Large-scale semi-supervised learning with online spectral graph sparsification

Daniele Calandriello, Alessandro Lazaric, Michal Valko

Large-scale semi-supervised learning with online spectral graph sparsification

Daniele Calandriello DANIELE.CALANDRIELLO@INRIA.FR
Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR
Michal Valko MICHAL.VALKO@INRIA.FR
Team SequeL INRIA Lille – Nord Europe, France

1. Introduction

In many classification and regression tasks, obtaining many good-quality labeled examples may be expensive. When the number of labeled examples is very small, traditional supervised learning algorithms fail in learning accurate predictors. *Semi-supervised learning* (SSL, Chapelle et al., 2006) deals with this problem by integrating the labeled examples with an additional set of unsupervised samples to make use of an underlying structure (e.g., a manifold) and reduced the need for labeling. In this paper we consider data whose similarity can be encoded in a *graph*, and the similarity between nodes is much easier to obtain than their label. Given the graph, SSL methods leverage the assumption that nodes which are similar according to the graph are more likely to be labeled similarly. Graph-based SSL propagates the labels from the labeled nodes to the unlabeled ones. For instance, the objective of *harmonic function solution* (HFS, Zhu et al., 2003; Belkin et al., 2004) is to find a solution where each node's value is the weighted average of its neighbors. The HFS solution can be found solving a linear system involving the graph Laplacian, but for dense graphs on n nodes this amounts to $\mathcal{O}(n^3)$ time complexity and a $\mathcal{O}(n^2)$ space complexity, which is infeasible for large n. In this paper, we consider a more realistic setting when *space and computational budgets are limited*. In particular, we only allow $\mathcal{O}(n\,\mathrm{polylog}(n))$ space for storing the graph structure and an amortized computational cost of $\mathcal{O}(\mathrm{polylog}(n))$ for each of the edges in the original graph. Notice that these constraints make it even impossible to store the full similarity matrix in memory. To this end we employ efficient online spectral graph sparsification techniques (Kelner & Levin, 2013) to incrementally process the stream. This, coupled with specific solvers for symmetric diagonally dominant (SDD) matrices (Koutis et al., 2011), allow us to never store the whole graph in memory and to control the computational complexity as the number of nodes grows. Using the approximation properties of spectral sparsifiers and with results from algorithmic stability theory (Bousquet & Elisseeff, 2002; Cortes et al., 2008) we will provide theoretical guarantees for the generalization error for this approximation.

2. SSL with Spectral Sparsification

Notation. We denote with lowercase letter a a scalar, with bold lowercase letter $\mathbf{a}$ a vector and with uppercase letters A a matrix. We consider the general transductive setting, where we assume that there exists a dictionary of labeled nodes $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where the nodes are organized over an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with n vertices $\mathcal{V} = \{1, \dots, n\}$ and m edges, and the labels are $y_i \in \mathbb{R}$. Given graphs $\mathcal{G}, \mathcal{A}$ defined on the same vertex set, the graph $\mathcal{G} + \mathcal{A}$ is obtained by adding the weights on the edges of $\mathcal{A}$ to $\mathcal{G}$. In a similar manner we define $\mathcal{G} + e$ for edge e. For $i \in \mathcal{V}$, we denote with χ_i the indicator vector, and with $\mathbf{b}_e$ the vector $\chi_i - \chi_j$. While the algorithm receives information on the features x_i of all nodes, only a limited (random) subset $\mathcal{S}$ of l nodes is actually labeled. The objective of the learning algorithm is to minimize the error on the complementary unlabeled set $\mathcal{T} = \mathcal{D} \backslash \mathcal{S}$. More precisely, the objective is to learn a function $\mathbf{f} : \mathcal{V} \to \mathbb{R}$ that minimizes the generalization error $R(\mathbf{f}) = \frac{1}{u} \sum_{i=1}^{u} (\mathbf{f}(x_i) - \mathbf{y}(x_i))^2$, where $u = |\mathcal{T}| = n - l$ is the number of unlabeled nodes. We indicate with $\mathbf{f}$ and $\mathbf{y} \in \mathbb{R}^n$ the vectors that contain the function and the labels evaluated at the n points x_i.

