A. Additional Background

A.1. Rademacher Process

We need the following classical results regarding comparison of Rademacher processes.

Lemma A.1 (Proposition 1 in (Ledoux & Talagrand, 1989))**.** *Let $F : [0, \infty) \to [0, \infty)$ be convex and increasing. Let $\sigma_1, \ldots, \sigma_n$ be a Rademacher sequence. Then for any bounded subset $T \subset \mathbb{R}^n$ it holds that*

$$\mathbb{E}\, F\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i|t_i|\right|\right) \le 2\,\mathbb{E}\, F\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i t_i\right|\right).$$

Lemma A.2 (Contraction Principle, Theorem 4.4 in (Ledoux & Talagrand, 1989))**.** *Let $F : [0, \infty) \to [0, \infty)$ be convex. Let $\sigma_1, \ldots, \sigma_n$ be a Rademacher sequence. Then for any finite sequence (x_i) and any real numbers (α_i) such that $|\alpha_i| \le 1$ for every i, it holds that*

$$\mathbb{E}\, F\left(\left|\sum_{i=1}^{n}\sigma_i a_i x_i\right|\right) \le \mathbb{E}\, F\left(\left|\sum_{i=1}^{n}\sigma_i x_i\right|\right).$$

The next one is a consequence of the contraction principle (see, e.g., Theorem 4.12 in (Ledoux & Talagrand, 1991)).

Lemma A.3. *Let $F : [0, \infty) \to [0, \infty)$ be convex and increasing. Suppose that $f(x), g(x)$ are nonnegative functions such that $|f(x) - f(y)| \le L|g(x) - g(y)|$ for all $x, y \in \mathbb{R}$. Let $\sigma_1, \ldots, \sigma_n$ be a Rademacher sequence. Then for any bounded subset $T \subset \mathbb{R}^n$ it holds that*

$$\mathbb{E}\, F\left(\frac{1}{2L}\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i f(t_i)\right|\right) \le \mathbb{E}\, F\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i g(t_i)\right|\right).$$

A.2. Convex Optimization

We recall some results in convex optimization.

Definition A.1. A function $f : \mathbb{R}^d \to \mathbb{R}$ is G-Lipschitz if for any $x, y \in \mathbb{R}^d$,

$$|f(x) - f(y)| \le G\|x - y\|_2.$$

We need the following result in (Allen-Zhu, 2017).

Theorem A.4 (Corollary 3.7 in (Allen-Zhu, 2017))**.** *For a given function $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, let $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. If each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and $\sqrt{G}$-Lipschitz, then there is an algorithm that receives an initial solution $x_0 \in \mathbb{R}^d$, and outputs a solution $x \in \mathbb{R}^d$ satisfying $\mathbb{E}[f(x)] - f(x^*)$ in*

$$T = O\left(n \log\frac{f(x_0) - f(x^*)}{\varepsilon} + \frac{\sqrt{nG}\|x_0 - x^*\|}{\varepsilon}\right)$$

stochastic subgradient iterations.

B. Missing Proofs

B.1. Proof of Theorem 2.3

Proof. By Lemma 2.4 in (Cohen & Peng, 2015), to calculate approximate ℓ_1 Lewis weights of rows of a matrix A, it suffices to calculate approximate leverage scores of rows of matrices of the form WA, for $O(\log n)$ different diagonal matrices $W \in \mathbb{R}^n$.

When A is the edge-vertex incidence matrix of a graph G, WA is the edge-vertex incidence matrix of a reweighted graph G'. In this case, leverage scores of rows of WA are the effective resistances of G' (cf. (Drineas & Mahoney, 2010)), which can be computed in $\widetilde{O}(m)$ time using the algorithm in (Spielman & Srivastava, 2011). □

B.2. Proof of Theorem 3.1

Proof of Theorem 3.1. Suppose that $i_1, i_2, \ldots, i_N$ are the indices of $\{e_1, \ldots, e_m\}$ randomly chosen in the construction of S. Since $\phi(\alpha t) = \alpha\phi(t)$ for $\alpha > 0$, for each coordinate of $\widetilde{A}x$, we have

$$\mathbb{E}\,\phi((\widetilde{A}x)_k) = \mathbb{E}\,\phi\left(\frac{1}{p_{i_k}}\langle A_{i_k}, x\rangle\right) = \sum_{j=1}^{n} \frac{1}{p_j}\phi(A_j^\top x)\cdot\frac{p_j}{N} = \frac{1}{N}\phi(Ax)$$

and

$$\mathbb{E}\,\phi(\widetilde{A}x) = \sum_{k=1}^{N} \mathbb{E}\,\phi((\widetilde{A}x)_k) = \phi(Ax).$$

