

# Statistical Translation, Heat Kernels, and Expected Distances

Joshua Dillon, Yi Mao, Guy Lebanon, Jian Zhang

Purdue University – West Lafayette, IN

# Outline

- 1 Outline
- 2 Motivation
- 3 Statistical Translation
- 4 Expected Distance
- 5 Large Deviation Interpretation
- 6 Experiments

# Motivation

## Traditional modeling of documents

- assume documents  $x \sim \text{Mult}(\theta_x^{\text{true}})$
- unknown  $\theta_x^{\text{true}}$  typically estimated by maximum likelihood (bow/tf)  $[\hat{\theta}_x^{\text{mle}}]_k = N^{-1} \sum_{i=1}^N \delta_{k,x_i}$
- Estimator  $\hat{\theta}^{\text{mle}}$  needs to be smoothed to reduce variance

Observation: smoothing  $\hat{\theta}^{\text{mle}}$  based on word correlation results in a new metric structure.

Example: documents containing UAI should contain also machine learning – even if they don't.

# Motivation

## Traditional modeling of documents

- assume documents  $x \sim \text{Mult}(\theta_x^{\text{true}})$
- unknown  $\theta_x^{\text{true}}$  typically estimated by maximum likelihood (bow/tf)  $[\hat{\theta}_x^{\text{mle}}]_k = N^{-1} \sum_{i=1}^N \delta_{k,x_i}$
- Estimator  $\hat{\theta}^{\text{mle}}$  needs to be smoothed to reduce variance

Observation: smoothing  $\hat{\theta}^{\text{mle}}$  based on word correlation results in a new metric structure.

Example: documents containing **UAI** should contain also **machine learning** – even if they don't.

## A related example: query expansion

Query expansion intends to solve the following type of problem:

- user submits the query term `uncertainty`
- standard retrieval: documents without `uncertainty` but with `probability` will not be retrieved
- query expansion: query is augmented with related terms e.g., `probability` and then documents are retrieved

Reduces query/document mismatch by expanding the query using words or phrases with a similar meaning or we obtain a new distance/geometry based on expansion/translation

## A related example: query expansion

Query expansion intends to solve the following type of problem:

- user submits the query term `uncertainty`
- standard retrieval: documents without `uncertainty` but with `probability` will not be retrieved
- query expansion: query is augmented with related terms e.g., `probability` and then documents are retrieved

Reduces query/document mismatch by expanding the query using words or phrases with a similar meaning or we obtain a new distance/geometry based on expansion/translation

# Statistical translation for document modeling

$$X \xrightarrow{p} Y$$

- document  $x$  translated to  $y$  with probability  $p$

$$\hat{\theta}_x^{\text{mle}} \xrightarrow{p} \hat{\theta}_y^{\text{mle}}$$

- bow  $\hat{\theta}_x^{\text{mle}}$  representation for  $x$  mapped to the **random variable**  $\hat{\theta}_y^{\text{mle}}$

# Statistical translation for document modeling

$$X \xrightarrow{p} Y$$

- document  $x$  translated to  $y$  with probability  $p$

$$\hat{\theta}_x^{\text{mle}} \xrightarrow{p} \hat{\theta}_y^{\text{mle}}$$

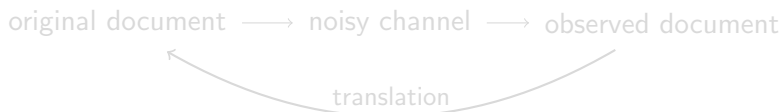
- bow  $\hat{\theta}_x^{\text{mle}}$  representation for  $x$  mapped to the **random variable**  $\hat{\theta}_y^{\text{mle}}$

# Interpretations of the model

## Regularization

- $\hat{\theta}_x^{\text{mle}}$  unbiased, high variance estimator of  $\theta_x^{\text{true}}$
- $\hat{\theta}_y^{\text{mle}}$  is slightly biased, lower variance estimator of  $\theta_x^{\text{true}}$
- Analogy: ridge and lasso regression, regularization

