Stat260/CS294: Spectral Graph Methods **Lecture 21 - 04/09/2015**

Lecture: Local Spectral Methods (4 of 4)

Lecturer: Michael Mahoney *Scribe: Michael Mahoney*

Warning: these notes are still very rough. They provide more details on what we discussed in class, but there may still be some errors, incomplete/imprecise statements, etc. in them.

21 Strongly and weakly locally-biased graph partitioning

Last time we introduced an objective function (LocalSpectral) that looked like the usual global spectral partitioning problem, except that it had a locality constraint, and we showed that its solution is of the form of a PPR vector. Today, we will do two things.

- We will introduce a locally-biased graph partitioning problem, we show that the solution to LocalSpectral can be used to compute approximate solutions to that problem.
- We describe the relationship between this problem and what the strongly-local spectral methods, e.g., the ACL push method, compute.

21.1 Locally-biased graph partitioning

We start with a definition.

Definition 1 (Locally-biased graph partitioning problem.)**.** *Given a graph $G = (V, E)$, an input node $u \in V$, a number $k \in \mathbb{Z}^+$, find a set of nodes $T \subset V$ s.t.*

$$\phi(u, k) = \min_{T \subset V : u \in T,\, Vol(T) \leq k} \phi(T),$$

i.e., find the best conductance set of nodes of volume not greater than k that contains the node u.

That is, rather than look for the best conductance cluster in the entire graph (which we considered before), look instead for the best conductance cluster that contains a specified seed node and that is not too large.

Before proceeding, let's state a version of Cheeger's Inequality that applies not just to the leading nontrivial eigenvector of L but instead to *any* "test vector."

Theorem 1. *Let $x \in \mathbb{R}^n$ s.t. $x^T D\vec{1} = 0$. Then there exists a $t \in [n]$ such that $S \equiv SweepCut_t(x) \equiv \{i : x_i \geq t\}$ satisfies $\frac{x^T L x}{x^T D x} \geq \frac{\phi(S)^2}{8}$.*

Remark. This form of Cheeger's Inequality provides additional flexibility in at least two ways. First, if one has computed an approximate Fiedler vector, e.g., by running a random walk many

steps but not quite to the asymptotic state, then one can appeal to this result to show that Cheeger-like guarantees hold for that vector, i.e., one can obtain a "quadratically-good" approximation to the global conductance objective function using that vector. Alternatively, one can apply this to *any* vector, e.g., a vector obtained by running a random walk just a few steps from a localized seed node. This latter flexibility makes this form of Cheeger's Inequality very useful for establishing bounds with both strongly and weakly local spectral methods.

Let's also recall the objective with which we are working; we call it $\mathsf{LocalSpectral}(G, s, \kappa)$ or LocalSpectral. Here it is.

$$\begin{aligned} \min \quad & x^T L_G x \\ \text{s.t.} \quad & x^T D_G x = 1 \\ & (x^T D_G 1)^2 = 0 \\ & (x^T D_G s)^2 \geq \kappa \\ & x \in \mathbb{R}^n \end{aligned}$$

Let's start with our first result, which says that LocalSpectral is a relaxation of the intractable combinatorial problem that is the locally-biased version of the global spectral paritioning problem (in a manner analogous to how the global spectral partitioning problem is a relaxation of the intractable problem of finding the best conductance partition in the entire graph). More precisely, we can choose the seed set s and correlation parameter κ such that $\mathsf{LocalSpectral}(G, s, \kappa)$ is a relaxation of the problem defined in Definition 1.

Theorem 2. *For $u \in V$, $\mathsf{LocalSpectral}(G, v_{\{u\}}, 1/k)$ is a relaxation of the problem of finding a minimum conductance cut T in G which contains the vertex u and is of volume at most k. In particular, $\lambda(G, v_{\{u\}}, 1/k) \leq \phi(u, k)$.*

Proof. If we let $x = v_T$ in $\mathsf{LocalSpectral}(G, v_{\{u\}}, 1/k)$, then $v_T^T L_G v_T = \phi(T)$, $v_T^T D_G 1 = 0$, and $v_T^T D_G v_T = 1$. Moreover, we have that

$$(v_T^T D_G v_{\{u\}})^2 = \frac{d_u(2m - \mathrm{vol}(T))}{\mathrm{vol}(T)(2m - d_u)} \geq 1/k,$$

which establishes the lemma. □

Next, let's apply sweep cut rounding to get locally-biased cuts that are quadratically good, thus establishing a locally-biased analogue of the hard direction of Cheeger's Inequality for this problem. In particular, we can apply Theorem 1 to the optimal solution for $\mathsf{LocalSpectral}(G, v_{\{u\}}, 1/k)$ and obtain a cut T whose conductance is quadratically close to the optimal value $\lambda(G, v_{\{u\}}, 1/k)$. By Theorem 2, this implies that $\phi(T) \leq O(\sqrt{\phi(u, k)})$, which essentially establishes the following theorem.