Stable-HFS. HFS exploits the graph structure to learn functions that predict similar values y for similar nodes. Given a weighted adjacency matrix $A_\mathcal{G}$, with edge weights a_e, and the degree matrix $D_\mathcal{G}$, the Laplacian is defined as $L_\mathcal{G} = D_\mathcal{G} - A_\mathcal{G}$. We assume the graph is connected. In this case, $L_\mathcal{G}$ is semi-definite positive (SDP) with $\mathrm{Ker}(L_\mathcal{G}) = \mathbf{1}$. Let $L_\mathcal{G}^+$ be the pseudoinverse of $L_\mathcal{G}$, and $L_\mathcal{G}^{-1/2} = (L_\mathcal{G}^+)^{1/2}$. The original HFS method (Zhu et al., 2003), when we allow the label of already labeled nodes to change, can be formulated as the Laplacian-regularized least-squares problem

$$\begin{aligned}\widehat{\mathbf{f}} &= \operatorname*{arg\,min}_{\mathbf{f} \in \mathbb{R}^n} \tfrac{1}{l} (\mathbf{f} - \mathbf{y})^\mathsf{T} I_\mathcal{S} (\mathbf{f} - \mathbf{y}) + \gamma \mathbf{f}^\mathsf{T} L_\mathcal{G} \mathbf{f} \\ &= (\gamma l L_\mathcal{G} + I_\mathcal{S})^+ (\mathbf{y}), \end{aligned} \tag{1}$$

where $I_\mathcal{S}$ is the identity matrix with zeros corresponding to the nodes not in $\mathcal{S}$, and γ is a regularizer. While HFS achieves interesting empirical results, it is not easy to provide theoretical guarantees for it due to the singularity of the Laplacian matrix. For this reason, we fo-

Algorithm 1 Sparse-HFS

input $\{x_i : i \in \mathcal{D}\}, \{y_i : i \in \mathcal{S}\}$, a stream of m edges $\mathcal{E}$
output $\widehat{\mathbf{f}}, \mathcal{H}$
  Initialize $\mathcal{H} = \emptyset, \mathcal{A} = \emptyset, t = 1$
  while $t \leq m$ **do**
    for $|\mathcal{A}| \leq n \log^2(n)/\varepsilon^2$ **do**
      Receive edge e_t and add it to $\mathcal{A}$
      $t = t + 1$
    end for
    Compute a new graph $\mathcal{H}$ using Alg. 2 on $\mathcal{H} + \mathcal{A}$
    Build Laplacian $L_{\mathcal{H}}$ and diag. matrix $\{I_{\mathcal{S}}(i,i) = 1 : i \in \mathcal{S}\}$
    Compute HFS $\widetilde{\mathbf{f}}$ using Eq. 1 and $L_{\mathcal{H}}$
    $\widetilde{\mathbf{f}} = \widetilde{\mathbf{f}} - \mu\mathbf{1}$ where μ is computed using Eq. 2
  end while

cus on the stable-HFS algorithm proposed by Belkin et al. (2004) where an additional regularization term is introduced to restrict the space of admissible hypothesis, so that $\mathcal{F} = \{\mathbf{f} : \langle \mathbf{f}, \mathbf{1} \rangle = 0\}$. This restriction can be enforced introducing an additional μ regularization term, that can be computed in closed form as

$$\mu = (\gamma l L_{\mathcal{G}} + I_{\mathcal{S}})^{+}\mathbf{y} / (\gamma l L_{\mathcal{G}} + I_{\mathcal{S}})^{+}\mathbf{1}, \qquad (2)$$

and subtracting $\mu\mathbf{1}$ from the unconstrained solution. It can be shown that this is equivalent to projecting the unregularized solution using the projection matrix $P_{\mathcal{F}} = L_{\mathcal{G}} L_{\mathcal{G}}^{+}$. While stable-HFS is more suited for theoretical analysis, its computational and space requirements remain polynomial. If the graph $\mathcal{G}$ has no particular property, solving the linear system takes $\mathcal{O}(n^2)$ space and $\mathcal{O}(n^3)$ time. To satisfy our resource constraints, we include spectral sparsification in stable-HFS. Computing the solution on a sparse graph $\mathcal{H}$ that approximates $\mathcal{G}$ removes the polynomial complexity.

