Supplement to "Graph Sparsification Approaches for Laplacian Smoothing"

Veeranjaneyulu Sadhanala[1] Yu-Xiang Wang[1,2] Ryan J. Tibshirani[1,2]

[1] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

This document contains proofs, supplementary details, and supplementary experiments for the paper "Graph Sparsification Approaches for Laplacian Smoothing". All section numbers, equation numbers, and figure numbers in this supplementary document are preceded by the letter A, to distinguish them from those from the main paper.

A.1 Proof of Theorem 1

Part (a). By optimality of $\hat{\theta}$ for problem (5),

$$\begin{aligned}\|y-\hat{\theta}\|_2^2 + \lambda \hat{\theta}^T \tilde{L} \hat{\theta} &\leq \|y-\hat{\beta}\|_2^2 + \lambda \hat{\beta}^T \tilde{L} \hat{\beta} \\ &\leq \|y-\hat{\beta}\|_2^2 + \lambda(1+\epsilon)\hat{\beta}^T L \hat{\beta},\end{aligned}$$

where we have used the spectral similarity of $L, \tilde{L}$. Rearranging, we see that

$$\|\hat{\theta}\|_2^2 - \|\hat{\beta}\|_2^2 \leq 2y^T(\hat{\theta}-\hat{\beta}) + \lambda(1+\epsilon)\hat{\beta}^T L \hat{\beta} - \lambda \hat{\theta}^T \tilde{L} \hat{\theta}.$$

Substituting $y = \hat{\beta} + y - \hat{\beta}$ on the right-hand side, and again rearranging,

$$\|\hat{\theta}-\hat{\beta}\|_2^2 \leq 2(y-\hat{\beta})^T(\hat{\theta}-\hat{\beta}) + \lambda(1+\epsilon)\hat{\beta}^T L \hat{\beta} - \lambda \hat{\theta}^T \tilde{L} \hat{\theta}.$$

Using $y - \hat{\beta} = \lambda L \hat{\beta}$ from the stationarity condition for (1),

$$\begin{aligned}\|\hat{\theta}-\hat{\beta}\|_2^2 &\leq 2\lambda \hat{\theta}^T L \hat{\beta} - \lambda(1-\epsilon)\hat{\beta}^T L \hat{\beta} - \lambda \hat{\theta}^T \tilde{L} \hat{\theta} \\ &\leq 2\lambda \|L^{1/2}\hat{\theta}\|_2 \|L^{1/2}\hat{\beta}\|_2 - \lambda(1-\epsilon)\hat{\beta}^T L \hat{\beta} - \lambda \hat{\theta}^T \tilde{L} \hat{\theta} \\ &\leq 2\lambda\sqrt{1+\epsilon}\|\tilde{L}^{1/2}\hat{\theta}\|_2 \|L^{1/2}\hat{\beta}\|_2 - \lambda(1-\epsilon)\hat{\beta}^T L \hat{\beta} - \lambda \hat{\theta}^T \tilde{L} \hat{\theta},\end{aligned}$$

where we have again used the spectral similarity of $L, \tilde{L}$. Now we examine two cases for last line above. If $\sqrt{1+\epsilon}\|\tilde{L}^{1/2}\hat{\theta}\|_2 \leq \|L^{1/2}\hat{\beta}\|_2$, then

$$\|\hat{\theta}-\hat{\beta}\|_2^2 \leq \lambda(1+\epsilon)\hat{\beta}^T L \hat{\beta} - \lambda \hat{\theta}^T \tilde{L} \hat{\theta}.$$

If $\sqrt{1+\epsilon}\|\tilde{L}^{1/2}\hat{\theta}\|_2 > \|L^{1/2}\hat{\beta}\|_2$, then

$$\|\hat{\theta}-\hat{\beta}\|_2^2 \leq \lambda(1+2\epsilon)\hat{\theta}^T \tilde{L} \hat{\theta} - \lambda(1-\epsilon)\hat{\beta}^T L \hat{\beta}.$$

Putting these together, we get the desired final bound.