Theorem 3 (Finding a Cut)**.** *Given an unweighted graph $G = (V, E)$, a vertex $u \in V$ and a positive integer k, we can find a cut in G of conductance at most $O(\sqrt{\phi(u, k)})$ by computing a sweep cut of the optimal vector for $\mathsf{LocalSpectral}(G, v_{\{u\}}, 1/k)$.*

Remark. What this theorem states is that we can perform a sweep cut over the vector that is the solution to $\mathsf{LocalSpectral}(G, v_{\{u\}}, 1/k)$ in order to obtain a locally-biased partition; and that

this partition comes with quality-of-approximation guarantees analogous to that provided for the global problem $\mathsf{Spectral}(G)$ by Cheeger's inequality.

We can also use the optimal value of $\mathsf{LocalSpectral}$ to provide lower bounds on the conductance value of other cuts, as a function of how well-correlated they are with the input seed vector s. In particular, if the seed vector corresponds to a cut U, then we get lower bounds on the conductance of other cuts T in terms of the correlation between U and T.

Theorem 4 (Cut Improvement)**.** *Let G be a graph and $s \in \mathbb{R}^n$ be such that $s^T D_G 1 = 0$, where D_G is the degree matrix of G. In addition, let $\kappa \geq 0$ be a correlation parameter. Then, for all sets $T \subseteq V$ such that $\kappa' \stackrel{\text{def}}{=} (s^T D_G v_T)^2$, we have that*

$$\phi(T) \geq \begin{cases} \lambda(G, s, \kappa) & \text{if } \kappa \leq \kappa' \\ \frac{\kappa'}{\kappa} \cdot \lambda(G, s, \kappa) & \text{if } \kappa' \leq \kappa. \end{cases}$$

In particular, if $s = s_U$ for some $U \subseteq V$, then note that $\kappa' = K(U, T)$.

Proof. It follows from the results that we established in the last class that $\lambda(G, s, \kappa)$ is the same as the optimal value of $\mathsf{SDP}_p(G, s, \kappa)$ which, by strong duality, is the same as the optimal value of $\mathsf{SDP}_d(G, s, \kappa)$. Let $\alpha^\star, \beta^\star$ be the optimal dual values to $\mathsf{SDP}_d(G, s, \kappa)$. Then, from the dual feasibility constraint $L_G - \alpha^\star L_{K_n} - \beta^\star (D_G s)(D_G s)^T \succeq 0$, it follows that

$$s_T^T L_G s_T - \alpha^\star s_T^T L_{K_n} s_T - \beta^\star (s^T D_G s_T)^2 \geq 0.$$

Notice that since $s_T^T D_G 1 = 0$, it follows that $s_T^T L_{K_n} s_T = s_T^T D_G s_T = 1$. Further, since $s_T^T L_G s_T = \phi(T)$, we obtain, if $\kappa \leq \kappa'$, that

$$\phi(T) \geq \alpha^\star + \beta^\star (s^T D_G s_T)^2 \geq \alpha^\star + \beta^\star \kappa = \lambda(G, s, \kappa).$$

If on the other hand, $\kappa' \leq \kappa$, then

$$\phi(T) \geq \alpha^\star + \beta^\star (s^T D_G s_T)^2 \geq \alpha^\star + \beta^\star \kappa \geq \frac{\kappa'}{\kappa} \cdot (\alpha^\star + \beta^\star \kappa) = \frac{\kappa'}{\kappa} \cdot \lambda(G, s, \kappa).$$

Finally, observe that if $s = s_U$ for some $U \subseteq V$, then $(s_U^T D_G s_T)^2 = K(U, T)$. Note that strong duality was used here. □

Remark. We call this result a "cut improvement" result since it is the spectral analogue of the flow-based "cut improvement" algorithms we mentioned when doing flow-based graph partitioning.