Algorithm 2 Kelner-Levin Sparsification Algorithm

input $\mathcal{H}, \mathcal{A}$, the previous probabilities $\widetilde{p}_e$ for all edges in $\mathcal{H}$ and the weights of the edges a_e.
output $\mathcal{H}'$, a $1 \pm \varepsilon$ sparsifier of $\mathcal{G}' = \mathcal{G} + \mathcal{A}$ and new prob. $\{\widetilde{p}'_e : e \in \mathcal{H}'\}$.
  $\alpha^2 = 1/(1-\varepsilon)^2, N = \alpha^2 n \log^2(n)/\varepsilon^2$
  Obtain estimates $\{\widetilde{R}'_e : e \in \mathcal{H} + \mathcal{A}\}$ such that
    $1/\alpha \leq \widetilde{R}'_e / R'_e \leq \alpha$ with an SDD solver (Koutis et al., 2011)
  Compute prob. $\widetilde{p}'_e = (a_e \widetilde{R}'_e)/(\alpha(n-1))$ and $w_e = a_e/(N\widetilde{p}'_e)$
  for all edges $e \in \mathcal{H}$ **do**
    $\widetilde{p}'_e \leftarrow \min\{\widetilde{p}_e, \widetilde{p}'_e\}$
  end for
  Initialize $\mathcal{H}' = \emptyset$
  for all edges $e \in \mathcal{H}$ **do**
    with probability $\widetilde{p}'_e / \widetilde{p}_e$ add edge e to $\mathcal{H}'$ with weight w_e
  end for
  for all edges $e \in \mathcal{A}$ **do**
    /*The inner loop is run implicitly by sampling a binomial*/
    for $i = 1$ to N **do**
      with probability $\widetilde{p}'_e$ add edge e to $\mathcal{H}'$ with weight w_e
    end for
  end for

Sparse-HFS. Spectral sparsifiers have been central in the development of efficient linear solvers (Koutis et al., 2011). Since their introduction by Spielman & Teng (2011), they were extended to insertion-only streams (Kelner & Levin, 2013).

Definition 1. *A $1 \pm \varepsilon$ spectral sparsifier of $\mathcal{G}$ is a graph $\mathcal{H} \subseteq \mathcal{G}$ such that for all $\mathbf{x}$*

$$(1-\varepsilon)\mathbf{x}^{\mathsf{T}} L_{\mathcal{G}} \mathbf{x} \leq \mathbf{x}^{\mathsf{T}} L_{\mathcal{H}} \mathbf{x} \leq (1+\varepsilon)\mathbf{x}^{\mathsf{T}} L_{\mathcal{G}} \mathbf{x}$$

In this paper, we propose to spectral sparsify $\mathcal{G}$ to reduce the complexity of HFS. A sparse graph $\mathcal{H}$ can be stored efficiently, but if the construction of the sparsifier requires access to the whole $\mathcal{G}$ graph at every moment, just storing the original graph in memory can be impossible. Moreover, traditional linear solvers for an $n \times n$ matrix with m nonzero entries have a time complexity of $\mathcal{O}(mn)$, which is already infeasible for $m = n$. To meet our space and time requirements, we propose to build the sparsifier incrementally using Alg. 2 (Kelner & Levin, 2013). Sparse-HFS (Alg. 1) receives as input a previous sparsifier $\mathcal{H}$ and a stream of edges insertions (i.e., from a disk or a network) and stores them in memory until a graph $\mathcal{A}$ with $\mathcal{O}(n \operatorname{polylog}(n))$ edges has formed. At this point, the sparsifier $\mathcal{H}$ gets updated, generating a new sparsifier that again occupies only $\mathcal{O}(n \operatorname{polylog}(n))$ space. The key component in generating the sparsifier is random sampling according to the effective resistances. The effective resistance of an edge e is defined as $R_e = \mathbf{b}_e^{\mathsf{T}} L_{\mathcal{H}}^{+} \mathbf{b}_e$. Computing R_e naïvely requires again $\mathcal{O}(n^2 \operatorname{polylog}(n))$ time and is not feasible in general. Using linear solvers for SDD matrices (Koutis et al., 2011) we can get R_e for all edges in $\mathcal{H}$ in $\mathcal{O}(n \operatorname{polylog}(n))$ time. Using the same solver, recomputing the updated solution $\widetilde{\mathbf{f}}$ for the updated sparsifier $\mathcal{H}$ takes the same time. Therefore, the whole update procedure takes $\mathcal{O}(n \operatorname{polylog}(n))$. Note that updating the solution at each step is still possible, but it will not meet the computational budget of a $\mathcal{O}(\operatorname{polylog}(n))$ amortized cost.