Proof of (b). Following the proof strategy for part (a), except with the regularization parameter denoted by

λ' for problem (5), we have

$$\begin{aligned}
\|\hat{\theta} - \hat{\beta}\|_2^2 &\leq 2\lambda \hat{\theta}^T L \hat{\beta} + ((1+\epsilon)\lambda' - 2\lambda)\hat{\beta}^T L \hat{\beta} - \lambda' \hat{\theta}^T \tilde{L} \hat{\theta} \\
&\leq 2\lambda \|L^{1/2}\hat{\theta}\|_2 \|L^{1/2}\hat{\beta}\|_2 + ((1+\epsilon)\lambda' - 2\lambda)\hat{\beta}^T L \hat{\beta} - \lambda' \hat{\theta}^T \tilde{L} \hat{\theta} \\
&\leq 2\lambda\sqrt{1+\epsilon} \|\tilde{L}^{1/2}\hat{\theta}\|_2 \|L^{1/2}\hat{\beta}\|_2 + ((1+\epsilon)\lambda' - 2\lambda)\hat{\beta}^T L \hat{\beta} - \lambda' \hat{\theta}^T \tilde{L} \hat{\theta} \\
&= \underbrace{2\lambda\sqrt{1+\epsilon} \|\tilde{L}^{1/2}\hat{\theta}\|_2 \Big(\|L^{1/2}\hat{\beta}\|_2 - \frac{\lambda'}{2\lambda\sqrt{1+\epsilon}} \|\tilde{L}^{1/2}\hat{\theta}\|_2 \Big)}_{a} + ((1+\epsilon)\lambda' - 2\lambda)\hat{\beta}^T L \hat{\beta}.
\end{aligned}$$

We now examine two cases for the first term a on the line above. If $\lambda'/(2\lambda\sqrt{1+\epsilon})\|\tilde{L}^{1/2}\hat{\theta}\|_2 \leq \|L^{1/2}\hat{\beta}\|_2$, then $a \leq (4\lambda^2(1+\epsilon)/\lambda')\hat{\beta}^T L \hat{\beta}$; if $\lambda'/(2\lambda\sqrt{1+\epsilon})\|\tilde{L}^{1/2}\hat{\theta}\|_2 > \|L^{1/2}\hat{\beta}\|_2$, then $a \leq 0$. Therefore,

$$\|\hat{\theta} - \hat{\beta}\|_2^2 \leq \Big(\frac{4\lambda^2(1+\epsilon)}{\lambda'} + (1+\epsilon)\lambda' - 2\lambda \Big) \hat{\beta}^T L \hat{\beta}.$$

An easy calculation shows that the optimal value of λ', making the leading factor above as small as possible, is $\lambda' = 2\lambda$. Plugging this in gives the desired result. □

A.2 Proof of Theorem 2

Let us first consider the univariate logistic function $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = -ax + \log(1 + e^x),$$

for a constant $a \in \mathbb{R}$. It is clear that g is convex, with first and second derivatives

$$g'(x) = -a + \pi(x), \quad g''(x) = \pi(x)(1 - \pi(x)),$$

where $\pi(x) = e^x/(1+e^x)$. If $|x| \leq R$, then $g''(x) \geq \delta(1-\delta)$, where $\delta = \pi(-R)$. Thus by strong convexity of g over the interval $[-R, R]$, we have

$$g(z) - g(x) - g'(x)(z-x) \geq \frac{\delta(1-\delta)}{2}(z-x)^2, \quad \text{for all } x, z \in [-R, R].$$

or equivalently, by the monotonicity of the inverse link function π,

$$g(z) - g(x) - g'(x)(z-x) \geq \frac{\delta(1-\delta)}{2}(z-x)^2, \quad \text{for all } \pi(x), \pi(z) \in [\delta, 1-\delta].$$

Now write the logistic loss in (3) as $f(\beta) = \sum_{i=1}^n g(\beta_i; y_i)$; then the above shows that

$$f(\theta) - f(\beta) - \nabla f(\beta)^T(\theta - \beta) \geq \frac{\delta(1-\delta)}{2}\|\theta - \beta\|_2^2, \quad \text{whenever } \pi(\beta_i), \pi(\theta_i) \in [\delta, 1-\delta],\ i = 1, \ldots n. \quad \text{(A.1)}$$

