# A Missing Proofs from Section 2

In this section, we give the full details of the statements in Section 2. Coreset constructions are known for a variety of problems, e.g., in computational geometry [FMSW10, FL11, BFL16, LK17, SW18, BLUZ19, HV20, Fel20], linear algebra [BDM$^+$20], machine learning [MSSW18, BLG$^+$19, MOB$^+$20]. We first show that coreset construction is adversarially robust by considering the merge and reduce framework. For example, consider the offline coreset construction through sensitivity sampling.

**Lemma A.1 (Lemma 2.3 in [LK17])** *Given $\varepsilon > 0$ and $\delta \in (0,1)$, let $P$ be a set of weighted points, with non-negative weight function $\mu : P \to \mathbb{R}_{\ge 0}$ and let $s : P \to \mathbb{R}^{\ge 0}$ denote an upper bound on the sensitivity of each point. For $S = \sum_{p \in P} \mu(p)s(p)$, let $m = \Omega\left(\frac{S^2}{\varepsilon^2}\left(d' + \log\frac{1}{\delta}\right)\right)$, where $d'$ is the pseudo-dimension of the query space. Let $C$ be a sample of $m$ points from $P$ with replacement, where each point $p \in P$ is sampled with probability $q(p) = \frac{\mu(p)s(p)}{S}$ and assigned the weight $\frac{\mu(p)}{m \cdot q(p)}$ if sampled. Then $C$ is an $\varepsilon$-coreset of $P$ with probability at least $1 - \delta$.*

We first observe that any streaming algorithm that uses linear memory is adversarially robust because intuitively, it can recompute an exact or approximate solution at each step.

**Lemma A.2** *Given a set of points $P$, there exists an offline adversarially robust construction that outputs an $\varepsilon$-coreset of $P$ with probability at least $1 - \delta$.*

**Proof :** Given an adversary $A$, let $P = p_1, \ldots, p_n$ be a set of points such that each $p_i$ with $i \in [n]$ is generated by $A$, possibly as a function of $p_1, \ldots, p_{i-1}$. For example, it may be possible that the points $p_1, \ldots, p_{n/2}$ are a coreset of some set of points $P_1$ and the points $p_{n/2+1}, \ldots, p_n$ (1) were either generated with full knowledge of $p_1, \ldots, p_{n/2}$ or (2) are a coreset of a set of points $P_2$ generated with full knowledge of $P_1$. Let $s(p)$ be an upper bound on the sensitivity of each point in $P$ and consider the sensitivity sampling procedure described in Lemma A.1. We would like to sample each point with probability $q(p)$. Each point in $C$ is chosen to be $p$ with probability $q(p)$. However, if our algorithm generates internal randomness to perform this sampling procedure, it may be possible for an adversary to either learn correlations with the internal randomness or even learn the internal randomness entirely (such as the seed of a pseudorandom generator). Thus the choice for each point of $C$ may no longer be independent, so we are no longer guaranteed that the resulting construction is a coreset.

Instead, suppose the randomness used by the algorithm at time $i$ in the sampling procedure is independent of the choices of $p_1, \ldots, p_{i-1}$, e.g., the algorithm has access to a source of fresh public randomness at each time in the data stream. Then the algorithm can generate $C$ independent of the choices of $p_1, \ldots, p_{i-1}$. Thus by Lemma A.1, $C$ is an $\varepsilon$-coreset of $P$ with probability at least $1 - \delta$. $\square$

We emphasize that Lemma A.2 shows that any offline coreset construction is adversarially robust; the example of sensitivity sampling is specifically catered to our applications of the merge and reduce framework to clustering.

We now prove our main statement.

**Proof of Theorem 1.1:** Let $\delta = \frac{1}{\mathrm{poly}(n)}$ and consider an $\varepsilon$-coreset construction with failure probability $\delta$. We prove that the merge and reduce framework gives an adversarially robust construction for an $\varepsilon$-coreset with probability at least $1 - 2n\delta$. We consider a proof by induction on an input set $P$ of $n$ points, supposing that $n = 2^k$ for some integer $k > 0$. Observe that $C_{0,j}$ is a coreset of $p_j$ for $j \in [n]$ since $C_{0,j} = p_j$. Let $\mathcal{E}_i$ be the event that for a fixed $i \in [k]$ that $C_{i-1,j}$ is an $\frac{\varepsilon}{2k}$-coreset of $C_{i-1,2j-1}$ and $C_{i-1,2j}$ for each $j \in \left[\frac{n}{2^{i-1}}\right]$. By Lemma A.2, it holds that for a fixed $j$, $C_{i,j}$ is an $\frac{\varepsilon}{2k}$-coreset of $C_{i-1,2j-1}$ and $C_{i-1,2j}$ with probability at least $1 - \delta$. By a union bound over $\frac{n}{2^{i-1}}$ possible indices $j$, we have that for a fixed $i$, all $C_{i,j}$ are $\frac{\varepsilon}{2k}$-coresets of $C_{i-1,2j-1}$ and $C_{i-1,2j}$ with probability at least $1 - \frac{n}{2^{i-1}} \cdot \delta$. Thus, $\mathbf{Pr}\left[\mathcal{E}_{i+1}\right] \ge 1 - \frac{n\delta}{2^{i-1}}$, which completes the induction. Hence with $\mathbf{Pr}\left[\cup_{i=0}^{k} \mathcal{E}_i\right]$, we have that the cost induced by $C_{k,1}$ is a $\left(1 + \frac{\varepsilon}{2k}\right)^k$-approximation to the cost induced by $P$. Since $\left(1 + \frac{\varepsilon}{2k}\right)^k \le e^{\varepsilon/2} \le 1 + \varepsilon$, then $C_{k,1}$ is an $\varepsilon$-coreset of $P$ with probability