- These flow-based cut improvement algorithms were originally used as a post-processing algorithm to improve partitions found by other algorithms. For example, GGT, LR (Lang-Rao), and AL (which we mentioned before).
- They provide guarantees of the form: for any cut $\left(C, \bar{C}\right)$ that is ϵ-correlated with the input cut, the cut output by the cut improvement algorithm has conductance $\leq$ some function of the conductance of $\left(C, \bar{C}\right)$ and ϵ.
- Theorem 4 shows that, while the cut value output by this spectral-based "improvement" algorithm might *not* be improved, relative to the input, as they are often guaranteed to do with flow-based cut-improvement algorithms, they do not decrease in quality too much, and in addition one can make claims about the cut quality of "nearby" cuts.

- Although we don't have time to discuss it, these two operations can be viewed as building blocks or "primitives" that can be combined in various ways to develop algorithms for other problems, e.g., finding minimum conductance cuts.

21.2 Relationship between strongly and weakly local spectral methods

So far, we have described two different ways to think about local spectral algorithms.

- **Operational.** This approach provides an algorithm, and one can prove locally-biased Cheeger-like guarantees. The exact statement of these results is quite complex, but the running time of these methods is extremely fast since they don't even need to touch all the nodes of a big graph.
- **Optimization.** This approach provides a well-defined optimization objective, and one can prove locally-biased Cheeger-like guarantees. The exact statement of these results is much simpler, but the running time is only moderately fast, since it involves computing eigenvectors or linear equations on sparse graphs, and this involves at least touching all the nodes of a big graph.

An obvious question here is the following.

- Shat is the precise relationship between these two approaches?

We'll answer this question by considering the weakly-local LocalSpectral optimization problem (that we'll call MOV below) and the PPR-based local spectral algorithm due to ACL (that we'll call ACL below). What we'll show is roughly the following.

- We'll show roughly that if MOV optimizes an ℓ_2 based penalty, then ACL optimizes an ℓ_1-regularized version of that ℓ_2 penalty.

That's interesting since ℓ_1 regularization is often introduced to enforce or encourage sparsity. Of course, there is no ℓ_1 regularization in the statement of the strongly local spectral methods like ACL, but clearly they enforce some sort of sparsity, since they don't even touch most of the nodes of a large graph. Thus, this result can be interpreted as providing an implicit regularization characterization of a fast approximation algorithm.

21.3 Setup for implicit ℓ_1 regularization in strongly local spectral methods

Recall that $L = D - A = B^T C B$, where B is the unweighted edge-incidence matrix. Then

$$\|Bx\|_{C,1} = \sum_{(ij)\in E} C_{(ij)}|x_i - x_j| = \text{cut}(S),$$

where $S = \{i : x_i = 1\}$. In addition, we can obtain a spectral problem by changing $\|\cdot\|_1 \to \|\cdot\|_2$ to get

$$\|Bx\|_{C,2}^2 = \sum_{(ij)\in E} C_{(ij)}\,(x_i - x_j)^2$$

Let's consider a specific (s,t)-cut problem that is inspired by the AL FlowImprove procedure. To do so, fix a set of vertices (like we did when we did the semi-supervised eigenvector construction), and define a *new* graph that we will call the "localized cut graph." Basically, this new graph will be the original graph augmented with two additional nodes, call them s and t, that are connected by weights to the nodes of the original graph. Here is the definition.

Definition 2 (localized cut graph)**.** *Let $G = (V, E)$ be a graph, let S be a set of vertices, possibly empty, let $\bar{S}$ be the complement set, and let α be a non-negative constant. Then the* localized cut graph *is the weighted, undirected graph with adjacency matrix:*

$$A_S = \begin{bmatrix} 0 & \alpha d_S^T & 0 \\ \alpha d_S & A & \alpha d_{\bar{S}} \\ 0 & \alpha d_{\bar{S}}^T & 0 \end{bmatrix}$$

where $d_S = De_S$ is a degree vector localized on the set S, A is the adjacency matrix of the original graph G, and $\alpha \geq 0$ is a non-negative weight. Note that the first vertex is s and the last vertex is t.

We'll use the α and S parameter to denote the matrices for the localized cut graph. For example, the *incidence matrix* $B(S)$ of the localized cut graph, which depends on the set S, is given by the following.

$$B(S) = \begin{bmatrix} e & -I_S & 0 \\ 0 & B & 0 \\ 0 & -I_{\bar{S}} & e \end{bmatrix},$$

where, recall, the variable I_S are the columns of the identity matrix corresponding to vertices in S. The edge-weights of the localized cut graph are given by the diagonal matrix $C(\alpha)$, which depends on the value α.