Hence, under the assumptions of the theorem, we may begin with the assertion that

$$f(\hat{\theta}) + \lambda' \hat{\theta}^T \tilde{L} \hat{\theta} \leq f(\hat{\beta}) + \lambda' \hat{\beta}^T \tilde{L} \hat{\beta},$$

by the optimality of $\hat{\theta}$ for its own problem, then rearrange, and use (A.1), to arrive at

$$\begin{aligned}
\frac{\delta(1-\delta)}{2}\|\hat{\beta} - \hat{\theta}\|_2^2 &\leq -\nabla f(\hat{\beta})^T(\hat{\theta} - \hat{\beta}) + \lambda' \hat{\beta}^T \tilde{L} \hat{\beta} - \lambda' \hat{\theta}^T \tilde{L} \hat{\theta} \\
&\leq -\nabla f(\hat{\beta})^T(\hat{\theta} - \hat{\beta}) + \lambda'(1+\epsilon)\hat{\beta}^T L \hat{\beta} - \lambda' \hat{\theta}^T \tilde{L} \hat{\theta}.
\end{aligned}$$

Using $-\nabla f(\hat{\beta}) = 2\lambda L \hat{\beta}$ from the stationarity condition for $\hat{\beta}$ in (1), we have

$$\frac{\delta(1-\delta)}{2}\|\hat{\beta} - \hat{\theta}\|_2^2 \leq 2\lambda \hat{\theta}^T L \hat{\beta} + ((1+\epsilon)\lambda' - 2\lambda)\hat{\beta}^T L \hat{\beta} - \lambda' \hat{\theta}^T \tilde{L} \hat{\theta}.$$

The right-hand side is precisely of the same form as that analyzed in parts (a) and (b) of Theorem 1, and the results follow. □

A.3 Estimation Error Bounds and Stability

The following is an estimation error bound derived for Laplacian smoothing in regression, where the error rate is shown to scale with λ/n.

Theorem A.1. *Assume that $y \sim N(\beta^*, \sigma^2 I)$, and denote by $\rho_1 \leq \rho_2 \leq \ldots \leq \rho_n$ the eigenvalues of the graph Laplacian matrix L. Let $i_0 \in \{1, \ldots n\}$, and consider a choice of tuning parameter*

$$\lambda = \Theta\left(\frac{\sqrt{\sum_{i=i_0+1}^n \frac{1}{\rho_i}}}{\|D\beta^*\|_2}\right),$$

where D the graph difference operator (so that $L = D^T D$). Then the graph Laplacian smoothing estimate $\hat{\beta}$ in (1)*, with the regression loss* (2)*, satisfies*

$$\frac{\|\hat{\beta} - \beta^*\|_2^2}{n} = O_{\mathbb{P}}\left(\frac{\text{nullity}(L)}{n} + \frac{i_0}{n} + \frac{1}{n}\sqrt{\sum_{i=i_0+1}^n \frac{1}{\rho_i}} \cdot \|D\beta^*\|_2\right).$$

Proof. Let $R = \text{row}(L)$, the row space of L, and $R^\perp = \text{null}(L)$, the null space of L. Also let P_R be the projection onto R, and $P_{R^\perp}$ be the projection onto $R^\perp$, and abbreviate $\|x\|_R = \|P_R x\|_2$, $\|x\|_{R^\perp} = \|P_{R^\perp} x\|_2$. We can decompose

$$\|\hat{\beta} - \beta^*\|_2^2 = \|\hat{\beta} - \hat{\beta}^*\|_{R^\perp}^2 + \|\hat{\beta} - \hat{\beta}^*\|_R^2.$$

The first term $\|\hat{\beta} - \hat{\beta}^*\|_{R^\perp}^2$ is on the order of nullity(L), which is the number of connected components in the underlying graph G. This contributes the first term in the error rate of the theorem. Now it suffices to consider $\|\hat{\beta} - \hat{\beta}^*\|_R^2$.

