7 Supplementary Material

```
Input: $\mathbf{V}^T = [\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_d] \in \mathbb{R}^{\ell \times d}$ with $\mathbf{v}_i \in \mathbb{R}^{\ell}$ and $r > \ell$.
Output: Matrices $\mathbf{S} \in \mathbb{R}^{d \times r}, \mathbf{D} \in \mathbb{R}^{r \times r}$.
1. Initialize $\mathbf{A}_0 = \mathbf{0}_{\ell \times \ell}, \mathbf{S} = \mathbf{0}_{d \times r}, \mathbf{D} = \mathbf{0}_{r \times r}$.
2. Set constants $\delta_L = 1$ and $\delta_U = \left(1 + \sqrt{\ell/r}\right) / \left(1 - \sqrt{\ell/r}\right)$.
3. for $\tau = 0$ to $r - 1$ do
     • Let $L_\tau = \tau - \sqrt{r\ell}; U_\tau = \delta_U \left(\tau + \sqrt{\ell r}\right)$.
     • Pick index $i \in \{1, 2, ..d\}$ and number $t_\tau > 0$, such that
           $\mathcal{U}\left(\mathbf{v}_i, \delta_U, \mathbf{A}_\tau, U_\tau\right) \leq \mathcal{L}\left(\mathbf{v}_i, \delta_L, \mathbf{A}_\tau, L_\tau\right)$.
     • Let $t_\tau^{-1} = \frac{1}{2}\left(\mathcal{U}\left(\mathbf{v}_i, \delta_U, \mathbf{A}_\tau, U_\tau\right) + \mathcal{L}\left(\mathbf{v}_i, \delta_L, \mathbf{A}_\tau, L_\tau\right)\right)$
     • Update $\mathbf{A}_{\tau+1} = \mathbf{A}_\tau + t_\tau \mathbf{v}_i \mathbf{v}_i^T$ ; set $\mathbf{S}_{i_\tau, \tau+1} = 1$ and $\mathbf{D}_{\tau+1, \tau+1} = 1/\sqrt{t_\tau}$.
4. end for
5. Multiply all the weights in $\mathbf{D}$ by $\sqrt{r^{-1}\left(1 - \sqrt{(\ell/r)}\right)}$.
6. Return $\mathbf{S}$ and $\mathbf{D}$.
```

Algorithm 1: Single-set Spectral Sparsification

Lemma 3. ***BSS*** *(Batson et al. (2009)): Given* $\mathbf{V} \in \mathbb{R}^{d \times \ell}$ *satisfying* $\mathbf{V}^T\mathbf{V} = \mathbf{I}_\ell$ *and* $r > \ell$*, we can deterministically construct sampling and rescaling matrices* $\mathbf{S} \in \mathbb{R}^{d \times r}$ *and* $\mathbf{D} \in \mathbb{R}^{r \times r}$ *with* $\mathbf{R} = \mathbf{SD}$*, such that, for all* $\mathbf{y} \in \mathbb{R}^\ell$: $\left(1 - \sqrt{\ell/r}\right)^2 \|\mathbf{V}\mathbf{y}\|_2^2 \leq \left\|\mathbf{V}^T\mathbf{R}\mathbf{y}\right\|_2^2 \leq \left(1 + \sqrt{\ell/r}\right)^2 \|\mathbf{V}\mathbf{y}\|_2^2$.

We now present a slightly modified version of Lemma 3 for our theorems.

Lemma 4. *Given* $\mathbf{V} \in \mathbb{R}^{d \times \ell}$ *satisfying* $\mathbf{V}^T\mathbf{V} = \mathbf{I}_\ell$ *and* $r > \ell$*, we can deterministically construct sampling and rescaling matrices* $\mathbf{S} \in \mathbb{R}^{d \times r}$ *and* $\mathbf{D} \in \mathbb{R}^{r \times r}$ *such that for* $\mathbf{R} = \mathbf{SD}$*,* $\left\|\mathbf{V}^T\mathbf{V} - \mathbf{V}^T\mathbf{R}\mathbf{R}^T\mathbf{V}\right\|_2 \leq 3\sqrt{\ell/r}$