3. Theoretical Analysis

By a good approximation of quadratic forms, spectral sparsifiers give many guarantees on eigenvalues, eigenvectors and solutions to linear systems. Let $P_{\mathcal{F}} = L_{\mathcal{G}} L_{\mathcal{G}}^{+}$ be the projection matrix on the $n-1$ dimensional space $\operatorname{Ker}(L_{\mathcal{G}})^{\mathsf{T}} = \mathcal{F}$. We derive a bound on the generalization error for sparse-HFS and compare it to the original stable-HFS. We start with the definition of a stable algorithm.

Definition 2 (Transduction β-stability)**.** *Let $\mathcal{L}$ be a transductive learning algorithm and let $\mathbf{f}$ denote the hypothesis returned by $\mathcal{L}$ for $\mathcal{D} = (\mathcal{S}, \mathcal{T})$ and $\mathbf{f}'$ the hypothesis returned for $\mathcal{D} = (\mathcal{S}', \mathcal{T}')$. $\mathcal{L}$ is uniformly β-stable with respect to the squared loss if there exists $\beta \geq 0$ such that for any two partitions $\mathcal{D} = (\mathcal{S}, \mathcal{T})$ and $\mathcal{D} = (\mathcal{S}', \mathcal{T}')$ that*

differ in exactly one training (and thus test one) point and for all $x \in \mathcal{D}$,

$$|(\mathbf{f}(x) - \mathbf{y}(x))^2 - (\mathbf{f}'(x) - \mathbf{y}(x))^2| \leq \beta.$$

The analysis of algorithmic stability (Bousquet & Elisseeff, 2002) has been extensively used in statistic for concentration inequalities in the transductive setting (El-Yaniv & Pechyony, 2006) and later for algorithmic guarantees (Cortes et al., 2008). Define the empirical error as $\widehat{R}(\mathbf{f}) = \frac{1}{l}\sum_{i=1}^{l}(\mathbf{f}(x_i) - \mathbf{y}(x_i))^2$ and the generalization error as $R(\mathbf{f}) = \frac{1}{u}\sum_{i=1}^{u}(\mathbf{f}(x_i) - \mathbf{y}(x_i))^2$.

Theorem 1. *Let $|\mathbf{f}(x) - \mathbf{y}(x)| \leq c$ and $|\mathbf{y}(x)| \leq k$ for all $x \in \mathcal{D}, \mathbf{f} \in \mathcal{F}$. Let $\widetilde{\mathbf{f}}$ be the hypothesis returned by sparse-HFS (Alg. 1) when trained on $\mathcal{D} = (\mathcal{S}, \mathcal{T})$, and $\widehat{\mathbf{f}}$ the solution returned by stable-HFS. Then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$R(\widetilde{\mathbf{f}}) \leq \widehat{R}(\widehat{\mathbf{f}}) + \frac{l^2\gamma^2\lambda_n^2 k^2 \varepsilon^2}{(l\gamma(1-\varepsilon)\lambda_1 - 1)^4} + \beta + \left(2\beta + \frac{c^2(l+u)}{lu}\right)\sqrt{\frac{\pi(l,u)\ln\frac{1}{\delta}}{2}},$$

where

$$\pi(l,u) = \frac{lu}{l+u-0.5}\frac{1}{1 - 1/(2\max\{l,u\})},$$

and

$$\beta \leq \frac{1.5k\sqrt{l}}{(l\gamma(1-\varepsilon)\lambda_1 - 1)^2} + \frac{\sqrt{2}k}{l\gamma(1-\varepsilon)\lambda_1 - 1}.$$

Theorem 1 shows how approximating $\mathcal{G}$ with $\mathcal{H}$ impacts the generalization error as the number of labeled samples l increases. If we compare the bound to the exact case ($\varepsilon = 0$), we see that for a fixed ε the rate of convergence remains unchanged. The first term $\varepsilon^2/l^2(1-\varepsilon)^4$ captures the increase of the empirical error due to the approximation. Since for a fixed ε this term scales as $1/l^2$, it is shadowed by the β term. The β term itself preserves the same order of convergence, and is only multiplied by a constant due to the presence of $(1-\varepsilon)$. In conclusion, for a fixed ε the approximated algorithm provides guarantees of the same order as the exact one. This allows us to freely choose ε to tradeoff precision and computational complexity.

