
The Cost of Compression: Tight Quadratic Black-Box Attacks on Sketches for ℓ_2 Norm Estimation

Sara Ahmadian

Google Research

sahmadian@google.com

Edith Cohen

Google Research and Tel Aviv University

edith@cohenwang.com

Uri Stemmer

Tel Aviv University and Google Research

u@uri.co.il

Abstract

Dimensionality reduction via linear sketching is a powerful and widely used technique, but it is known to be vulnerable to adversarial inputs. We study the *black-box adversarial setting*, where a fixed, hidden sketching matrix $A \in \mathbb{R}^{k \times n}$ maps high-dimensional vectors $v \in \mathbb{R}^n$ to lower-dimensional sketches $Av \in \mathbb{R}^k$, and an adversary can query the system to obtain approximate ℓ_2 -norm estimates that are computed from the sketch.

We present a *universal, nonadaptive attack* that, using $\tilde{O}(k^2)$ queries, either causes a failure in norm estimation or constructs an adversarial input on which the optimal estimator for the query distribution (used by the attack) fails. The attack is completely agnostic to the sketching matrix and to the estimator—it applies to *any* linear sketch and *any* query responder, including those that are randomized, adaptive, or tailored to the query distribution.

Our lower bound construction tightly matches the known upper bounds of $\tilde{\Omega}(k^2)$, achieved by specialized estimators for Johnson–Lindenstrauss transforms and AMS sketches. Beyond sketching, our results uncover structural parallels to adversarial attacks in image classification, highlighting fundamental vulnerabilities of compressed representations.

1 Introduction

Dimensionality reduction is a fundamental technique in data analysis, algorithm design, and machine learning. A common paradigm is to apply a *sketching map*, a compressive transformation $C : \mathbb{R}^n \rightarrow \mathbb{R}^k$, which maps a high-dimensional input vector to a lower-dimensional representation. The map is typically sampled from a known distribution (e.g., The Johnson–Lindenstrauss (JL) transform [29]) or learned during training, and is designed to support estimation of specific properties of the input v , such as norms, inner products, or distances, using only the sketch $C(v)$.

Once constructed, the sketching map C typically remains *fixed* across all inputs. This is true when compression layers are part of a trained model, and it is necessary in algorithmic contexts that require *composability*—the ability to sketch distributed datasets independently and combine the results—without revisiting the original data. A fixed sketching map also supports downstream tasks that operate directly in sketch space.

However, this compression introduces an intrinsic vulnerability: small input perturbations can produce large changes in the sketch, even when the true property (such as the norm) is nearly unchanged. For linear sketching maps $A \in \mathbb{R}^{k \times n}$, this vulnerability arises from structural facts such as the existence

of nontrivial null-space vectors and directions along which small-norm perturbations yield large distortions.

Such *adversarial inputs* can be constructed easily in the white-box setting, where the sketching map C is known. In this work, we focus on the *black-box attack model*, which captures realistic situations where the adversary does not see the sketching map directly but can access it interactively via a *responder* using queries of the following form:

- (i) The adversary chooses a query vector $\mathbf{v} \in \mathbb{R}^n$.
- (ii) The system receives \mathbf{v} and computes a sketch $C(\mathbf{v})$.
- (iii) The responder selects a map ψ of sketches to distributions, receives the sketch $C(\mathbf{v})$ from system and returns a (possibly randomized) value $s \sim \psi(C(\mathbf{v}))$ to the adversary.

The goal of the adversary, in general terms, is to compromise the sketching map by causing a failure: If responses are correct, construct an adversarial input. We measure the efficiency of such *attacks* by the *number of queries* they require, as a function of the sketch size k .

We distinguish between two settings. In the *nonadaptive* case, all queries (including the final candidate) are chosen without regard to responses from earlier queries. The adversary performs a blind search, and success depends on the density of adversarial directions in the input space. In contrast, in the *adaptive* setting, each query can depend on prior responses, allowing the adversary to extract information about the sketching map and potentially converge to an adversarial input more efficiently.

The black-box model in the adaptive setting is well studied across multiple areas, including statistical queries [19, 28, 31, 25, 17, 6], sketching and streaming algorithms [33, 26, 9, 27, 40, 5, 8, 13, 15, 1, 22, 16], dynamic graph algorithms [36, 2, 20, 24, 39, 7], and machine learning [37, 21, 4, 34, 32, 35].

In this study, we aim to better understand this vulnerability for the task of ℓ_2 -norm estimation and to gain more general insights through this lens. Specifically, we consider linear transformations specified by a *sketching matrix* $A \in \mathbb{R}^{k \times n}$ which maps $\mathbf{v} \in \mathbb{R}^n$ to a *sketch* $A\mathbf{v} \in \mathbb{R}^k$.

Two classic methods for this task are the Johnson–Lindenstrauss (JL) transform [29] and the AMS sketch [3]. They define *distributions* over sketching matrices that approximately preserve Euclidean norms (and therefore support sketch-based approximation of distances). The provided guarantees are probabilistic: for any input vector, with high probability over the random choice of the sketching matrix $A \in \mathbb{R}^{k \times n}$, the scaled norm of the sketch $\|A\mathbf{v}\|_2$ closely approximates $\|\mathbf{v}\|_2$ (within relative error ϵ with probability $1 - \delta$ when $k = O(\epsilon^{-2} \log(1/\delta))$).

However, as said, in practice, the sketching matrix is typically *fixed* over all input vectors, and importantly, no fixed matrix can preserve approximate norms for *all* inputs: every sketching matrix inevitably has inputs on which it fails. In the nonadaptive setting, the guarantees of JL and AMS apply via a union bound: with high probability (at least $1 - \delta$) over the sampling of matrices, the fixed sampled sketching matrix supports up to $\delta e^{O(k\epsilon^2)}$ approximate norm queries, in the sense that with probability $1 - \delta$, they are all accurate to within relative error ϵ . Therefore, finding a bad input requires a number of queries that is exponential in k .

The adaptive setting was studied in multiple works. In terms of positive results, it is known that JL and AMS sketches can effectively trade off baseline (non-adaptive) accuracy and robustness within a fixed sketch size budget of k . By using carefully designed “robust” estimators [27, 10], the sketch can support for a fixed ϵ , a number of adaptive queries that is *quadratic* in k . The idea, based on [18, 6], is to protect information on the sketching matrix by adding noise to the returned “best possible” estimate (or by subsampling a part of the sketch for each response).