Proof. From Lemma 3, it follows, $\sigma_\ell\left(\mathbf{V}^T\mathbf{R}\mathbf{R}^T\mathbf{V}\right) \geq \left(1 - \sqrt{\ell/r}\right)^2$, $\sigma_1\left(\mathbf{V}^T\mathbf{R}\mathbf{R}^T\mathbf{V}\right) \leq \left(1 + \sqrt{\ell/r}\right)^2$. Thus, $\lambda_{max}\left(\mathbf{V}^T\mathbf{V} - \mathbf{V}^T\mathbf{R}\mathbf{R}^T\mathbf{V}\right) \leq \left(1 - \left(1 - \sqrt{\ell/r}\right)^2\right) \leq 2\sqrt{\ell/r}$. Similarly, $\lambda_{min}\left(\mathbf{V}^T\mathbf{V} - \mathbf{V}^T\mathbf{R}\mathbf{R}^T\mathbf{V}\right) \geq \left(1 - \left(1 + \sqrt{\ell/r}\right)^2\right) \geq 3\sqrt{\ell/r}$. Combining these two results, we have $\left\|\mathbf{V}^T\mathbf{V} - \mathbf{V}^T\mathbf{R}\mathbf{R}^T\mathbf{V}\right\|_2 \leq 3\sqrt{\ell/r}$. □

7.1 Proof That the Data Radius is preserved by Unsupervised BSS-Feature Selection.

Theorem 3. *Let* $r_2 = O\left(n/\epsilon^2\right)$*, where* $\epsilon > 0$ *is an accuracy parameter,* n *is the number of training points and* r_2 *is the number of features selected. Let* B *be the radius of the minimum ball enclosing all points in the full-dimensional space, and let* $\tilde{B}$ *be the radius of the ball enclosing all points in the sampled subspace obtained by using BSS in an unsupervised manner. For* $\mathbf{R}$ *as in Lemma 4,* $\tilde{B}^2 \leq (1 + \epsilon)B^2$.

Proof. We consider the matrix $\mathbf{X}_B \in \mathbb{R}^{(n+1) \times d}$ whose first n rows are the rows of $\mathbf{X^{tr}}$ and whose last row is the vector $\mathbf{x}_B^T$; here $\mathbf{x}_B$ denotes the center of the minimum radius ball enclosing all n points. Then, the SVD of $\mathbf{X}_B$ is equal to $\mathbf{X}_B = \mathbf{U}_B\mathbf{\Sigma}_B\mathbf{V}_B^T$, where $\mathbf{U}_B \in \mathbb{R}^{(n+1) \times \rho_B}$, $\mathbf{\Sigma}_B \in \mathbb{R}^{\rho_B \times \rho_B}$, and $\mathbf{V} \in \mathbb{R}^{d \times \rho_B}$. Here ρ_B is the rank of the matrix $\mathbf{X}_B$ and clearly $\rho_B \leq \rho + 1$. (Recall that ρ is the rank of the matrix $\mathbf{X^{tr}}$.) Let B be the radius of the minimal radius ball enclosing all n points in the original space. Then, for any $i = 1, \ldots, n$,

$$B^2 \geq \|\mathbf{x}_i - \mathbf{x}_B\|_2^2 = \left\|\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)^T \mathbf{X}_B\right\|_2^2. \qquad (14)$$

Now consider the matrix $\mathbf{X}_B\mathbf{R}$ and notice that

$$\begin{aligned} & \left| \left\|\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)^T \mathbf{X}_B\right\|_2^2 - \left\|\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)^T \mathbf{X}_B\mathbf{R}\right\|_2^2 \right| \\ = & \left|\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)^T \left(\mathbf{X}_B\mathbf{X}_B^T - \mathbf{X}_B\mathbf{R}\mathbf{R}^T\mathbf{X}_B^T\right)\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)\right| \\ = & \left|\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)^T \mathbf{U}_B\mathbf{\Sigma}_B\mathbf{E}_B\mathbf{\Sigma}_B\mathbf{U}_B^T\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)\right| \\ \leq & \|\mathbf{E}_B\|_2 \left\|\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)^T \mathbf{U}_B\mathbf{\Sigma}_B\right\|_2^2 \\ = & \|\mathbf{E}_B\|_2 \left\|\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)^T \mathbf{X}_B\right\|_2^2. \end{aligned}$$