## De-noising

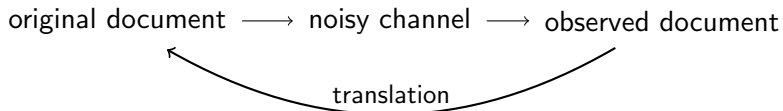


# Interpretations of the model

## Regularization

- $\hat{\theta}_x^{\text{mle}}$  unbiased, high variance estimator of  $\theta_x^{\text{true}}$
- $\hat{\theta}_y^{\text{mle}}$  is slightly biased, lower variance estimator of  $\theta_x^{\text{true}}$
- Analogy: ridge and lasso regression, regularization

## De-noising



# Assumption about document translation

- translation  $x \mapsto y$  is done word by word independently

Problem: estimate word translation model  $T$

$$T_{ij} = P(w_i \rightarrow w_j)$$

- utilize large external corpus
- can be done in an unsupervised manner

# Assumption about document translation

- translation  $x \mapsto y$  is done word by word independently

Problem: estimate word translation model  $T$

$$T_{ij} = P(w_i \rightarrow w_j)$$

- utilize large external corpus
- can be done in an unsupervised manner

# Estimating $T_{ij} = P(w_i \rightarrow w_j)$

General approach: diffusion kernel  $K_t(q_u, q_v)$  on graph  $(V, E)$  whose nodes are distributions that correspond to words

- $V$ : each vertex is a contextual distribution  $q_v(w) = P(w|v)$  corresponding to a word  $v$
- $E$ : graph edge weights are the Fisher diffusion kernel on multinomial simplex
- $T$  is from diffusion kernel on  $(V, E)$

# Estimating $T_{ij} = P(w_i \rightarrow w_j)$

General approach: diffusion kernel  $K_t(q_u, q_v)$  on graph  $(V, E)$  whose nodes are distributions that correspond to words

- $V$ : each vertex is a contextual distribution  $q_v(w) = P(w|v)$  corresponding to a word  $v$

$$\hat{q}_v(w) \propto \sum_d \text{tf}(w, d) \text{tf}(v, d)$$

- $E$ : graph edge weights are the Fisher diffusion kernel on multinomial simplex
- $T$  is from diffusion kernel on  $(V, E)$

# Estimating $T_{ij} = P(w_i \rightarrow w_j)$

General approach: diffusion kernel  $K_t(q_u, q_v)$  on graph  $(V, E)$  whose nodes are distributions that correspond to words

- $V$ : each vertex is a contextual distribution  $q_v(w) = P(w|v)$  corresponding to a word  $v$
- $E$ : graph edge weights are the Fisher diffusion kernel on multinomial simplex

$$e(u, v) = \exp \left( -\frac{1}{t} \arccos^2 \left( \sum_w \sqrt{q_u(w)q_v(w)} \right) \right)$$

- $T$  is from diffusion kernel on  $(V, E)$

# Estimating $T_{ij} = P(w_i \rightarrow w_j)$

General approach: diffusion kernel  $K_t(q_u, q_v)$  on graph  $(V, E)$  whose nodes are distributions that correspond to words

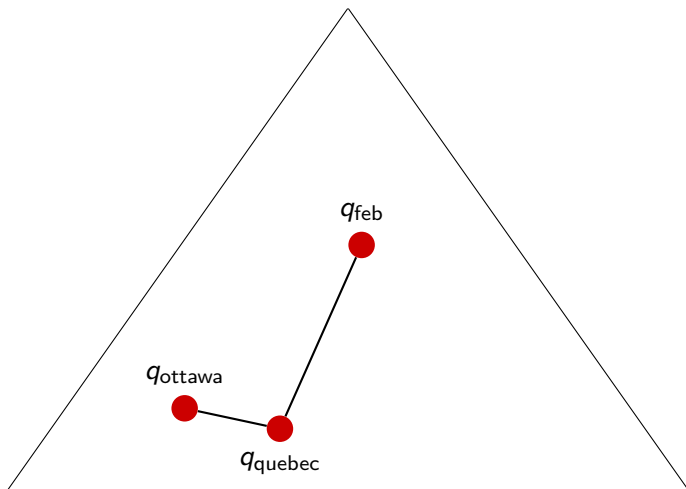
- $V$ : each vertex is a contextual distribution  $q_v(w) = P(w|v)$  corresponding to a word  $v$
- $E$ : graph edge weights are the Fisher diffusion kernel on multinomial simplex
- $T$  is from diffusion kernel on  $(V, E)$