As for negative results, Hardt and Woodruff [26] constructed an attack of size polynomial in k that for any sketching matrix constructs a distribution over inputs under which any estimator would fail. The idea in the attack is to identify vectors that lie close to the null space of A and hence are transparent to the sketch. Cherapanamjeri and Nelson [12], Ben-Eliezer et al. [9] presented an attack of size linear in k on the JL and AMS sketches with the standard estimator (which returns a scaled norm of the sketch). The product of these attack is a vector on which the standard estimator fails. Cohen et al. [15] constructed an attack of quadratic size on the AMS sketch that is *universal* (applies against any query responder), that constructs a vector on which the standard estimator fails. These negative results vary by (i) the scope of the sketching matrices compromised (general or JL/AMS), (ii) the

power of the query responder (strategic and adaptive or the standard estimator), and (iii) the product of the attack (a distribution that fails any query responder or a single vector with an out-of-distribution sketch that fails the “optimal” estimator).

An intriguing gap remains, however, between the established quadratic guarantee for the number of adaptive queries with correct responses and the super-quadratic sizes of the known attacks for general sketching matrices.

1.1 Overview of contributions

Our primary contribution in this work, is a construction of an attack of quadratic size in k that applies against *any* sketching matrix $A \in \mathbb{R}^{k \times n}$ and with any query responder for ℓ_2 norm estimation. The attack produces a vector with an out-of-distribution sketch:

Theorem 1.1 (Attack Properties). *There exist a universal constant $C > 0$, and families of distributions \mathcal{F}_n over \mathbb{R}^n , such that the following holds.*

For every sketching matrix $A \in \mathbb{R}^{k \times n}$ with $n = \Omega(k^2)$, any query responder, and deviation $\gamma \geq 1$ and accuracy $\alpha \in (0, 1)$ parameters, with probability at least 0.9 over the choice of a distribution $\mathcal{D} \sim \mathcal{F}$, after $r = C\gamma^2\alpha^2k^2\log^2 k$ i.i.d. queries $\mathbf{v} \sim \mathcal{D}$, one of the following outcomes occurs:

- (i) *At least $\delta(\alpha) > 0^1$ fraction of the responses have relative error greater than α , or*
- (ii) *We construct a query vector on which the optimal estimator (with respect to A and \mathcal{D}) returns a value that deviates from the true norm by at least a multiplicative factor of γ .*

Our attack deploys a simple and natural query distribution that combines a weighted sparse signal with additive dense noise:

$$\mathbf{v} = w\mathbf{e}_h + \mathbf{u},$$

where \mathbf{e}_h is the standard basis vector corresponding to index $h \in [n]$, and \mathbf{u} is Gaussian noise supported on a set $M \subset [n] \setminus \{h\}$ of size $m > k^2$.

Importantly, sparse queries alone are insufficient for attack: under the mild assumption that the sketching matrix A has its columns in general position, the ℓ_2 norm of any k -sparse vector can be exactly recovered from its sketch (e.g., [11]). This implies that if the attack is restricted to sparse inputs, the response leaks no information about the sketching matrix and full robustness is preserved.

To sample from our query distribution, we first choose the *signal coordinate* $h \sim [n]$ uniformly at random, and then independently sample a *noise support* $M \subset [n] \setminus \{h\}$ of size $m = \Omega(k^2)$.

The goal of the query responder is to estimate the ℓ_2 norm of \mathbf{v} from its sketch $A\mathbf{v}$, with relative error at most α . Our attack remains effective even against the simpler *norm gap* task: return -1 if $\|\mathbf{v}\|_2 \leq 1$ and return 1 if $\|\mathbf{v}\|_2 \geq 1 + \alpha$, with either output permitted when $\|\mathbf{v}\|_2 \in (1, 1 + \alpha)$.

Note that the norm gap task reveals strictly less information than norm estimation: a norm estimate with relative error $\Theta(\alpha)$ trivially yields a correct solution to the norm gap task with parameter $\Theta(\alpha)$, but not vice versa.

Our attack is described in Algorithm 1. Query vectors $(\mathbf{v}^{(t)})_{t \in [r]}$ are constructed by sampling a signal value $w^{(t)} \sim W$, where W is a probability distribution over \mathbb{R} , and Gaussian noise $\mathbf{u}^{(t)}$ to form $\mathbf{v}^{(t)} = w^{(t)}\mathbf{e}_h + \mathbf{u}^{(t)}$. The adversary collects the responses $s^{(t)}$ for the sketch $A\mathbf{v}^{(t)}$. We establish that if the responses are correct for the norm gap (except for a small fraction of queries) then the normalized signed sum of the noise vectors $\sum_t s^{(t)}\mathbf{u}^{(t)}$ is adversarial.

Universality and Limitations Our attack is *universal* in that it applies with *any* query responder. The analysis allows the responder to be strategic and adaptive, with full knowledge of the query distribution \mathcal{D} and the internal state of the attacker. Notably, the attack is *single batch* – it uses adaptivity in the minimal possible way: all queries are generated independently and only the final adversarial vector is constructed adaptively from the responses.²

¹A constant that depends on α

²Single batch attacks were constructed in prior works for JL, Count-Sketch, AMS, and Cardinality sketches [12–14, 1].

A limitation of our result is that the attack guarantees failure only for the *optimal estimator* tailored to \mathcal{D} and A , rather than for *every* possible query responder. The stronger goal—constructing a distribution over inputs that defeats all responders with high probability—remains open.

Despite this, we believe our result is both theoretically meaningful and practically relevant. Theoretically, we obtain a *tight* quadratic bound in the batch-query model, whereas known attacks with the stronger guarantee require significantly more queries (a higher degree polynomial in k). Moreover, any attack achieving the stronger guarantee would require at least $\tilde{\Omega}(k)$ *adaptive batches*, and thus cannot be realized within our single-batch setting. In this sense, the adversarial vector we construct is the strongest outcome achievable in our setting.

Our attack product is practically relevant because the optimization process in a model training tends to converge to an (at least locally) optimal estimator for the training distribution. Therefore, compromising the model means compromising this specific implemented estimator, coded in the model parameter values, rather than any possible query responder.

Roadmap Our attack algorithm is described in Section 3 with the analysis presented in Section 4. An empirical study of our attack on JL and AMS sketching matrices is included in Section 5. We conclude in Section 6 with a discussion of open directions and implications to image classifiers.

2 Preliminaries

We denote vectors in boldface \mathbf{v} , scalars v in non boldface, and inner product of two vectors by $\langle \mathbf{v}, \mathbf{u} \rangle = \sum_i v_i u_i$. For a vector $\mathbf{v} \in \mathbb{R}^n$, we refer to $i \in [n]$ as a *key* and to v_i as the value of the *i*th key. For $M \subset [n]$ let v_M be the projection of \mathbf{v} on entries $(v_i)_{i \in M}$. To streamline the presentation, we interchangeably use the same notation to refer to both a random variable and a distribution.