Using the optimality of $\hat{\beta}$ in (1),

$$\|y - \hat{\beta}\|_2^2 + \lambda \hat{\beta}^T L \hat{\beta} \leq \|y - \beta^*\|_2^2 + \lambda (\beta^*)^T L \beta^*.$$

After setting $y = \beta^* + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I)$, expanding, and rearranging, we arrive at the basic inequality

$$\|\hat{\beta} - \beta^*\|_R^2 \leq 2\langle \hat{\beta} - \beta^*, P_R \epsilon \rangle + \lambda (\beta^*) T L \beta^* - \lambda \hat{\beta}^T L \hat{\beta}. \tag{A.2}$$

Abbreviating $\delta = \hat{\beta} - \beta^*$, let us write

$$\epsilon^T P_R \delta = \epsilon^T P_{i_0} P_R \delta + \epsilon^T (I - P_{i_0}) P_R \delta,$$

where P_{i_0} is the projection onto the first i_0 eigenvectors of L, i.e., the eigenvectors associated with eigenvalues $\rho_1 \leq \rho_2 \leq \ldots \leq \rho_{i_0}$. The first term above satisfies

$$\epsilon^T P_{i_0} P_R \delta \leq \|P_{i_0} \epsilon\|_2 \|\delta\|_R = O_{\mathbb{P}}\big(\sqrt{i_0}\big) \|\delta\|_R,$$

whereas the second term satisfies

$$\epsilon^T (I - P_{i_0}) P_R \delta = \epsilon^T (I - P_{i_0}) D^+ D\delta = \frac{\epsilon^T (I - P_{i_0}) D^+}{\sqrt{\lambda}} \sqrt{\lambda} D\delta \leq \frac{\|(D^+)^T (I - P_{i_0}) \epsilon\|_2^2}{2\lambda} + \frac{\lambda}{2} \|D\delta\|_2^2. \tag{A.3}$$

(In the last line here, we used the inequality $2a^T b \leq \|a\|_2^2 + \|b\|_2^2$.) Directly from the singular value decomposition of D, it is easy to verify that

$$\|(D^+)^T (I - P_{i_0}) \epsilon\|_2^2 = O_{\mathbb{P}}\left(\sum_{i=i_0+1}^n \frac{1}{\rho_i}\right).$$

Plugging these bounds into the basic inequality (A.2), we see that

$$\|\hat{\beta} - \beta^*\|_R^2 \leq O_{\mathbb{P}}\big(\sqrt{i_0}\big)\|\hat{\beta} - \beta^*\|_R + O_{\mathbb{P}}\bigg(\sum_{i=i_0+1}^{n} \frac{1}{\rho_i}\bigg)\frac{1}{2\lambda} + \frac{\lambda}{2}\|D\hat{\beta} - D\hat{\beta}_0\|_2^2 + \lambda\|D\beta^*\|_2^2 - \lambda\|D\hat{\beta}\|_2^2$$
$$\leq O_{\mathbb{P}}\big(\sqrt{i_0}\big)\|\hat{\beta} - \beta^*\|_R + O_{\mathbb{P}}\bigg(\sum_{i=i_0+1}^{n} \frac{1}{\rho_i}\bigg)\frac{1}{2\lambda} + 2\lambda\|D\beta^*\|_2^2.$$

(In the last line, we have again used $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$.) Plugging in the value of λ as described in the statement of the theorem yields

$$\|\hat{\beta} - \beta^*\|_R^2 \leq O_{\mathbb{P}}\big(\sqrt{i_0}\big)\|\hat{\beta} - \beta^*\|_R + O_{\mathbb{P}}\bigg(\sqrt{\sum_{i=i_0+1}^{n} \frac{1}{\rho_i}}\bigg)\|D\beta^*\|_2.$$

We can view this as a quadratic equation of form $ax^2 - bx - c \leq 0$ in $x = \|\hat{\beta} - \beta^*\|_R$. As $a > 0$, the larger of its two roots serves as a bound for x, i.e., $x \leq (b + \sqrt{b^2 + 4ac})/(2a) \leq b/a + \sqrt{c/a}$, or $x^2 \leq 2b^2/a^2 + 2c/a$, which means that

$$\|\hat{\beta} - \beta^*\|_R^2 = O_{\mathbb{P}}\bigg(i_0 + \sqrt{\sum_{i=i_0+1}^{n} \frac{1}{\rho_i}} \cdot \|D\beta^*\|_2\bigg).$$