Given this, recall that the 1-norm formulation of the LP for the min-s,t-cut problem, i.e., the minimum weighted s,t cut in the flow graph, is given by the following.

$$\begin{aligned} &\min && \|Bx\|_{C(\alpha),1} \\ &\text{s.t. } x_s = 1, x_t = 0, x \geq 0. \end{aligned}$$

Here is a theorem that shows that PageRank implicitly solves a 2-norm variation of the 1-norm formulation of the s,t-cut problem.

Theorem 5. *Let $B(S)$ be the incidence matrix for the localized cut graph, and $C(\alpha)$ be the edge-weight matrix. The PageRank vector z that solves*

$$(\alpha D + L)z = \alpha v$$

with $v = d_S/\mathrm{vol}(S)$ is a renormalized solution of the 2-norm cut computation:

$$\begin{aligned} &min && \|B(S)x\|_{C(\alpha),2} \\ &s.t. && x_s = 1, x_t = 0. \end{aligned} \tag{1}$$

Specifically, if $x(\alpha, S)$ is the solution of Prob. (1)*, then*

$$x(\alpha, S) = \begin{bmatrix} 1 \\ \mathrm{vol}(S)z \\ 0 \end{bmatrix}.$$

Proof. The key idea is that the 2-norm problem corresponds with a quadratic objective, which PageRank solves. The quadratic objective for the 2-norm approximate cut is:

$$\begin{aligned}\|B(S)x\|^2_{C(\alpha),2} &= x^T B(S)^T C(\alpha) B(S) x \\ &= x^T \begin{bmatrix} \alpha\mathrm{vol}(S) & -\alpha d_S^T & 0 \\ -\alpha d_S & L+\alpha D & -\alpha d_{\bar{S}} \\ 0 & -\alpha d_{\bar{S}} & \alpha\mathrm{vol}(\bar{S}) \end{bmatrix} x.\end{aligned}$$

If we apply the constraints that $x_s = 1$ and $x_t = 0$ and let x_G be the free set of variables, then we arrive at the unconstrained objective:

$$\begin{bmatrix} 1 & x_G^T & 0 \end{bmatrix} \begin{bmatrix} \alpha\mathrm{vol}(S) & -\alpha d_S^T & 0 \\ -\alpha d_S & L+\alpha D & -\alpha d_{\bar{S}} \\ 0 & -\alpha d_{\bar{S}} & \alpha\mathrm{vol}(\bar{S}) \end{bmatrix} \begin{bmatrix} 1 \\ x_G \\ 0 \end{bmatrix}$$
$$= x_G^T(L+\alpha D)x_G - 2\alpha x_G^T d_S + \alpha\mathrm{vol}(S).$$

Here, the solution x_G solves the linear system

$$(\alpha D + L)x_G = \alpha d_S.$$

The vector $x_G = \mathrm{vol}(S)z$, where z is the solution of the PageRank problem defined in the theorem, which concludes the proof. □

Theorem 5 essentially says that for each PR problem, there is a related cut/flow problem that "gives rise" to it. One can also establish the reverse relationship that extracts a cut/flow problem from *any* PageRank problem.

To show this, first note that the proof of Theorem 5 works since the edges we added had weights proportional to the degree of the node, and hence the increase to the degree of the nodes was proportional to their current degree. This causes the diagonal of the Laplacian matrix of the localized cut graph to become $\alpha D + D$. This idea forms the basis of our subsequent analysis. For a general PageRank problem, however, we require a slightly more general definition of the localized cut graph, which we call a *PageRank cut graph*. Here is the definition.

Definition 3. *Let $G = (V, E)$ be a graph, and let $s \geq 0$ be a vector such that $d - s \geq 0$. Let s connect to each node in G with weights given by the vector αs, and let t connect to each node in G with weights given by $\alpha(d - s)$. Then the* PageRank cut graph *is the weighted, undirected graph with adjacency matrix:*

$$A(s) = \begin{bmatrix} 0 & \alpha s^T & 0 \\ \alpha s & A & \alpha(d-s) \\ 0 & \alpha(d-s)^T & 0 \end{bmatrix}.$$

We use $B(s)$ to refer to the incidence matrix of this PageRank cut graph. Note that if $s = d_S$, then this is simply the original construction.