$\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean 0 and variance σ^2 with density function

$$\varphi_{\mu, \sigma^2}(x) := \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}. \quad (1)$$

$\mathcal{N}_\ell(0, \sigma^2)$ is the ℓ -dimensional Gaussian distribution with covariance matrix I_ℓ (ℓ i.i.d. $\mathcal{N}(0, \sigma^2)$). Its probability density function is

$$f_\sigma(\mathbf{u}) = \frac{1}{(\sigma\sqrt{2\pi})^\ell} e^{-\frac{\|\mathbf{u}\|_2^2}{2\sigma^2}}. \quad (2)$$

A *linear sketching map* is defined by a *sketching matrix* $A \in \mathbb{R}^{k \times n}$ where $k \ll n$. The input is represented as a vector $\mathbf{v} \in \mathbb{R}^n$ and the *sketch* of \mathbf{v} is the product $A\mathbf{v} \in \mathbb{R}^k$.

3 Attack description

Definition 3.1 ((y, α)-Gap Problem). Given width parameter $\alpha > 0$ and $y \in \mathbb{R}$, a *gap problem* is, for input $x \in \mathbb{R}$ to return -1 when $x \leq y$ and to return 1 when $x \geq y + \alpha$. The output may be arbitrary in $\{-1, 1\}$ if $\|\mathbf{v}\|_2 \in (y, y + \alpha)$.

Observe that an additive approximation of $\alpha/2$ or a multiplicative approximation of $\alpha/(2(y + \alpha))$ for x yields a correct solution to the (ℓ, α) -gap problem for x . Hence an attack that is effective with the weaker norm gap responses is more powerful.

Our attack is described in Algorithm 1. The signal density and parameter settings are described in Definition 3.3.

In each of r attack steps,

1. The adversary samples an independent query vector $\mathbf{v} := w\mathbf{e}_h + c\mathbf{z}$ by sampling Gaussian noise with support $M \subset [n]$ $\mathbf{z} \sim \mathcal{N}_{n,M}(0, \sigma^2 = 1/m)$, sampling signal weight $w \sim W$.
2. The responder chooses an estimator map $\psi : \mathbb{R}^k \rightarrow \mathcal{P}(\{-1, 1\})$, obtains the sketch $A\mathbf{v}$ from the system, and returns $s \sim \psi(A\mathbf{v})$ to the adversary.

Algorithm 1: Universal Attack on Sketching Matrix A ; ℓ_2 norm gap responder

Input: $A \in \mathbb{R}^{k \times n}$, accuracy parameter α , number of queries r , signal index $h \in [n]$, support $M \subset [n]$ of size $m = |M|$, noise scale factor c

for $t \in [r]$ **do** // Main loop

$$\begin{cases} z_i^{(t)} \sim \begin{cases} 0 & \text{if } i \notin M \\ \mathcal{N}(0, 1/m) & \text{if } i \in M \end{cases} & \text{// sample noise vector} \\ w^{(t)} \sim W & \text{// Sample signal weight from } W \text{ (Definition 3.3)} \\ v^{(t)} = w^{(t)} \cdot e_h + c z^{(t)} & \text{// Query vector} \\ \text{Responder chooses an estimator } \psi^{(t)} : \mathbb{R}^k \rightarrow \mathcal{P}\{-1, 1\} & \text{// Map sketches to responses} \\ s^{(t)} \sim \psi^{(t)}(A v^{(t)}) & \text{// Responder receives sketch, returns } s^{(t)} \in \{-1, 1\} \end{cases}$$

return $z^{(\text{adv})} \leftarrow \frac{\sum_{t \in [r]} s^{(t)} z^{(t)}}{\|\sum_{t \in [r]} s^{(t)} z^{(t)}\|_2}$ // Adversarial noise

3. The adversary then adds $s z$ to the accumulated output.

The product of the attack $z^{(\text{adv})}$ is the normalized accumulated output.

Theorem 3.2 (Attack efficacy). *Let $A \in \mathbb{R}^{k \times n}$ be a sketching matrix. Consider applying Algorithm 1 with a randomly selected $h \in [n]$ and support $M \subset [n] \setminus \{h\}$ of size $m = \Omega(k^2)$ and $r = O(\gamma^2 \alpha^2 k^2 \log^2 k)$ queries. Then with constant probability one of the following holds. Either the error rate of responses $s^{(t)}$ for the $(1, \alpha)$ -gap problem on the input norm $\|v^{(t)}\|_2$ (see Definition 3.1) exceeded some constant $\delta(\alpha) > 0$ or the vector $z^{(\text{adv})}$ is a γ -adversarial noise vector (see Definition 4.3) for A , signal h and noise support M .*

We define a vector as γ -adversarial if, under the query distribution specified by h and M , the optimal estimator returns a value that deviates from the true norm by a multiplicative factor of at least γ .

3.1 Density and parameter setting

Definition 3.3 (Signal density and parameters). The distribution W we use for w is parametrized by $a < 1 < 1 + \alpha < b$ and has density function:

$$C = \frac{2}{b - a + \alpha}, \quad \nu(w) = \begin{cases} 0, & w < a \text{ or } w > b, \\ C \frac{w - a}{1 - a}, & a \leq w \leq 1, \\ C, & 1 < w < 1 + \alpha, \\ C \frac{b - w}{b - 1 - \alpha}, & 1 + \alpha \leq w \leq b. \end{cases}$$

We set the parameters as follows according to the gap $\alpha > 0$ and the error rate $\delta > 0$ allowed for the responder. We use $\delta > 0$ that satisfies $\delta / \log(1/\delta) = O(\alpha^2)$ and c that is a small constant (that does not depend on α and is selected according to other constants).

$$\begin{aligned} a &= 1 - 10\alpha/c \\ b &= 1 + \alpha + 10\alpha/c \end{aligned}$$

Observe from our settings that each of the intervals $[a, 1]$ and $[1 + \alpha, b]$ has at least a constant fraction of the probability mass.

4 Analysis of the attack

This section presents the key components and outlines the proof of Theorem 3.2, with full details deferred to the appendix.

We introduce notation for the query and noise distributions. For indices $M \subset [n]$, let $\mathcal{N}_{n,M}$ be the distribution over $\mathbf{u} \in \mathbb{R}^n$ in which $u_{[n] \setminus M} = \mathbf{0}$ and the coordinates indexed by M are sampled from the $m = |M|$ -dimensional Gaussian distribution $u_M \sim \mathcal{N}(0, \frac{1}{m} I_m)$. Note that $\mathcal{N}_{n,M}$ is the distribution of noise vectors selected in Algorithm 1.