We just assume that $p_i \geq C_s\varepsilon^{-2}w_i \log N$ for now and shall rescale ε in the end. The ℓ_1 Lewis weight sampling result in (Cohen & Peng, 2015) implies that with probability at least $1 - 1/\operatorname{poly}(N)$ (over $i_1, \ldots, i_N$),

$$\|\widetilde{A}x\|_1 \leq C_1\|Ax\|_1, \quad \forall x \in \mathbb{R}^d. \tag{7}$$

We shall condition on this event in the rest of the proof.

Our goal is to derive a tail inequality for

$$\sup_{x:\phi(Ax)=1} \left|\phi(\widetilde{A}x) - 1\right| = \sup_{x:\phi(Ax)=1} \left|\sum_{k=1}^{N} \frac{\phi(\langle A_{i_k}, x\rangle)}{p_{i_k}} - 1\right|.$$

We shall look at a higher moment of the deviation and apply Markov's inequality. To this end, we investigate

$$M = \mathop{\mathbb{E}}_{S}\left(\sup_{\phi(Ax)=1}\left|\phi(\widetilde{A}x) - 1\right|\right)^{\ell} = \mathop{\mathbb{E}}_{i}\left(\sup_{\phi(Ax)=1}\left|\sum_{k=1}^{N}\frac{\phi(\langle A_{i_k}, x\rangle)}{p_{i_k}} - 1\right|\right)^{\ell},$$

where $i = (i_1, i_2, \ldots, i_N)$.

A standard symmetrization argument gives that

$$M \leq 2^{\ell} \mathop{\mathbb{E}}_{i,\sigma}\left[\left(\sup_{\phi(Ax)=1}\left|\sum_{k=1}^{N}\sigma_k\frac{\phi(\langle A_{i_k}, x\rangle)}{p_{i_k}}\right|\right)^{\ell}\right], \tag{8}$$

where $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ is a Rademacher sequence. It follows that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{i,\sigma}\left(\sup_{\phi(Ax)=1}\left|\sum_{k=1}^{N}\sigma_k\frac{\phi(\langle A_{i_k}, x\rangle)}{p_{i_k}}\right|\right)^{\ell} \\
&= \mathop{\mathbb{E}}_{i,\sigma}\sup_{\phi(Ax)=1}\left|a\sum_{k=1}^{N}\sigma_k\frac{|\langle A_{i_k}, x\rangle|}{p_{i_k}} + b\sum_{k=1}^{N}\sigma_k\frac{\langle A_{i_k}, x\rangle}{p_{i_k}}\right|^{\ell} \\
&\leq \mathop{\mathbb{E}}_{i,\sigma}\sup_{\phi(Ax)=1}\left(\left|a\sum_{k=1}^{N}\sigma_k\frac{|\langle A_{i_k}, x\rangle|}{p_{i_k}}\right| + b\left|\sum_{k=1}^{N}\sigma_k\frac{\langle A_{i_k}, x\rangle}{p_{i_k}}\right|\right)^{\ell} \\
&\leq 2^{\ell-1}\cdot\mathop{\mathbb{E}}_{i,\sigma}\sup_{\phi(Ax)=1}\left[\left|a\sum_{k=1}^{N}\sigma_k\frac{|\langle A_{i_k}, x\rangle|}{p_{i_k}}\right|^{\ell} + \left(b\left|\sum_{k=1}^{N}\sigma_k\frac{\langle A_{i_k}, x\rangle}{p_{i_k}}\right|\right)^{\ell}\right] \\
&\leq (2a^{\ell} + b^{\ell})\cdot 2^{\ell-1}\cdot\mathop{\mathbb{E}}_{i,\sigma}\sup_{\phi(Ax)=1}\left|\sum_{k=1}^{N}\sigma_k\frac{\langle A_{i_k}, x\rangle}{p_{i_k}}\right|^{\ell} \\
&\leq 3a^{\ell}2^{\ell-1}\mathop{\mathbb{E}}_{i,\sigma}\sup_{\phi(Ax)=1}\left|\sum_{k=1}^{N}\sigma_k\frac{\langle A_{i_k}, x\rangle}{p_{i_k}}\right|^{\ell},
\end{aligned} \tag{9}$$

where the penultimate inequality follows from Lemma A.1.