$\mathbf{Pr}\left[\cup_{i=0}^{k}\mathcal{E}_i\right]$. By a union bound, we have that $\mathbf{Pr}\left[\cup_{i=0}^{k}\mathcal{E}_i\right] \ge 1 - \sum_{i=0}^{k}\left(1 - \mathbf{Pr}\left[\mathcal{E}_{i+1}\right]\right) \ge 1 - 2n\delta$.
$\square$

## B Missing Proofs from Section 3

**Theorem B.1 (Freedman's inequality)** *[Fre75] Suppose $Y_0, Y_1, \ldots, Y_n$ is a scalar martingale with difference sequence $X_1, \ldots, X_n$. Specifically, we initiate $Y_0 = 0$ and set $Y_i = Y_{i-1} + X_i$ for all $i \in [n[$ Let $R \ge |X_t|$ for all $t \in [n]$ with high probability. We define the predictable quadratic variation process of the martingale by $w_k := \sum_{t=1}^{k} \underset{t-1}{\mathbb{E}}\left[X_t^2\right]$, for $k \in [n]$. Then for all $\varepsilon \ge 0$ and $\sigma^2 > 0$, and every $k \in [n]$,*

$$\mathbf{Pr}\left[\max_{t\in[k]} |Y_t| > \varepsilon \text{ and } w_k \le \sigma^2\right] \le 2\exp\left(-\frac{\varepsilon^2/2}{\sigma^2 + R\varepsilon/3}\right).$$

We first show robustness of our algorithm by justifying correctness of approximation for $L_p$ norms.

**Lemma B.2 ($L_p$ subspace embedding)** *Suppose $\varepsilon > \frac{1}{n}$, $p \in \{1,2\}$, and $C > \kappa^p$, where $\kappa$ is an upper bound on the condition number of the stream. Then Algorithm 1 returns a matrix $\mathbf{M}$ such that for all $\mathbf{x} \in \mathbb{R}^d$,*

$$|\,\|\mathbf{Mx}\|_p - \|\mathbf{Ax}\|_p\,| \le \varepsilon \|\mathbf{Ax}\|_p,$$

*with high probability.*

**Proof :** Consider an arbitrary $\mathbf{x} \in \mathbb{R}^d$ and suppose $\varepsilon \in (0, 1/2)$ with $\varepsilon > \frac{1}{n}$. We claim through induction the stronger statement that $|\|\mathbf{M}_j\mathbf{x}\|_p^p - \|\mathbf{A}_j\mathbf{x}\|_p^p| \le \varepsilon\|\mathbf{A}_j\mathbf{x}\|_p^p$ for all times $j \in [n]$ with high probability. Here $\mathbf{M}_j$ is the matrix consisting of the rows of the input matrix $\mathbf{A}$ that have already been sampled at time $j$ and $\mathbf{A}_j = \mathbf{a}_1 \circ \ldots \circ \mathbf{a}_j$. Note that either $\mathbf{a}_1$ is the zero vector or $p_1 = 1$, so that either way, we have $\mathbf{M}_1 = \mathbf{A}_1$ for our base case. We assume the statement holds for all $j \in [n-1]$ and prove it must hold for $j = n$. We implicitly define a martingale $Y_0, Y_1, \ldots, Y_n$ through the difference sequence $X_1, \ldots, X_n$, where for $j \ge 1$, we set $X_j = 0$ if $Y_{j-1} > \varepsilon\|\mathbf{A}_{j-1}\mathbf{x}\|_p^p$ and otherwise if $Y_{j-1} \le \varepsilon\|\mathbf{A}_{j-1}\mathbf{x}\|_p^p$, we set

$$X_j = \begin{cases} \left(\frac{1}{p_j} - 1\right)|\mathbf{a}_j^\top\mathbf{x}|^p & \text{if } \mathbf{a}_j \text{ is sampled in } \mathbf{M} \\ -|\mathbf{a}_j^\top\mathbf{x}|^p & \text{otherwise.} \end{cases} \tag{1}$$

Since $\mathbb{E}\left[Y_j | Y_1, \ldots, Y_{j-1}\right] = Y_{j-1}$, then the sequence $Y_0, \ldots, Y_n$ induced by the differences is indeed a valid martingale. Furthermore, by the design of the difference sequence, we have that $Y_j = \|\mathbf{M}_j\mathbf{x}\|_p^p - \|\mathbf{A}_j\mathbf{x}\|_p^p$.

If $p_j = 1$, then $\mathbf{a}_j$ is sampled in $\mathbf{M}_j$, so we have that $X_j = 0$. Otherwise, we have that

$$\mathbb{E}\left[X_j^2 | Y_1, \ldots, Y_{j-1}\right] = p_j\left(\frac{1}{p_j} - 1\right)^2 |\mathbf{a}_j^\top\mathbf{x}|^{2p} + (1 - p_j)|\mathbf{a}_j^\top\mathbf{x}|^{2p} \le \frac{1}{p_j}|\mathbf{a}_j^\top\mathbf{x}|^{2p}.$$

