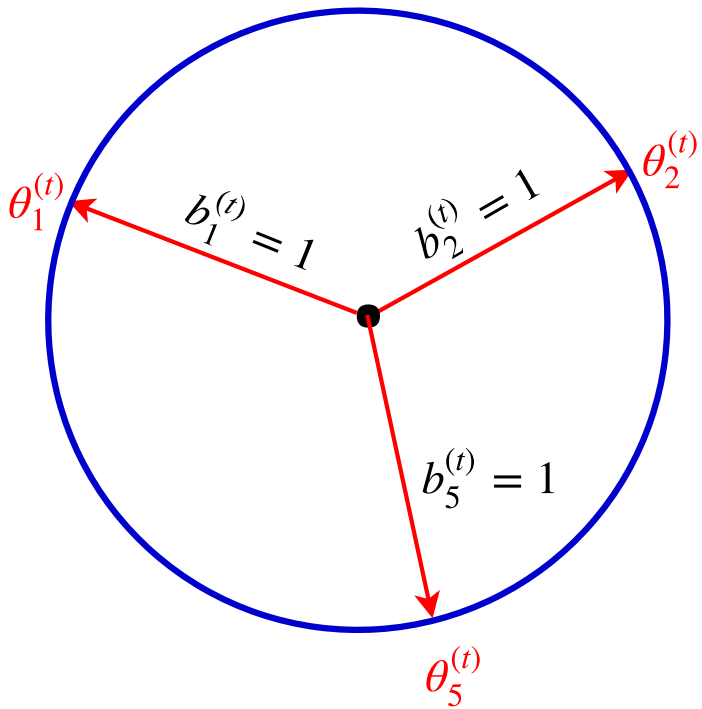
Dynamic Beta process: Local topics



Local topics $T^{(t)}|Q_t \sim \mathrm{BeP}(Q_t)$,

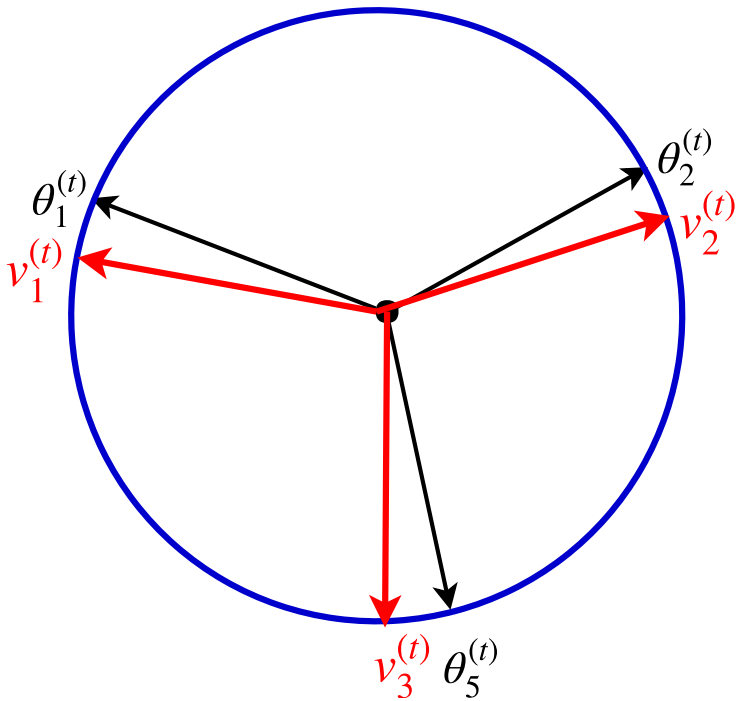
$$T^{(t)} := \sum_i b_i^{(t)} \delta_{\theta_i^{(t)}}, \quad b_i^{(t)}|q_i \sim \mathrm{Bern}(q_i).$$