$$T \propto \exp(-t\mathcal{L})$$

where  $\mathcal{L}$  is the normalized Laplacian

- $t$  controls the amount of translation
- $\lim_{t \rightarrow 0} T = I$  and  $\lim_{t \rightarrow \infty} T = \text{uniform}$

# Simplex



# Word translation result

jan	databas	nbc	wang	ottawa
feb	intranet	abc	chen	quebec
nov	server	cnn	liu	montreal
dec	softwar	hollywood	beij	toronto
oct	internet	tv	wu	ontario
aug	netscap	viewer	china	vancouv
apr	onlin	movi	chines	canada
mar	web	audienc	peng	canadian
sep	browser	fox	hui	calgari

# Expected Distance

Two documents  $x, w$  stochastically translate into documents  $y, z$  and are represented by bow random variables  $\hat{\theta}_y^{\text{mle}}, \hat{\theta}_z^{\text{mle}}$ .

Distance  $d(\hat{\theta}_y^{\text{mle}}, \hat{\theta}_z^{\text{mle}})$  is a random variable, summarized by its expectation (given in closed form)

$$\begin{aligned} E_{p(y|x)p(z|w)} \|\hat{\theta}_y^{\text{mle}} - \hat{\theta}_z^{\text{mle}}\|_2^2 &= N_1^{-2} \sum_{i=1}^{N_1} \sum_{j \in \{1, \dots, N_1\} \setminus \{i\}} (TT^\top)_{x_i, x_j} \\ &+ N_2^{-2} \sum_{i=1}^{N_2} \sum_{j \in \{1, \dots, N_2\} \setminus \{i\}} (TT^\top)_{w_i, w_j} \\ &- 2N_1^{-1}N_2^{-2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} (TT^\top)_{x_i, w_j} + N_1^{-1} + N_2^{-1}. \end{aligned}$$

# Expected Distance

- If  $T = I$ ,  $E_{p(y|x)p(z|w)} \|\hat{\theta}_y^{\text{mle}} - \hat{\theta}_z^{\text{mle}}\|_2^2 = \|\hat{\theta}_x^{\text{mle}} - \hat{\theta}_w^{\text{mle}}\|_2^2$ 
  - The distance remains the same under permutation of the words within a document
- Pre-compute  $TT^\top$  to speed up the distance computation
- Unsupervised metric learning

# Large Deviation Interpretation

By the Chernoff-Stein lemma, KL-divergence is the best exponent in the probability of type II error (and bounded type I error), i.e.,

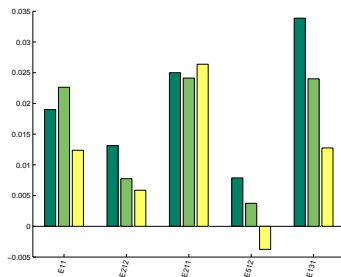
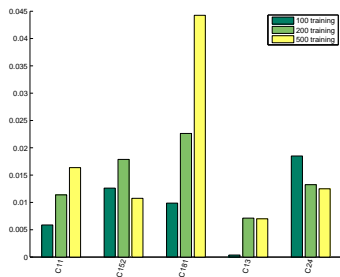
$$\beta_n^{\text{opt}} \approx \exp(-\gamma n D(q_u || q_v)).$$