This completes the proof. □

Remark. Assuming that the number of connected components nullity(L) stays bounded as n grows, the optimal i_0 in the theorem would be chosen to balance i_0/n with the last term in the rate. This depends, of course, on the decay of eigenvalues of L; for different graphs G, the eigenvalues will decay at different rates, leading to different error bounds. But in any case, the result in the theorem reduces to

$$\frac{\|\hat{\beta} - \beta^*\|_2^2}{n} = O_{\mathbb{P}}\Big(\frac{\lambda}{n}\|D\beta^*\|_2^2\Big),$$

which, when $\|D\beta^*\|_2^2 = O(1)$, is precisely as claimed in (7) in the main paper.

Our next result shows that when the solution $\hat{\beta}$ is tuned as in Theorem A.1, the achieved penalty is on the same order as that for β^*.

Lemma A.1. *Under the same conditions as in Theorem A.1, if* nullity$(L) = O(1)$*, and we were to choose* $i_0 = O(\lambda\|D\beta^*\|_2^2)$*, then the achieved penalty term satisfies* $\|D\hat{\beta}\|_2^2 = O_{\mathbb{P}}(\|D\beta^*\|_2^2)$*. Hence in particular if* $\|D\beta^*\|_2^2 = O(1)$*, then* $\|D\hat{\beta}\|_2^2 = O_{\mathbb{P}}(1)$*.*

Proof. Returning to the proof of Theorem A.1, consider replacing the step in (A.3) by

$$\epsilon^T(I - P_{i_0})P_R\delta = \epsilon^T(I - P_{i_0})D^+D\delta = \frac{\epsilon^T(I - P_{i_0})D^+}{\sqrt{0.5\lambda}}\sqrt{0.5\lambda}D\delta \leq \frac{\|(D^+)^T(I - P_{i_0})\epsilon\|_2^2}{\lambda} + \frac{\lambda}{4}\|D\delta\|_2^2.$$

Carrying on as in the proof of Theorem A.1, we arrive at

$$\|\hat{\beta} - \beta^*\|_R^2 \leq O_{\mathbb{P}}\big(\sqrt{i_0}\big)\|\hat{\beta} - \beta^*\|_R + O_{\mathbb{P}}\bigg(\sum_{i=i_0+1}^{n} \frac{1}{\rho_i}\bigg)\frac{1}{\lambda} + \frac{3\lambda}{2}\|D\beta^*\|_2^2 - \frac{\lambda}{2}\|D\hat{\beta}\|_2^2.$$

Using $\|\hat{\beta} - \beta^*\|_R^2 \geq 0$, and rearranging, we have

$$\|D\hat{\beta}\|_2^2 \leq O_{\mathbb{P}}\Big(\frac{\sqrt{i_0}}{\lambda}\Big)\|\hat{\beta} - \beta^*\|_R + 3\|D\beta^*\|_2^2 = O_{\mathbb{P}}\big(\|D\beta^*\|_2^2\big),$$

where we have used the choice of i_0, and the corresponding rate of $\|\hat{\beta} - \beta^*\|_R$ from Theorem A.1. □

Lastly, we show that under the conditions of the last lemma, both $\hat{\beta}$ and $\hat{\theta}$, the latter being the solution of the sparsified problem (5), achieve the same error rate.

Corollary A.1. *Under the same conditions as in Lemma A.1, assuming that $\tilde{L}$ is a $(1+\epsilon)$ approximation of L, the solution $\hat{\theta}$ of the sparsified problem (5), with the regression loss (2), and tuning parameter*

$$\lambda' = \Theta\bigg(\frac{\sqrt{\sum_{i=i_0+1}^{n} \frac{1}{\rho_i}}}{\|D\beta^*\|_2}\bigg),$$

satisfies

$$\frac{\|\hat{\theta} - \beta^*\|_2^2}{n} = O_{\mathbb{P}}\Big(\frac{\lambda}{n}\|D\beta^*\|_2^2\Big),$$

just as $\hat{\beta}$ does.