With this, we state the following theorem, which is a sort of converse to Theorem 5. The proof is similar to that of Theorem 5 and so it is omitted.

Theorem 6. *Consider any PageRank problem that fits the framework of*

$$(I - \beta P^T)x = (1-\beta)v.$$

The PageRank vector z that solves

$$(\alpha D + L)z = \alpha v$$

is a renormalized solution of the 2-norm cut computation:

$$\begin{aligned} min \quad & \|B(s)x\|_{C(\alpha),2} \\ & x_s = 1, x_t = 0 \end{aligned} \tag{2}$$

with $s = v$. Specifically, if $x(\alpha, S)$ is the solution of the 2-norm cut, then

$$x(\alpha, s) = \begin{bmatrix} 1 \\ z \\ 0 \end{bmatrix}.$$

Two things are worth noting about this result.

- A corollary of this result is the following: if $s = e$, then the solution of a 2-norm cut is a reweighted, renormalized solution of PageRank with $v = e/n$. That is, as a corollary of this approach, the *standard* PageRank problem with $v = e/n$ gives rise to a cut problem where s connects to each node with weight α and t connects to each node v with weight $\alpha(d_v - 1)$.
- This also holds for the semi-supervised learning results we discussed. In particular, e.g., the procedure of Zhou et al. for semi-supervised learning on graphs solves the following:

  $$(I - \beta D^{-1/2} A D^{-1/2})^{-1} Y.$$

  (The other procedures solve a very similar problem.) This is exactly a PageRank equation for a degree-based scaling of the labels, and thus the construction from Theorem 6 is directly applicable.

21.4 Implicit ℓ_1 regularization in strongly local spectral methods

In light of these results, let's now move onto the ACL procedure. We will show a connection between it and an ℓ_1 regularized version of an ℓ_2 objective, as established in Theorem 6. In particular, we will show that the ACL procedure for *approximating* a PPR vector *exactly* computes a hybrid 1-norm 2-norm variant of the min-cut problem. The balance between these two terms (the ℓ_2 term from Problem 2 and an additional ℓ_1 term) has the effect of producing sparse PageRank solutions that also have sparse truncated residuals, and it also provides an interesting connection with ℓ_1-regularized ℓ_2-regression problems.

We start by reviewing the ACL method and describing it in such a way to make these connections easier to establish.

Consider the problem $(I - \beta A D^{-1})x = (1 - \beta)v$, where $v = e_i$ is localized onto a single node. In addition to the PageRank parameter β, the procedure has two parameters: $\tau > 0$ is a accuracy parameter that determines when to stop, and $0 < \rho \leq 1$ is an additional approximation term that we introduce. As $\tau \to 0$, the computed solution x goes to the PPR vector that is non-zero everywhere. The value of ρ has been $1/2$ in most previous implementations of the procedure; and here we present a modified procedure that makes the effect of ρ explicit.

1. $x^{(1)} = 0, r^{(1)} = (1-\beta)e_i$, $k = 1$
2. *while* any $r_j > \tau d_j$ *(where d_j is the degree of node j)*
3. $x^{(k+1)} = x^{(k)} + (r_j - \tau d_j \rho)e_j$
4. $$r_i^{(k+1)} = \begin{cases} \tau d_j \rho & i = j \\ r_i^{(k)} + \beta(r_j - \tau d_j \rho)/d_j & i \sim j \\ r_i^{(k)} & \text{otherwise} \end{cases}$$
5. $k \leftarrow k + 1$

As we have noted previously, one of the important properties of this procedure is that the algorithm maintains the invariant $r = (1-\beta)v - (I - \beta A D^{-1})x$ throughout. For any $0 \leq \rho \leq 1$, this algorithm converges because the sum of entries in the residual always decreases monotonically. At the solution we will have

$$0 \leq r \leq \tau d,$$

which provides an ∞-norm style worst-case *approximation* guarantee to the exact PageRank solution.

Consider the following theorem. In the same way that Theorem 6 establishes that a PageRank vector can be interpreted as optimizing an ℓ_2 objective involving the edge-incidence matrix, the following theorem establishes that, in the case that $\rho = 1$, the ACL procedure to approximate this vector can be interpreted as solving an ℓ_1-regularized ℓ_2 objective. That is, in addition to *approximating* the solution to the objective function that is optimized by the PPR, this algorithm also *exactly* computes the solution to an ℓ_1 regularized version of the same objective.