For $p_j < 1$, then we have $p_j = \alpha\tau_j$ and thus $\mathbb{E}\left[X_j^2 | Y_1, \ldots, Y_{j-1}\right] \le \frac{1}{\alpha\tau_j}|\mathbf{a}_j^\top\mathbf{x}|^{2p}$. By the definition of $\tau_j$ and the inductive hypothesis that $|\|\mathbf{M}_{j-1}\mathbf{x}\|_p^p - \|\mathbf{A}_{j-1}\mathbf{x}\|_p^p\| \le \varepsilon\|\mathbf{A}_{j-1}\mathbf{x}\|_p^p < \frac{1}{2}\|\mathbf{A}_{j-1}\mathbf{x}\|_p^p$, then we have

$$\tau_j \ge \frac{2|\mathbf{a}_j^\top\mathbf{x}|^p}{\|\mathbf{M}_{j-1}\mathbf{x}\|_p^p + |\mathbf{a}_j^\top\mathbf{x}|^p} \ge \frac{|\mathbf{a}_j^\top\mathbf{x}|^p}{\|\mathbf{A}_{j-1}\mathbf{x}\|_p^p + |\mathbf{a}_j^\top\mathbf{x}|^p} = \frac{|\mathbf{a}_j^\top\mathbf{x}|^p}{\|\mathbf{A}_j\mathbf{x}\|_p^p} \ge \frac{|\mathbf{a}_j^\top\mathbf{x}|^p}{\|\mathbf{Ax}\|_p^p}.$$

Thus, $\sum_{j=1}^{n} \mathbb{E}\left[X_j^2 | Y_1, \ldots, Y_{j-1}\right] \le \sum_{j=1}^{n} \frac{\|\mathbf{Ax}\|_p^p \cdot |\mathbf{a}_j^\top\mathbf{x}|^p}{\alpha} \le \frac{\|\mathbf{Ax}\|_p^{2p}}{\alpha}$.

Moreover, we have that $|X_j| \le \frac{1}{p_j}|\mathbf{a}_j^\top\mathbf{x}|^p$. For $p_j = 1$, we have $\frac{1}{p_j}|\mathbf{a}_j^\top\mathbf{x}|^p \le \|\mathbf{A}_j\mathbf{x}\|_p^p \le \|\mathbf{Ax}\|_p^p$. For $p_j < 1$, we have $p_j = \alpha\tau_j < 1$. Again by the definition of $\tau_j$ and by the inductive hypothesis that $|\|\mathbf{M}_{j-1}\mathbf{x}\|_p^p - \|\mathbf{A}_{j-1}\mathbf{x}\|_p^p\| \le \varepsilon\|\mathbf{A}_{j-1}\mathbf{x}\|_p^p < \frac{1}{2}\|\mathbf{A}_{j-1}\mathbf{x}\|_p^p$, we have that

$$\frac{|\langle\mathbf{a}_j, \mathbf{x}\rangle|^p}{2\|\mathbf{A}_j\mathbf{x}\|_p^p} \le \frac{|\langle\mathbf{a}_j, \mathbf{x}\rangle|^p}{|\mathbf{M}_{j-1}\mathbf{x}|_p^p + |\langle\mathbf{a}_j, \mathbf{x}\rangle|^p} \le \tau_j.$$

Hence for $\alpha = \frac{Cd}{\varepsilon^2}\log n$, it follows that

$$|X_j| \le \frac{1}{p_j}|\mathbf{a}_j^\top \mathbf{x}|^p \le \frac{2}{\alpha}\|\mathbf{A}_j\mathbf{x}\|_p^p \le \frac{2\varepsilon^2}{Cd\log n}\|\mathbf{A}_j\mathbf{x}\|_p^p \le \frac{2\varepsilon^2}{Cd\log n}\|\mathbf{A}\mathbf{x}\|_p^p.$$

We would like to apply Freedman's inequality (Theorem B.1) with $\sigma^2 = \frac{\|\mathbf{A}\mathbf{x}\|_p^{2p}}{\alpha}$ for $\alpha = \mathcal{O}\left(\frac{d}{\varepsilon^2}\log n\right)$ and $R \le \frac{2\varepsilon^2}{d\log n}\|\mathbf{A}\mathbf{x}\|_p^p$, as in [BDM⁺20]. However, in the adversarial setting we won't be able bound the probability that $|Y_n|$ exceeds $\varepsilon\|\mathbf{A}\mathbf{x}|_p^p$ using Freedman's inequality as the latter is a random variable. Thus we instead assume that $\kappa_1, \kappa_2$ are constants so that for $p = 1$, we have $\kappa_1$ and $\kappa_2$ are lower and upper bounds on $\|\mathbf{A}\|_1$ and for $p = 2$, we have that $\kappa_1$ and $\kappa_2$ are lower and upper bounds on the singular values of $\mathbf{A}$. We are now ready to apply Freedman's inequality with $\sigma^2 \le \frac{\kappa_2^{2p}\|\mathbf{x}\|_p^{2p}}{\alpha}$ for $\alpha = \mathcal{O}\left(\frac{d}{\varepsilon^2}\log n\right)$ and $R \le \frac{2\varepsilon^2}{d\log n}\kappa_2^p\|\mathbf{x}\|_p^p$. By Freedman's inequality, we have that

$$\mathbf{Pr}\left[|Y_n| > \varepsilon\kappa_1^p\|\mathbf{x}\|_p^p\right] \le 2\exp\left(-\frac{\kappa_1^{2p}\varepsilon^2\|\mathbf{x}\|_p^{2p}/2}{\sigma^2 + R\kappa_1^p\varepsilon\|\mathbf{x}\|_p^p/3}\right) \le 2\exp\left(-\frac{3Cd\kappa_1^{2p}\log n/2}{6\kappa_2^{2p} + 2\kappa_1^p\kappa_2^p}\right) \le \frac{1}{2^d\ \mathrm{poly}(n)},$$

for sufficiently large $C > (\kappa_1/\kappa_2)^p$. Note for $p = 2$, we have the upper bound on the condition number $\kappa \ge \kappa_1/\kappa_2$ so it suffices to set $C = \kappa^2$. Since $\kappa_1^p\|x\|_p^p \le \|\mathbf{A}\mathbf{x}\|_p^p$, then we have

$$\mathbf{Pr}\left[|Y_n| > \varepsilon\|A\mathbf{x}\|_p^p\right] \le \mathbf{Pr}\left[|Y_n| > \varepsilon\kappa_1^p\|\mathbf{x}\|_p^p\right].$$