Proof. There are two ways to prove this result. The first strategy is simply to note that

$$\frac{\|\hat{\theta} - \beta^*\|_2^2}{n} \leq \frac{2\|\hat{\theta} - \hat{\beta}\|_2^2}{n} + \frac{2\|\hat{\beta} - \beta^*\|_2^2}{n},$$

and both terms are $O_{\mathbb{P}}(\lambda\|D\beta^*\|_2^2/n)$, the first term controlled by part (b) of Theorem 1 in the main paper (in combination with Lemma A.1), and the second term handled by Theorem A.1.

The second strategy is to replicate the proof of Theorem A.1, applied to the problem (5) with $\tilde{L}$ in place of L, and then conclude that the resulting error rate is not changed, since all eigenvalues of $\tilde{L}$ are within a multiplicative factor between $1/(1+\epsilon)$ and $(1+\epsilon)$ of those of L. □

A.4 Proof of Lemma 1

We use Hoeffding's inequality as a main tool for proving the results for both the uniform and kN samplers, first treating the uniform sampling case. For $i = 1, 2, \ldots q$, let $e_i = (u_i, v_i)$ be the ith edge sampled, and denote by X_i the term added to $x^T\tilde{L}x$ due to sampling e_i. Then

$$X_i = \frac{W}{q}(x_u - x_v)^2 \quad \text{with probability } w_e/W \text{ for each } e = (u,v) \in E,$$

recalling $W = \sum_{e\in E} w_e$. Note that $\sum_{i=1}^q X_i = x^T\tilde{L}x$. The variables $X_1, X_2, \ldots X_q$ are independent, and lie in an interval of length Wr/q, where $r = \max_{(u,v)\in E}(x_u - x_v)^2 - \min_{(u,v)\in E}(x_u - x_v)^2$. Using Hoeffding's inequality,

$$\mathbb{P}\bigg(\Big|\sum_{i=1}^{q}\Big(X_i - \mathbb{E}(X_i)\Big)\Big| > t\bigg) \leq 2\exp\bigg(\frac{-2t^2}{\sum_{i=1}^q (W^2r^2/q^2)}\bigg) = 2\exp\bigg(\frac{-2t^2q}{W^2r^2}\bigg). \tag{A.4}$$

As $\tilde{L}$ is an unbiased estimator of L, we have $\sum_{i=1}^q \mathbb{E}(X_i) = \mathbb{E}(x^T\tilde{L}x) = x^TLx$. Abbreviating $\mu = x^TLx$, and substituting $t = \delta\mu$ into the Hoeffding bound (A.4), we infer that

$$|x^T\tilde{L}x - \mu| \leq \delta\mu, \tag{A.5}$$

with probability at least $1 - 2\exp(-2\delta^2\mu^2q/(W^2r^2))$. Now we use the smoothness assumption (8) on x, i.e., $\|Dx\|_\infty^2 \leq \mu s w_{\min}/W$, as well as

$$r \leq \max_{(u,v)\in E}(x_u - x_v)^2 \leq \max_{(u,v)\in E}\frac{w_{uv}}{w_{\min}}(x_u - x_v)^2 = \frac{\|Dx\|_\infty^2}{w_{\min}},$$

recalling $w_{\min} = \min_{e\in E} w_e$. Therefore $r \leq \mu s/W$, and the result in (A.5) holds with probability at least $1 - 2\exp(-2\delta^2q/s^2)$. To make this probability at least $1 - 1/n$, we need to choose $q \geq s^2/(2\delta^2)\cdot\log(2n)$ samples. Lastly, to give the result as written in the lemma, we simply substitute $\delta = \epsilon/(1+\epsilon)$ and observe that (A.5) then implies

$$\frac{1}{1+\epsilon}\mu \leq x^T\tilde{L}x \leq \Big(1 + \frac{\epsilon}{1+\epsilon}\Big)\mu \leq (1+\epsilon)\mu.$$