Proof. **Step 1 (generalization of stable algorithms).** When $\mathcal{L}$ is a transductive algorithm with stability β, then for any $\delta > 0$, with probability at least $1 - \delta$ (w.r.t. the randomness of the partition of the graph in labeled and unlabeled sets $\mathcal{S}, \mathcal{T}$) the hypotesis $\widetilde{\mathbf{f}}$ returned by the algorithm satisfies

$$R(\widetilde{\mathbf{f}}) \leq \widehat{R}(\widetilde{\mathbf{f}}) + \beta + \left(2\beta + \frac{c^2(l+u)}{lu}\right)\sqrt{\frac{\pi(l,u)\ln\frac{1}{\delta}}{2}},$$

hence it is sufficient to study the stability of sparse-HFS and relate its empirical loss to the result of stable-HFS to obtain the final result.

Step 2 (stability). Let $\mathcal{S}$ and $\mathcal{S}'$ be two different realizations differing only in one label. For simplicity we will assume that $I_{\mathcal{S}}(l,l) = 1$ and $I_{\mathcal{S}}(l+1,l+1) = 0$, and the opposite for $I_{\mathcal{S}'}$. The original proof (Cortes et al., 2008) showed that for the two hypotheses returned by stable-HFS, $\beta \leq \|\widehat{\mathbf{f}} - \widehat{\mathbf{f}}'\|$. Similarly, for our algorithm $\beta \leq \|\widetilde{\mathbf{f}} - \widetilde{\mathbf{f}}'\|$. All that is left is to upper bound the norm. The spectral radius of $I_{\mathcal{S}}$ is 1. On the other hand, while $\lambda_0 = 0$, the smallest eigenvalue of $L_{\mathcal{H}}$ restricted to $\mathcal{F}$ is λ_1. This reduction of the spectral radius of the Laplacian over the restricted space $\mathcal{F}$ plays a critical role in the proof, and motivates the choice of this particular constraint. Let $\mathbf{y}_{\mathcal{S}} = I_{\mathcal{S}}\mathbf{y}$, $A = P_{\mathcal{F}}(l\gamma L_{\mathcal{H}} + I_{\mathcal{S}})$ and $B = P_{\mathcal{F}}(l\gamma L_{\mathcal{H}} + I_{\mathcal{S}'})$. The hypotheses $\widetilde{\mathbf{f}}$ and $\widetilde{\mathbf{f}}'$ returned by sparse-HFS are given by $\widetilde{\mathbf{f}} = A^{-1}\mathbf{y}_{\mathcal{S}}$ and $\widetilde{\mathbf{f}}' = B^{-1}\mathbf{y}_{\mathcal{S}'}$. We have

$$\begin{aligned}\widetilde{\mathbf{f}} - \widetilde{\mathbf{f}}' &= A^{-1}\mathbf{y}_{\mathcal{S}} - B^{-1}\mathbf{y}_{\mathcal{S}'} \\ &= A^{-1}(\mathbf{y}_{\mathcal{S}} - \mathbf{y}_{\mathcal{S}'}) + A^{-1}\mathbf{y}_{\mathcal{S}'} - B^{-1}\mathbf{y}_{\mathcal{S}'}\end{aligned}$$

Therefore,

$$\|\widetilde{\mathbf{f}} - \widetilde{\mathbf{f}}'\| \leq \|A^{-1}(\mathbf{y}_{\mathcal{S}} - \mathbf{y}_{\mathcal{S}'})\| + \|A^{-1}\mathbf{y}_{\mathcal{S}'} - B^{-1}\mathbf{y}_{\mathcal{S}'}\|.$$