For now, condition on the choices of $i_1, i_2, \ldots, i_N$. By Lemma B.1 in (Cohen & Peng, 2015), there exists a matrix A' of $O(d^2)$ rows such that

$$\|A'x\|_1 \le C_1\|Ax\|_1, \quad \forall x \in \mathbb{R}^d, \tag{10}$$

and the Lewis weights of A' are $O(1/d)$. Append A' to the matrix $\widetilde{A}$, obtaining a new matrix A'' of $N' = N + O(d^2)$ rows, and it is a direct consequence of the contraction principle (Lemma A.2) that

$$\mathop{\mathbb{E}}_{\sigma} \sup_{\phi(Ax)=1} \left| \sum_{k=1}^{N} \sigma_k \frac{\langle A_{i_k}, x\rangle}{p_{i_k}} \right|^{\ell} \le \mathop{\mathbb{E}}_{\sigma} \sup_{\phi(Ax)=1} \left| \sum_{k=1}^{N'} \sigma_k \langle A''_k, x\rangle \right|^{\ell}, \tag{11}$$

since the sum on the right-hand side has more terms.

Furthermore, it can be proved that the Lewis weights of A'' are all $O(\varepsilon^2/\log N')$ (see (Cohen & Peng, 2015)). In this case, for an appropriate choice of N (and thus N'), it is implicitly shown in the proof of Theorem 2.3 in (Cohen & Peng, 2015) that

$$\mathop{\mathbb{E}}_{\sigma} \sup_{\|A''x\|_1=1} \left| \sum_{k=1}^{N'} \sigma_k \langle A''_k, x\rangle \right|^{\ell} \le \frac{\varepsilon^{\ell}}{\operatorname{poly}(N)} \tag{12}$$

for some $\ell = \Theta(\log N')$. The next step is to relate the right-hand side of (11) to the left-hand side of (12). Note that

$$\begin{aligned} \|A''x\|_1 &= \|\widetilde{A}x\|_1 + \|A'x\|_1 \\ &\le C_1\|Ax\|_1 + C_2\|Ax\|_1 \quad \text{(by (7) and (10))} \\ &\le (C_1 + C_2)B\phi(Ax) \end{aligned}$$

and thus

$$\sup_{\phi(Ax)=1} \left| \sum_{k=1}^{N'} \sigma_k \langle A''_k, x\rangle \right|^{\ell} \le (C_1+C_2)^{\ell} B^{\ell} \sup_{\|A''x\|_1=1} \left| \sum_{k=1}^{N'} \sigma_k \langle A''_k, x\rangle \right|^{\ell}.$$

Taking expectation over σ on both sides, we obtain using (12) that

$$\mathop{\mathbb{E}}_{\sigma} \sup_{\|Ax\|_1=1} \left| \sum_{k=1}^{N'} \sigma_k \langle A''_k, x\rangle \right|^{\ell} \le ((C_1+C_2)B)^{\ell} \frac{\varepsilon^{\ell}}{\operatorname{poly}(N)}.$$

Taking expectation over $i_1, \ldots, i_N$ on both sides subject to the conditioning (7) and combining with (8), (9) and (11), we obtain that

$$M \le C_3(C_4 aB)^{\ell} \frac{\varepsilon^{\ell}}{\operatorname{poly}(N)}$$

The result follows from Markov's inequality with a rescaling of ε by a factor of $1/(C_5 aB)$. □

B.3. Proof of Theorem 1.1

Proof. Recall that $\rho_\tau(x) = -\tau x$ if $x \le 0$ and $\rho_\tau(x) = x$ if $x \ge 0$. We can rewrite $\rho_\tau(x) = (1+\tau)|x|/2 + (1-\tau)x/2$. Also $\rho_\tau(x) \le \|x\|_1/\tau$, and thus we can take $B = 1/\tau$ in Theorem 3.1. By Theorem 2.2, we can obtain $\{\overline{w}_i\}_{i=1}^n$, such that with probability $1 - 1/\operatorname{poly}(d)$, for all $i \in [n]$, $w_i \le \overline{w}_i \le 2w_i$, where $\{w_i\}_{i=1}^n$ are the ℓ_1 Lewis weights of A. Now we invoke the row sampling algorithm in Theorem 3.1 and the fact in Lemma 2.1 to finish the proof. □