The arguments for kN sampling are similar. We sample only from nodes with degree greater than k; call this set U. When sampling edges incident to node $u \in U$, for $i = 1, \dots k$, let $e_{u,i}$ denote the ith edge sampled, and let $X_{u,i}$ be its contribution to $x^T \tilde{L} x$. Then

$$X_{u,i} = \frac{W_u}{2k}(x_u - x_v)^2 \quad \text{with probability } w_e/W_u \text{ for each } v \in N(u),$$

recalling $W_u = \sum_{v \in N(u)} w_{u,v}$ and $N(u)$ denotes the neighbors of u. The variables $X_{u,i}$, $u \in U$, $i = 1, \dots k$ are independent, and lie in an interval of length at most $W_{\max} r/(2k)$, where $W_{\max} = \max_{u \in V} W_u$, and r is as above (in the proof for the uniform sampling part). Using Hoeffding's inequality once again,

$$\mathbb{P}\left(\left| \sum_{u \in U} \sum_{i=1}^{k} \Big(X_{u,i} - \mathbb{E}(X_{u,i}) \Big) \right| > t \right) \leq 2 \exp\left(\frac{-2t^2}{\sum_{u \in U} \sum_{i=1}^{k} \big(W_{\max} r/(2k) \big)^2} \right) = 2 \exp\left(\frac{-8t^2 k}{r^2 \sum_{u \in U} W_{\max}^2} \right). \tag{A.6}$$

Denoting $z_u = \sum_{v \in N(u)} w_{u,v}(x_u - x_v)^2$ for all $u \in V$, note that $\mathbb{E}(X_{u,i}) = z_u/(2k)$ for all $u \in U$. Thus,

$$\sum_{u \in U} \sum_{i=1}^{k} X_{u,i} = x^T \tilde{L} x - \frac{1}{2} \sum_{u \in V \setminus U} z_u,$$

and

$$\sum_{u \in U} \sum_{i=1}^{k} \mathbb{E}(X_{u,i}) = x^T L x - \frac{1}{2} \sum_{u \in V \setminus U} z_u.$$

This means that the Hoeffding bound (A.6), setting $t = \delta\mu$, implies the statement in (A.5), with probability $1 - 2\exp(-8\delta^2\mu^2 k/(r^2 \sum_{u \in U} W_{\max}^2)) \geq 1 - 2\exp(-8\delta^2\mu^2 k/(nr^2 W_{\max}^2))$. Bounding $r \leq s\mu/W$, from the smoothness condition (8) (as argued in the proof for the uniform sampling case) this is further lower bounded by $1 - 2\exp(-8\delta^2 W^2 k/(ns^2 W_{\max}^2))$. To make this probability at least $1 - 1/n$, we need to choose $k \geq n(sW_{\max}/W)^2/(4\delta^2) \cdot \log(2n)$. The rest follows as in the uniform sampling case, i.e., to get the result as stated in the lemma, we take $\delta = \epsilon/(1+\epsilon)$. □

A.5 Proof of Lemma 2

The proof is simple. Denote by L the Laplacian matrix of $G \times H$, and $\tilde{L}$ the Laplacian matrix of $\tilde{G} \times \tilde{H}$. Also denote the edge weights of G, H by $w_{u,u'}, w_{v,v'}$, and the edge weights of $\tilde{G}, \tilde{H}$ by $\tilde{w}_{u,u'}, \tilde{w}_{v,v'}$. Observe,

$$x^T L x = \sum_{u \in V_G} \sum_{\{v,v'\} \in E_H} w_{v,v'}(x_{u,v} - x_{u,v'})^2 + \sum_{v \in V_H} \sum_{\{u,u'\} \in E_G} w_{u,u'}(x_{u,v} - x_{u',v})^2,$$

$$x^T \tilde{L} x = \sum_{u \in V_{\tilde{G}}} \sum_{\{v,v'\} \in E_{\tilde{H}}} \tilde{w}_{v,v'}(x_{u,v} - x_{u,v'})^2 + \sum_{v \in V_{\tilde{H}}} \sum_{\{u,u'\} \in E_{\tilde{G}}} \tilde{w}_{u,u'}(x_{u,v} - x_{u',v})^2.$$

Applying the appropriate spectral sparsification bounds to the inner sums gives the result. □

A.6 Proof of Theorem 3

Let us denote

$$F(\beta) = \sum_{i \in \Omega} \ell(x_i^T \beta; y_i) + \mu \|\beta - v\|_2^2.$$