Dynamic Beta process: Local topics



$$T^{(t)} = \{\theta_i^{(t)} : b_i^{(t)} = 1, i = 1, 2, \ldots\}.$$

Dynamic Beta process: Local topics

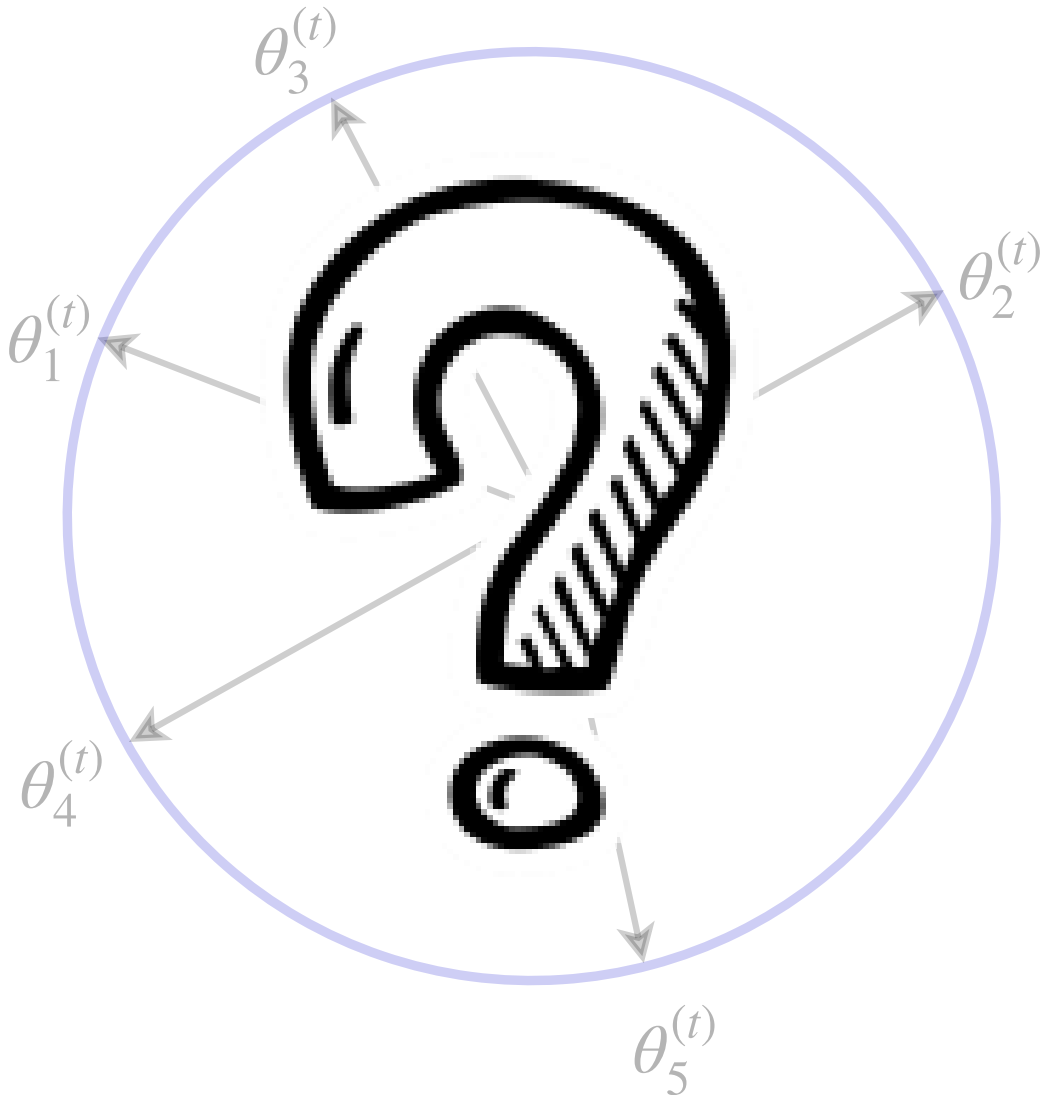


$$T^{(t)} = \{\theta_i^{(t)} : b_i^{(t)} = 1, i = 1, 2, \ldots\}.$$

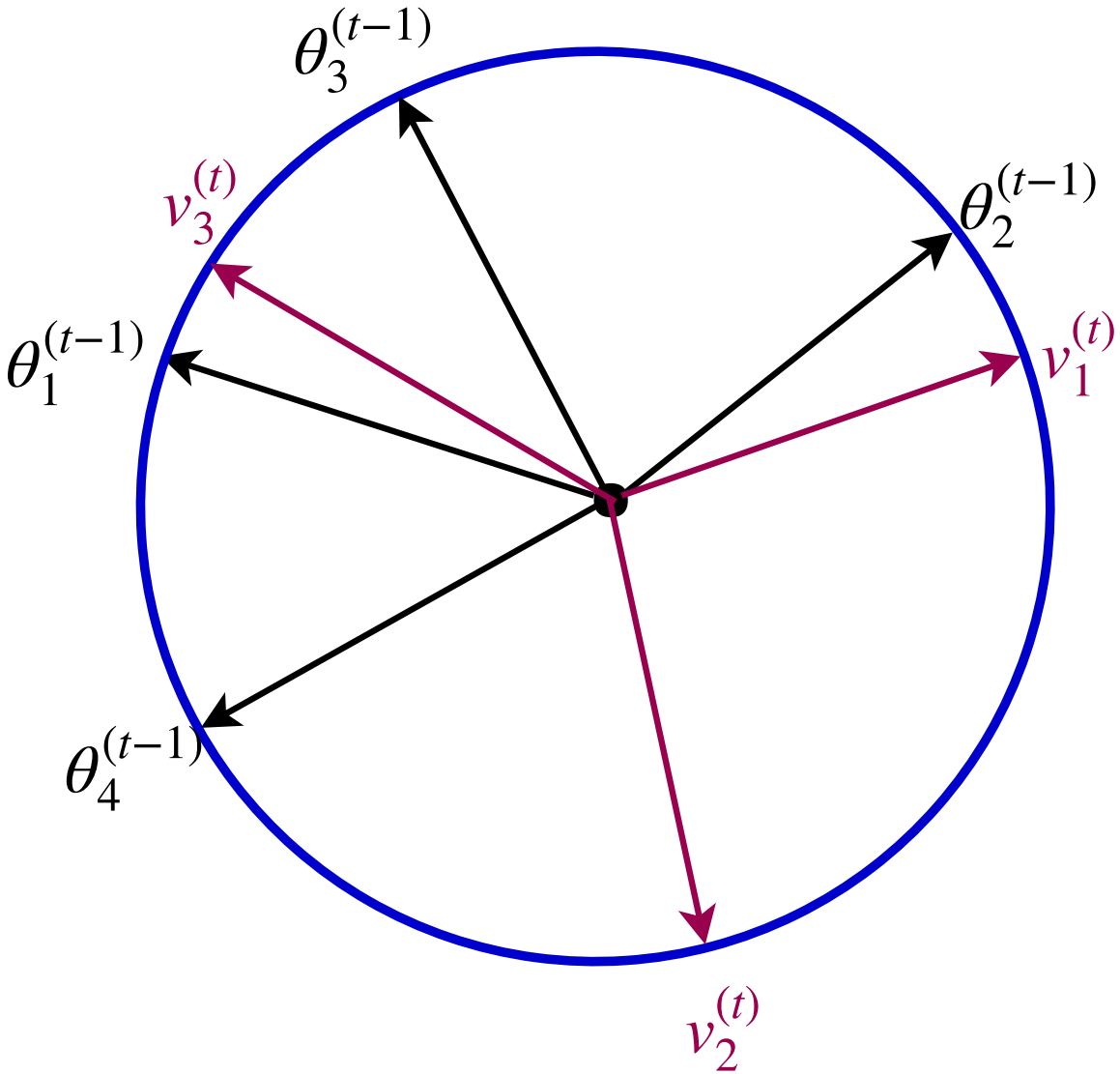
We can estimate noisy measurements:

$$v_k^{(t)} | T^{(t)} \sim \text{vMF}(T_k^{(t)}, \tau_1),\ k = 1, \ldots, K^{(t)}.$$