For $h \in [n]$, $M \subset [n] \setminus \{h\}$, and noise scale c let

$$F_{h,M,c}[w] := w\mathbf{e}_h + c\mathcal{N}_{n,M} \quad (3)$$

be the distribution of vectors formed by adding a scaled noise vector sampled from $\mathcal{N}_{n,M}$ to a *signal* $w\mathbf{e}_h$, where $\mathbf{e}_h \in \mathbb{R}^n$ is the standard basis vector at index h . $F_{h,M,c}[w]$ is the distribution of query vectors selected in Algorithm 1 for signal value w .

Our analysis is in terms of *signal estimation*. In order to facilitate it, we establish that a correct norm gap output yields a correct signal gap output with similar parameters:

Lemma 4.1 (Norm gap to signal gap). *With the choice of parameters for our attack and $m = \Omega((k+r)\log((k+r)/\delta))$, with probability close to 1, a correct $(1-c^2, 1.1\alpha)$ -norm gap output implies a correct $(1, \alpha)$ -signal gap output on all queries.*

The proof is included in Appendix A. The norm gap in the statement of Theorem 3.2 is with parameters $(1, \alpha)$. To reduce clutter, we treat it in the sequel as signal gap of $(1, \alpha)$.

4.1 Signal estimation and the optimal estimator

For a fixed sketching matrix A , $h \in [n]$, and noise support M , we express the optimal estimator on the signal w from a sketch Av when $\mathbf{v} \sim F_{h,M,c}[w]$. For this purpose we may assume that the distributions (and A , h , and M) are given to the responder. Note that if it holds that $A_{\bullet,h} = \mathbf{0}$, then the sketch carries no information on the signal w . When this is not the case, we can express the optimal unbiased estimator. The proof is included in Appendix B.

Lemma 4.2 (Estimator for the Signal). *Fix $h \in [n]$, a noise support set $M \subset [n] \setminus \{h\}$, and a noise scale factor c . Consider the distributions $F_{h,M,c}[w]$ parametrized by w .*

If the column $A_{\bullet,h}$ is nonzero, then there exists an unbiased, complete, and sufficient statistic $T_{h,M} : \mathbb{R}^k \rightarrow \mathbb{R}$ for the signal w based on the sketch Av .

Furthermore, the deviation of this estimator from its mean, defined as

$$\Delta_{h,M}(cA\mathbf{u}) := T_{h,M}(Av) - w,$$

depends only on the sketch of the noise $\mathbf{u} \sim \mathcal{N}_{n,M}$, and is distributed as a Gaussian random variable $\mathcal{N}(0, c^2\sigma_T^2(h, M))$.

The estimator $T_{h,M}(Av)$ also minimizes the mean squared error (MSE) [30].

We define an adversarial noise vector to be one that causes a large deviation in this optimal estimator:

Definition 4.3 (γ adversarial noise). A unit vector \mathbf{u} with support M is γ -adversarial for A , h , M if $|\Delta_{h,M}(\mathbf{u})| > \gamma$

Adversity of γ means that the value is $\gamma/(c\sigma_T)$ standard deviations off. We will see that the attack size needed for certain adversity depends on σ_T .

4.2 Lower bounding the error

Since $T_{h,M}(Av)$ minimizes the MSE for estimating the signal w from the sketch, it implies a lower bound on the error that applies for any query responder on the query distribution of Algorithm 1:

Corollary 4.4. *The mean squared error (MSE) on any estimator, even one that is tailored to h , M , and c , on queries of the form $\mathbf{v} \sim F_{h,M,c}[w]$ when $w \sim W$ is $\Omega(c^2\sigma_T(h, M)^2)$.*

We next establish that for any sketching matrix $A \in \mathbb{R}^{k \times n}$, a random choice of signal index $h \in [n]$ and noise support (of size M that is slightly superlinear in k), it is likely that either column h is all zeros (and hence any estimator must fail with constant probability) or $\sigma_T^2(h, M) = \tilde{\Omega}(1/k)$. The proof is in Appendix C.

Lemma 4.5 (Lower Bound on Error). *Let $A \in \mathbb{R}^{k \times (m+1)}$ be a matrix with $m \geq 20k \log^2 k$. Then, for at least 0.9 fraction of columns $h \in [m+1]$, it holds that either $A_{\bullet h}$ is all zeros or $\sigma_T^2(h, M = [m+1] \setminus \{h\}) = \Omega\left(\frac{1}{k \log k}\right)$.*

Recall that when the input vectors are k -sparse, exact recovery of the norm is possible, and hence there is no estimation error. Therefore, there is potential vulnerability only when the sparsity of the query vectors exceeds k and thus our slightly super linear sparsity is necessary.

As a corollary of Lemma 4.5, we obtain a lower bound on the relative error guarantees of any estimator:

Corollary 4.6. Consider a fixed sketching matrix Let $A \in \mathbb{R}^{k \times n}$ be a sketching matrix with $n = \Omega(k \log^2 k)$. Consider the distribution of queries that samples h and M (of size $m = \Omega(k \log^2 k)$) randomly from $[n]$ and then sample a query from $F_{h, M, c}[w]$ where $w \sim W$ (and the density of W is at least a constant in range of size $> 100/\sqrt{k}$). Then with at least constant probability, the MSE of any sketch-based signal (and hence norm) estimator is $\Omega(1/(k \log k))$.

Remark 4.7 (Error for JL matrices). Sampled matrices from the JL distributions meet this upper bound: For fixed noise support M of size $m > k \log k$, with high probability over the sampling of A , for all $h \in [n]$, $\sigma_T^2(h, M) = O(1/k)$ (the property needed for the sampled A is that the projected rows A_{iM} are close to orthogonal).

4.3 The gain lemma

We quantify the expected progress (towards an adversarial vector) in each step of the attack in terms of $\sigma_T^2(h, M)$. What we show is that when the estimator has a low error rate, then queries for which the noise component \mathbf{u} has a higher deviation $\Delta_{h, M}(A\mathbf{u})$ are more likely to have $s^{(t)} = 1$. Specifically, even though $\mathbb{E}[\Delta_{h, M}(A\mathbf{u})] = 0$ (as the deviation has distribution $\mathcal{N}(0, \sigma_T^2)$), the response $s^{(t)}$ is correlated with it and we will show that $\mathbb{E}[\Delta_{h, M}(s^{(t)} A\mathbf{u})] \propto \sigma_T^2(h, M)$.