Noticing that $\mathcal{F}$ is invariant under $L_{\mathcal{H}}$ and that for any vector $P_{\mathcal{F}}$ is an orthogonal projection operator, then by the triangle inequality we immediately obtain that for any $\mathbf{f} \in \mathcal{F}$

$$\begin{aligned}\|P_{\mathcal{F}}(l\gamma L_{\mathcal{H}} + I_{\mathcal{S}})\mathbf{f}\| &\geq \|P_{\mathcal{F}} l\gamma L_{\mathcal{H}}\mathbf{f}\| - \|P_{\mathcal{F}} I_{\mathcal{S}}\mathbf{f}\| \\ &\geq (l\gamma(1-\varepsilon)\lambda_1 - 1)\|\mathbf{f}\|\end{aligned}$$

It follows that the spectral radius of the inverse operator $(P_{\mathcal{F}}(l\gamma L_{\mathcal{H}} + I_{\mathcal{S}}))^{-1}$ and therefore of A^{-1} and B^{-1} does not exceed $1/(l\gamma(1-\varepsilon)\lambda_1 - 1)$ when restricted to $\mathcal{F}$ (the inverse is not even defined outside of $\mathcal{F}$). This together with $\|\mathbf{y}_{\mathcal{S}} - \mathbf{y}_{\mathcal{S}'}\| \leq \sqrt{2}k$ gives us

$$\|A^{-1}(\mathbf{y}_{\mathcal{S}} - \mathbf{y}_{\mathcal{S}'})\| \leq \frac{\sqrt{2}k}{l\gamma(1-\varepsilon)\lambda_1 - 1}$$

On the other hand, it can be checked that $\|\mathbf{y}_{\mathcal{S}'}\| \leq \sqrt{l}k$. Noticing that the spectral radius of $P_{\mathcal{F}}(I_{\mathcal{S}} - I_{\mathcal{S}'})$ cannot exceed $\sqrt{2} < 1.5$, we obtain:

$$\begin{aligned}&\|A^{-1}\mathbf{y}_{\mathcal{S}'} - B^{-1}\mathbf{y}_{\mathcal{S}'}\| = \|B^{-1}(B-A)A^{-1}\mathbf{y}_{\mathcal{S}'}\| \\ &= \|B^{-1}P_{\mathcal{F}}(I_{\mathcal{S}} - I_{\mathcal{S}'})A^{-1}\mathbf{y}_{\mathcal{S}'}\| \leq \frac{1.5k\sqrt{l}}{(l\gamma(1-\varepsilon)\lambda_1 - 1)^2}\end{aligned}$$

Putting it all together

$$\|\widetilde{\mathbf{f}} - \widetilde{\mathbf{f}}'\| \leq \frac{1.5k\sqrt{l}}{(l\gamma(1-\varepsilon)\lambda_1 - 1)^2} + \frac{\sqrt{2}k}{l\gamma(1-\varepsilon)\lambda_1 - 1}$$

Step 3 (empirical error). We can now proceed with the proof of Thm. 1. We have already bounded β when using $\mathcal{H}$ instead of $\mathcal{G}$. We can also provide guarantees for the difference in the empirical error using the sparsifier. Given $\widetilde{Q} = P_{\mathcal{F}}(l\gamma L_{\mathcal{H}} + I_{\mathcal{S}})$, $\widehat{Q} = P_{\mathcal{F}}(l\gamma L_{\mathcal{G}} + I_{\mathcal{S}})$ we have

$$\begin{aligned}
\widehat{R}(\widetilde{\mathbf{f}}) &= \frac{1}{l}\sum_{i=1}^{l}\left(\widetilde{\mathbf{f}}(x_i) - \mathbf{y}(x_i)\right)^2 \\
&= \tfrac{1}{l}\|I_{\mathcal{S}}\widetilde{\mathbf{f}} - I_{\mathcal{S}}\widehat{\mathbf{f}} + I_{\mathcal{S}}\widehat{\mathbf{f}} - \mathbf{y}_{\mathcal{S}}\|^2 \\
&\leq \tfrac{1}{l}\|I_{\mathcal{S}}\widehat{\mathbf{f}} - \mathbf{y}_{\mathcal{S}}\|^2 + \frac{1}{l}\|I_{\mathcal{S}}\widetilde{\mathbf{f}} - I_{\mathcal{S}}\widehat{\mathbf{f}}\|^2 \\
&\leq \widehat{R}(\widehat{\mathbf{f}}) + \tfrac{1}{l}\|I_{\mathcal{S}}(\widetilde{Q}^{-1} - \widehat{Q}^{-1})\mathbf{y}_{\mathcal{S}}\|^2 \\
&\leq \widehat{R}(\widehat{\mathbf{f}}) + \tfrac{1}{l}\|\widehat{Q}^{-1}(\widehat{Q} - \widetilde{Q})\widetilde{Q}^{-1}\mathbf{y}_{\mathcal{S}}\|^2 \\
&\leq \widehat{R}(\widehat{\mathbf{f}}) + \frac{lk^2}{l(l\gamma(1-\varepsilon)\lambda_1 - 1)^4}\|\widehat{Q} - \widetilde{Q}\|^2
\end{aligned}$$