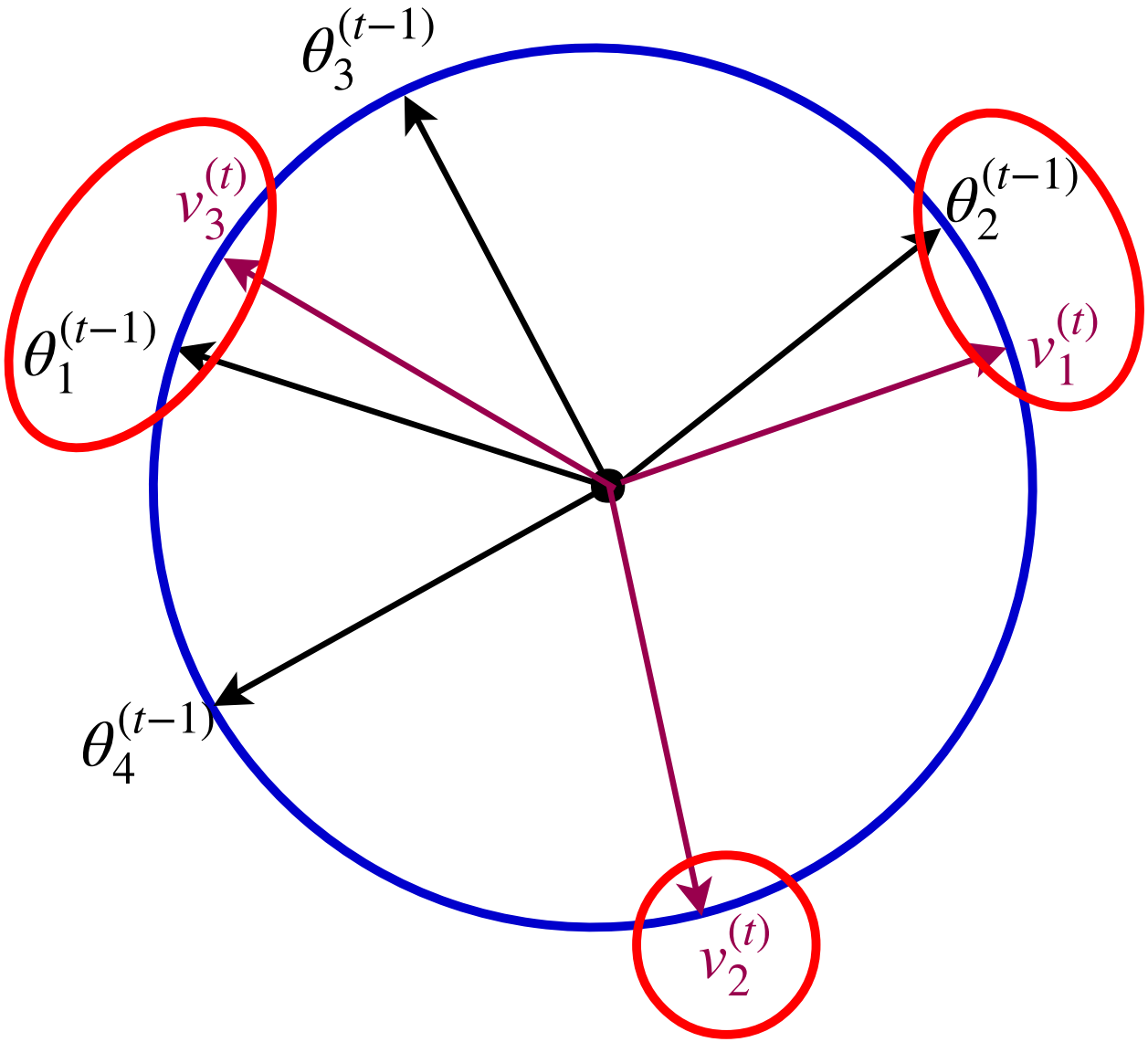
Inference: Matching problem



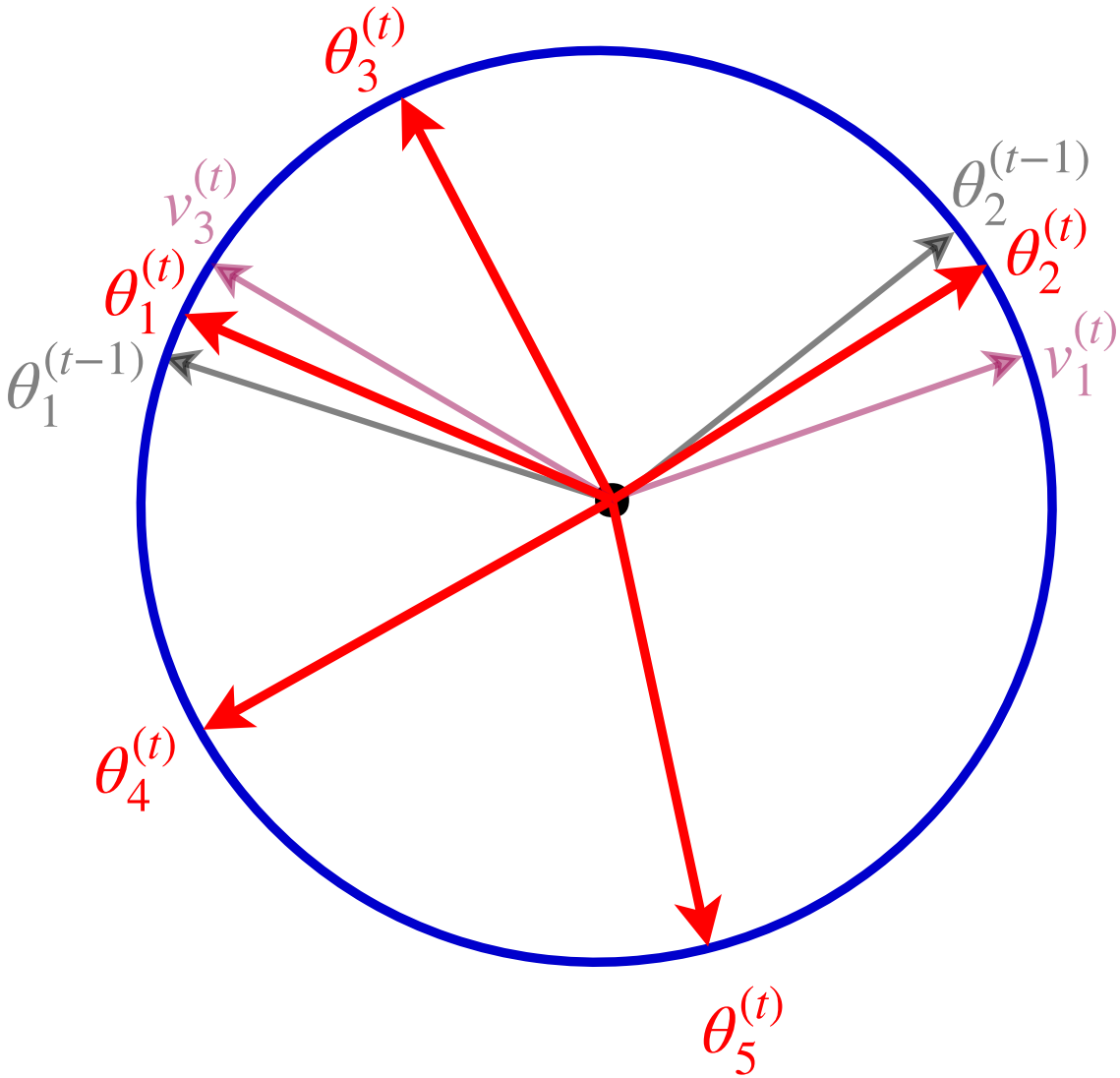
Inference: Matching problem



Inference: Matching problem



Inference: Matching problem



Streaming Dynamic Matching

MAP streaming estimation: $\underset{\theta^{(t)}, B^{(t)}}{\operatorname{argmax}} \mathbb{P}(\theta^{(t)}, B^{(t)} | \theta^{(t-1)}, v^{(t)})$.

$B_{ik}^{(t)} = 1$ iff $v_k^{(t)}$ *matched* to global topic i.

Streaming Dynamic Matching

MAP streaming estimation: $\underset{\theta^{(t)}, B^{(t)}}{\operatorname{argmax}} \mathbb{P}(\theta^{(t)}, B^{(t)} | \theta^{(t-1)}, v^{(t)})$.

Given matching solution for $\theta^{(t)}$ is in closed form:

$$\hat{\theta}_i^{(t)} | B_{ik}^{(t)} = 1, \theta_i^{(t-1)} = \frac{\tau_0 \theta_i^{(t-1)} + \tau_1 v_k^{(t)}}{\|\tau_0 \theta_i^{(t-1)} + \tau_1 v_k^{(t)}\|_2}.$$

Streaming Dynamic Matching

MAP streaming estimation: $\underset{\theta^{(t)}, B^{(t)}}{\operatorname{argmax}} \mathbb{P}(\theta^{(t)}, B^{(t)} | \theta^{(t-1)}, v^{(t)})$.

$$\underset{B^{(t)}}{\operatorname{argmax}} \sum_{i,k} B_{ik}^{(t)} \mathrm{R}(i,k), \text{ where}$$

$$\mathrm{R}(i,k) = \log \frac{m_i^{(t-1)}}{t - m_i^{(t-1)}} + \|\tau_0 \theta_i^{(t-1)} + \tau_1 v_k^{(t)}\|_2,$$

$m_i^{(t-1)}$ — popularity of topic i before time t.

