Supplemental Material for "A Practical Algorithm for Topic Modeling with Provable Guarantees"

Sanjeev Arora ARORA@CS.PRINCETON.EDU
Rong Ge RONGGE@CS.PRINCETON.EDU
Yoni Halpern HALPERN@CS.NYU.EDU
David Mimno MIMNO@CS.PRINCETON.EDU
Ankur Moitra MOITRA@IAS.EDU
David Sontag DSONTAG@CS.NYU.EDU
Yichen Wu YICHENWU@PRINCETON.EDU
Michael Zhu MHZHU@PRINCETON.EDU

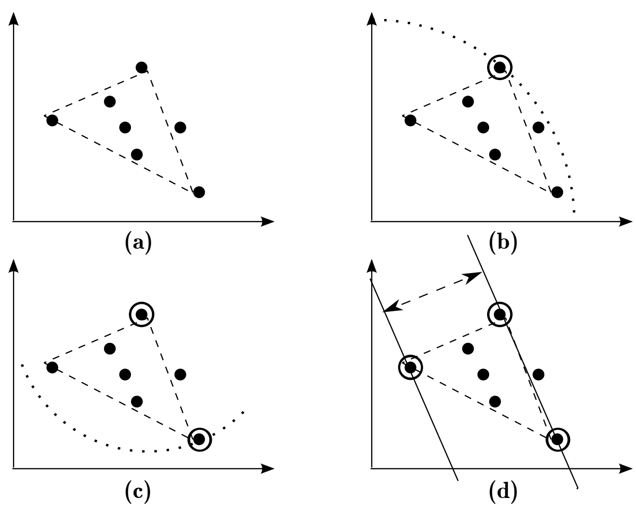

Figure 1. Illustration of the Algorithm

1. Proof for Anchor-Words Finding Algorithm

Recall that the correctness of the algorithm depends on the following Lemma:

Lemma 1.1. *The point d_j found by the algorithm must be $\delta = O(\epsilon/\gamma^2)$ close to some vertex v_i. In particular, the corresponding a_j $O(\epsilon/\gamma^2)$-covers v_i.*

In order to prove this Lemma, we first show that even if previously found vertices are only δ close to some vertices, there is still another vertex that is far from the span of previously found vertices.

Lemma 1.2. *Suppose all previously found vertices are $O(\epsilon/\gamma^2)$ close to distinct vertices, there is a vertex v_i whose distance from* $\mathrm{span}(S)$ *is at least $\gamma/2$.*

In order to prove Lemma 1.2, we use a volume argument. First we show that the volume of a robust simplex cannot change by too much when the vertices are perturbed.

Lemma 1.3. *Suppose $\{v_1, v_2, ..., v_K\}$ are the vertices of a γ-robust simplex S. Let S' be a simplex with vertices $\{v_1', v_2', ..., v_K'\}$, each of the vertices v_i' is a perturbation of v_i and $\|v_i' - v_i\|_2 \le \delta$. When $10\sqrt{K}\delta < \gamma$ the volume of the two simplices satisfy*

$$vol(S)(1-2\delta/\gamma)^{K-1} \le vol(S') \le vol(S)(1+4\delta/\gamma)^{K-1}.$$

Proof: As the volume of a simplex is proportional to the determinant of a matrix whose columns are the edges of the simplex, we first show the following perturbation bound for determinant.

Claim 1.4. *Let A, E be $K \times K$ matrices, the smallest eigenvalue of A is at least γ, the Frobenius norm $\|E\|_F \le \sqrt{K}\delta$, when $\gamma > 5\sqrt{K}\delta$ we have*

$$\det(A+E)/\det(A) \ge (1-\delta/\gamma)^K.$$

Proof: Since $\det(AB) = \det(A)\det(B)$, we can multiply both A and $A+E$ by A^{-1}. Hence $\det(A+E)/\det(A) = \det(I + A^{-1}E)$.

The Frobenius norm of $A^{-1}E$ is bounded by

$$\left\|A^{-1}E\right\|_F \le \left\|A^{-1}\right\|_2 \|E\|_F \le \sqrt{K}\delta/\gamma.$$

Let the eigenvalues of $A^{-1}E$ be $\lambda_1, \lambda_2, ..., \lambda_K$, then by definition of Frobenius Norm $\sum_{i=1}^K \lambda_i^2 \le \left\|A^{-1}E\right\|_F^2 \le K\delta^2/\gamma^2$.

The eigenvalues of $I + A^{-1}E$ are just $1 + \lambda_1, 1 + \lambda_2, ..., 1 + \lambda_K$, and the determinant $\det(I + A^{-1}E) = \prod_{i=1}^{K}(1 + \lambda_i)$. Hence it suffices to show

$$\min \prod_{i=1}^{K}(1 + \lambda_i) \geq (1 - \delta/\gamma)^K \text{ when } \sum_{i=1}^{K} \lambda_i^2 \leq K\delta^2/\gamma^2.$$

To do this we apply Lagrangian method and show the minimum is only obtained when all λ_i's are equal. The optimal value must be obtained at a local optimum of

$$\prod_{i=1}^{K}(1 + \lambda_i) + C\sum_{i=1}^{K}\lambda_i^2.$$

Taking partial derivatives with respect to λ_i's, we get the equations $-\lambda_i(1 + \lambda_i) = -\prod_{i=1}^{K}(1 + \lambda_i)/2C$ (here using $\sqrt{K}\delta/\gamma$ is small so $1 + \lambda_i > 1/2 > 0$). The right hand side is a constant, so each λ_i must be one of the two solutions of this equation. However, only one of the solution is larger than $1/2$, therefore all the λ_i's are equal.

■

For the lower bound, we can project the perturbed subspace to the $K - 1$ dimensional space. Such a projection cannot increase the volume and the perturbation distances only get smaller. Therefore we can apply the claim directly, the columns of A are just $v_{i+1} - v_1$ for $i = 1, 2, ..., K - 1$; columns of E are just $v'_{i+1} - v_{i+1} - (v'_1 - v_1)$. The smallest eigenvalue of A is at least γ because the polytope is γ robust, which is equivalent to saying after orthogonalization each column still has length at least γ. The Frobenius norm of E is at most $2\sqrt{K-1}\delta$. We get the lower bound directly by applying the claim.

For the upper bound, swap the two sets S and S' and use the argument for the lower bound. The only thing we need to show is that the smallest eigenvalue of the matrix generated by points in S' is still at least $\gamma/2$. This follows from Wedin's Theorem(Wedin, 1972) and the fact that $\|E\| \leq \|E\|_F \leq \sqrt{K}\delta \leq \gamma/2$. ■

Now we are ready to prove Lemma 1.2.