B.4. Proof of Lemma 4.1

Proof. By Theorem 1.1, with probability at least 0.9, for all $x \in \mathbb{R}^d$,

$$\left(1 - \frac{\varepsilon}{9}\right) \rho_\tau(Ax - b) \le \rho_\tau(\widetilde{A}x - \widetilde{b}) \le \left(1 + \frac{\varepsilon}{9}\right) \rho_\tau(Ax - b).$$

Let $x^{\mathrm{opt}} = \operatorname{argmin}_{x \in \mathbb{R}^d} \rho_\tau(Ax - b)$, we have

$$\begin{aligned}
\rho_\tau(Ax^* - b) &= \rho_\tau(AR^{-1}\overline{x} - b) \\
&\le \frac{1}{1-\varepsilon/9}\rho_\tau(\widetilde{A}R^{-1}\overline{x} - \widetilde{b}) \\
&\le \frac{1+\varepsilon/3}{1-\varepsilon/9}\rho_\tau(\widetilde{A}R^{-1}(Rx^{\mathrm{opt}}) - \widetilde{b}) && \text{(By Equation (6))} \\
&= \frac{1+\varepsilon/3}{1-\varepsilon/9}\rho_\tau(\widetilde{A}x^{\mathrm{opt}} - \widetilde{b}) \\
&\le \frac{(1+\varepsilon/3)(1+\varepsilon/9)}{1-\varepsilon/9}\rho_\tau(Ax^{\mathrm{opt}} - b) \\
&\le (1+\varepsilon)\rho_\tau(Ax^{\mathrm{opt}} - b).
\end{aligned}$$

□

B.5. Proof of Lemma 4.2

We need the following claim in our proof.

Claim 1. *For a matrix $A \in \mathbb{R}^{n \times d}$ and a vector $y \in \mathbb{R}^n$,*

$$\|A^\top y\|_2 \le \frac{1}{\tau}\rho_\tau(y) \max_{1 \le i \le N} \|A_i\|_2.$$

Proof. Observe that $\sum_{i=1}^n |y_i| \le \frac{1}{\tau}\rho_\tau(y)$. It follows that

$$\|A^\top y\|_2 = \left\| \sum_{i=1}^N y_i A_i \right\|_2 \le \sum_{i=1}^N |y_i| \|A_i\|_2 \le \max_{1 \le i \le N} \|A_i\|_2 \cdot \sum_{i=1}^N |y_i| \le \frac{1}{\tau}\rho_\tau(y) \max_{1 \le i \le N} \|A_i\|_2.$$

□

Now we are ready to prove Lemma 4.2.

Proof. By Lemma 30 in (Durfee et al., 2018), with probability at least 0.9, the leverage score of each row of $[\widetilde{A}, \widetilde{b}]$ satisfies $\tau_i([\widetilde{A}, \widetilde{b}]) = O(d/N)$. We condition on this event in the remaining part of the proof. By Lemma 2 in (Cohen et al., 2015),

$$\tau_i([\widetilde{A}, \widetilde{b}]) = \min_{[\widetilde{A}, \widetilde{b}]^\top x = ([\widetilde{A}, \widetilde{b}])_i} \|x\|_2^2,$$

and

$$\tau_i(\widetilde{A}) = \min_{\widetilde{A}^\top x = \widetilde{A}_i} \|x\|_2^2,$$

which implies $\tau_i(\widetilde{A}) \le \tau_i([\widetilde{A}, \widetilde{b}])$. Thus, $\tau_i(\widetilde{A}) = O(d/N)$. Since $\widetilde{A}R^{-1} = Q$ has orthonormal columns, for each row Q_i of Q,

$$\|Q_i\|_2^2 = \tau_i(\widetilde{A}) = O(d/N).$$

Here we used the standard fact of the relation between leverage scores and the QR decomposition. See Definition 2.6 in (Woodruff, 2014) for details.