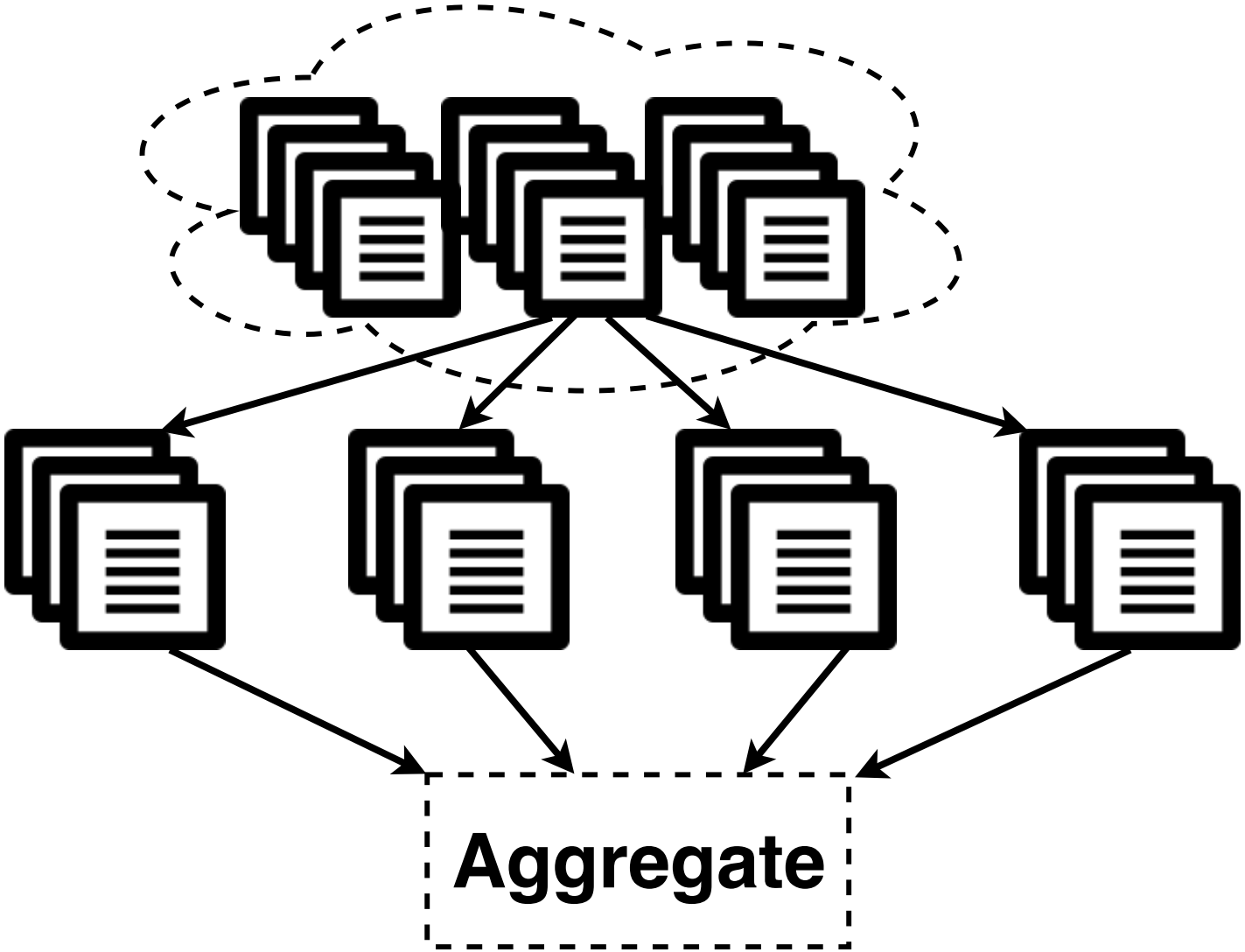
Streaming Dynamic Matching

MAP streaming estimation: $\underset{\theta^{(t)}, B^{(t)}}{\operatorname{argmax}} \mathbb{P}(\theta^{(t)}, B^{(t)} | \theta^{(t-1)}, v^{(t)})$.

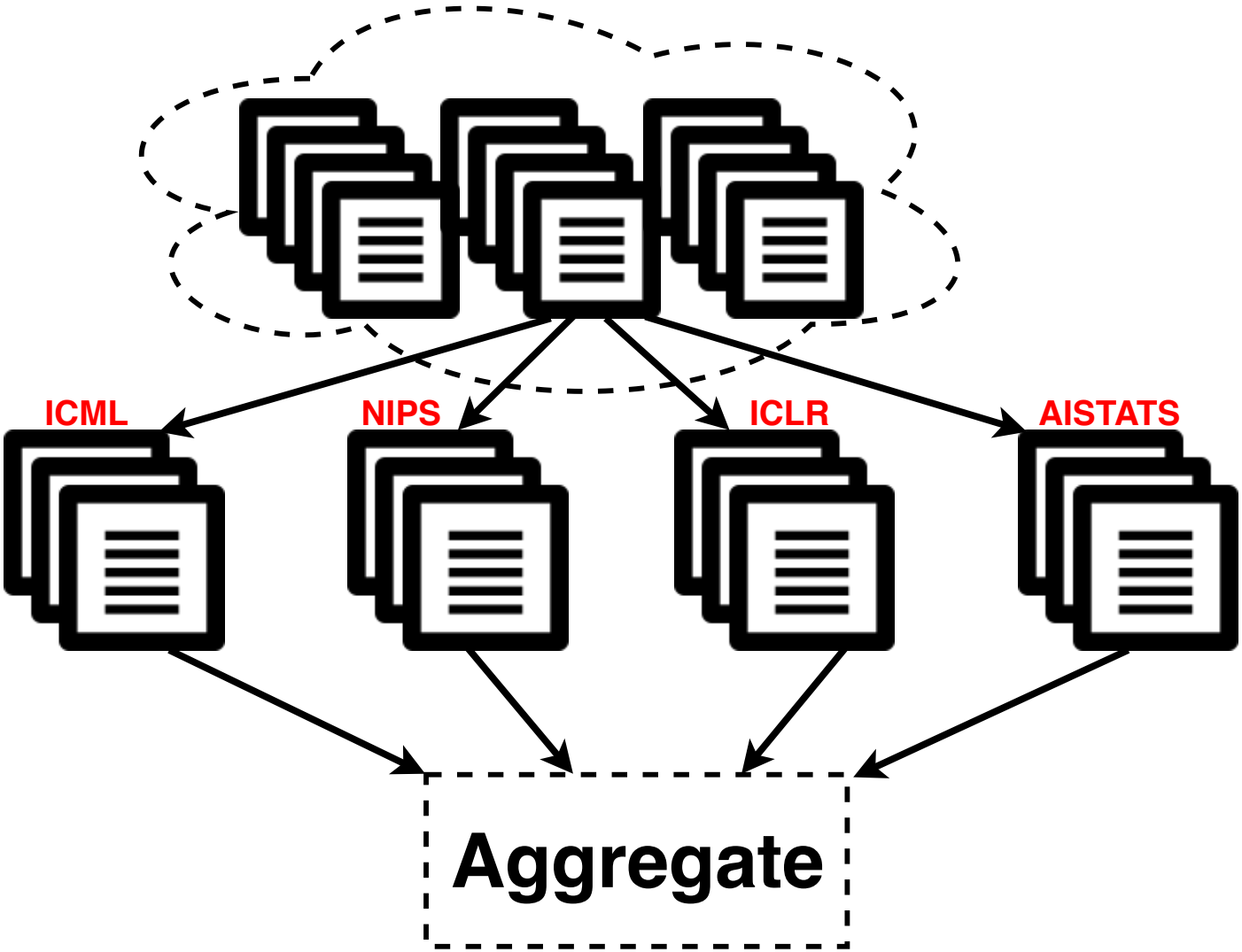
$$\underset{B^{(t)}}{\operatorname{argmax}} \sum_{i,k} B^{(t)}_{ik} \mathrm{R}(i,k)$$

Solve with Hungarian algorithm!