Proof: The first case is for the first step of the algorithm, when we try to find the farthest point to the origin. Here essentially $S = \{\vec{0}\}$. For any two vertices v_1, v_2, since the simplex is γ robust, the distance between v_1 and v_2 is at least γ. Which means $\text{dis}(\vec{0}, v_1) + \text{dis}(\vec{0}, v_2) \geq \gamma$, one of them must be at least $\gamma/2$.

For the later steps, recall that S contains vertices of a perturbed simplex. Let S' be the set of original vertices corresponding to the perturbed vertices in S. Let v be any vertex in $\{v_1, v_2, ..., v_K\}$ which is not in S. Now we know the distance between v and S is equal to $\text{vol}(S \cup \{v\})/(|S| - 1)\text{vol}(S)$. On the other hand, we know $\text{vol}(S' \cup \{v\})/(|S'| - 1)\text{vol}(S') \geq \gamma$. Using Lemma 1.3 to bound the ratio between the two pairs $\text{vol}(S)/\text{vol}(S')$ and $\text{vol}(S \cup \{v\})/\text{vol}(S' \cup \{v\})$, we get

$$\text{dis}(v, S) \geq (1 - 4\epsilon'/\gamma)^{2|S|-2}\gamma > \gamma/2$$

when $\gamma > 20K\epsilon'$.

■

Lemma 1.1 is based on the following observation: in a simplex the point with largest ℓ_2 is always a vertex. Even if two vertices have the same norm if they are not close to each other the vertices on the edge connecting them will have significantly lower norm.

Proof: (Lemma 1.1)

Since d_j is the point found by the algorithm, let us consider the point a_j before perturbation. The point a_j is inside the simplex, therefore we can write a_j as a convex combination of the vertices:

$$a_j = \sum_{t=1}^{K} c_t v_t$$

Let v_t be the vertex with largest coefficient c_t. Let Δ be the largest distance from some vertex to the space spanned by points in S ($\Delta = \max_l \text{dis}(v_l, \text{span}(S))$. By Lemma 1.2 we know $\Delta > \gamma/2$. Also notice that we are not assuming $\text{dis}(v_t, \text{span}(S)) = \Delta$.

Now we rewrite a_j as $c_t v_t + (1 - c_t)w$, where w is a vector in the convex hull of vertices other than v_t.

Observe that a_j must be far from $\text{span}(S)$, because d_j is the farthest point found by the algorithm. Indeed

$$\begin{aligned}\text{dis}(a_j, \text{span}(S)) &\geq \text{dis}(d_j, \text{span}(S)) - \epsilon \\ &\geq \text{dis}(v_l, \text{span}(S)) - 2\epsilon \geq \Delta - 2\epsilon.\end{aligned}$$

The second inequality is because there must be some point d_l that correspond to the farthest vertex v_l and have $\text{dis}(d_l, \text{span}(S)) \geq \Delta - \epsilon$. Thus as d_j is the farthest point $\text{dis}(d_j, \text{span}(S)) \geq \text{dis}(d_l, \text{span}(S)) \geq \Delta - \epsilon$.

The point a_j is on the segment connecting v_t and w, the distance between a_j and $\text{span}(S)$ is not much

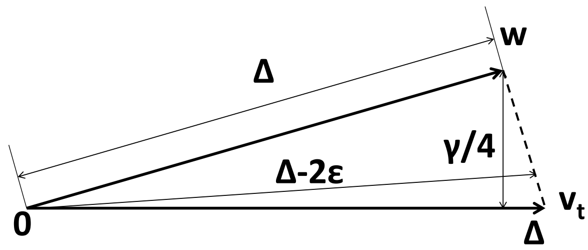

Figure 2. Proof of Lemma 1.1, after projecting to the orthogonal subspace of span(S).

smaller than that of v_t and w. Following the intuition in ℓ_2 norm when v_t and w are far we would expect a_j to be very close to either v_t or w. Since $c_t \geq 1/K$ it cannot be really close to w, so it must be really close to v_t. We formalize this intuition by the following calculation (see Figure 2):

Project everything to the orthogonal subspace of span(S) (points in span(S) are now at the origin). After projection distance to span(S) is just the ℓ_2 norm of a vector. Without loss of generality we assume $\|v_t\|_2 = \|w\|_2 = \Delta$ because these two have length at most Δ, and extending these two vectors to have length Δ can only increase the length of d_j.

The point v_t must be far from w by applying Lemma 1.2: consider the set of vertices $V' = \{v_i : v_i \text{ does not correspond to any point in } S \text{ and } i \neq t\}$. The set $V' \cup S$ satisfy the assumptions in Lemma 1.2 so there must be one vertex that is far from span($V' \cup S$), and it can only be v_t. Therefore even after projecting to orthogonal subspace of span(S), v_t is still far from any convex combination of V'. The vertices that are not in V' all have very small norm after projecting to orthogonal subspace (at most δ_0) so we know the distance of v_t and w is at least $\gamma/2 - \delta_0 > \gamma/4$.

Now the problem becomes a two dimensional calculation. When c_t is fixed the length of a_j is strictly increasing when the distance of v_t and w decrease, so we assume the distance is $\gamma/4$. Simple calculation (using essentially just pythagorean theorem) shows

$$c_t(1 - c_t) \leq \frac{\epsilon}{\Delta - \sqrt{\Delta^2 - \gamma^2/16}}.$$

The right hand side is largest when $\Delta = 2$ (since the vectors are in unit ball) and the maximum value is $O(\epsilon/\gamma^2)$. When this value is smaller than $1/K$, we must have $1 - c_t \leq O(\epsilon/\gamma^2)$. Thus $c_t \geq 1 - O(\epsilon/\gamma^2)$ and $\delta \leq (1 - c_t) + \epsilon \leq O(\epsilon/\gamma^2)$. ■

The cleanup phase tries to find the farthest point to a subset of $K-1$ vertices, and use that point as the K-th vertex. This will improve the result because when we have $K-1$ points close to $K-1$ vertices, only one of the vertices can be far from their span. Therefore the farthest point must be close to the only remaining vertex. Another way of viewing this is that the algorithm is trying to greedily maximize the volume of the simplex, which makes sense because the larger the volume is, the more words/documents the final LDA model can explain.

The following lemma makes the intuitions rigorous and shows how cleanup improves the guarantee of Lemma 1.1.