In the above, we let $\mathbf{E}_B \in \mathbb{R}^{\rho_B \times \rho_B}$ be the matrix that satisfies $\mathbf{V}_B^T\mathbf{V}_B = \mathbf{V}_B^T\mathbf{R}\mathbf{R}^T\mathbf{V}_B + \mathbf{E}_B$, and we also used $\mathbf{V}_B^T\mathbf{V}_B = \mathbf{I}$. Now consider the ball whose center is the $(n+1)$-th row of the matrix $\mathbf{X}_B\mathbf{R}$ (essentially, the center of the minimal radius enclosing ball for the original points in the sampled space). Let $\tilde{i} = \arg\max_{i=1...n} \left\|\left(\mathbf{e}_i - \mathbf{e}_{n+1}\right)^T \mathbf{X}_B\mathbf{R}\right\|_2^2$; then, using the above bound and eqn. (14), we get $\left\|\left(\mathbf{e}_{\tilde{i}} - \mathbf{e}_{n+1}\right)^T \mathbf{X}_B\mathbf{R}\right\|_2^2 \leq (1 + \|\mathbf{E}_B\|_2) \left\|\left(\mathbf{e}_{\tilde{i}} - \mathbf{e}_{n+1}\right)^T \mathbf{X}_B\right\|_2^2 \leq (1 + \|\mathbf{E}_B\|_2) B^2$. Thus, there exists a ball centered at $\mathbf{e}_{n+1}^T\mathbf{X}_B\mathbf{R}$ (the projected center of the minimal radius ball in the original space) with radius at most $\sqrt{1 + \|\mathbf{E}_B\|_2}B$

that encloses all the points in the sampled space. Recall that $\tilde{B}$ is defined as the radius of the minimal radius ball that encloses all points in sampled subspace; clearly, $\tilde{B}^2 \leq (1 + \|\mathbf{E}_B\|_2)\, B^2$. We can now use Lemma 4 on $\mathbf{V}_B$ to conclude that (using $\rho_B \leq \rho + 1$) $\|\mathbf{E}_B\|_2 \leq \epsilon$. □

Theorem 4. *Given $\epsilon \in (0,1)$, run supervised Leverage-score sampling based feature selection on $\mathbf{X}^{sv}$ with $r_1 = \tilde{O}(p/\epsilon^2)$, to obtain the feature sampling and rescaling matrix $\mathbf{R}$. Let γ^* and $\tilde{\gamma}^*$ be the margins obtained by solving the SVM dual (2) with $(\mathbf{X}^{sv}, \mathbf{Y}^{sv})$ and $(\mathbf{X}^{sv}\mathbf{R}, \mathbf{Y}^{sv})$ respectively. Then with probability at least 0.99, $\tilde{\gamma}^{*2} \geq (1-\epsilon)\,\gamma^{*2}$.*

Theorem 5. *Given $\epsilon \in (0,1)$, run unsupervised Leverage-score feature selection on the full data $\mathbf{X}^{tr}$ with $r_2 = \tilde{O}\left(\rho/\epsilon^2\right)$, where $\rho = \mathrm{rank}(\mathbf{X}^{tr})$, to obtain the feature sampling and rescaling matrix $\mathbf{R}$. Let γ^* and $\tilde{\gamma}^*$ be the margins obtained by solving the SVM dual (2) with $(\mathbf{X}^{tr}, \mathbf{Y}^{tr})$ and $(\mathbf{X}^{tr}\mathbf{R}, \mathbf{Y}^{tr})$ respectively; and, let B and $\tilde{B}$ be the radii for the data matrices $\mathbf{X}^{tr}$ and $\mathbf{X}^{tr}\mathbf{R}$ respectively. Then with probability at least 0.99,*

$$\frac{\tilde{B}^2}{\tilde{\gamma}^{*2}} \leq \frac{(1+\epsilon)}{(1-\epsilon)} \frac{B^2}{\gamma^{*2}} = (1 + O(\epsilon)) \frac{B^2}{\gamma^{*2}}.$$

Proofs of Theorems 4 and 5 follow directly from Theorems 1 and 2. In Theorems 1 and 2, we make use of Lemma 1. For Theorems 4 and 5, we make use of Lemma 2 to obtain the proof.

Proof of Lemma 2 can be found in Rudelson and Vershynin (2007).

7.2 Other Feature Selection Methods

In this section, we describe other feature-selection methods with which we compare BSS.