Definition 4.8 (Error rate of sketch-based signal gap estimator). An estimator $\psi : \mathbb{R}^k \rightarrow \mathcal{P}\{-1, 1\}$ is a map from a sketch to a probability distribution on $\{-1, 1\}$. For a sketching map $A \in \mathbb{R}^{k \times n}$, h, M, c and signal distribution W , the *error rate* of ψ for the α -signal gap problem (Definition 3.1) is the probability over the query distribution of an incorrect $(1, \alpha)$ -signal gap response:

$$\text{err}(\psi) := \mathbb{E}_{w \sim W, \mathbf{v} \sim F_{h, M, c}[w]} [\Pr[\psi(A\mathbf{v}) = 1] \cdot \mathbf{1}\{w \leq 1\} + \Pr[\psi(A\mathbf{v}) = -1] \cdot \mathbf{1}\{w \geq 1 + \alpha\}] .$$

If $A_{\bullet h} = \mathbf{0}$ then any estimator would have at least a constant error rate. Additionally, from standard tail bounds, for error rate that is $O(\delta)$, it must hold that $\alpha > c\Omega(\sqrt{\log(1/\delta)})\sigma_T$ and therefore $\sigma_T < \alpha/(c\sqrt{\log(1/\delta)})$.

We therefore consider the case where $A_{\bullet h} \neq \mathbf{0}$, and therefore unbiased $T_{h, M}$ exists and $\sigma_T^2(h, M)$ is well defined, and assume that $\sigma_T < \alpha/(c\sqrt{\log(1/\delta)})$. We quantify the per-step expected gain (see Appendix D for the proof):

Lemma 4.9 (Gain Lemma). *If $\text{err}[\psi] < \delta$ and attack parameters are as in Definition 3.3, and $\sigma_T < \alpha/(c\sqrt{\log(1/\delta)})$ then*

$$\mathbb{E}_{\mathbf{v}} [\psi(A\mathbf{v})(T_{h, M}(A\mathbf{v}) - w)] = \Omega(\alpha^{-1}c^3\sigma_T^2(h, M)^2)$$

4.4 Proof of Attack efficacy Theorem

Proof of Theorem 3.2. The proof idea is to bound the deviation and the norm of the sum $\sum_{t \in [r]} s^{(t)} \mathbf{z}^{(t)}$ and show that the deviation increases faster than the norm.

We first bound norm. We show that if the noise support is sufficiently wide with $m > r^2$, then for *any* choice of $s^{(t)}$, the norm of $\sum_{t \in [r]} s^{(t)} \mathbf{z}^{(t)}$ can not be much higher than that of the sum $\sum_{t \in [r]} \mathbf{z}^{(t)}$ of independent Gaussians:

$$\left\| \sum_{t \in [r]} s^{(t)} \mathbf{z}^{(t)} \right\|_2 = O(\sqrt{r}) \tag{4}$$

This follows as an immediate corollary of Lemma E.1.

We now consider the deviation of the sum. From Lemma 4.9 we obtain that for each t , $\mathbb{E}_{\mathbf{v}}[s^{(t)}\Delta_{h,M}(A\mathbf{z}^{(t)})] = \Omega(\alpha^{-1}\sigma_T^2(h, M))$.

From concentration of $\Delta_{h,M}(A\mathbf{z}^{(t)})$ (that are independent $\mathcal{N}(0, \sigma_T^2)$) we obtain that with high probability

$$\sum_{i=1}^r s^{(t)}\Delta_{h,M}(A\mathbf{z}^{(t)}) = \Omega(r\alpha^{-1}\sigma_T^2(h, M)). \quad (5)$$

From Lemma 4.5, for any sketching matrix $A \in \mathbb{R}^{k \times n}$, when we sample h , and $M \subset [n]$ of size $m = \Omega(k \log^2 k)$, then with constant probability we have $\sigma_T^2(h, M) = \Omega(\frac{1}{k \log k})$.

Combining with (5) we obtain that

$$\sum_{i=1}^r s^{(t)}\Delta_{h,M}(A\mathbf{z}^{(t)}) = \Omega\left(\frac{r\alpha^{-1}}{k \log k}\right). \quad (6)$$

Combining (4) and (6) we obtain that with constant probability

$$\Delta_{h,M}(A\mathbf{z}^{(\text{adv})}) = \Omega\left(\frac{\sqrt{r}\alpha^{-1}}{k \log k}\right).$$

For the deviation to exceed γ , solving for $\frac{\sqrt{r}\alpha^{-1}}{k \log k} > \gamma$, we obtain $r = O(\gamma^2 \alpha^2 k^2 \log^2 k)$.

This concludes the proof. \square

5 Empirical study

We implemented Algorithm 1 and evaluated its effectiveness on two families of sketching matrices: Gaussian Johnson–Lindenstrauss (JL) transforms [29] with $A \in \mathbb{R}^{k \times n}$ with i.i.d. entries $A_{ij} \sim \mathcal{N}(0, 1/k)$ and AMS sketches [3] with $A \in \{\pm 1\}^{k \times n}$ consisting of i.i.d. Rademacher entries ($A_{ij} = \pm 1$ with equal probability). Our evaluation used different configurations of sketch dimension k and ambient dimension n (effectively, the noise support). For each configuration we applied the attack against the corresponding *standard* norm estimator and its *robustified* variant, and measured how rapidly the adversarial bias grows.

Standard Estimators. For Gaussian JL matrices, the *standard* norm estimator is simply the sketch norm, which coincides with the minimum-variance unbiased estimator for $\|v\|_2^2$ under Gaussian projections:

$$\widehat{\|v\|_2^2} := \|A\mathbf{v}\|_2^2.$$

For AMS sketches, we implemented the *median-of-means* (MoM) estimator. Specifically, we partition the k rows of A into $g = \max\{5, \lfloor \sqrt{k} \rfloor\}$ groups of equal size $b = \lceil k/g \rceil$, compute the mean of y_i^2 within each group, and take the median of the g group means:

$$\widehat{\|v\|_2^2} := \text{median}\left(\frac{1}{b} \sum_{i \in G_1} y_i^2, \frac{1}{b} \sum_{i \in G_2} y_i^2, \dots, \frac{1}{b} \sum_{i \in G_g} y_i^2\right).$$

Our final estimate is $\widehat{\|v\|_2} := \sqrt{\widehat{\|v\|_2^2}}$.

