Statistical Translation, Heat Kernels, and Expected Distances

Joshua Dillon[†] Yi Mao[†] Guy Lebanon[‡†] Jian Zhang[‡]

[†]School of Electrical and Computer Engineering

[‡]Department of Statistics

Purdue University - West Lafayette

Motivation

Traditional modeling of documents

- assume documents $x \sim \text{Mult}(\theta_x^{\text{true}})$
- unknown θ_x^{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

Motivation

Traditional modeling of documents

- assume documents $x \sim \text{Mult}(\theta_x^{\text{true}})$
- unknown θ_x^{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 $\boxed{\text{NIPS}}$ should contain also $\boxed{\text{machine learning}}$ - 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 $\boxed{\text{metric}}$
- standard retrieval: documents without $\boxed{\text{metric}}$ but with $\boxed{\text{distance}}$ will not be retrieved
- query expansion: query is augmented with related terms e.g., $\boxed{\text{distance}}$ and then documents are retrieved

A related example: query expansion

Query expansion intends to solve the following type of problem:

- user submits the query term $\boxed{\text{metric}}$
- standard retrieval: documents without $\boxed{\text{metric}}$ but with $\boxed{\text{distance}}$ will not be retrieved
- query expansion: query is augmented with related terms e.g., $\boxed{\text{distance}}$ 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

$$x \xrightarrow{\mathsf{P}} y$$

- document x translated to y with probability P

Statistical translation for document modeling

$$x \xrightarrow{\text{P}} y \qquad\qquad \hat{\theta}_x^{\text{mle}} \xrightarrow{\text{P}} \hat{\theta}_y^{\text{mle}}$$

- document x translated to y with probability P
- bow $\hat{\theta}_x^{\text{mle}}$ representation for x mapped to the random variable $\hat{\theta}_y^{\text{mle}}$

Interpretations of the model

Regularization

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

Interpretations of the model

Regularization

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

Denoising

$$\begin{tikzcd} \text{original document} \arrow[r] & \text{noisy channel} \arrow[r] & \text{observed document} \arrow[ll, bend left, "\text{translation}"] \end{tikzcd}$$