Let $f(x) = \rho_\tau(\widetilde{A}R^{-1}x - \widetilde{b})$, we can write

$$f(x) = \sum_{i=1}^n \rho_\tau(\langle (\widetilde{A}R^{-1})_i, x \rangle - b_i) = \frac{1}{N}\sum_{i=1}^n f_i(x),$$

where $f_i(x) = N \cdot \rho_\tau(\langle (\widetilde{A}R^{-1})_i, x \rangle - b_i)$. Let $g_i(x) = \langle (\widetilde{A}R^{-1})_i, x \rangle - b_i$, then

$$\|\nabla g_i(x)\|_2 \le \|(\widetilde{A}R^{-1})_i\|_2 = \|Q_i\|_2 = O(\sqrt{d/N}),$$

which implies that each $g_i(x)$ is $O(\sqrt{d/N})$-Lipschitz. Observe that $\rho_\tau(\cdot)$ is 1-Lipschitz, it follows that $f_i(x) = N\rho_\tau(g_i(x))$ is $O(\sqrt{Nd})$-Lipschitz.

By Claim 1, we have

$$\begin{aligned}
\|x_0 - \widetilde{x^{\mathrm{opt}}}\|_2 &= \|(\widetilde{A}R^{-1})^\top b - \widetilde{x^{\mathrm{opt}}}\|_2 \\
&= \|(\widetilde{A}R^{-1})^\top b - \widetilde{x^{\mathrm{opt}}}\|_2 \\
&= \|(\widetilde{A}R^{-1})^\top (b - (\widetilde{A}R^{-1})\widetilde{x^{\mathrm{opt}}})\|_2 \\
&\le \sqrt{\frac{d}{N\tau^2}} \cdot \rho_\tau((\widetilde{A}R^{-1})\widetilde{x^{\mathrm{opt}}} - b).
\end{aligned}$$

Finally, for any vector $y \in \mathbb{R}^N$, we have

$$\tau\|y\|_2 \le \tau\|y\|_1 \le \rho_\tau(y) \le \|y\|_1 \le \sqrt{N}\|y\|_2.$$

Since $\widetilde{A}R^{-1} = Q$ has orthonormal columns, $x_0 = (\widetilde{A}R^{-1})^\top b$ is the optimal solution to $\min_{x\in\mathbb{R}^d} \|\widetilde{A}R^{-1}x_0 - \widetilde{b}\|_2$. Thus, we have $\|\widetilde{A}R^{-1}x_0 - \widetilde{b}\|_2 \le \|\widetilde{A}R^{-1}\widetilde{x^{\mathrm{opt}}} - \widetilde{b}\|_2$. Therefore,

$$\begin{aligned}
\rho_\tau(\widetilde{A}R^{-1}x_0 - \widetilde{b}) &\le \sqrt{N}\|\widetilde{A}R^{-1}x_0 - \widetilde{b}\|_2 \\
&\le \sqrt{N}\|\widetilde{A}R^{-1}\widetilde{x^{\mathrm{opt}}} - \widetilde{b}\|_2 \\
&\le \sqrt{\frac{N}{\tau^2}}\rho_\tau(\widetilde{A}R^{-1}\widetilde{x^{\mathrm{opt}}} - \widetilde{b}).
\end{aligned}$$

□

B.6. Proof of Lemma 4.3

Proof. The initial solution $x_0 = (\widetilde{A}R^{-1})^\top b$ can be calculated in $O(Nd)$ time. We condition on the event in Lemma 4.2. By Theorem A.4, with probability at least 0.9, after

$$T = O\left(N \log \frac{f(x_0) - f(\widetilde{x^{\mathrm{opt}}})}{\varepsilon \cdot f(\widetilde{x^{\mathrm{opt}}})} + \frac{\sqrt{N^2 d}\|x_0 - x^*\|_2}{\varepsilon \cdot f(\widetilde{x^{\mathrm{opt}}})}\right) = O\left(N \log \frac{N}{\varepsilon\tau^2} + \frac{dN^{1/2}}{\varepsilon\tau}\right) = \widetilde{O}\left(\frac{d^{1.5}}{\tau^2\varepsilon^2}\right)$$

stochastic subgradient iterations, we can find a solution $\overline{x}$ such that

$$\mathbb{E}\left[\rho_\tau(\widetilde{A}R^{-1}\overline{x} - \widetilde{b}) - \min_{x\in\mathbb{R}^d} \rho_\tau(\widetilde{A}R^{-1}x - \widetilde{b})\right] \le \varepsilon/30 \cdot \rho_\tau(\widetilde{A}R^{-1}x - \widetilde{b}).$$