Thus $|\,\|\mathbf{M}\mathbf{x}\|_p^p - \|\mathbf{A}\mathbf{x}\|_p^p\,| \le \varepsilon\,\|\mathbf{A}\mathbf{x}\|_p^p$ with probability at least $1 - \frac{1}{2^d\ \mathrm{poly}(n)}$. By a rescaling of $\varepsilon$ since $p \le 2$, we thus have that $|\,\|\mathbf{M}\mathbf{x}\|_p - \|\mathbf{A}\mathbf{x}\|_p\,| \le \varepsilon\,\|\mathbf{A}\mathbf{x}\|_p$ with probability at least $1 - \frac{1}{2^d\ \mathrm{poly}(n)}$.

We now show that we can union bound over an $\varepsilon$-net. We first define the unit ball $B = \{\mathbf{A}\mathbf{y} \in \mathbb{R}^n \mid \|\mathbf{A}\mathbf{y}\|_p = 1\}$. We also define $\mathcal{N}$ to be a greedily constructed $\varepsilon$-net of $B$. Since balls of radius $\frac{\varepsilon}{2}$ around each point cannot overlap, but must all fit into a ball of radius $1+\frac{\varepsilon}{2}$, then it follows that $\mathcal{N}$ has at most $\left(\frac{3}{\varepsilon}\right)^d$ points. Therefore, by a union bound for $\frac{1}{\varepsilon} < n$, we have $|\,\|\mathbf{M}\mathbf{y}\|_p - \|\mathbf{A}\mathbf{y}\|_p\,| \le \varepsilon\,\|\mathbf{A}\mathbf{y}\|_p$ for all $\mathbf{A}\mathbf{y} \in \mathcal{N}$, with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$.

We now argue that accuracy on this $\varepsilon$-net implies accuracy everywhere. Indeed, consider any vector $\mathbf{z} \in \mathbb{R}^d$ normalized to $\|\mathbf{A}\mathbf{z}\|_p = 1$. We shall inductively define a sequence $\mathbf{A}\mathbf{y}_1, \mathbf{A}\mathbf{y}_2, \ldots$ such that $\left\|\mathbf{A}\mathbf{z} - \sum_{j=1}^{i}\mathbf{A}\mathbf{y}_j\right\|_p \le \varepsilon^i$ and there exists some constant $\gamma_i \le \varepsilon^{i-1}$ with $\frac{1}{\gamma_i}\mathbf{A}\mathbf{y}_i \in \mathcal{N}$ for all $i$. Define our base point $\mathbf{A}\mathbf{y}_1$ to be the closest point to $\mathbf{A}\mathbf{z}$ in the $\varepsilon$-net $\mathcal{N}$. Then since $\mathcal{N}$ is a greedily constructed $\varepsilon$-net, we have that $\|\mathbf{A}\mathbf{z} - \mathbf{A}\mathbf{y}_1\|_p \le \varepsilon$. Given a sequence $\mathbf{A}\mathbf{y}_1, \ldots, \mathbf{A}\mathbf{y}_{i-1}$ such that $\gamma_i := \left\|\mathbf{A}\mathbf{z} - \sum_{j=1}^{i-1}\mathbf{A}\mathbf{y}_j\right\|_p \le \varepsilon^{i-1}$, note that $\frac{1}{\gamma_i}\left\|\mathbf{A}\mathbf{z} - \sum_{j=1}^{i-1}\mathbf{A}\mathbf{y}_j\right\|_p = 1$. Thus we inductively define the point $\mathbf{A}\mathbf{y}_i \in \mathcal{N}$ so that $\mathbf{A}\mathbf{y}_i$ is within distance $\varepsilon$ of $\mathbf{A}\mathbf{z} - \sum_{j=1}^{i-1}\mathbf{A}\mathbf{y}_j$. Therefore,

$$|\,\|\mathbf{M}\mathbf{z}\|_p - \|\mathbf{A}\mathbf{z}\|_p\,| \le \sum_{i=1}^{\infty}|\,\|\mathbf{M}\mathbf{y}_i\|_p - \|\mathbf{A}\mathbf{y}_i\|_p\,| \le \sum_{i=1}^{\infty}\varepsilon^i\,\|\mathbf{A}\mathbf{y}_i\|_p = \mathcal{O}(\varepsilon)\,\|\mathbf{A}\mathbf{z}\|_p,$$

which completes the induction for time $n$. □

### B.1 Adversarially Robust Spectral Approximation

We observe that Lemma B.2 provides adversarial robustness for free.

**Lemma B.3 (Adversarially robust spectral approximation )** *Algorithm 1 is adversarially robust.*

**Proof :** Let us inspect the proof of Lemma B.2. Observe that since the adversary can observe the past data and the past randomness of Algorithm 1, then the rows $\mathbf{a}_i$ are random variables that depend on the history and the randomness of the algorithm. In other words, $\mathbf{a}_i$ is measurable with respect to the sigma algebra generated by $\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, B_1, \ldots, B_{i-1}, C_i$ where $B_i$ is the indicator of the event that we sample row $a_i$ in Algorithm 1 and $C_i$ is the random vector generated by the adversary at step $i$ to create row $a_i$.