Simplified attack We implemented a simplified attack that sufficed for the special case of JL/AMS matrices (distribution that is invariant under column permutations) and query responders that are not adaptive and not tailored to the input distribution). Our query vectors are sampled i.i.d. from $F_{n, [n-1], 1}[w = 1]$ (see Eq. (3)), that is, use a a fixed signal value $w = 1$ and fixed $h = n$ and $M = [n - 1]$. Observe that for our query vectors, the norm $\|v\|_2$ is concentrated around $\sqrt{w^2 + 1} = \sqrt{2}$.

Robust Estimators. The robust variants are parameterized by a noise scale σ and add Gaussian noise to the squared-norm standard estimate before taking the square root:

$$\hat{s}_\sigma(v) := \sqrt{\widehat{\|v\|_2^2} + \mathcal{N}(0, \sigma^2)}.$$

In our attack implementation the responder outputs $s = 1$ when $\hat{s}_\sigma \geq \sqrt{w^2 + 1}$ and $s = -1$ otherwise.³

On freshly sampled (non-adaptive) inputs, the estimator has standard deviation $1/\sqrt{k}$ in the JL case, whereas for the robustified estimator the standard deviation increases to $\sqrt{\frac{1}{k} + \sigma^2}$. Therefore, the non-adaptive accuracy decreases with the robustness noise σ .

Experiments For each (k, n) configuration of sketch size k and support size n , we performed $\text{rep} = 20$ trials. For each trial we sample a fresh sketching matrix $A \in \mathbb{R}^{k \times n}$. We sample query vectors as described and apply estimators with different values of the robustness noise $\sigma \in \{0, 0.1, 0.2, 0.4\}$ to the same query vectors (observe that $\sigma = 0$ is the standard estimator). We study the effectiveness of our attack by tracking the ratio $\hat{s}_\sigma(z^{(\text{adv})})/\|z^{(\text{adv})}\|_2$, where \hat{s}_σ is the (respective) standard estimator that is applied to the sketch and $z^{(\text{adv})}$ is the adversarial vector.

Results Fig. 1 and Fig. 2 report, for different configurations with $k \in \{100, 250, 1000\}$, the ratio $\hat{s}_\sigma(z^{(\text{adv})})/\|z^{(\text{adv})}\|_2$ (the y -axis) as a function of the number of attack queries (the x -axis).

We observe that the attack is effective even in configurations with small support size of $n/k \in [2, 5]$. The standard and lower- σ robust estimators are initially more accurate but also accumulate bias faster and incur a much higher error as the attack progresses. A larger sketch size k results in higher robustness (slower increase in the bias) for the same noise σ . The plots show a quadratic pattern where the bias induced by the attack increases like square root of the number of queries. Results for additional configurations for the same k are reported in Fig. 3. We observe that the effectiveness of the attack tends to increase with a larger support size n .

Discussion. Our empirical results demonstrate that the attack is substantially more effective in practice than our current analysis predicts. In particular, we observe strong empirical success even when the noise support size is very small—on the order of $O(k)$ —whereas our analysis requires $n = \Omega(k^2)$. The attack is also effective with relatively few queries, again beyond what is guaranteed by our theoretical bounds.

Moreover, our analysis only establishes that the attack product compromises the *optimal* estimator for the sketching matrix, which is a linear statistic of the sketch. For JL sketches, the standard estimator is very close (up to row orthogonality) to this optimal estimator, so our theory is largely predictive. In contrast, for AMS sketches we employ a median-of-means estimator, which is non-linear, yet the attack empirically compromises it with comparable efficiency. This suggests that the vulnerability extends beyond linear estimators and may be more fundamental than our current proofs capture.

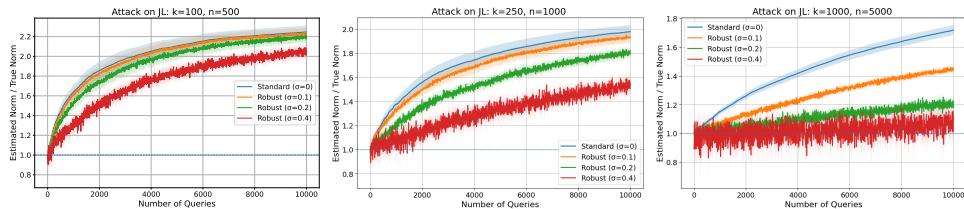


Figure 1: Attack on JL with $(k, n) \in \{(100, 500), (250, 1000), (1000, 5000)\}$. Standard and robust estimators. Mean ratio of estimate to actual norm with 95% confidence intervals over 20 repetitions.

³Our simplified attack would fail against a fully strategic responder that is tuned to the input distribution and always returns the deterministic value $\sqrt{w^2 + 1} = \sqrt{2}$. Such a responder leaks no information about the sketching matrix and is correct with high probability on all queries.

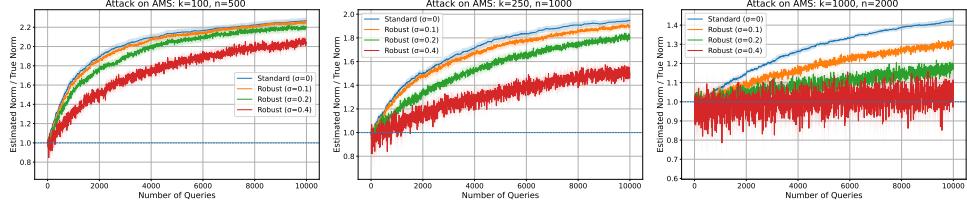


Figure 2: Attack on AMS with $(k, n) \in \{(100, 500), (250, 1000), (1000, 2000)\}$. Standard and robust estimators. Mean ratio of estimate to actual norm with 95% confidence intervals over 20 repetitions.

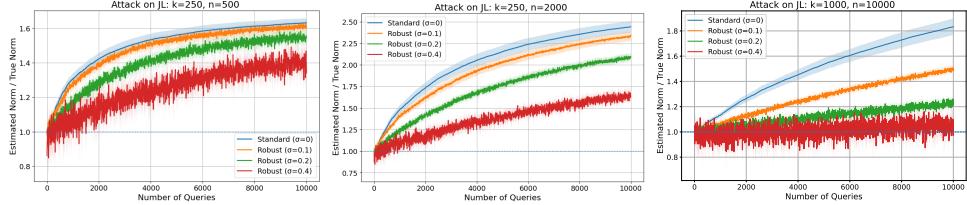


Figure 3: Attack on JL with $(k, n) \in \{(250, 500), (250, 2000), (1000, 10000)\}$. Mean ratio of estimate to actual norm with 95% confidence intervals over 20 repetitions.