Lemma 1.5. *Suppose $|S| = K-1$ and each point in S is $\delta = O(\epsilon/\gamma^2) < \gamma/20K$ close to distinct vertices v_i's, the farthest point found by the algorithm is d_j, then the corresponding a_j $O(\epsilon/\gamma)$-covers the remaining vertex.*

Proof: We still look at the original point a_j and express it as $\sum_{t=1}^{K} c_t v_t$. Without loss of generality let v_1 be the vertex that does not correspond to anything in S. By Lemma 1.2 v_1 is $\gamma/2$ far from span(S). On the other hand all other vertices are at least $\gamma/20r$ close to span(S). We know the distance $\text{dis}(a_j, \text{span}(S)) \geq \text{dis}(v_1, \text{span}(S)) - 2\epsilon$, this cannot be true unless $c_1 \geq 1 - O(\epsilon/\gamma)$. ■

These lemmas directly lead to the following theorem:

Theorem 1.6. *FastAnchorWords algorithm runs in time $\tilde{O}(V^2 + VK/\epsilon^2)$ and outputs a subset of $\{d_1, ..., d_V\}$ of size K that $O(\epsilon/\gamma)$-covers the vertices provided that $20K\epsilon/\gamma^2 < \gamma$.*

Proof: In the first phase of the algorithm, do induction using Lemma 1.1. When $20K\epsilon/\gamma^2 < \gamma$ Lemma 1.1 shows that we find a set of points that $O(\epsilon/\gamma^2)$-covers the vertices. Now Lemma 1.5 shows after cleanup phase the points are refined to $O(\epsilon/\gamma)$-cover the vertices. ■

2. Proof for Nonnegative Recover Procedure

In order to show RecoverL2 learns the parameters even when the rows of $\bar{Q}$ are perturbed, we need the following lemma that shows when columns of $\bar{Q}$ are close to the expectation, the posteriors c computed by the algorithm is also close to the true value.

Lemma 2.1. *For a γ robust simplex S with vertices $\{v_1, v_2, ..., v_K\}$, let v be a point in the simplex that can be represented as a convex combination $v = \sum_{i=1}^{K} c_i v_i$. If the vertices of S are perturbed to $S' = \{..., v_i', ...\}$*

where $\|v_i' - v_i\| \leq \delta_1$ and v is perturbed to v' where $\|v - v'\| \leq \delta_2$. Let v^* be the point in $conv\{S'\}$ that is closest to v', and $v^* = \sum_{i=1}^K c_i' v_i$, when $10\sqrt{K}\delta_1 \leq \gamma$ for all $i \in [K]$ $|c_i - c_i'| \leq 4(\delta_1 + \delta_2)/\gamma$.

Proof: Consider the point $u = \sum_{i=1}^K c_i v_i'$, by triangle inequality: $\|u - v\| \leq \sum_{i=1}^K c_i \|v_i - v_i'\| \leq \delta_1$. Hence $\|u - v'\| \leq \|u - v\| + \|v - v'\| \leq \delta_1 + \delta_2$, and u is in S'. The point v^* is the point in conv$\{S'\}$ that is closest to v', so $\|v^* - v'\| \leq \delta_1 + \delta_2$ and $\|v^* - u\| \leq 2(\delta_1 + \delta_2)$.

Then we need to show when a point (u) moves a small distance, its representation also changes by a small amount. Intuitively this is true because S is γ robust. By Lemma 1.2 when $10\sqrt{K}\delta_1 < \gamma$, the simplex S' is also $\gamma/2$ robust. For any i, let $Proj_i(v^*)$ and $Proj_i(u)$ be the projections of v^* and u in the orthogonal subspace of span$(S' \backslash v_i')$, then

$$\begin{aligned} |c_i - c_i'| &= \|Proj_i(v^*) - Proj_i(u)\| / \mathrm{dis}(v_i, \mathrm{span}(S' \backslash v_i')) \\ &\leq 4(\delta_1 + \delta_2)/\gamma \end{aligned}$$

and this completes the proof. ■

With this lemma it is not hard to show that RecoverL2 has polynomial sample complexity.

Theorem 2.2. *When the number of documents M is at least*

$$\max\{O(aK^3 \log V / D(\gamma p)^6 \epsilon), O((aK)^3 \log V / D\epsilon^3 (\gamma p)^4)\}$$

our algorithm using the conjunction of FastAnchorWords and RecoverL2 learns the A matrix with entry-wise error at most ϵ.

Proof: (sketch) We can assume without loss of generality that each word occurs with probability at least $\epsilon/4aK$ and furthermore that if M is at least $50 \log V / D\epsilon_Q^2$ then the empirical matrix $\tilde{Q}$ is entry-wise within an additive ϵ_Q to the true $Q = \frac{1}{M}\sum_{d=1}^M A W_d W_d^T A^T$ see (Arora et al., 2012) for the details. Also, the K anchor rows of $\bar{Q}$ form a simplex that is γp robust.

The error in each column of $\bar{Q}$ can be at most $\delta_2 = \epsilon_Q \sqrt{4aK/\epsilon}$. By Theorem 1.6 when $20K\delta_2/(\gamma p)^2 < \gamma p$ (which is satisfied when $M = O(aK^3 \log V / D(\gamma p)^6 \epsilon)$) , the anchor words found are $\delta_1 = O(\delta_2/(\gamma p))$ close to the true anchor words. Hence by Lemma 2.1 every entry of C has error at most $O(\delta_2/(\gamma p)^2)$.

With such number of documents, all the word probabilities $p(w = i)$ are estimated more accurately than the entries of $C_{i,j}$, so we omit their perturbations here for simplicity. When we apply the Bayes rule, we know $A_{i,k} = C_{i,k} p(w = i)/p(z = k)$, where $p(z = k)$ is α_k which is lower bounded by $1/aK$. The numerator and denominator are all related to entries of C with positive coefficients sum up to at most 1. Therefore the errors δ_{num} and δ_{denom} are at most the error of a single entry of C, which is bounded by $O(\delta_2/(\gamma p)^2)$. Applying Taylor's Expansion to $(p(z = k, w = i) + \delta_{num})/(\alpha_k + \delta_{denom})$, the error on entries of A is at most $O(aK\delta_2/(\gamma p)^2)$. When $\epsilon_Q \leq O((\gamma p)^2 \epsilon^{1.5}/(aK)^{1.5})$, we have $O(aK\delta_2/(\gamma p)^2) \leq \epsilon$, and get the desired accuracy of A. The number of document required is $M = O((aK)^3 \log V / D\epsilon^3 (\gamma p)^4)$.

The sample complexity for R can then be bounded using matrix perturbation theory. ■

For RecoverKL, we observe that the dimension and minimum values of v_i's are all bounded by polynomials of ϵ, a, r (see Section 3.5 Reducing Dictionary Size of (Arora et al., 2012)). In this case, when distance δ is small enough, we know the KL-divergence is both upper and lowerbounded by some polynomial factor times ℓ_2 norm squared.