Examining the Taylor series expansion of KL-divergence for nearby  $q_u, q_v$ , one also finds that for the Fisher geodesic distance,  $d(p, q)$ ,

$$d^2(q_u, q_v) \approx 2D(q_u || q_v).$$

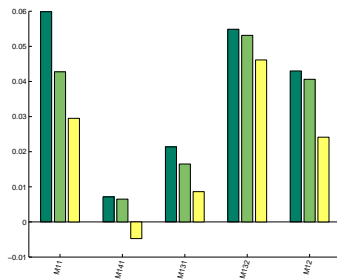
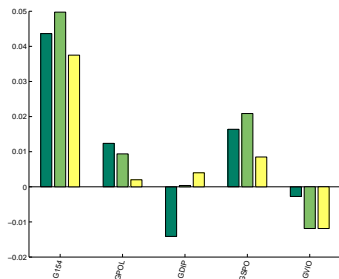
Thus one may interpret the heat kernel translation model as being based on a graph whose edge weights approximate the optimal error rate between a test of  $Q = q_u$  vs.  $Q = q_v$ .

# RCV1 Document classification results



Average error reduction of expected  $L_2$  from  $L_2$  over 40 realizations of balanced training and testing documents for 20 RCV1 topic categories with a nearest neighbor classifier.

# RCV1 Document classification results



Average error reduction of expected  $L_2$  from  $L_2$  over 40 realizations of balanced training and testing documents for 20 RCV1 topic categories with a nearest neighbor classifier.

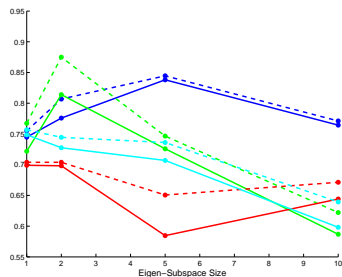
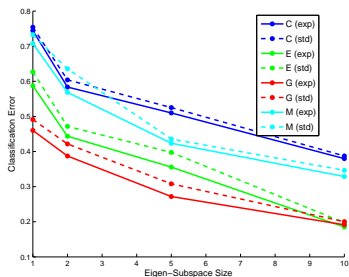
## RCV1 Document PCA (topics)

task	topics from RCV1
C	C152 C181 C13 C24 C21
E	E212 E211 E512 E131 E11
M	M11 M12 M131 M132 M141 M142 M143
G	GPOL GDIP GSPO GVIO GCRIM

Topics from RCV1 that are involved in the kernel PCA experiments.

# RCV1 Document PCA (classification error)

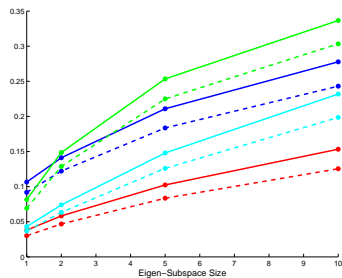
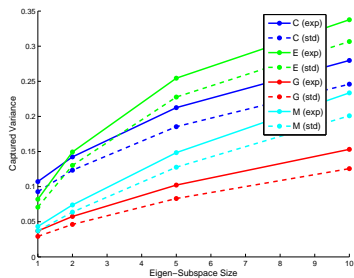
Classification error rate of Fishers linear discriminant when documents are first projected to the  $k$ -dimensional eigen-subspace.



Expected/standard linear kernels (left) and expected/standard RBF kernels (right).

# RCV1 Document PCA (captured variance)

The percentage of the overall variance captured by the first  $k$  eigenvalues for training data.



Expected/standard linear kernels (left) and expected/standard RBF kernels (right).

# Conclusion

- translation-based estimate for multinomial parameters results in a new random geometry
- learned geometry realizes bias-variance trade-off in direct analogy with ridge regression, lasso and regularization
- diffusion kernel on a graph embedded in the Fisher simplex
- expected distance has closed form
- works for document classification

## Related Work

- Baker and McCallum, 1993.
- Berger and Lafferty, 1999.
- Smola and Kondor, 2003.
- Lafferty and Lebanon, 2005.
- Lafferty and Zhai, 2001.
- Collins-Thompson and Callan, 2005.