We now need to bound $\|\widehat{Q} - \widetilde{Q}\|^2 = \|P_{\mathcal{F}} l\gamma(L_{\mathcal{G}} - L_{\mathcal{H}})\|^2$. Let $y = L_{\mathcal{G}}^{1/2}x$ and $\widetilde{P}_{\mathcal{F}} = L_{\mathcal{G}}^{-1/2}L_{\mathcal{H}}L_{\mathcal{G}}^{-1/2}$. Def. 1 implies

$$(1-\varepsilon)P_{\mathcal{F}} \leq \widetilde{P}_{\mathcal{F}} \leq (1+\varepsilon)P_{\mathcal{F}}.$$

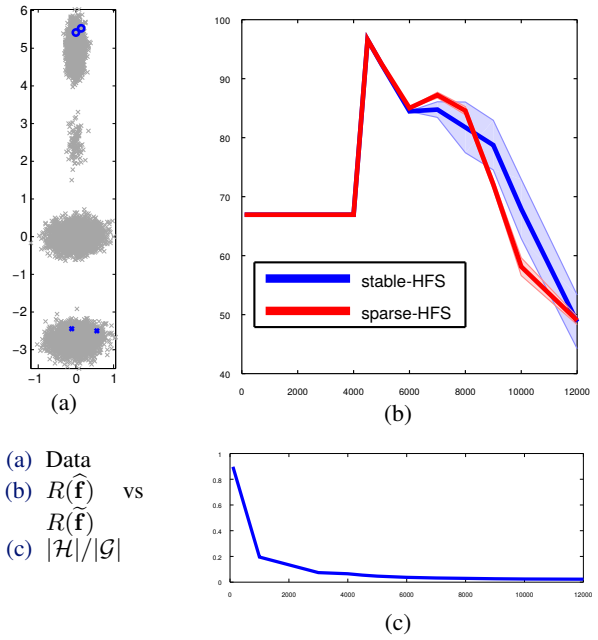
Since $\mathcal{H}$ is a sparsifier of $\mathcal{G}$, by definition we get

$$\begin{aligned}
\|P_{\mathcal{F}} l\gamma(L_{\mathcal{G}} - L_{\mathcal{H}})\|^2 &\leq l^2\gamma^2\|L_{\mathcal{G}} - L_{\mathcal{H}}\|^2 \\
&\leq l^2\gamma^2\|L_{\mathcal{G}}^{1/2}(P_{\mathcal{F}} - \widetilde{P}_{\mathcal{F}})L_{\mathcal{G}}^{1/2}\|^2 \leq l^2\gamma^2\lambda_n^2\varepsilon^2.
\end{aligned}$$