6 Conclusion

Our results further suggest that vulnerability to black-box attacks is an inherent consequence of dimensionality reduction. Nonetheless, it can be partly mitigated by sacrificing some non-adaptive accuracy in exchange for increased robustness, underscoring a fundamental trade-off. We close with several directions for further investigation.

Theory questions. We proved a tight, quadratic-size universal attack that succeeds against any query responder, but our analysis only guarantees compromise of the *optimal* (linear) estimator for the sketch and attack distribution. Empirically we observed that our attack compromised also the non-linear median-of-means estimator for AMS. This suggests exploring broader applicability of our attack. Another natural question is whether our technique can be extended to other norms, as in Gribelyuk et al. [23]. A more ambitious goal is to construct a quadratic-size attack that generates a *single distribution* capable of compromising *every* query responder. Such an attack would necessarily require $\tilde{\Omega}(k)$ adaptive batches and could plausibly build on an enhanced version of our construction that incorporates the discovered adversarial directions into the input distribution [16]. Another open question concerns the required size of the noise support, which directly corresponds to the attacker’s storage cost. Our current proof requires $n = \Omega(k^2)$ because it relies on a general bound (Lemma E.1) for arbitrary signed sums of Gaussian vectors. Empirically, however, we observe that our attack remains effective with support as small as $2k$, and we conjecture that $\tilde{O}(k)$ support should suffice.

Connections to Image Classifiers. Our findings on the vulnerability of linear sketches may shed light on the phenomenon of adversarial examples in image classification, which has been extensively studied in prior work [37, 21, 4, 34, 32, 35]. Many of these attacks exhibit a striking structural similarity to our setting: they construct adversarial examples by aggregating small, randomly oriented perturbations that consistently push the system’s output in a particular direction. When normalized and accumulated, these perturbations yield a large, adversarial deviation that fools the model. This additive, alignment-based mechanism appears to be remarkably effective across domains.

Although image classifiers such as CNNs are highly nonlinear, they contain significant linear components, particularly in early layers. This observation raises a natural question: could adversarial susceptibility in image models arise, at least in part, from the same linearity-driven accumulation of aligned noise? If so, it may be possible to improve robustness by introducing carefully designed randomness—such as injecting noise that is unknown to the attacker into internal representations or into the input.

Acknowledgments

Edith Cohen was partially supported by Israel Science Foundation (grant 1156/23). Uri Stemmer was Partially supported by the Israel Science Foundation (grant 1419/24) and the Blavatnik Family foundation.

References

- [1] Sara Ahmadian and Edith Cohen. Unmasking vulnerabilities: Cardinality sketches under adaptive inputs. In *ICML*, 2024. URL <https://doi.org/10.48550/arXiv.2405.17780>.
- [2] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Analyzing graph structure via linear measurements. In *Proceedings of the 2012 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 459–467, 2012. doi: 10.1137/1.9781611973099.40. URL <https://pubs.siam.org/doi/abs/10.1137/1.9781611973099.40>.
- [3] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58:137–147, 1999.
- [4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.
- [5] Idan Attias, Edith Cohen, Moshe Shechner, and Uri Stemmer. A framework for adversarial streaming via differential privacy and difference estimators. *CoRR*, abs/2107.14527, 2021.
- [6] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan R. Ullman. Algorithmic stability for adaptive data analysis. *SIAM J. Comput.*, 50(3), 2021. doi: 10.1137/16M1103646. URL <https://doi.org/10.1137/16M1103646>.
- [7] Amos Beimel, Haim Kaplan, Yishay Mansour, Kobbi Nissim, Thatchaphol Saranurak, and Uri Stemmer. Dynamic algorithms against an adaptive adversary: Generic constructions and lower bounds. *CoRR*, abs/2111.03980, 2021.
- [8] Omri Ben-Eliezer, Talya Eden, and Krzysztof Onak. Adversarially robust streaming via dense-sparse trade-offs. *CoRR*, abs/2109.03785, 2021.
- [9] Omri Ben-Eliezer, Rajesh Jayaram, David P. Woodruff, and Eylon Yogev. A framework for adversarially robust streaming algorithms. *SIGMOD Rec.*, 50(1):6–13, 2021.
- [10] Guy Blanc. Subsampling suffices for adaptive data analysis. In Barna Saha and Rocco A. Servedio, editors, *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pages 999–1012. ACM, 2023. doi: 10.1145/3564246.3585226. URL <https://doi.org/10.1145/3564246.3585226>.
- [11] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
- [12] Yeshwanth Cherapanamjeri and Jelani Nelson. On adaptive distance estimation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [13] Edith Cohen, Xin Lyu, Jelani Nelson, Tamás Sarlós, Moshe Shechner, and Uri Stemmer. On the robustness of countsketch to adaptive inputs. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 4112–4140. PMLR, 2022.
- [14] Edith Cohen, Jelani Nelson, Tamas Sarlos, and Uri Stemmer. Tricking the hashing trick: A tight lower bound on the robustness of countsketch to adaptive inputs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7235–7243, Jun. 2023. doi: 10.1609/aaai.v37i6.25882. URL <https://ojs.aaai.org/index.php/AAAI/article/view/25882>.

[15] Edith Cohen, Jelani Nelson, Tamas Sarlos, and Uri Stemmer. Tricking the hashing trick: A tight lower bound on the robustness of countsketch to adaptive inputs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7235–7243, Jun. 2023. doi: 10.1609/aaai.v37i6.25882. URL <https://ojs.aaai.org/index.php/AAAI/article/view/25882>.

[16] Edith Cohen, Jelani Nelson, Tamás Sarlós, Mihir Singhal, and Uri Stemmer. One Attack to Rule Them All: Tight Quadratic Bounds for Adaptive Queries on Cardinality Sketches. *arXiv preprint arXiv:2411.06370*, 2024.

[17] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *STOC*, pages 117–126. ACM, 2015.

[18] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC ’15, page 117–126, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746580. URL <https://doi.org/10.1145/2746539.2746580>.

[19] David A. Freedman. A note on screening regression equations. *The American Statistician*, 37(2):152–155, 1983. doi: 10.1080/00031305.1983.10482729. URL <https://www.tandfonline.com/doi/abs/10.1080/00031305.1983.10482729>.

[20] Paweł Gawrychowski, Shay Mozes, and Oren Weimann. Minimum cut in $o(m \log^2 n)$ time. In *ICALP*, volume 168 of *LIPICS*, pages 57:1–57:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[22] Elena Gribelyuk, Honghao Lin, David P. Woodruff, Huacheng Yu, and Samson Zhou. A strong separation for adversarially robust ℓ_0 estimation for linear sketches. In *Proceedings of the 65th Annual Symposium on Foundations of Computer Science (FOCS)*, 2024. URL <https://arxiv.org/pdf/2409.16153.pdf>.