Lemma 2.3. *When all values in the vectors $\{v_i'\}$ are at least $l = \epsilon^2/20a^2r^2$, if u is one of v_i', and v is in the convex hull of perturbed vertices $\{v_1', v_2', ..., v_K'\}$, $\|u - v\| \leq \epsilon^2/100a^2r^2$, then $D_{KL}(u\|v) \leq 2\|u - v\|^2/l$.*

Proof: Let $s_i = u_i - v_i$, apply Taylor's expansion on $\log(v_i + s_i)/v_i$, we know in the range of parameters $s_i + s_i^2/2v_i \leq \log(v_i + s_i)/v_i \leq s_i + 2s_i^2/v_i$.

Adding this up, using the fact $\sum s_i = \sum u_i - \sum v_i = 0$, we know the KL-divergence is bounded by

$$D_{KL}(u\|v) \leq 2\sum s_i^2/v_i \leq 2\|u - v\|^2/l.$$

■

On the other hand, by Pinsker's inequality, we know $D_{KL}(u\|v) \geq 2|u - v|_1^2 \geq 2\|u - v\|^2$.

Using these two bounds we can easily prove a replacement for Lemma 2.1.

Lemma 2.4. *For a γ robust simplex S with vertices $\{v_1, v_2, ..., v_K\}$, let v be a point in the simplex that can be represented as a convex combination $v = \sum_{i=1}^K c_i v_i$. If the vertices of S are perturbed to $S' = \{..., v_i', ...\}$ where $\|v_i' - v_i\| \leq \delta_1$ and v is perturbed to v' where $\|v - v'\| \leq \delta_2$. Further assume all entries of v' and v_i' are at least $l = \epsilon^2/20a^2r^2$. Let v^{KL} be the point in $conv\{S'\}$ that has smallest $D_{KL}(v'\|v^{KL})$, and $v^{KL} = \sum_{i=1}^K c_i' v_i$, when $10\sqrt{K}\delta_1 \leq \gamma$, $(\delta_1 + \delta_2) < l/5$, for all $i \in [K]$ $|c_i - c_i'| \leq 4(\delta_1 + \delta_2)/\gamma\sqrt{l}$.*

Proof: Let v^* be the closest point (in ℓ_2 distance) of v' in $\mathrm{conv}\{S'\}$. By proof of Lemma 2.1 we know $\|v^* - v'\| \le \delta_1 + \delta_2$. Hence by Lemma 2.3 $D_{KL}(v'\|v^*) \le 2(\delta_1 + \delta_2)^2/l$.

Since v^{KL} is the point with smallest divergence, we know in particular $D_{KL}(v'\|v^{KL}) \le 2(\delta_1 + \delta_2)^2/l$. On the other hand, by Pinkser's inequality $D_{KL}(v'\|v^{KL}) \ge 2\|v' - v^{KL}\|^2$, therefore we know $\|v' - v^{KL}\| \le (\delta_1 + \delta_2)/\sqrt{l}$.

Now we follow the proof of Lemma 2.1 and define $u = \sum_{i=1}^K c_i v_i'$, then we know $\|u - v^{KL}\| \le \|u - v'\| + \|v' - v^{KL}\| \le 2(\delta_1 + \delta_2)/\sqrt{l}$, and similar to Lemma 2.1 we know $|c_i - c_i'| \le 4(\delta_1 + \delta_2)/\gamma\sqrt{l}$. ■

We can simply replace Lemma 2.1 with this Lemma and get provable guarantee of RecoverKL. However, the argument here is not tight (in particular it gives worse bound than ℓ_2).

3. Empirical Results

This section contains plots for ℓ_1, held-out probability, coherence, and uniqueness for all semi-synthetic data sets. Up is better for all metrics except ℓ_1 error. The advantage of the non-negative recovery methods over the original Recover method on the real data is consistent with the results observed on the semi-synthetic data. For example, one can compare the mean log likelihood on real NY Times data from Figure 5 of the main paper (100 topics; 236k docs) with the semi-synthetic NY Times data shown in Figure 3 of the supplementary materials (100 topics; 250k docs). The values for the real data are [Recover: -8.42, RecoverL2: -8.16, RecoverKL: -8.09, Gibbs -7.93] and for semi-synthetic are [Recover: -8.23, RecoverL2: -8.08, RecoverKL: -8.08, Gibbs: -8.076].

3.1. Sample Topics

Tables 1, 2, and 3 show 100 topics trained on real NY Times articles using the RecoverL2 algorithm. Each topic is followed by the most similar topic (measured by ℓ_1 distance) from a model trained on the same documents with Gibbs sampling. When the anchor word is among the top six words by probability it is highlighted in bold. Note that the anchor word is frequently not the most prominent word.

4. Algorithmic Details

4.1. Generating Q matrix

For each document, let H_d be the vector in $\mathbb{R}^V$ such that the i-th entry is the number of times word i ap-

Table 1. Example topic pairs from NY Times sorted by ℓ_1 distance, anchor words in bold.

RecoverL2	run inning game hit season zzz_anaheim_angel
Gibbs	run inning hit game ball pitch
RecoverL2	king goal game team games season
Gibbs	point game team play season games
RecoverL2	yard game play season team touchdown
Gibbs	yard game season team play quarterback
RecoverL2	point game team season games play
Gibbs	point game team play season games
RecoverL2	zzz_laker point **zzz_kobe_bryant** zzz_o_neal game team
Gibbs	point game team play season games
RecoverL2	point game team season player **zzz_clipper**
Gibbs	point game team season play zzz_usc
RecoverL2	ballot election court votes vote zzz_al_gore
Gibbs	election ballot zzz_florida zzz_al_gore votes vote
RecoverL2	game zzz_usc team play point season
Gibbs	point game team season play zzz_usc
RecoverL2	company billion companies percent million stock
Gibbs	company million percent billion analyst deal
RecoverL2	car race team season driver point
Gibbs	race car driver racing zzz_nascar team
RecoverL2	zzz_dodger season run inning right game
Gibbs	season team baseball game player yankees
RecoverL2	palestinian zzz_israeli zzz_israel official attack **zzz_palestinian**
Gibbs	palestinian zzz_israeli zzz_israel attack zzz_palestinian zzz_yasser_arafat
RecoverL2	zzz_tiger_wood shot round player par play
Gibbs	zzz_tiger_wood shot golf tour round player
RecoverL2	percent stock market companies fund quarter
Gibbs	percent economy market stock economic growth
RecoverL2	zzz_al_gore **zzz_bill_bradley** campaign president zzz_george_bush vice
Gibbs	zzz_al_gore zzz_george_bush campaign presidential republican zzz_john_mccain
RecoverL2	zzz_george_bush **zzz_john_mccain** campaign republican zzz_republican voter
Gibbs	zzz_al_gore zzz_george_bush campaign presidential republican zzz_john_mccain
RecoverL2	net team season point player **zzz_jason_kidd**
Gibbs	point game team play season games
RecoverL2	yankees run team season inning hit
Gibbs	season team baseball game player yankees
RecoverL2	zzz_al_gore zzz_george_bush percent president campaign zzz_bush
Gibbs	zzz_al_gore zzz_george_bush campaign presidential republican zzz_john_mccain
RecoverL2	zzz_enron company firm **zzz_arthur_andersen** companies lawyer
Gibbs	zzz_enron company firm accounting zzz_arthur_andersen financial
RecoverL2	team play game yard season player
Gibbs	yard game season team play quarterback
RecoverL2	film movie show director play character
Gibbs	film movie character play minutes hour
RecoverL2	zzz_taliban zzz_afghanistan official zzz_u_s government military
Gibbs	zzz_taliban zzz_afghanistan zzz_pakistan afghan zzz_india government
RecoverL2	palestinian zzz_israel israeli peace zzz_yasser_arafat leader
Gibbs	palestinian zzz_israel peace israeli zzz_yasser_arafat leader
RecoverL2	point team game shot play zzz_celtic
Gibbs	point game team play season games
RecoverL2	zzz_bush **zzz_mccain** campaign republican tax zzz_republican
Gibbs	zzz_al_gore zzz_george_bush campaign presidential republican zzz_john_mccain
RecoverL2	zzz_met run team game hit season
Gibbs	season team baseball game player yankees
RecoverL2	team game season play games win
Gibbs	team coach game player season football
RecoverL2	government war **zzz_slobodan_milosevic** official court president
Gibbs	government war country rebel leader military
RecoverL2	game set player **zzz_pete_sampras** play won
Gibbs	player game match team soccer play
RecoverL2	zzz_al_gore campaign **zzz_bradley** president democratic zzz_clinton
Gibbs	zzz_al_gore zzz_george_bush campaign presidential republican zzz_john_mccain
RecoverL2	team zzz_knick player season point play
Gibbs	point game team play season games
RecoverL2	com web www information sport question
Gibbs	palm beach com statesman daily american