By Markov's inequality, with probability at least 0.9, we have

$$\rho_\tau(\widetilde{A}R^{-1}\overline{x} - \widetilde{b}) \le (1 + \varepsilon/3) \cdot \min_{x\in\mathbb{R}^d} \rho_\tau(\widetilde{A}R^{-1}x - \widetilde{b}).$$

Furthermore, each stochastic subgradient can be calculated in $O(d)$ time, since

$$\nabla f_i(x) = \begin{cases} \mathrm{sign}(\langle A_i, x\rangle - b_i) \cdot A_i & \text{if } \langle A_i, x\rangle - b_i \ge 0 \\ \tau \cdot \mathrm{sign}(\langle A_i, x\rangle - b_i) \cdot A_i & \text{otherwise} \end{cases}.$$

where we choose a subgradient $\nabla f_i(x) = 0$ at the nondifferentiable points x. □

B.7. Proof of Theorem 4.4

Proof. Finding the QR decomposition of the concatenated matrix $[\widetilde{A}, \widetilde{b}]$ can be done in $\widetilde{O}(d^\omega/(\varepsilon^2\tau^2))$ time (see Lemma 33 in (Durfee et al., 2018)). By Lemma 4.3, we can obtain $\overline{x}$ such that $\rho_\tau(\widetilde{A}R^{-1}\overline{x} - \widetilde{b}) \le (1+\varepsilon/3) \cdot \min_{x\in\mathbb{R}^d} \rho_\tau(\widetilde{A}R^{-1}x - \widetilde{b})$ in $\widetilde{O}(d^{2.5}/(\tau^2\varepsilon^2))$ time and succeeds with probability at least 0.8. By Lemma 4.1, with probability at least 0.9, the obtained solution $x^* \in \mathbb{R}^d$ satisfies $\rho_\tau(Ax^* - b) \le (1+\varepsilon)\min_{x\in\mathbb{R}^d} \rho_\tau(Ax - b)$. We complete the proof of the theorem by taking a union bound over the two events mentioned above. □

B.8. Proof of Lemma 5.1

Proof. By permuting the columns we may assume without loss of generality that $x_1 \leq x_2 \leq \cdots \leq x_n$. Suppose that $a = x_1 < x_n = b$, otherwise $Ax = 0$ and the desired inequality holds automatically.

Let $I = \{i \in [n] : (Ax)_i \geq 0\}$. Let $u = \sum_{i \in I} A_i$ and $v = \sum_{i \in I} A_i - \sum_{i \notin I} A_i$, then $\rho_0(Ax) = \langle u, x \rangle$ and $\rho_1(Ax) = \langle v, x \rangle$ for all $x \in P$, where $P = \{x \in [a,b]^n : a = x_1 \leq x_2 \leq \cdots \leq x_n = b\}$ is a polytope.

Observe that $\rho_1(Ax) \neq 0$ on P and thus the function $f(x) = \langle u, x\rangle / \langle v, x \rangle$ attains its minimum value λ on the compact set P. Suppose that $\langle u, x^* \rangle = \lambda \langle v, x^* \rangle$ for some $x^* \in P$. We claim that x^* is also the minimizer of $\langle u - \lambda v, x \rangle$ on P. Indeed, if $\langle u - \lambda v, x' \rangle < \langle u - \lambda v, x^* \rangle$ for some $x' \in P$, then $\langle u, x' \rangle / \langle v, x' \rangle < \lambda$, contradicting the minimality of x'.

Now, since x^* is a minimizer of $\langle u - \lambda v, x \rangle$ on the polytope P, it must be some vertex of P, that is, there exists k such that $x_1^* = \cdots = x_k^* = a$ and $x_{k+1}^* = \cdots = x_n^* = b$. Let $S = \{x_1, \ldots, x_k\}$, then

$$\frac{\rho_0(Ax^*)}{\rho_1(Ax^*)} = \frac{w(S, V \setminus S)}{w(S, V \setminus S) + w(V \setminus S, S)} \geq \frac{1}{\alpha + 1},$$

where the last step follows from the definition of the α-balanced graph. We complete the proof by noticing that

$$\frac{\rho_0(Ax)}{\rho_1(Ax)} \geq \frac{\rho_0(Ax^*)}{\rho_1(Ax^*)} \geq \frac{1}{\alpha + 1}.$$