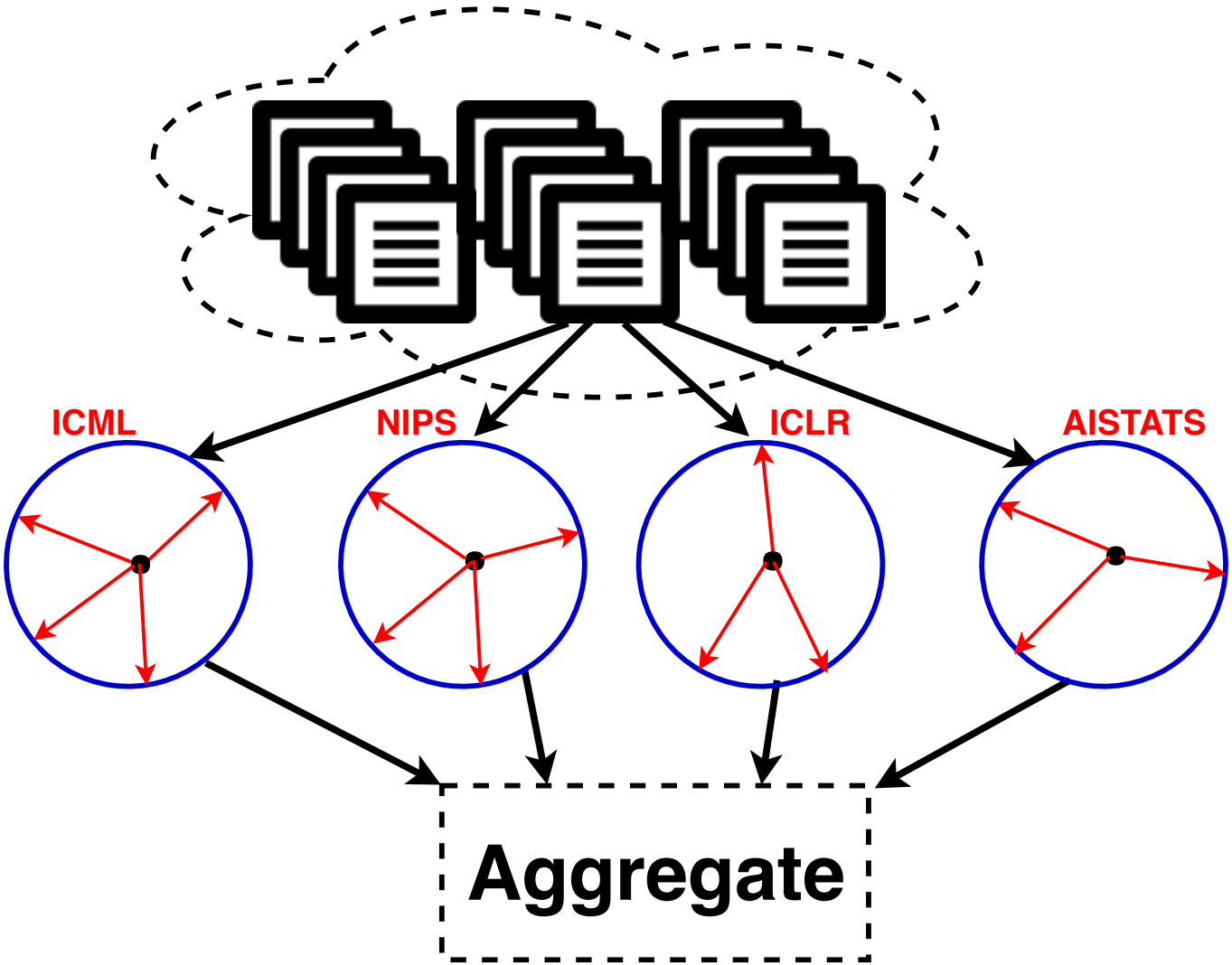
Distributed Matching



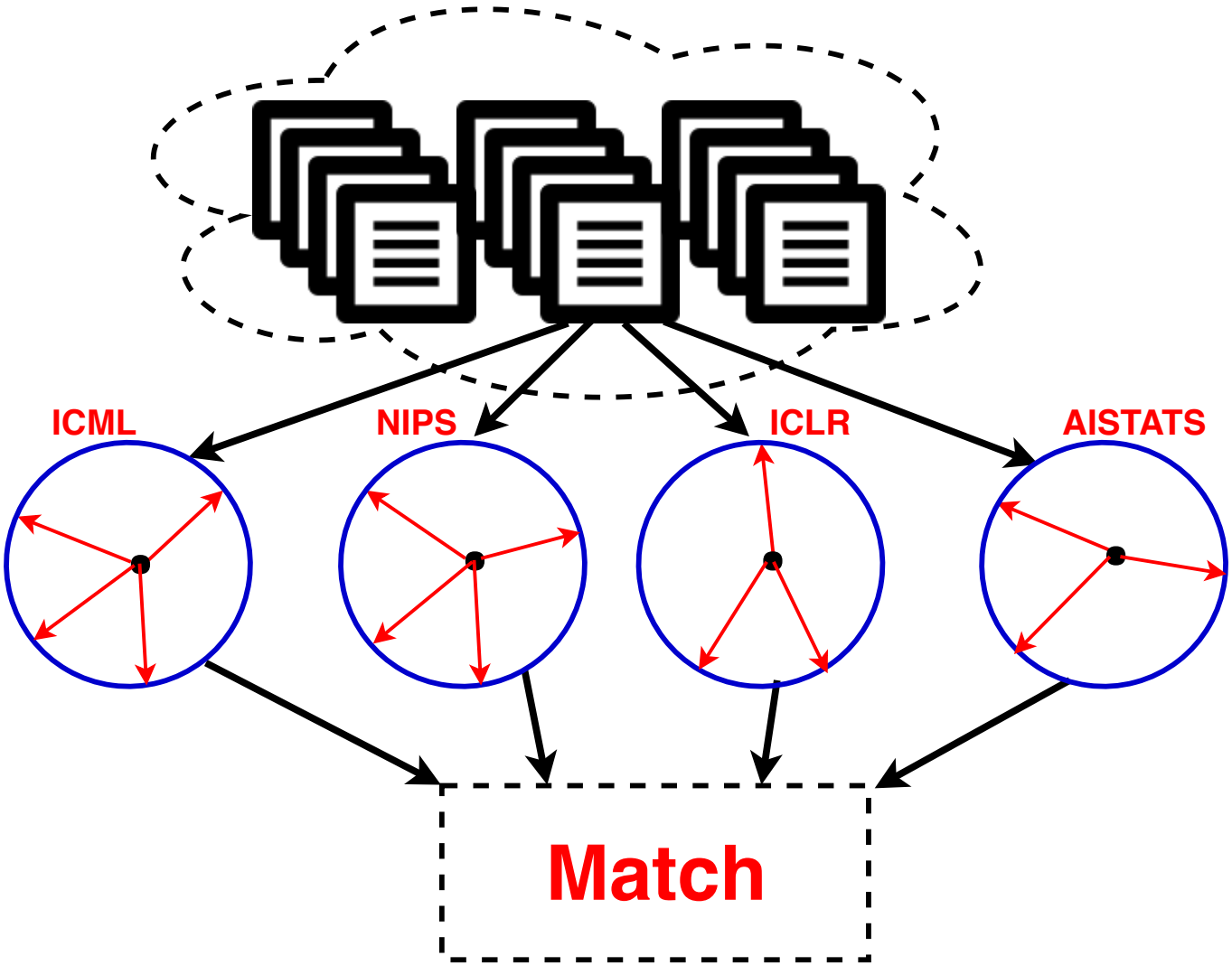
Distributed Matching



Distributed Matching



Distributed Matching



Streaming Dynamic Distributed Matching

Key ideas:

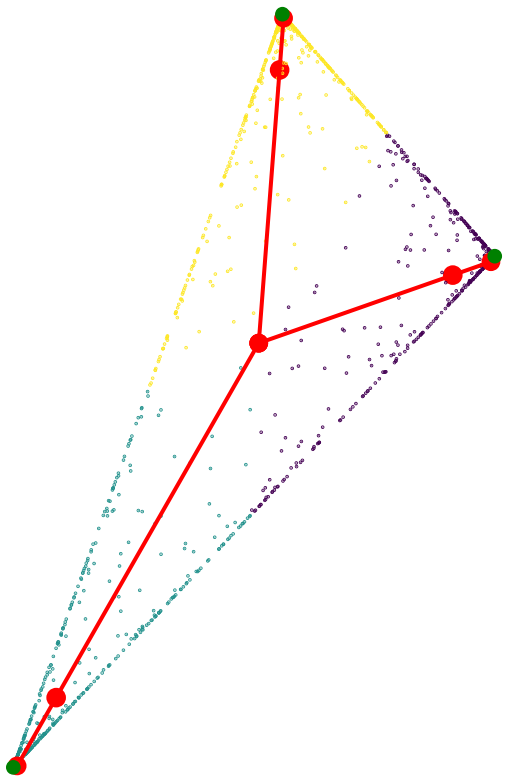
- Model with Dynamic Hierarchical Beta Process
- Estimate using Hungarian algorithm fixing all but one group
- Iterate in coordinate descent style

Modeling 3 million Wikipedia articles

	Perplexity	Time	Topics	Cores used
SDM	1236	35min	182	20
DM	1260	14min	183	20
SDDM	**1228**	**12min**	184	20

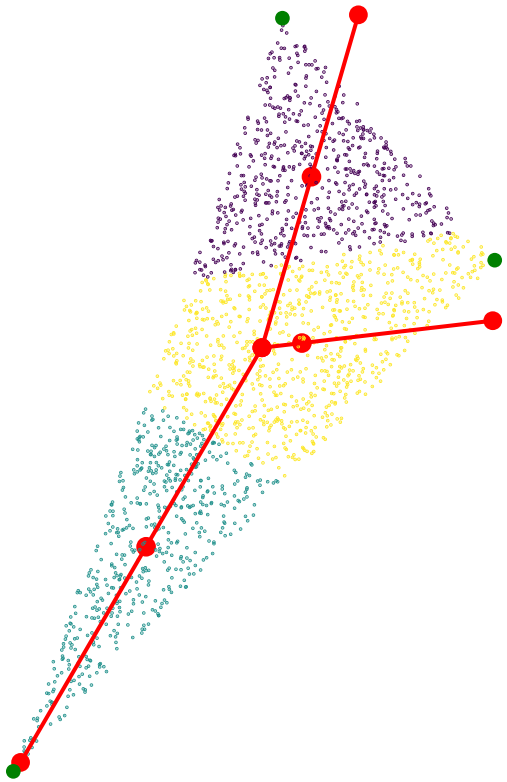
Generalized GDM

GDM: small α *or* equilateral simplex



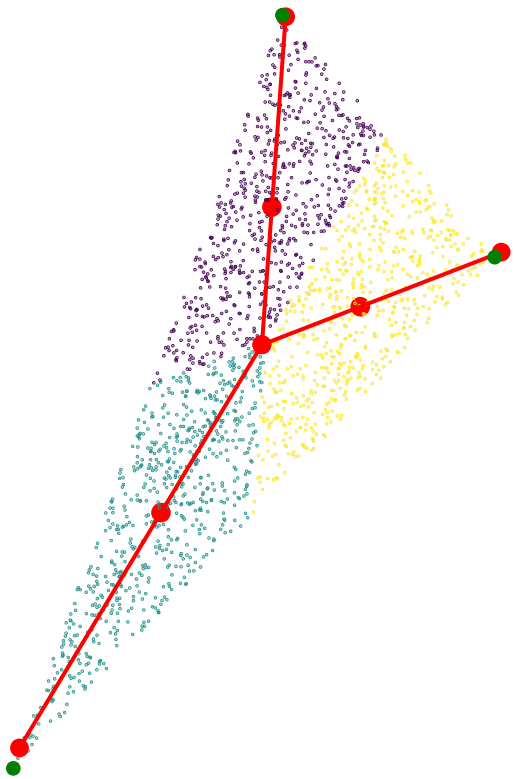
Generalized GDM

GDM assumption violation: $\alpha = 1$ *and* skewed simplex



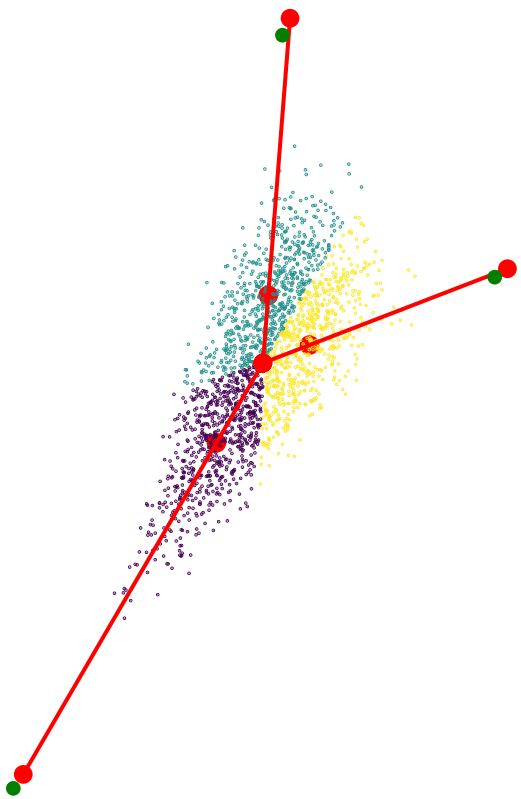
Generalized GDM

New approach: *any* α and *any* simplex



Generalized GDM

New approach: $\alpha = 5$



Some interesting statistics:

Some interesting statistics:

- This presentation has 54 images
- Among them, 31 triangle and 23 circles
- Thousands of triangles has been drawn inside West Hall over the past 4 years

THANK YOU!

References

Arora, Sanjeev, Ge, Rong, Halpern, Yonatan, Mimno, David, Moitra, Ankur, Sontag, David, Wu, Yichen, and Zhu, Michael. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 280–288, 2013.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022, March 2003.

Griffiths, Thomas L and Ghahramani, Zoubin. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

Himmelstein, Daniel S, Greene, Casey S, and Moore, Jason H. Evolving hard problems: generating human genetics datasets with a complex etiology. *BioData mining*, 4(1):1, 2011.

Ho, N., Nguyen, X., Yurochkin, M., Bui, H., Huynh, V., and Phung, D. Multilevel clustering via Wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Rendle, Steffen. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 995–1000. IEEE, 2010.

Thibaux, Romain and Jordan, Michael I. Hierarchical Beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pp. 564–571, 2007.

Yurochkin, M., Guha, A., and Nguyen, X. Conic Scan-and-Cover algorithms for nonparametric topic modeling. In *Advances in Neural Information Processing Systems*, pp. 3881–3890, 2017a.

Yurochkin, M., Nguyen, X., and Vasiloglou, N. Multi-way interacting regression via factorization machines. In *Advances in Neural Information Processing Systems*, pp. 2595–2603, 2017b.

Yurochkin, M., Fan, Z., Guha, A., Nguyen, X., and Koutris, P. Streaming dynamic and distributed inference of latent geometric structures. In *Under revision in ICML 2018*, 2018a.