Note that, by construction, F is strongly convex with parameter $2\mu > 0$, which implies that

$$F(\theta) - F(\beta) \geq g^T(\theta - \beta) + \mu \|\theta - \beta\|_2^2, \tag{A.7}$$

for any subgradient g of F at β. Therefore, by the optimality of $\hat{\theta}$ for problem (10),

$$F(\hat{\theta}) + \lambda' \hat{\theta}^T \tilde{L} \hat{\theta} \leq F(\hat{\beta}) + \lambda' \hat{\beta}^T \tilde{L} \hat{\beta},$$

and after rearranging and applying (A.7), we have

$$\begin{aligned}\mu\|\hat{\beta}-\hat{\theta}\|_2^2 &\leq -g^T(\hat{\theta}-\hat{\beta})+\lambda'\hat{\beta}^T\tilde{L}\hat{\beta}-\lambda'\hat{\theta}^T\tilde{L}\hat{\theta} \\ &\leq -g^T(\hat{\theta}-\hat{\beta})+\lambda'(1+\epsilon)\hat{\beta}^T L\hat{\beta}-\lambda'\hat{\theta}^T\tilde{L}\hat{\theta},\end{aligned}$$

for any subgradient g of F at β. As there exists a subgradient such that $-g = 2\lambda L\hat{\beta}$ from the stationarity condition for $\hat{\beta}$ in (9), we have

$$\mu\|\hat{\beta}-\hat{\theta}\|_2^2 \leq 2\lambda\hat{\theta}^T L\hat{\beta}+((1+\epsilon)\lambda'-2\lambda)\hat{\beta}^T L\hat{\beta}-\lambda'\hat{\theta}^T\tilde{L}\hat{\theta},$$

and the remainder of the analysis proceeds exactly as in parts (a) and (b) of Theorem 1. □

A.7 Gaussian Smoothing with the Google+ Data

In this experiment, we added i.i.d. $N(0, 5.5)$ noise to the components of β^*, the smooth signal constructed by a simulated diffusion over the Google+ graph, as described in the main paper. Figure A.1 shows the results, when considering a Gaussian loss (2) in the Laplacian smoothing problems (1), (5).

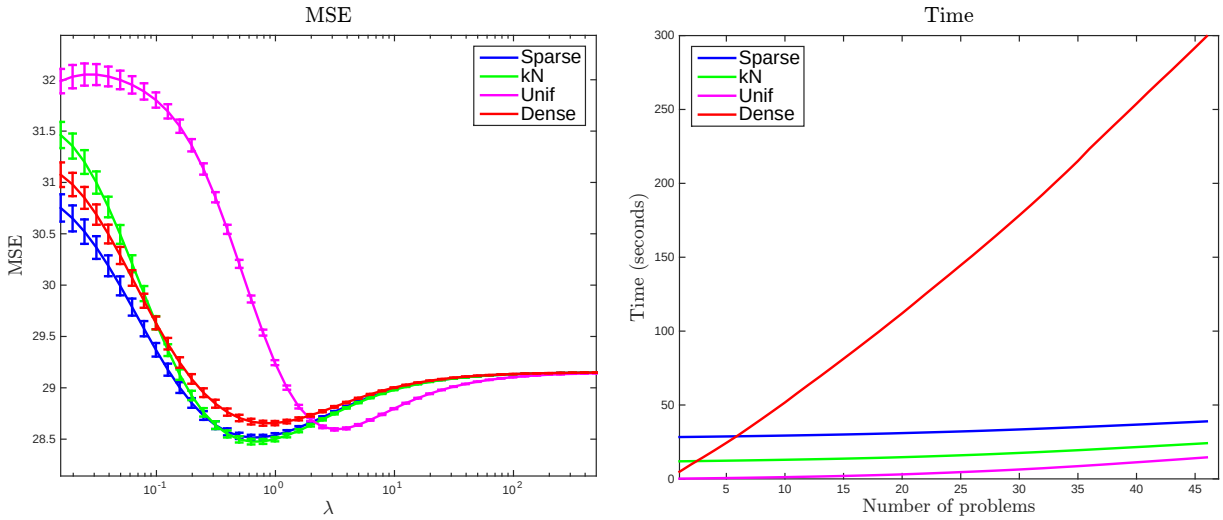

Figure A.1: *MSE and timing results for a Gaussian smoothing problem with the Google+ data.*