Denote $\mathfrak{F}_i$ the sigma algebra generated by $a_1, \dots, a_{i-1}, B_1, \dots B_{i-1}, C_i$. Then $a_i$ is measurable with respect to $\mathfrak{F}_i$. Let us remind that $Y_j = Y_{j-1} + X_j$ and let us observe that the definition of $X_j$ in Equation (1) can be rewritten as

$$X_j = \left( \left( \frac{1}{p_j} - 1 \right) |\mathbf{a}_j^\top \mathbf{x}|^p B_j - |\mathbf{a}_j^\top \mathbf{x}|^p (1 - B_j) \right) \mathbb{I}_{Y_{j-1} < \varepsilon}. \tag{2}$$

It can be easily checked that

$$\mathbb{E}\left[X_j | \mathfrak{F}_j\right] = \left( \left( \frac{1}{p_j} - 1 \right) |\mathbf{a}_j^\top \mathbf{x}|^p p_j - |\mathbf{a}_j^\top \mathbf{x}|^p (1 - p_j) \right) \mathbb{I}_{Y_{j-1} < \varepsilon} = 0.$$

This is because $Y_{j-1}, p_j, a_j$ are measurable with respect to $\mathfrak{F}_j$. This implies that in the adversarial setting sequence $Y_j$ is a martingale with respect to the filtration $\mathfrak{F}_0 \subset \mathfrak{F}_1 \subset \dots \subset \mathfrak{F}_n$.

The remainder of the proof of Lemma B.2 goes through as is for arbitrary rows $\mathbf{a}_i$'s. Thus, the algorithm is indeed adversarially robust. ☐

We note the established upper bounds on the sum of the online $L_p$ sensitvities, e.g., Theorem 2.2 in [CMP16], Lemma 2.2 and Lemma 4.7 in [BDM+20].

**Lemma B.4 (Bound on Sum of Online $L_p$ Sensitivities)** *[CMP16, BDM+20] Let the rows of $\mathbf{A} = \mathbf{a}_1 \circ \ldots \circ \mathbf{a}_n \in \mathbb{R}^{n \times d}$ arrive in a stream with condition number at most $\kappa$ and let $\ell_i$ be the online $L_p$ sensitivity of $\mathbf{a}_i$. Then $\sum_{i=1}^n \ell_i = \mathcal{O}\left(d \log n \log \kappa\right)$ for $p = 1$ and $\sum_{i=1}^n \ell_i = \mathcal{O}\left(d \log \kappa\right)$ for $p = 2$.*

We note that $\kappa$ is an adversarially chosen parameter, since the rows of the input matrix $\mathbf{A}$ are generated by an adversary. One can mitigate possible adversarial space attacks by tracking $\kappa$ and aborting if $\log \kappa$ exceeds a desired threshold.

**Proof of Lemma 3.7:** Algorithm 1 is adversarially robust by Lemma B.3. It remains to analyze the space complexity of Algorithm 1. By Lemma B.3 and a union bound over the $n$ rows in the stream, each row $\mathbf{a}_i$ is sampled with probability at most $4\alpha\tau_i$, where $\tau_i$ is the online leverage score of row $\mathbf{a}_i$. By Lemma B.4, we have $\sum_{i=1}^n \tau_i = \mathcal{O}\left(d \log \kappa\right)$ and we also set $\alpha = \mathcal{O}\left(\frac{d\kappa}{\varepsilon^2} \log n\right)$. Let $\gamma > 0$ be a sufficiently large constant such that $\sum_{i=1}^n \alpha\tau_i \le \frac{d^2\gamma\kappa\log\kappa}{\varepsilon^2} \log n$.

We use a martingale argument to bound the number of rows that are sampled. Consider a martingale $U_0, U_1, \dots, U_n$ with difference sequence $W_1, \dots, W_n$, where for $j \ge 1$, we set $W_j = 0$ if $U_{j-1} > \frac{d^2\gamma\kappa\log\kappa}{\varepsilon^2} \log n$ and otherwise if $U_{j-1} \le \frac{d^2\gamma\kappa\log\kappa}{\varepsilon^2} \log n$, we set

$$W_j = \begin{cases} 1 - p_j & \text{if } \mathbf{a}_j \text{ is sampled in } \mathbf{M} \\ -p_j & \text{otherwise.} \end{cases} \tag{3}$$

We have $\mathbb{E}\left[U_j | U_1, \dots, U_{j-1}\right] = U_{j-1}$, then the sequence $U_0, \dots, U_n$ induced by the differences is indeed a valid martingale. Note that intuitively, $U_n$ is the difference between the number of sampled rows and $\sum_{j=1}^n p_j$.

Since $\mathbf{a}_j$ is sampled with probability $p_j \in [0, 1]$,

$$\mathbb{E}\left[W_j^2 | U_1, \dots, U_{j-1}\right] \le \sum_{j=1}^n p_j \le \sum_{j=1}^n \alpha\tau_j.$$

Moreover, we have $\mathbb{E}\left[|W_j| \,|\, U_1, \dots, U_{j-1}\right] \le 1$. Thus by Freedman's inequality (Theorem B.1) with $\sigma^2 = \sum_{j=1}^n \alpha\tau_j \le \frac{d^2\gamma\kappa\log\kappa}{\varepsilon^2} \log n$ and $R \le 1$,

$$\mathbf{Pr}\left[|U_n| > \frac{d^2\gamma\kappa\log\kappa}{\varepsilon^2} \log n\right] \le 2\exp\left(-\frac{d^4\gamma^2\kappa^2\log^2\kappa\log^2 n/(2\varepsilon^4)}{\sigma^2 + Rd^2\gamma\kappa\log\kappa\log n/(3\varepsilon^2)}\right) \le \frac{1}{\text{poly}(n)}.$$

Hence we have that with high probability, the number of rows sampled is $\mathcal{O}\left(\frac{1}{\varepsilon^2} d^2\kappa\log\kappa\log n\right)$. ☐

We remark that the space bounds for Lemma 3.7 could similarly be shown (with constant probability of success) using Markov's inequality though analysis Freedman's inequality provides much higher guarantees in terms of probability of success.