Yurochkin, M., Thai, D., Bui, H., and Nguyen, X. UPS: optimizing undirected positive sparse graph for neural graph filtering. In *Unpublished*, 2018b.

Yurochkin, Mikhail and Nguyen, XuanLong. Geometric Dirichlet Means Algorithm for topic inference. In *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2016.

Regression with Interactions

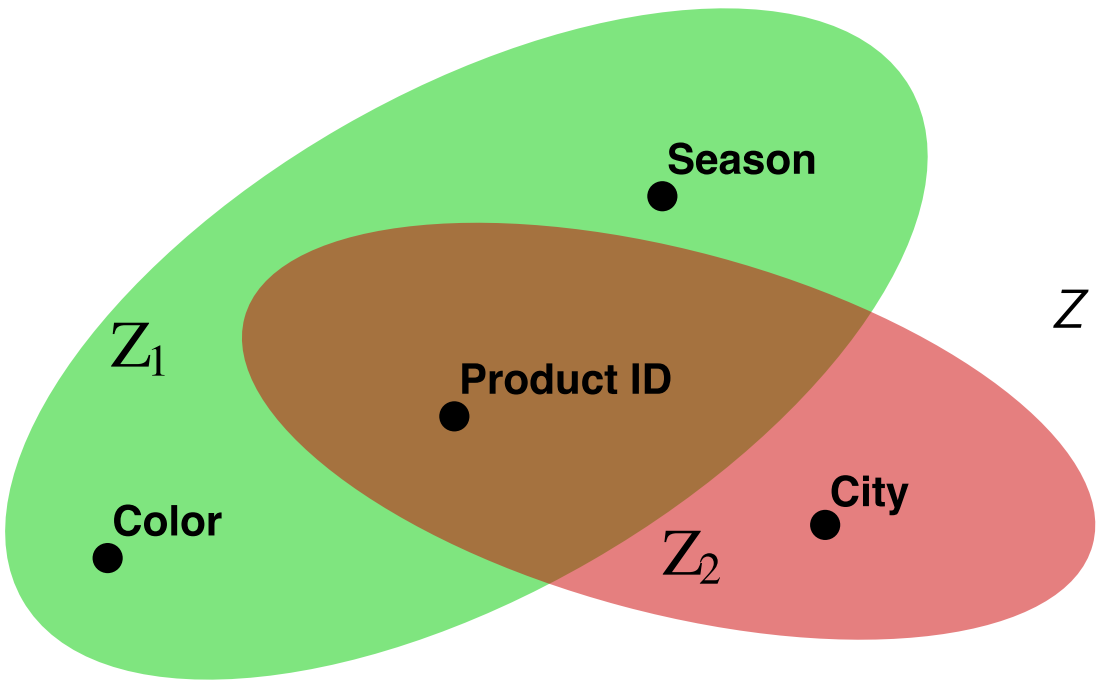
$$\mathbb{E}(Y|x) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{j=1}^{J} \beta_j \prod_{i \in Z_j} x_i.$$

- $w_0, \ldots, w_D$ are bias and linear coefficients
- $\beta_1, \ldots, \beta_J$ are coefficients of the J interactions
- $Z_1, \ldots, Z_J$ are sets of indices of interacting variables

Challenges:

- $2^D - D - 1$ of possible interactions - infeasible even for $D = 30$
- how to model coefficients β_js for all interactions

Hypergraph of Interactions example



$$Z = \left[\begin{array}{c|cc} & Z_1 & Z_2 \\ \hline \text{Season} & 1 & 0 \\ \text{Product ID} & 1 & 1 \\ \text{Color} & 1 & 0 \\ \text{City} & 0 & 1 \end{array}\right]$$

Hypergraph of Interactions

We model collection of interactions $Z_1, \ldots, Z_J$ as a hypergraph.

Definition

Let $S = \{e_1, \ldots, e_D\}$ be a set of D objects and $Z = \{Z_1, \ldots, Z_J\}$ set of J subsets of S: $Z_j \subset S$, for $j = 1, \ldots, J$. Then we say that $G = (S, Z)$ is a hypergraph with D vertices and J hyperedges.

- let $S = \{1, \ldots, D\}$, then $G = (S, Z)$ is a hypergraph of interactions
- Z - incidence matrix of interactions: $Z \in \{0, 1\}^{D \times J}$, where $Z_{i_1 j} = Z_{i_2 j} = 1$ iff i_1 and i_2 are part of a hyperedge indexed by column/interaction j

MiFM: Multi-way interacting Factorization Machine

Factorization Machines (FM) (Rendle, 2010)

- $\mathbb{E}(Y|x) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{1 \le i < j \le D} \beta_{i,j} x_i x_j$
- factorize interaction weights:
  $\beta_{i,j} := \sum_{k=1}^{K} v_{ik} v_{jk}$, where $V \in \mathbb{R}^{D \times K}$ and $K \ll D$

Combining with Hypergraph of Interactions Z:

$$\text{MiFM: } \hat{y} := w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{j=1}^{J} \sum_{k=1}^{K} \prod_{i \in Z_j} x_i v_{ik}$$