Theorem 7. *Fix a subset of vertices S. Let x be the output from the ACL procedure with $\rho = 1$, $0 < \beta < 1$, $v = d_S/\mathrm{vol}(S)$, and τ fixed. Set $\alpha = \frac{1-\beta}{\beta}$, $\kappa = \tau \mathrm{vol}(S)/\beta$, and let z_G be the solution on graph vertices of the sparsity-regularized cut problem:*

$$\begin{array}{ll} \mathit{min} & \frac{1}{2}\|B(s)z\|^2_{C(\alpha),2} + \kappa\|Dz\|_1 \\ \mathit{s.t.} & z_s = 1, z_t = 0, z \geq 0, \end{array} \tag{3}$$

where $z = \begin{bmatrix} 1 \\ z_G \\ 0 \end{bmatrix}$ *as above. Then* $x = Dz_G/\mathit{vol}(S)$.

Proof. If we expand the objective function and apply the constraint $z_s = 1, z_t = 0$, then Prob. (3) becomes:

$$\begin{array}{ll} \min & \frac{1}{2} z_G^T(\alpha D + L) z_G - \alpha z_G^T d_S + \alpha^2 \mathrm{vol}(S) + \kappa d^T z_G \\ \text{s.t.} & z_G \geq 0 \end{array} \tag{4}$$

Consider the optimality conditions of this quadratic problem (where s are the Lagrange multipliers):

$$\begin{aligned} 0 &= (\alpha D + L) z_G - \alpha d_{\bar{S}} + \kappa d - s \\ s &\geq 0 \\ z_G &\geq 0 \\ z_G^T s &= 0. \end{aligned}$$

These are both necessary and sufficient because $(\alpha D + L)$ is positive definite. In addition, and for the same reason, the solution is unique.

In the remainder of the proof, we demonstrate that vector x produced by the ACL method satisfies these conditions. To do so, we first translate the optimality conditions to the equivalent PageRank normalization:

$$\begin{aligned} 0 &= (I - \beta AD^{-1})Dz_G/\mathrm{vol}(S) - (1-\beta)d_S/\mathrm{vol}(S) + \beta\kappa/\mathrm{vol}(S)d - \beta s/\mathrm{vol}(S) \\ s &\geq 0 \qquad z_G \geq 0 \qquad z_G^T s = 0. \end{aligned}$$

When the ACL procedure finishes with β, ρ, and τ as in the theorem, the vectors x and r satisfy:

$$\begin{aligned} r &= (1-\beta)v - (I - \beta AD^{-1})x \\ x &\geq 0 \\ 0 &\leq r \leq \tau d = \beta\kappa/\mathrm{vol}(S)d. \end{aligned}$$

Thus, if we set s such that $\beta s/\mathrm{vol}(S) = \beta\kappa/\mathrm{vol}(S)d - r$, then we satisfy the first condition with $x = Dz_G/\mathrm{vol}(S)$. All of these transformations preserve $x \geq 0$ and $z_G \geq 0$. Also, because $\tau d \geq r$, we also have $s \geq 0$. What remains to be shown is $z_G^T s = 0$.

Here, we show $x^T(\tau d - r) = 0$, which is equivalent to the condition $z_G^T s = 0$ because the non-zero structure of the vectors is identical. Orthogonal non-zero structure suffices because $z_G s = 0$ is equivalent to either $x_i = 0$ or $\tau d_i - r_i = 0$ (or both) for all i. If $x_i \neq 0$, then at some point in the execution, the vertex i was chosen at the step $r_j > \tau d_j$. In that iteration, we set $r_i = \tau d_i$. If any other step increments r_i, we must revisit this step and set $r_i = \tau d_i$ again. Then at a solution, $x_i \neq 0$ requires $r_i = \tau d_i$. For such a component, $s_i = 0$, using the definition above. For $x_i = 0$, the value of s_i is irrelevant, and thus, we have $x^T(\tau d - r) = 0$. □

Remark. Finally, a comment about ρ, which is set to 1 in this theorem but equals $1/2$ in most prior uses of the ACL push method. The proof of Theorem 7 makes the role of ρ clear. If $\rho < 1$, then the output from ACL is *not* equivalent to the solution of Prob. (3), i.e., the renormalized solution will *not* satisfy $z_G^T s = 0$; but setting $\rho < 1$, however, *will* compute a solution much more rapidly. It is a nice open problem to get a clean statement of implicit regularization when $\rho < 1$.