Table 2. Example topic pairs from NY Times sorted by ℓ_1 distance, anchor words in bold.

RecoverL2	season team game coach play school
Gibbs	team coach game player season football
RecoverL2	air shower rain wind storm front
Gibbs	water fish weather storm wind air
RecoverL2	book film **beginitalic** enditalic look movie
Gibbs	film movie character play minutes hour
RecoverL2	zzz_al_gore campaign election zzz_george_bush zzz_florida president
Gibbs	zzz_al_gore zzz_george_bush campaign presidential republican zzz_john_mccain
RecoverL2	race won horse **zzz_kentucky_derby** win winner
Gibbs	horse race horses winner won zzz_kentucky_derby
RecoverL2	company companies **zzz_at** percent business stock
Gibbs	company companies business industry firm market
RecoverL2	company million companies percent business customer
Gibbs	company companies business industry firm market
RecoverL2	team coach season player jet job
Gibbs	team player million season contract agent
RecoverL2	season team game play player zzz_cowboy
Gibbs	yard game season team play quarterback
RecoverL2	zzz_pakistan zzz_india official group attack zzz_united_states
Gibbs	zzz_taliban zzz_afghanistan zzz_pakistan afghan zzz_india government
RecoverL2	show network night television zzz_nbc program
Gibbs	film movie character play minutes hour
RecoverL2	com information question zzz_eastern commentary daily
Gibbs	com question information zzz_eastern daily commentary
RecoverL2	power plant company percent million energy
Gibbs	oil power energy gas prices plant
RecoverL2	cell stem research zzz_bush human patient
Gibbs	cell research human scientist stem genes
RecoverL2	**zzz_governor_bush** zzz_al_gore campaign tax president plan
Gibbs	zzz_al_gore zzz_george_bush campaign presidential republican zzz_john_mccain
RecoverL2	cup minutes add tablespoon water oil
Gibbs	cup minutes add tablespoon teaspoon oil
RecoverL2	family home book right com children
Gibbs	film movie character play minutes hour
RecoverL2	zzz_china chinese zzz_united_states **zzz_taiwan** official government
Gibbs	zzz_china chinese zzz_beijing zzz_taiwan government official
RecoverL2	death court law case lawyer zzz_texas
Gibbs	trial death prison case lawyer prosecutor
RecoverL2	company percent million sales business companies
Gibbs	company companies business industry firm market
RecoverL2	dog jump show quick brown **fox**
Gibbs	film movie character play minutes hour
RecoverL2	**shark** play team attack water game
Gibbs	film movie character play minutes hour
RecoverL2	anthrax official mail letter worker attack
Gibbs	anthrax official letter mail nuclear chemical
RecoverL2	president zzz_clinton zzz_white_house zzz_bush official zzz_bill_clinton
Gibbs	zzz_bush zzz_george_bush president administration zzz_white_house zzz_dick_cheney
RecoverL2	father family **zzz_elian** boy court zzz_miami
Gibbs	zzz_cuba zzz_miami cuban zzz_elian boy protest
RecoverL2	oil prices percent million market zzz_united_states
Gibbs	oil power energy gas prices plant
RecoverL2	**zzz_microsoft** company computer system window software
Gibbs	zzz_microsoft company companies cable zzz_at zzz_internet
RecoverL2	government election zzz_mexico political zzz_vicente_fox president
Gibbs	election political campaign zzz_party democratic voter
RecoverL2	fight **zzz_mike_tyson** round right million champion
Gibbs	fight zzz_mike_tyson ring fighter champion round
RecoverL2	right law president zzz_george_bush zzz_senate **zzz_john_ashcroft**
Gibbs	election political campaign zzz_party democratic voter
RecoverL2	com home look found show www
Gibbs	film movie character play minutes hour
RecoverL2	car driver race **zzz_dale_earnhardt** racing zzz_nascar
Gibbs	night hour room hand told morning
RecoverL2	book women family called author woman
Gibbs	film movie character play minutes hour

Table 3. Example topic pairs from NY Times sorted by ℓ_1 distance, anchor words in bold.