The statement of the theorem is obtained by the combination of the above. □

4. Experiments

We evaluate the proposed algorithm on the $\mathbb{R}^2$ data reported in Fig. (a) which we designed to show the effect of the spectral sparsification in the case when a rather dense graph is needed for a good performance. The dataset is composed of $n = 12100$ points, where the two upper clusters belong to one class and the two lower to the other. We build an unweighted, k-nn graph $\mathcal{G}$ for $k = 100, \ldots, 12000$. This gives us values for m ranging from 1.21×10^6 to 1.38×10^8 edges. After constructing the graph, we randomly select two points from the uppermost and and two from lowermost cluster as our labeled set $\mathcal{S}$. We then run sparse-HFS to compute $\mathcal{H}$ and $\widetilde{\mathbf{f}}$, and run stable-HFS on $\mathcal{G}$ to compute $\widehat{\mathbf{f}}$, both with $\gamma = 1$. For sparse-HFS we set $\varepsilon = 0.8$ Using the labels in $\mathcal{T}$ we compute the generalization error R, which corresponds to the accuracy. Fig. (b) reports the performance of the two algorithms. Both algorithms fail to recover a good solution until $k > 4000$. This is due to the fact that until a certain threshold of neighbours is not surpassed, each cluster remains separated and the labels cannot propagate. Even after this threshold, sparse-HFS cannot consistently outperform stable-HFS in accuracy. This



(a) (b) (c)

(a) Data
(b) $R(\widehat{\mathbf{f}})$ vs $R(\widetilde{\mathbf{f}})$
(c) $|\mathcal{H}|/|\mathcal{G}|$

is because they are both trying to approximate the stable-HFS solution, but sparse-HFS uses an approximated matrix $\mathcal{H}$. Nonetheless, the difference in performance is not large, especially near the optimum. This is in line with the theoretical analysis that shows that the contribution due to the approximation error has the same order of magnitude as the other elements in the bound. Furthermore, in Fig. (c) we report the ratio of the number of edges in the sparsifier $\mathcal{H}$ over the number of edges in the orginal graph $\mathcal{G}$. Since $\mathcal{H} \subseteq \mathcal{G}$, this quantity is always smaller than one, but we can see that for $k = 4500$, where the accuracy is at its maximum, the sparsifier contains only about 10% as many edges as the original graph, with similar accuracy.

5. Conclusions and Future Work

We introduced Sparse-HFS, a scalable algorithm that can compute solutions to SSL problems using only $\mathcal{O}(n\,\mathrm{polylog}(n))$ space and $\mathcal{O}(m\,\mathrm{polylog}(n))$ time. This is achieved in the semi-streaming setting, where a stream of edges insertion is presented to the algorithm. Extending this approach to also deal with edge removals in the stream may not be trivial. The approach taken in (Kapralov et al., 2014) resorts to sketches to keep track of all the updates, but this implicit representation requires $\mathcal{O}(n^2\,\mathrm{polylog}(n))$ time to compute the final SSL solution. In the large scale setting we target an $\mathcal{O}(n^2)$ operation is too costly to meet our amortized cost, so we limit our attention to insertion-only streaming setting. Extending sparsification techniques to the full dynamic setting in a computationally efficient manner is an interesting open problem.

Acknowledgments We would like to thank Ioannis Koutis for many useful discussions.

References

Belkin, Mikhail, Matveeva, Irina, and Niyogi, Partha. Regularization and semi-supervised learning on large graphs. In *Learning theory*, pp. 624–638. Springer, 2004.

Bousquet, Olivier and Elisseeff, André. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Chapelle, O, Schölkopf, B, and Zien, A (eds.). *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

Cortes, Corinna, Mohri, Mehryar, Pechyony, Dmitry, and Rastogi, Ashish. Stability of transductive regression algorithms. In *Proceedings of the 25th international conference on Machine learning*, pp. 176–183. ACM, 2008.

El-Yaniv, Ran and Pechyony, Dmitry. Stable transductive learning. In *Learning theory*, pp. 35–49. Springer, 2006.

Kapralov, Michael, Lee, Yin Tat, Musco, Cameron, Musco, Christopher, and Sidford, Aaron. Single Pass Spectral Sparsification in Dynamic Streams. *arXiv:1407.1289 [cs]*, July 2014. arXiv: 1407.1289.

Kelner, Jonathan A. and Levin, Alex. Spectral Sparsification in the Semi-streaming Setting. *Theory of Computing Systems*, 53(2):243–262, August 2013. ISSN 1432-4350, 1433-0490.

Koutis, Ioannis, Miller, Gary L., and Peng, Richard. Solving SDD linear systems in timeO (mlognlog (1/ϙ)). *arXiv preprint arXiv:1102.4842*, 2011.

Spielman, Daniel A. and Teng, Shang-Hua. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4): 981–1025, 2011.

Zhu, Xiaojin, Ghahramani, Zoubin, Lafferty, John, and others. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pp. 912–919, 2003.