Real Data Application: Retail

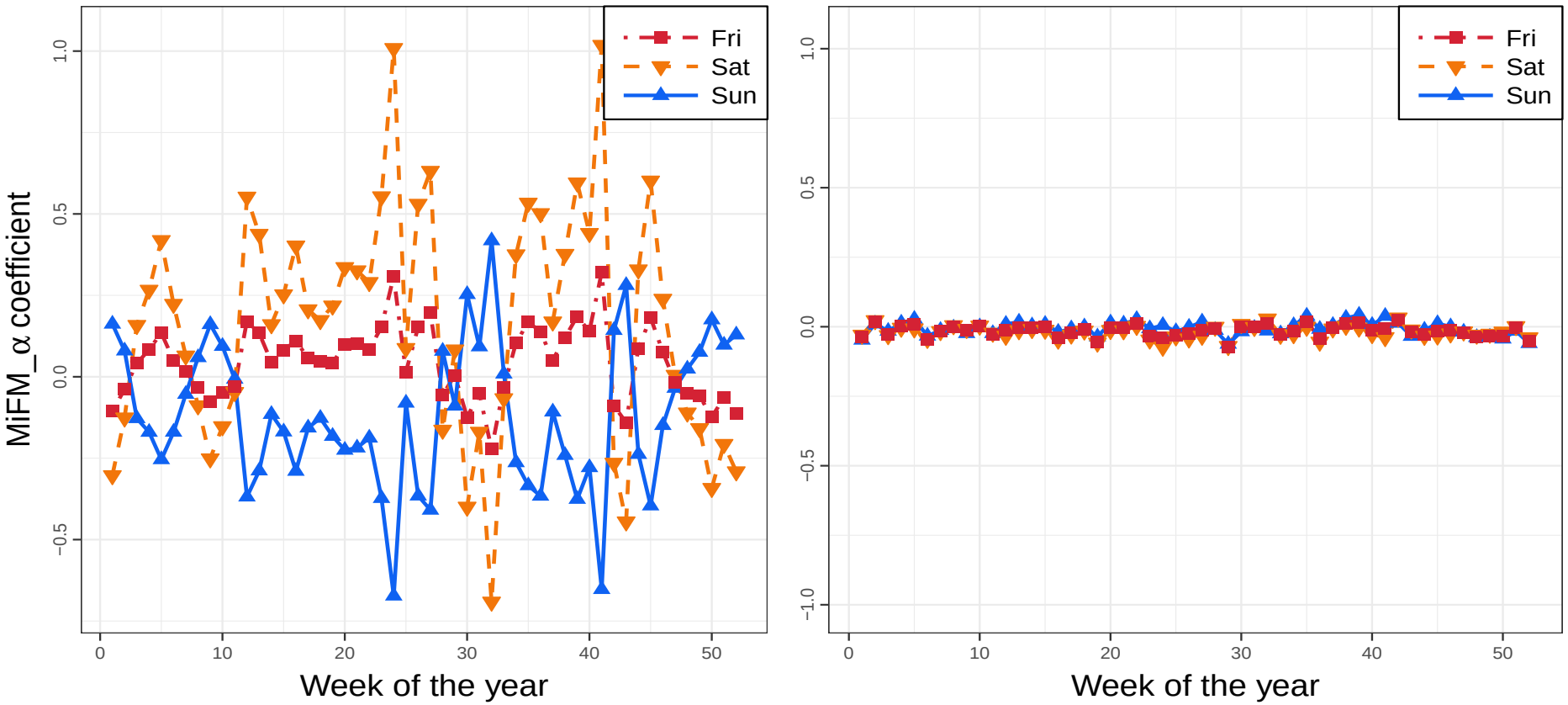


Figure: Coefficients for city - store ID - day of week - week of year interaction: (left) store in Merignac; (right) store in Perols

Real Data Application: Genetics

Finding interactions between genes (i.e. epistasis) based on the data from Himmelstein et al. (2011).

	5-way	5-way no low
MiFM$_1$	0.775	**0.649**
MiFM$_\alpha$	0.771	**0.645**
FM	0.501	0.500
Logistic MiFM$_1$	**0.883**	0.628
Logistic MiFM$_\alpha$	**0.860**	0.623
Logistic	0.460	0.461
MLP	**0.870**	0.625
SVM	0.473	0.451
RF	**0.887**	0.628

THANK YOU!

References

Arora, Sanjeev, Ge, Rong, Halpern, Yonatan, Mimno, David, Moitra, Ankur, Sontag, David, Wu, Yichen, and Zhu, Michael. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 280–288, 2013.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022, March 2003.

Griffiths, Thomas L and Ghahramani, Zoubin. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

Himmelstein, Daniel S, Greene, Casey S, and Moore, Jason H. Evolving hard problems: generating human genetics datasets with a complex etiology. *BioData mining*, 4(1):1, 2011.

Ho, N., Nguyen, X., Yurochkin, M., Bui, H., Huynh, V., and Phung, D. Multilevel clustering via Wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Rendle, Steffen. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 995–1000. IEEE, 2010.

Thibaux, Romain and Jordan, Michael I. Hierarchical Beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pp. 564–571, 2007.

Yurochkin, M., Guha, A., and Nguyen, X. Conic Scan-and-Cover algorithms for nonparametric topic modeling. In *Advances in Neural Information Processing Systems*, pp. 3881–3890, 2017a.

Yurochkin, M., Nguyen, X., and Vasiloglou, N. Multi-way interacting regression via factorization machines. In *Advances in Neural Information Processing Systems*, pp. 2595–2603, 2017b.

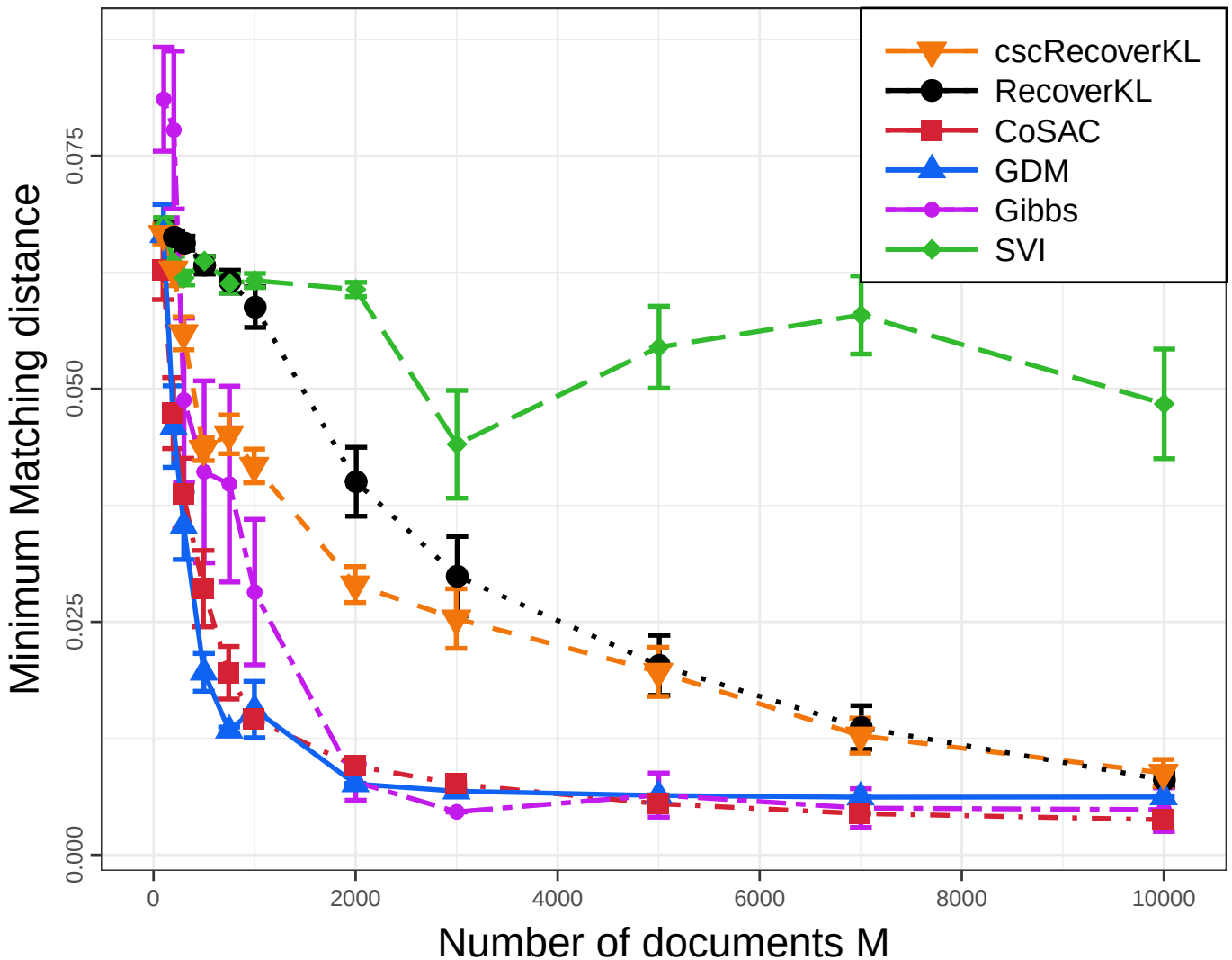
Yurochkin, M., Fan, Z., Guha, A., Nguyen, X., and Koutris, P. Streaming dynamic and distributed inference of latent geometric structures. In *Under revision in ICML 2018*, 2018a.

Yurochkin, M., Thai, D., Bui, H., and Nguyen, X. UPS: optimizing undirected positive sparse graph for neural graph filtering. In *Unpublished*, 2018b.

Yurochkin, Mikhail and Nguyen, XuanLong. Geometric Dirichlet Means Algorithm for topic inference. In *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2016.

Simulations: corpora size

$\alpha = 0.1,\ \eta = 0.1,\ V = 2000,\ K = 15,\ N_m = 500.$



Simulations: run-time