RecoverL2	tax bill zzz_senate billion plan zzz_bush
Gibbs	bill zzz_senate zzz_congress zzz_house legislation zzz_white_house
RecoverL2	company **francisco** san com food home
Gibbs	palm beach com statesman daily american
RecoverL2	team player season game **zzz_john_rocker** right
Gibbs	season team baseball game player yankees
RecoverL2	zzz_bush official zzz_united_states zzz_u_s president zzz_north_korea
Gibbs	zzz_united_states weapon zzz_iraq nuclear zzz_russia zzz_bush
RecoverL2	zzz_russian zzz_russia official military war attack
Gibbs	government war country rebel leader military
RecoverL2	wine **wines** percent zzz_new_york com show
Gibbs	film movie character play minutes hour
RecoverL2	police **zzz_ray_lewis** player team case told
Gibbs	police officer gun crime shooting shot
RecoverL2	government group political tax leader money
Gibbs	government war country rebel leader military
RecoverL2	percent company million airline flight deal
Gibbs	flight airport passenger airline security airlines
RecoverL2	book ages children school boy web
Gibbs	book author writer word writing read
RecoverL2	**corp** group president energy company member
Gibbs	palm beach com statesman daily american
RecoverL2	team tour **zzz_lance_armstrong** won race win
Gibbs	zzz_olympic games medal gold team sport
RecoverL2	priest church official abuse bishop sexual
Gibbs	church religious priest zzz_god religion bishop
RecoverL2	human drug company companies million scientist
Gibbs	scientist light science planet called space
RecoverL2	music **zzz_napster** company song com web
Gibbs	palm beach com statesman daily american
RecoverL2	death government case federal official **zzz_timothy_mcveigh**
Gibbs	trial death prison case lawyer prosecutor
RecoverL2	million shares offering public company initial
Gibbs	company million percent billion analyst deal
RecoverL2	buy **panelist** thought flavor product ounces
Gibbs	food restaurant chef dinner eat meal
RecoverL2	school student program teacher public children
Gibbs	school student teacher children test education
RecoverL2	security official government airport federal bill
Gibbs	flight airport passenger airline security airlines
RecoverL2	company member credit card money mean
Gibbs	zzz_enron company firm accounting zzz_arthur_andersen financial
RecoverL2	million percent bond tax debt bill
Gibbs	million program billion money government federal
RecoverL2	million company zzz_new_york business art percent
Gibbs	art artist painting museum show collection
RecoverL2	percent million number official group black
Gibbs	palm beach com statesman daily american
RecoverL2	company tires million car zzz_ford percent
Gibbs	company companies business industry firm market
RecoverL2	article zzz_new_york **misstated** company percent com
Gibbs	palm beach com statesman daily american
RecoverL2	company million percent companies government official
Gibbs	company companies business industry firm market
RecoverL2	official million train car system plan
Gibbs	million program billion money government federal
RecoverL2	**test** student school look percent system
Gibbs	patient doctor cancer medical hospital surgery
RecoverL2	con una mas dice las anos
Gibbs	fax syndicate article com information con
RecoverL2	**por** con una mas millones como
Gibbs	fax syndicate article com information con
RecoverL2	las como **zzz_latin_trade** articulo telefono fax
Gibbs	fax syndicate article com information con
RecoverL2	**los** con articulos telefono representantes zzz_america_latina
Gibbs	fax syndicate article com information con
RecoverL2	**file** sport read internet email zzz_los_angeles
Gibbs	web site com www mail zzz_internet

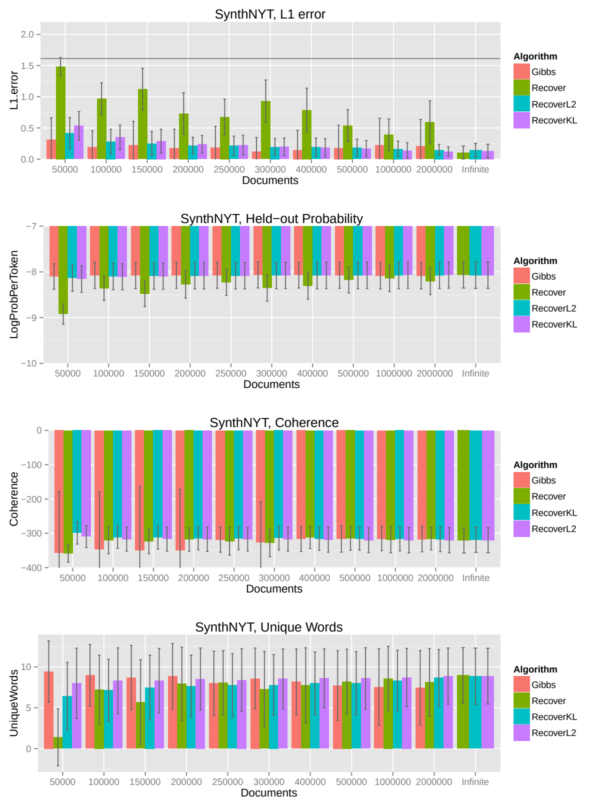

Figure 3. Results for a semi-synthetic model generated from a model trained on NY Times articles with $K = 100$.

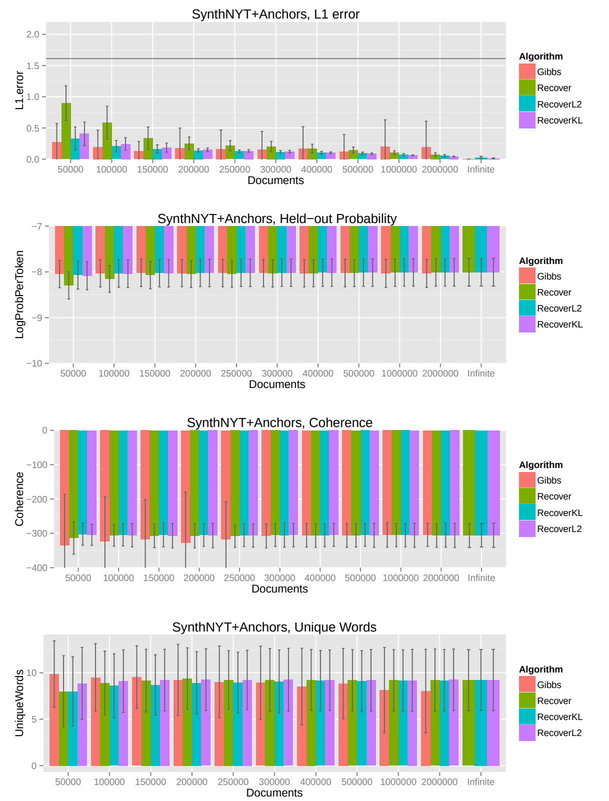

Figure 4. Results for a semi-synthetic model generated from a model trained on NY Times articles with $K = 100$, with a synthetic anchor word added to each topic.

pears in document d, n_d be the length of the document and W_d be the topic vector chosen according to Dirichlet distribution when the documents are generated. Conditioned on W_d's, our algorithms require the expectation of Q to be $\frac{1}{M}\sum_{d=1}^{M} AW_dW_d^TA^T$.