Assumption about document translation

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

Assumption about document translation

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

Problem: estimate word translation model T

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

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

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

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

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

G eneral 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

$$\hat{q}_v(w) \propto \sum_{d: v \in d} \mathrm{tf}(w, d)$$

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

G eneral 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

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

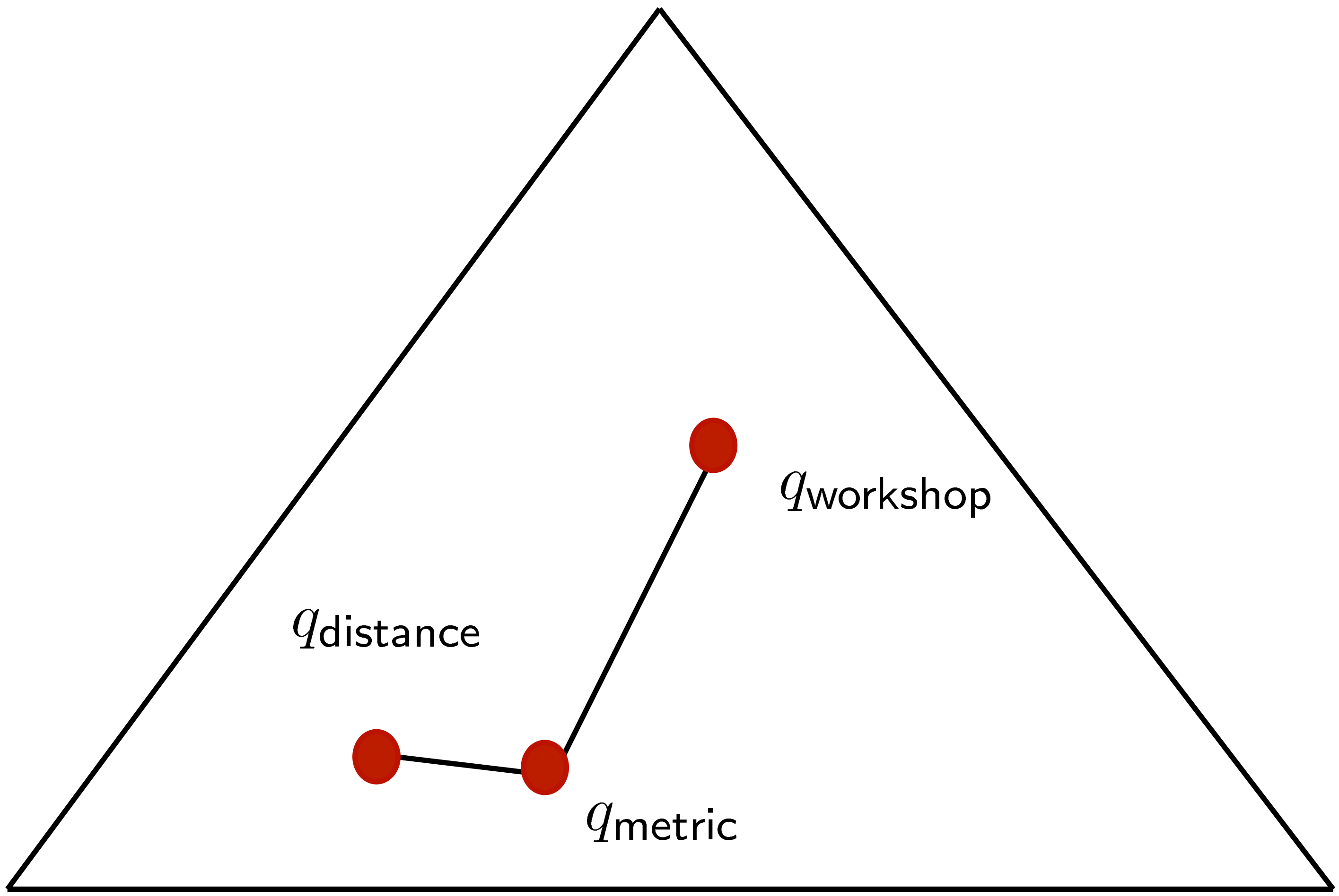
Estimating $T_{ij} = P(w_i \to w_j)$

G eneral 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
  - small $t \to T \approx I$, large $t \to$ approximately uniform T


q_{workshop}
q_{distance}
q_{metric}

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

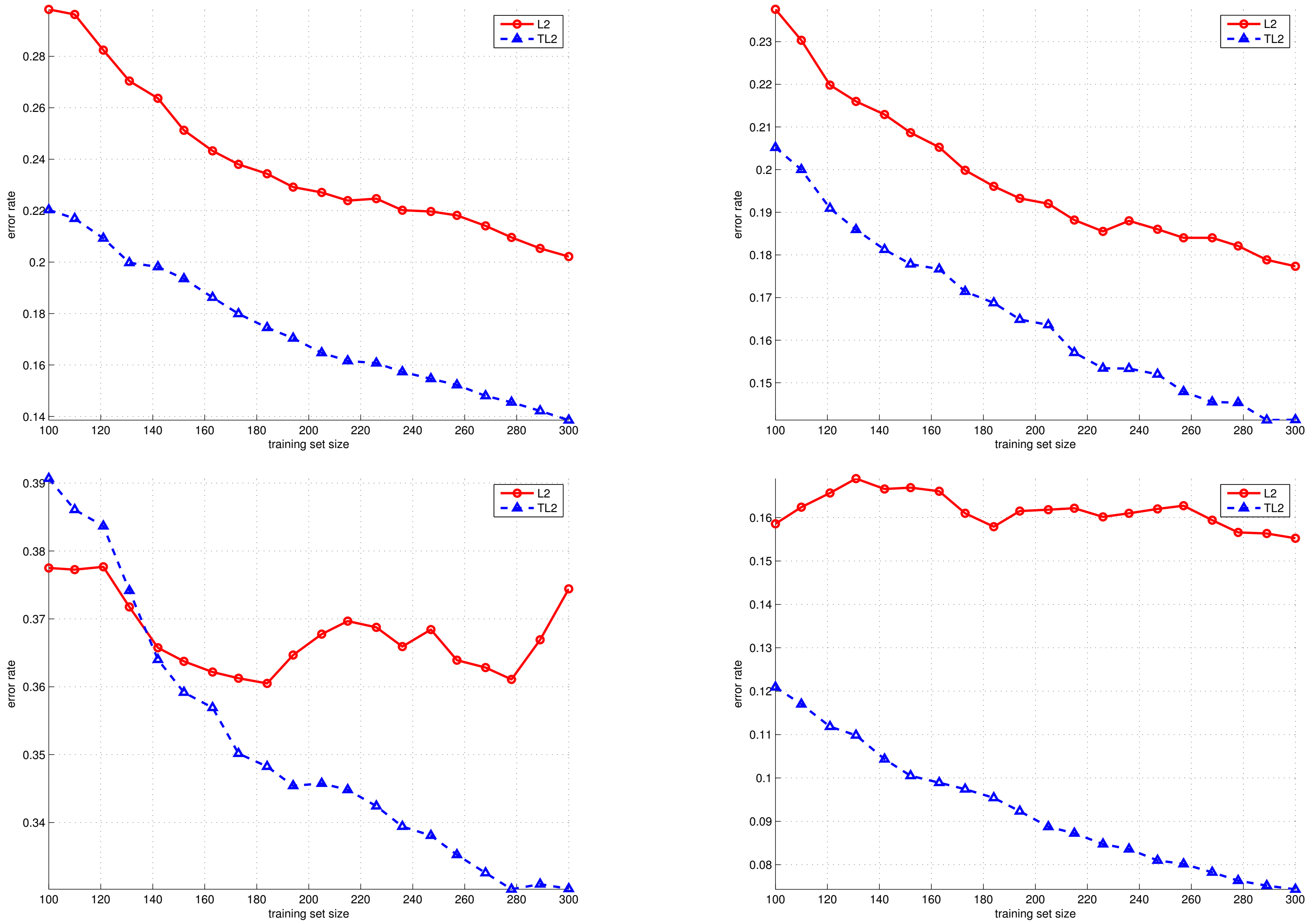
Two documents x, w stochastically translate into documents y, z and are represented by bow random variables $\hat{\theta}_y^{\mathsf{mle}}, \hat{\theta}_z^{\mathsf{mle}}$. Distance $d(\hat{\theta}_y^{\mathsf{mle}}, \hat{\theta}_z^{\mathsf{mle}})$ is a random variable, summarized by its expectation (given in closed form)