On the other hand, it is not clear how to execute a similar strategy using the Matrix Freedman's Inequality rather than using Freedman's inequality. This is because to obtain the desired spectral bound, we must define a martingale at time $j$ in terms of both the matrix $\mathbf{A}_j$ and whether the rows $\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}$ were previously sampled. However, since $\mathbf{A}_j$ is itself a function of whether $\mathbf{a}_1, \ldots \mathbf{a}_{j-1}$ were previously sampled, the resulting sequence is not a valid martingale.

We first require the following bound on the sum of the online ridge leverage scores, e.g., Theorem 2.12 from [BDM$^+$20], which results from considering Lemma 2.11 in [BDM$^+$20] at $\mathcal{O}\left(\log n\right)$ different scales.

**Lemma B.5 (Bound on Sum of Online Ridge Leverage Scores)** *[BDM$^+$20] Let the rows of* $\mathbf{A} = \mathbf{a}_1 \circ \ldots \circ \mathbf{a}_n \in \mathbb{R}^{n \times d}$ *arrive in a stream with condition number at most* $\kappa$, *let* $\lambda_i = \frac{\|\mathbf{A}_i - (\mathbf{A}_i)_{(k)}\|_F^2}{k}$, *where* $\mathbf{A}_i = \mathbf{a}_1 \circ \ldots \circ \mathbf{a}_i$ *and* $(\mathbf{A}_i)_{(k)}$ *is the best rank* $k$ *approximation to* $\mathbf{A}_i$. *Let* $\ell_i$ *be the online ridge leverage score of* $\mathbf{a}_i$ *with regularization* $\lambda_i$. *Then* $\sum_{i=1}^n \ell_i = \mathcal{O}\left(k \log n \log \kappa\right)$.

From Lemma B.5 and a similar argument to Lemma B.2, we also obtain adversarially robust projection-cost preservation and therefore low-rank approximation. Namely, [CMM17, BDM$^+$20] showed that projection-cost preservation essentially reduces to sampling a weighted submatrix $\mathbf{M}$ of $\mathbf{A}$ such that $\|\mathbf{M}\mathbf{x}\|_2^2 + \lambda\|x\|_2^2 \in (1 \pm \varepsilon)(\|\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|x\|_2^2)$ for a ridge parameter $\lambda$. Since the online ridge leverage score of each row $\mathbf{a}_i$ can be rewritten as $\max_{\mathbf{x} \in \mathbb{R}^d} \frac{\langle \mathbf{a}_i, x\rangle^2 + \lambda\|\mathbf{x}\|_2^2}{\|\mathbf{A}_i\mathbf{x}\|_2^2 + \lambda\|x\|_2^2}$, then the same concentration argument of Lemma B.2 gives Lemma 3.8.

### B.2 Adversarially Robust Linear Regression

We first give the formal definition of linear regression:

**Problem B.6 (Linear Regression)** *Given a matrix* $\mathbf{A} \in \mathbb{R}^{n \times d}$, *a vector* $\mathbf{b} \in \mathbb{R}^n$ *and an approximation parameter* $\varepsilon > 0$, *the goal is to output a vector* $\mathbf{y}$ *such that* $\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2 \le (1+\varepsilon) \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$.

**Lemma B.7 (Adversarially Robust Linear Regression)** *Given* $\varepsilon > 0$ *and a matrix* $\mathbf{A} \in \mathbb{R}^{n \times d}$ *whose rows* $\mathbf{a}_1, \ldots, \mathbf{a}_n$ *arrive sequentially in a stream with condition number at most* $\kappa$, *there exists an adversarially robust streaming algorithm that outputs a* $(1+\varepsilon)$ *approximation to linear regression and uses* $\mathcal{O}\left(\frac{d^3}{\varepsilon^2} \log^2 n \log \kappa\right)$ *bits of space, with high probability.*

**Proof :** Suppose each row of $\mathbf{A}$ arrives sequentially, along with the corresponding entry in $\mathbf{b}$. Let $\mathbf{B} = \mathbf{A} \circ \mathbf{b}$ so that the effectively, the rows of $\mathbf{B}$ arrive sequentially. Note that if $\mathbf{M}$ is a spectral approximation to $\mathbf{B}$, then we have

$$(1-\varepsilon)\left\|\mathbf{B}\mathbf{v}\right\|_2 \le \left\|\mathbf{M}\mathbf{v}\right\|_2 \le (1+\varepsilon)\left\|\mathbf{B}\mathbf{v}\right\|_2$$

for all vectors $\mathbf{v} \in \mathbb{R}^{d+1}$. In particular, let $\mathbf{w} \in \mathbb{R}^{d+1}$ be the vector that minimizes $\left\|\mathbf{M}\mathbf{v}\right\|_2$ subject to the constraint that the last coordinate of $\mathbf{w}$ is 1, and let $\mathbf{w} = \begin{bmatrix}\mathbf{y}\\1\end{bmatrix}$. Then we have

$$\left\|\mathbf{A}\mathbf{y} - \mathbf{b}\right\|_2 = \left\|\mathbf{B}\mathbf{w}\right\|_2 \le \frac{1}{1-\varepsilon}\left\|\mathbf{M}\mathbf{w}\right\|_2.$$