Rank-Revealing QR Factorization (RRQR): Within the numerical linear algebra community, subset selection algorithms use the so-called Rank Revealing QR (RRQR) factorization. Let $\mathbf{A}$ be a $n \times d$ matrix with $(n < d)$ and an integer $k\,(k < d)$ and assume partial QR factorizations of the form

$$\mathbf{AP} = \mathbf{Q}\begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{R}_{22} \end{pmatrix},$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix, $\mathbf{P} \in \mathbb{R}^{d \times d}$ is a permutation matrix, $\mathbf{R}_{11} \in \mathbb{R}^{k \times k}, \mathbf{R}_{12} \in \mathbb{R}^{k \times (d-k)}, \mathbf{R}_{22} \in \mathbb{R}^{(d-k)\times(d-k)}$ The above factorization is called a RRQR factorization if $\sigma_{min}(\mathbf{R}_{11}) \geq \sigma_k(\mathbf{A})/p(k,d)$, $\sigma_{max}(\mathbf{R}_{22}) \leq \sigma_{min}(\mathbf{A})p(k,d)$, where $p(k,d)$ is a function bounded by a low-degree polynomial in k and d. The important columns are given by $\mathbf{A}_1 = \mathbf{Q}\begin{pmatrix} \mathbf{R}_{11} \\ \mathbf{0} \end{pmatrix}$ and $\sigma_i(\mathbf{A}_1) = \sigma_i(\mathbf{R}_{11})$ with $1 \leq i \leq k$. We perform feature selection using RRQR by picking the important columns which preserve the rank of the matrix.

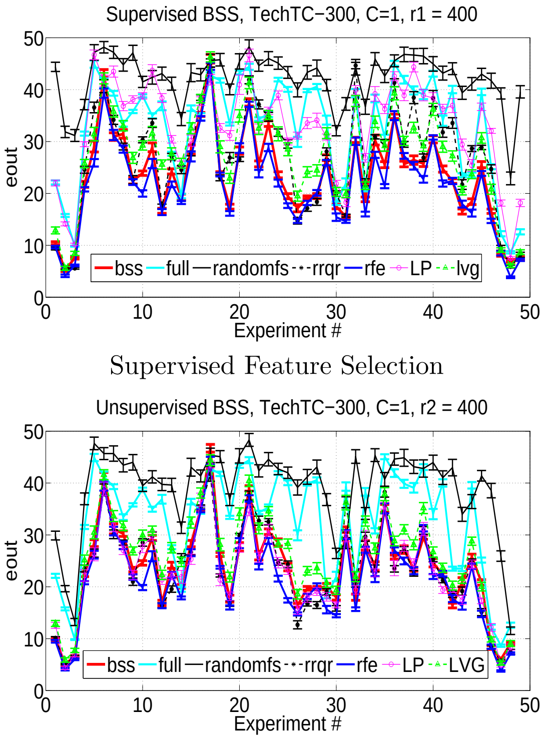
Random Feature Selection: We select features uniformly at random without replacement which serves as a baseline method. To get around the randomness, we repeat the sampling process five times.

Recursive Feature Elimination: Recursive Feature Elimination (RFE), Guyon et al. (2002) tries to find the best subset of features which leads to the largest margin of class separation using SVM. At each iteration, the algorithm greedily removes the feature that decreases the margin the least, until the required number of features remain. At each step, it computes the weight vector and removes the feature with smallest weight. RFE is computationally expensive for high-dimensional datasets. Therefore, at each iteration, multiple features are removed to avoid the computational bottleneck.

LPSVM: The feature selection problem for SVM can be formulated in the form of a linear program. LPSVM Fung and Mangasarian (2004) uses a fast Newton method to solve this problem and obtains a sparse solution of the weight vector, which is used to select the features.

```
Input: Support vector matrix $\mathbf{X} \in \mathbb{R}^{p \times d}$, $t, r$.
Output: Matrices $\mathbf{S} \in \mathbb{R}^{d \times r}, \mathbf{D} \in \mathbb{R}^{r \times r}$.

  1. Generate a random Gaussian matrix,
     $\mathbf{G} \in \mathbb{R}^{t \times p}$.

  2. Compute $\hat{\mathbf{X}} = \mathbf{G}\mathbf{X}$.

  3. Compute right singular vectors $\mathbf{V}$ of $\hat{\mathbf{X}}$
     using SVD.

  4. Run Algorithm 1 using $\mathbf{V}$ and $r$ as inputs
     and get matrices $\mathbf{S}$ and $\mathbf{D}$ as outputs.

  5. Return $\mathbf{S}$ and $\mathbf{D}$.
```

Algorithm 2: Approximate BSS



Supervised Feature Selection

Unsupervised Feature Selection

Figure 2: Plots of out-of-sample error of Supervised and Unsupervised BSS and leverage-score compared with other methods for 49 TechTC-300 documents averaged over ten ten-fold cross validation experiments. Vertical bars represent standard deviation.