In order to achieve this, similar to (Anandkumar et al., 2012), let the normalized vector $\tilde{H}_d = \frac{H_d}{\sqrt{n_d(n_d-1)}}$ and diagonal matrix $\hat{H}_d = \frac{\text{Diag}(H_d)}{n_d(n_d-1)}$. Compute the matrix

$$\tilde{H}_d\tilde{H}_d^T - \hat{H}_d = \frac{1}{n_d(n_d-1)} \sum_{i\neq j, i,j\in[n_d]} e_{z_{d,i}} e_{z_{d,j}}^T.$$

Here $z_{d,i}$ is the i-th word of document d, and $e_i \in \mathbb{R}^V$ is the basis vector. From the generative model, the expectation of all terms $e_{z_{d,i}} e_{z_{d,j}}^T$ are equal to $AW_dW_d^TA^T$, hence by linearity of expectation we know $\mathbf{E}[\tilde{H}_d\tilde{H}_d^T - \hat{H}_d] = AW_dW_d^TA^T$.

If we collect all the column vectors $\tilde{H}_d$ to form a large sparse matrix $\tilde{H}$, and compute the sum of all $\hat{H}_d$ to get the diagonal matrix $\hat{H}$, we know $Q = \tilde{H}\tilde{H}^T - \hat{H}$ has the desired expectation. The running time of this step is $O(MD^2)$ where D^2 is the expectation of the length of the document squared.

4.2. Applying Recover to Small Datasets

The original Recover algorithm from Arora et al. (2012) can fail on small datasets if the $Q_{\mathbf{S},\mathbf{S}}$ matrix which holds the anchor-anchor co-occurrence counts is rank deficient due to sparsity. When Recover fails, we use a modified version of the algorithm, solving for $\vec{z}$ by finding a least squares solution to $Q_{\mathbf{S},\mathbf{S}}\vec{z} = \vec{p}_{\mathbf{S}}$ and solving for A^T with a pseudoinverse: $A^T = (Q_{\mathbf{S},\mathbf{S}}\text{Diag}(\vec{z}))^{\dagger} Q_{\mathbf{S}}^T)$. This procedure can return an A matrix in which some columns contain all 0s. In that case we replace columns of 0s with a uniform distribution over the vocabulary words, $\frac{1}{V}\mathbf{1}$.

Negative values also often occur in the A matrix returned by the original Recover method. To project back onto the simplex, we clip all negative values to 0 and normalize the columns before evaluating the learned model.

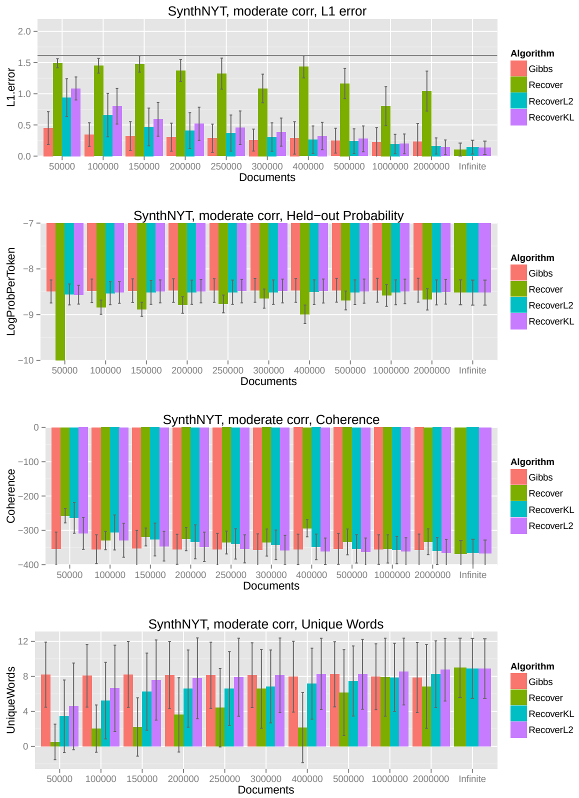

Figure 5. Results for a semi-synthetic model generated from a model trained on NY Times articles with $K = 100$, with moderate correlation between topics.

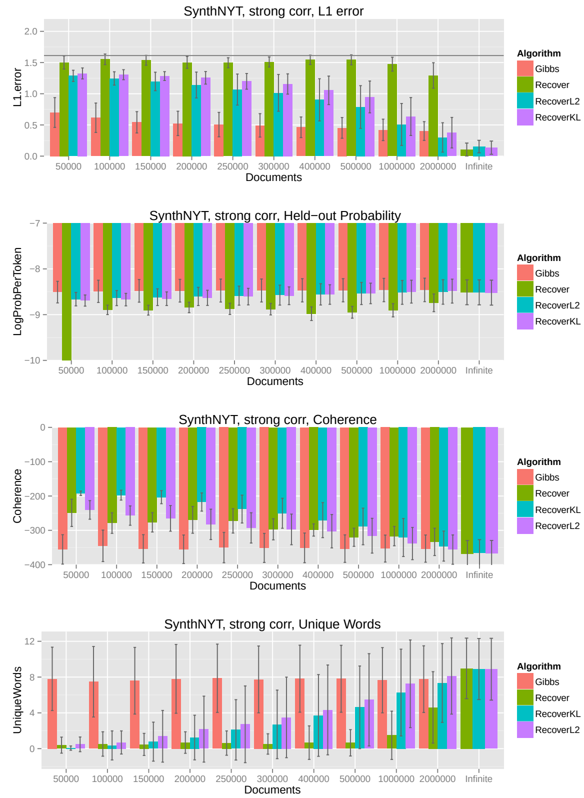

Figure 6. Results for a semi-synthetic model generated from a model trained on NY Times articles with $K = 100$, with stronger correlation between topics.

4.3. Exponentiated gradient algorithm

The optimization problem that arises in RecoverKL and RecoverL2 has the following form:

$$\min_{\vec{x}} d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x})$$

$$\text{subject to: } \vec{x} \geq 0 \text{ and } \sum_{i=1}^{K} x_i = 1,$$

where $d(\cdot, \cdot)$ is a Bregman divergence (in particular it is squared Euclidean distance for RecoverL2 and KL divergence for RecoverKL), $\vec{x}$ is a column vector of size K, $\mathbf{S}$ is the set of K anchor indices, $\bar{Q}_i$ is a row vector of size V, and $\bar{Q}_{\mathbf{S}}$ is the $K \times V$ matrix formed by stacking the rows of $\bar{Q}$ corresponding to the indices in $\mathbf{S}$.

This is a convex optimization problem with simplex constraints, which can be solved with the Exponentiated Gradient algorithm (Kivinen & Warmuth, 1995), described in Algorithm 1. The Exponentiated Gradient algorithm iteratively generates values of $\vec{x}$ which are feasible and converge to the optimal value $\vec{x}^*$. In our experiments we show results using both squared Euclidean distance and KL divergence for the divergence measure.

To determine whether the algorithm has converged, we test whether the KKT conditions (which are sufficient for optimality in this problem) hold to within some tolerance, ϵ. In our experiments ϵ varies between 10^{-6} and 10^{-9} depending on the data set.