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

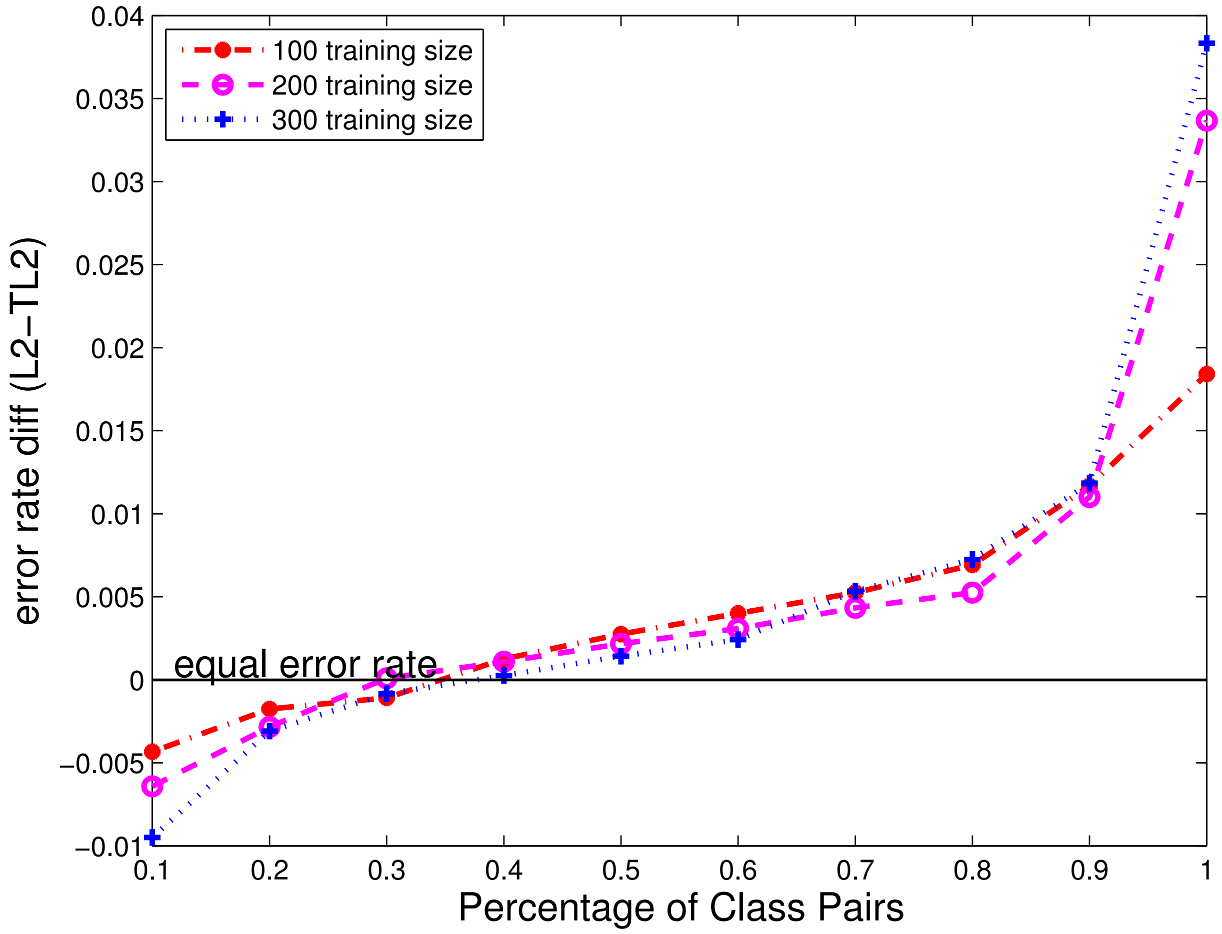
Expected Distance

- 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

RCV1 Document classification results



RCV1 Document classification results



Conclusion

- translation-based estimate for multinomial parameters results in a new random geometry
- learned geometry realizes bias-variance tradeoff 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

- Distributional clustering of english words, 1993
- Diffusion kernels on statistical manifolds, 2005
- Kernels and regularization on graphs, 2003
- Spectral graph theory, 1997
- Query expansion using random walk models, 2005
- Information retrieval as statistical translation, 1999