□

B.9. Proof of Corollary 5.2

Proof. Observe that $\rho_0(x) = \frac{1}{2}|x| + \frac{1}{2}x$. Moreover, by Lemma 5.1, we have $\|Bx\|_1 \leq (1+\alpha)\rho_0(Bx)$ for all $x \in \mathbb{R}^n$. Now we invoke Theorem 3.1 with $a = b = \frac{1}{2}$ and $B = \alpha + 1$, which states that with probability at least $1 - 1/\operatorname{poly}(n)$, for all $x \in \mathbb{R}^n$, $(1-\varepsilon)\rho_0(Bx) \leq \rho_0(B'x) \leq (1+\varepsilon)\rho_0(Bx)$. Moreover, the rows of B' are reweighted rows of B, which implies B' is the edge-vertex matrix of a graph G', whose edges are reweighted edges of G. The running time of Algorithm 2 directly follows from Theorem 2.3. □

C. Proof of Theorem 6.1

Similar to the proof of Theorem 3.1 and below we shall only highlight the changes. Instead of the comparison result of Lemma A.1, we shall use Lemma A.3.

Note that we have here that

$$\mathbb{E}\, \phi^w(\widetilde{A}x) = \phi(Ax)$$

and want to upper bound

$$M := \mathbb{E}_S \left[\sup_{x \neq 0} \left| \frac{\phi^w(\widetilde{A}x)}{\phi(Ax)} - 1 \right|^\ell \right].$$

Again by a standard symmetrization argument,

$$M = \mathbb{E}_S \left[\sup_{x \neq 0} \left| \sum_{i=1}^N \frac{\phi((\widetilde{A}x)_i)}{p_{i_k}\phi(Ax)} - 1 \right|^\ell \right] \leq 2^\ell \mathbb{E}_{i,\sigma} \left[\sup_{x \neq 0} \left| \sum_{k=1}^N \sigma_k \frac{\phi(\langle A_{i_k}, x \rangle)}{p_{i_k}\phi(Ax)} \right|^\ell \right],$$

where $\sigma_1, \sigma_2, \ldots$ is a Rademacher sequence and $i_1, i_2, \ldots$ are the indices of the rows chosen randomly during the construction of S.

Invoking Lemma A.3 we have for fixed $i_1, i_2, \ldots$ that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma}\left[\sup_{x\neq 0}\left|\sum_{k=1}^{N}\sigma_k\frac{\phi(\langle A_{i_k}, x\rangle)}{p_{i_k}\phi(Ax)}\right|^{\ell}\right] \\
&\leq (2L)^{\ell}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{x\neq 0}\left|\sum_{k=1}^{N}\sigma_k\frac{|\langle A_{i_k}, x\rangle|^p}{p_{i_k}\phi(Ax)}\right|^{\ell}\right] \\
&\leq \left(\frac{2L}{\gamma}\right)^{\ell}\mathop{\mathbb{E}}_{\sigma}\left[\left(\sup_{x\neq 0}\left|\sum_{k=1}^{N}\sigma_k\frac{|\langle A_{i_k}, x\rangle|^p}{p_{i_k}\|Ax\|_p^p}\right|\right)^{\ell}\right] \\
&\leq \left(\frac{2L}{\gamma}\right)^{\ell}\mathop{\mathbb{E}}_{\sigma}\left[\left(\sup_{\|Ax\|_p=1}\left|\sum_{k=1}^{N}\sigma_k\frac{|\langle A_{i_k}, x\rangle|^p}{p_{i_k}}\right|\right)^{\ell}\right]
\end{aligned}$$

The problem is thus reduced to the analysis of ℓ_p Lewis weight sampling and it has been proved implicitly in (Cohen & Peng, 2015) that

$$\mathop{\mathbb{E}}_{i,\sigma}\left[\left(\sup_{\|Ax\|_p=1}\left|\sum_{k=1}^{N}\sigma_k\frac{|\langle A_{i_k}, x\rangle|^p}{p_{i_k}}\right|\right)^{\ell}\right] \leq \frac{(C_1\varepsilon)^{\ell}}{\operatorname{poly}(N)}.$$

and hence

$$M \leq 2^{2\ell}\cdot\left(\frac{2L}{\gamma}\right)^{\ell}\cdot\frac{(C_1\varepsilon)^{\ell}}{\operatorname{poly}(N)} \leq \frac{(C_2L/\gamma\cdot\varepsilon)^{\ell}}{\operatorname{poly}(N)}.$$

The result follows from Markov's inequality with a rescaling of ε by a factor of $\gamma/(C_2L)$.