Let $\mathbf{z}$ be the vector that minimizes $\left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|_2$ and let $\mathbf{u} = \begin{bmatrix}\mathbf{z}\\1\end{bmatrix}$. Then we have

$$\left\|\mathbf{A}\mathbf{z} - \mathbf{b}\right\|_2 = \left\|\mathbf{B}\mathbf{u}\right\|_2 \ge \frac{1}{1+\varepsilon}\left\|\mathbf{M}\mathbf{u}\right\|_2 \ge \frac{1}{1+\varepsilon}\left\|\mathbf{M}\mathbf{w}\right\|_2,$$

where the last inequality follows from the minimality of $\mathbf{w}$. Thus we have that $\left\|\mathbf{A}\mathbf{y} - \mathbf{b}\right\|_2 \le (1 + \mathcal{O}\left(\varepsilon\right))\left\|\mathbf{A}\mathbf{z} - \mathbf{b}\right\|_2$. □

## C Missing Proofs from Section 4

**Other Related Works** Note that there is an alternate streaming algorithm for graph sparsification given in [GKK10] which has the same guarantees but is computationally faster. However, we choose to analyze the algorithm of [AG09] since its core argument is sampling based. Nevertheless, it is possible that the algorithm from [GKK10] is also adversarially robust. Lastly, we recall that our model is the streaming model where edges arrive one at a time. There is also related work in the dynamic streaming model (see [KNST19] and references therein) where previously shown edges can be deleted but this is not the scope of our work.

The notion of the connectivity of an edge is needed to in the algorithm of [AG09].

**Definition C.1 (Connectivity [BK96])** *A graph is $k$-strong connected iff every cut in the graph has value at least $k$. A $k$-strong connected component is a maximal node-induced subgraph which is $k$-strong connected. The connectivity of an edge $e$ is the maximum $k$ such that there exists a $k$-strong connected component that contains $e$.*

**Algorithm 2** Graph sparsification algorithm from [AG09].

```
Input: A stream of edges $e_1, \cdots, e_m$ and an accuracy parameter $\varepsilon > 0$
Output: Sparified graph $H$
$H \leftarrow \emptyset$
$\rho \leftarrow C(\log n + \log m)/\varepsilon^2$ for sufficiently large constant $C > 0$
for each new edge $e$ do
    compute the connectivity $c_e$ of $e$ in $H$
    $p_e = \min(\rho/c_e, 1)$                    ▷Importance of edge $e$, see Definition C.1
    Add $e$ to $H$ with probability $p_e$ and weight $1/p_e$ times its original weight
return $H$
```

We begin by providing a brief overview of our proof. The first step is to show that for a cut in $G$ of value $c$, the same cut in the sparsified graph $H$ has value that concentrates around $c$. Note that in [AG09], the concentration inequality they obtain *depends* roughly on $\exp(-c)$. In other words, they get a stronger concentration for larger cuts in the original graph. However, their concentration inequality is not valid in our setting since the value $c$ is *random*. Therefore, we employ a different concentration inequality, namely Freedman's inequality (Theorem Theorem B.1) in conjunction with an assumption about the sizes of cuts in the graph to obtain concentration for a fixed cut. The second step is to use a standard worst-case union bound strategy to bound the total number of cuts with a particular size in the original graph. This uses the standard fact that the number of cuts in a graph that is at most $\alpha$ times the minimum cut is at most $n^{2\alpha}$. Then the final result for the property $(1)$ in Problem 4.1 follows by combining the union bound with the previously mentioned concentration inequality. The bound for the total number of edges (condition $(2)$ in Problem 4.1) is a "worst case" calculation in [AG09] so it automatically ports over to our setting. Note that we assume $\kappa_1$ and $\kappa_2$ to be deterministic lower and upper bounds on the size of any cut in $G$ and define $\kappa$ to be their ratio.

**Theorem 1.3** *Given a weighted graph $G = (V, E)$ with $|V| = n$ whose edges $e_1, \ldots, e_m$ arrive sequentially in a stream, there exists an adversarially robust streaming algorithm that outputs a $1 \pm \varepsilon$ cut sparsifier with $\mathcal{O}\left(\frac{\kappa^2 n \log n}{\varepsilon^2}\right)$ edges with probability $1 - 1/\operatorname{poly}(n)$.*

**Proof :** We claim through induction the stronger statement that the value $C_H$ of any cut in $H$ is a $(1+\varepsilon)$-approximation of the value $C_G$ of the corresponding cut in $G$ for all times $j \in [m]$ with high probability. Consider a fixed set $S \subseteq V$ and the corresponding cut $C = (S, V \setminus S)$. Let $e_1, \ldots, e_m$ be the edges of the stream in the order that they arrive. We emphasize that $e_1, \ldots, e_m$ are possibly random variables given by the adversary rather than fixed edges. For each $j \in [m]$, let $G_j$ be the graph consisting of the edges $e_1, \ldots, e_j$ and let $H_j$ be the corresponding sampled weighted subgraph. We abuse notation and define $p_j := p_{e_j}$ to denote the probability of sampling the edge $e_j$ that arrives at time $j$. We use $C_G^{(j)}$ and $C_H^{(j)}$ to denote the value of the cut at time $j$ in graphs $G$ and $H$, respectively. Note that $p_1 = 1$, so we have $H_1 = G_1$ for our base case.

We assume the statement holds for all $j \in [m-1]$ and prove it must hold for $j = m$. We define a martingale $Y_0, Y_1, \ldots, Y_m$ through its difference sequence $X_1, \ldots, X_m$, where for $j \geq 1$, we set

$X_j = 0$ if $C_H^{(j-1)} \notin (1 \pm \varepsilon) C_G^{(j-1)}$. Otherwise if $(1-\varepsilon)C_G^{(j-1)} \le C_H^{(j-1)} \le (1+\varepsilon)C_G^{(j-1)}$, then we set

$$X_j = \begin{cases} 0 & \text{if } e_j \text{ does not cross the cut } C \\ \left(\frac{1}{p_j} - 1\right) & \text{if } e_j \text{ crosses the cut and is sampled in } H \\ -1 & \text{if } e_j \text{ crosses the cut and is not sampled in } H. \end{cases} \tag{4}$$

Because $\mathbb{E}\left[Y_j | Y_1, \ldots, Y_{j-1}\right] = Y_{j-1}$, then we have that the sequence $Y_0, \ldots, Y_n$ is indeed a valid martingale and that $Y_j = C_H^{(j)} - C_G^{(j)}$. (We abuse notation and use $Y_1, \ldots, Y_i$ to indicate the similar filtration to the one in Lemma Lemma B.3).