The KKT conditions for our constrained minimization problem are:

1. Stationarity: $\nabla_{\vec{x}} d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x}) - \vec{\lambda} + \mu\mathbf{1} = 0$.
2. Primal Feasibility: $\vec{x} \geq 0$, $\sum_{i=1}^{K} x_i = 1$.
3. Dual Feasibility: $\lambda_i \geq 0$ for $i \in \{1, 2, ..., K\}$.
4. Complementary Slackness: $\lambda_i x_i = 0$ for $i \in \{1, 2, ..., K\}$.

We define the following approximation to Condition 4:

4′. ϵ-Complementary Slackness: $0 \leq \vec{\lambda}^T\vec{x} < \epsilon$.

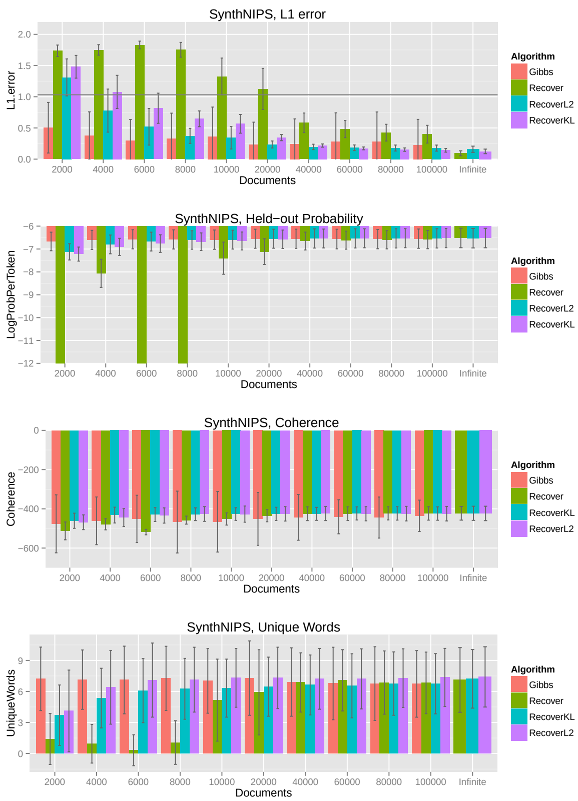

Figure 7. Results for a semi-synthetic model generated from a model trained on NIPS papers with $K = 100$. For $D \in \{2000, 6000, 8000\}$, Recover produces log probabilities of $-\infty$ for some held-out documents.

Let $\vec{x}_t$ be the t^{th} value generated by Exponentiated Gradient. $\vec{x}_t$ is ϵ-optimal if there exist $\vec{\lambda}$ and μ such that Conditions 1-3 and $4'$ are satisfied.

We initialize $\vec{x}_0 = \frac{1}{K}\mathbf{1}$ and Exponentiated Gradient preserves primal feasibility, so $\vec{x}_t$ satisfies Condition 2. The following $\vec{\lambda}_t$ and μ_t minimize $\vec{\lambda}_t^T\vec{x}_t$ while satisfying conditions 1 and 3:

$$\mu_t = -\min\left(\nabla_{\vec{x}} d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x})\big|_{\vec{x}_t}\right)$$
$$\vec{\lambda}_t = \nabla_{\vec{x}} d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x})\big|_{\vec{x}_t} + \mu_t\mathbf{1}.$$

The algorithm converges when Condition $4'$ is satisfied (i.e. $\vec{\lambda}_t^T\vec{x}_t < \epsilon$).

$\vec{\lambda}_t^T\vec{x}_t$ can also be understood as the gap between an upper and lower bound on the objective. To see this, note that the Lagrangian function is:

$$L(\vec{x}, \vec{\lambda}, \mu) = d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x}) - \vec{\lambda}^T\vec{x} + \mu(\vec{x}^T\mathbf{1} - 1),$$

The first term in the Lagrangian is exactly the primal objective, and $({\vec{x}_t}^T\mathbf{1} - 1)$ is zero at every iteration.

Algorithm 1. Exponentiated Gradient

Input: Matrix $\bar{Q}_{\mathbf{S}}$, vector $\bar{Q}_i{}^T$, divergence measure $d(\cdot,\cdot)$, tolerance parameter ϵ
Output: non-negative normalized vector $\vec{x}$ close to $\vec{x}^*$, the minimizer of $d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x}))$
$\vec{x}_0 \leftarrow \frac{1}{K}\mathbf{1}$
$t \leftarrow 0$
Converged $\leftarrow$ False
while not Converged **do**
  $t \leftarrow t + 1$
  $\vec{g}_t = \nabla_{\vec{x}} d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x})\big|_{\vec{x}_{t-1}}$
  Choose a step size η_t
  $\vec{x}_t \leftarrow \vec{x}_{t-1}e^{-\eta_t\vec{g}_t}$ (Gradient step)
  $\vec{x}_t \leftarrow \frac{\vec{x}}{|\vec{x}_t|_1}$ (Projection onto the simplex)
  $\mu_t \leftarrow -\min\left(\nabla_{\vec{x}} d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x})\big|_{\vec{x}_t}\right)$
  $\vec{\lambda}_t \leftarrow \nabla_{\vec{x}} d(\bar{Q}_i^T, \bar{Q}_{\mathbf{S}}\vec{x})\big|_{\vec{x}_t} + \mu_t\mathbf{1}$
  Converged $\leftarrow \vec{\lambda}_t^T\vec{x}_t < \epsilon$
end while
return x_t

Since the Lagrangian lower bounds the objective, $\vec{\lambda}_t^T\vec{x}_t$ is the value of the gap. Strong duality holds for this problem, so at optimality, this gap is 0. Testing that the gap is less than ϵ is an approximate optimality test.

Stepsizes at each iteration are chosen with a line search to find an η_t that satisfies the Wolfe and Armijo conditions (For details, see Nocedal & Wright (2006)).

The running time of RecoverL2 is the time of solving V small $(K \times K)$ quadratic programs. When using Exponentiated Gradient to solve the quadratic program, each word requires $O(KV)$ time for preprocessing and $O(K^2)$ per iteration. The total running time is $O(KV^2 + K^2VT)$ where T is the average number of iterations. The value of T is about $100 - 1000$ depending on data sets.

References

Anandkumar, A., Foster, D., Hsu, D., Kakade, S., and Liu, Y. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. In *NIPS*, 2012. 4.1

Arora, S., Ge, R., and Moitra, A. Learning topic models – going beyond svd. In *FOCS*, 2012. 2, 4.2

Kivinen, Jyrki and Warmuth, Manfred K. Exponentiated gradient versus gradient descent for linear predictors. *Inform. and Comput.*, 132, 1995. 4.3

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, New York, 2nd edition, 2006. 4.3

Wedin, P. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972. 1