If $p_j = 1$, then $e_j$ is sampled in $H_j$, so we have that $X_j = 0$. Otherwise,

$$\mathbb{E}\left[X_j^2 | Y_1, \ldots, Y_{j-1}\right] = p_j \left(\frac{1}{p_j} - 1\right)^2 + (1 - p_j) \le \frac{1}{p_j}.$$

For $p_j < 1$, then we have $p_j = \rho / c_{e_j}$ and thus $\mathbb{E}\left[X_j^2 | Y_1, \ldots, Y_{j-1}\right] \le \frac{c_{e_j}}{\rho}$. Thus, $\sum_{j=1}^n \mathbb{E}\left[X_j^2 | Y_1, \ldots, Y_{j-1}\right] \le \sum_{j : e_j \in C} \frac{c_{e_j}}{\rho}$. Recall that $c_{e_j}$ is the connectivity of $e_j$ in $H$ rather than $G$. However, by the definition of $c_{e_j}$ and the inductive hypothesis that $H_{j-1}$ is a $(1+\varepsilon)$ cut sparsifier of $G_{j-1}$, then we have that for $\varepsilon < \frac{1}{2}$, the connectivity of $c_{e_j}$ in $H$ is within a factor of two of the connectivity of $c_{e_j}$ in $G$. By definition of connectivity, we have that the connectivity of $c_{e_j}$ at time $j$ in $G$ is at most $C_G^{(j)} \le C_G^{(m)}$ if $e_j$ crosses the cut $C$. Hence,

$$\sum_{j=1}^m \mathbb{E}\left[X_j^2 | Y_1, \ldots, Y_{j-1}\right] \le \sum_{j : e_j \in C} \frac{C_G^{(j)}}{\rho} \le \frac{2(C_G^{(m)})^2}{\rho}.$$

By similar reasoning, we have $|X_j| \le \frac{1}{p_j} \le \frac{c_{e_j}}{\rho} \le \frac{2(C_G^{(m)})}{\rho}$. Now we would like to apply Freedman's inequality (Theorem B.1) with $\sigma^2 = \frac{2(C_G^{(m)})^2}{\rho}$ and $R \le \frac{2(C_G^{(m)})}{\rho}$ for $\rho = C(\log n + \log m)/\varepsilon^2$. However, we cannot bound the probability that $|Y_n|$ exceeds $\varepsilon C_G^{(m)}$, as the latter is a random variable. Thus we instead assume that $\kappa_1$ and $\kappa_2$ are lower and upper bounds on $C_G^{(m)}$. By Freedman's inequality,

$$\mathbf{Pr}\left[|Y_n| > \varepsilon \kappa_1\right] \le 2 \exp\left(-\frac{\kappa_1^2 \varepsilon^2 / 2}{\sigma^2 + R\kappa_1 \varepsilon / 3}\right) \le 2 \exp\left(-\frac{3C\kappa_1^2 \log n / 2}{6\kappa_2^2 + 2\kappa_1 \kappa_2}\right) \le n^{-O(C/\kappa^2)},$$

where we define $\kappa := \kappa_2 / \kappa_1$. Since $\kappa_1 \le C_G^{(m)}$, then we have

$$\mathbf{Pr}\left[|Y_n| > \varepsilon C_G^{(m)}\right] \le \mathbf{Pr}\left[|Y_n| > \varepsilon \kappa_1\right].$$

Thus $|C_H^{(m)} - C_G^{(m)}| \le \varepsilon C_G^{(m)}$ with probability at least $1 - n^{-O(C/\kappa^2)}$.

We now union bound over all cuts $C$. Based on our assumption that every cut in $G$ has value at least $\kappa_1$, it follows that for any $\alpha \ge 1$, the number of cuts in $G$ of size $\alpha \kappa_1$ is at most $n^{2\alpha}$ [BK96, AG09]. Note that we are using a *deterministic* upper bound on the number of cuts that holds for any graph. Due to our assumption, on the size of cuts, we know that $\alpha$ ranges from $1 \le \alpha \le \kappa_2 / \kappa_1 = \kappa$. Then using our concentration result derived above, it follows by a union bound that the probability that there exists some $C$ such that $|C_H^{(m)} - C_G^{(m)}| \le \varepsilon C_G^{(m)}$ is at most

$$\int_1^{\kappa_2/\kappa_1} n^{2\alpha} \cdot n^{-O(C/\kappa^2)} \, d\alpha \le \frac{n^{2\kappa}}{2\log(\kappa)} \cdot n^{-O(C/\kappa^2)} \le \frac{1}{\mathrm{poly}(n)}$$

where the last inequality follows by setting $C = c'\kappa^2$ for some large enough constant $c' > 1$. This verifies part $(1)$ of Problem 4.1.

We now need to check the number of edges in $H$. For this, we note that the proof of Theorem 3.2 in [AG09] carries over to our setting since the proof there only relies on the fact that if an edge has strong connectivity at most $z$ in $G$, its weight in $H$ is at most $z/\rho$ in $H$ which is true for us as well. The extra $\kappa^2$ factor in the number of edges comes from our setting of the parameter $C$ in $\rho$. □