[23] Elena Gribelyuk, Honghao Lin, David P. Woodruff, Huacheng Yu, and Samson Zhou. Lifting linear sketches: Optimal bounds and adversarial robustness. In *Proceedings of the 57th ACM Symposium on Theory of Computing (STOC ’25)*, pages 395–406, 2025. doi: 10.1145/3717823.3718227.

[24] Maximilian Probst Gutenberg and Christian Wulff-Nilsen. Decremental SSSP in weighted digraphs: Faster and against an adaptive adversary. In *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’20, page 2542–2561, USA, 2020. Society for Industrial and Applied Mathematics.

[25] M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 454–463. IEEE Computer Society, 2014. doi: 10.1109/FOCS.2014.55. URL <https://doi.ieeecomputersociety.org/10.1109/FOCS.2014.55>.

[26] Moritz Hardt and David P. Woodruff. How robust are linear sketches to adaptive inputs? In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC ’13, page 121–130, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi: 10.1145/2488608.2488624. URL <https://doi.org/10.1145/2488608.2488624>.

[27] Avinatan Hassidim, Haim Kaplan, Yishay Mansour, Yossi Matias, and Uri Stemmer. Adversarially robust streaming algorithms via differential privacy. In *Annual Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[28] John P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, (2):8, 2005.

[29] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

- [30] E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998.
- [31] Paul M. Lukacs, Kenneth P. Burnham, and David R. Anderson. Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*, 62(1):117, 2009.
- [32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL <https://arxiv.org/abs/1706.06083>. arXiv:1706.06083.
- [33] Ilya Mironov, Moni Naor, and Gil Segev. Sketching in adversarial environments. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, page 651–660, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580470. doi: 10.1145/1374376.1374471. URL <https://doi.org/10.1145/1374376.1374471>.
- [34] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.
- [35] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3a24b25a7b092a252166a1641ae953e7-Paper.pdf.
- [36] Yossi Shiloach and Shimon Even. An on-line edge-deletion problem. *J. ACM*, 28(1):1–4, jan 1981. ISSN 0004-5411. doi: 10.1145/322234.322235. URL <https://doi.org/10.1145/322234.322235>.
- [37] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [38] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK, 2018. ISBN 9781108415194. doi: 10.1017/9781108231596.
- [39] David Wajc. *Rounding Dynamic Matchings against an Adaptive Adversary*. Association for Computing Machinery, New York, NY, USA, 2020. URL <https://doi.org/10.1145/3357713.3384258>.
- [40] David P. Woodruff and Samson Zhou. Tight bounds for adversarially robust streams and sliding windows via difference estimators. In *Proceedings of the 62nd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2021.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer **[Yes]** , **[No]** , or **[NA]** .
- **[NA]** means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "**[Yes]**" is generally preferable to "**[No]**", it is perfectly acceptable to answer "**[No]**" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "**[No]**" or "**[NA]**" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer **[Yes]** to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”.**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: **[Yes]**

Justification: The claims reflect the contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: **[Yes]**

Justification: We explain that what we are doing is what we can hope to do in a single-batch attack.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Most of the proofs are in the supplementary but ideas and sketches are contained in the main body.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We included very light weight experiments as a demonstration. This is a theory paper. We disclose everything needed to reproduce the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: There is no data and the code is very minimal and easy to redo.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: We do not do any model training. We only implemented a version of our Algorithm 1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: We report results of few repetitions in our demonstration.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: Experiments are simple and light weight

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: We believe we did.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: This is a theory paper exposing some vulnerabilities in dimensionality reduction. There is no direct societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[NA\]](#)

Justification: Not applicable

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We only used standard python libraries.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new datasets or code. This is a theory paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Not applicable

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not applicable

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We did not use LLM beyond editing, formatting, and such.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Norm estimation to signal estimation

This section contains a restatement and proof of Lemma 4.1:

Lemma 4.1 (Norm gap to signal gap). *With the choice of parameters for our attack and $m = \Omega((k+r)\log((k+r)/\delta))$, with probability close to 1, a correct $(1-c^2, 1.1\alpha)$ -norm gap output implies a correct $(1, \alpha)$ -signal gap output on all queries.*

Proof. The query vectors $\mathbf{v} = w\mathbf{e}_h + c\mathbf{u}$ sampled in Algorithm 1 are from distributions $F_{h,M,c}[w]$, for $w \in [a, b]$. The squared norm is $\|\mathbf{v}\|_2^2 = w^2 + c^2\|\mathbf{u}\|_2^2$. We show that $\|\mathbf{u}\|_2^2$ is tightly concentrated around its expectation 1:

Claim A.1 (Concentration of $\|\mathbf{u}\|_2^2$).

$$\Pr[|\|\mathbf{u}\|_2^2 - 1| \geq \epsilon] \leq 2e^{-\frac{m}{2}(\epsilon^2/2)}$$

Proof. The sum of the squares of m i.i.d. $\mathcal{N}(0, \sigma^2)$ has distribution $\sigma^2\chi_m^2$ and expected value $m\sigma^2$. Applying tail bounds on χ_m^2 (Gaussian concentration of measure) we obtain that for $\epsilon \in (0, 1)$

$$\Pr\left[\frac{|\|\mathbf{u}\|_2^2 - m\sigma^2|}{m\sigma^2} \geq \epsilon\right] \leq 2e^{-\frac{m}{2}(\epsilon^2/2)}.$$

Substituting $\sigma^2 = 1/m$ we obtain the claim. \square

It follows that if we choose $m = \Omega((k+r)\log((k+r)/\delta))$ then with probability $1 - \delta$, on all our queries, the squared norm of the noise is within $\|\mathbf{u}\|_2^2 \in (1 \pm 1/(10\sqrt{k}))$. Therefore, a correct $(1, \alpha)$ -gap output on the norm yields a correct $(1, \alpha^2 + 2\alpha)$ -gap on the squared norm $\|\mathbf{v}\|_2^2$. This gives an $(1 - c^2(1 + 1/(10\sqrt{k})), \alpha^2 + 2\alpha - c^2/(10\sqrt{k}))$ -gap output on the squared signal w^2 .

We now note that we can assume $\alpha > 1/\sqrt{k}$ because otherwise, any responder would be incorrect with constant probability. Using our parameter setting of ℓ close to 1 and fixed $c \ll 1$ we obtain the claim in the statement of the lemma. \square

B Estimator for the Signal

This section contains a restatement and proof of Lemma 4.2.

Lemma 4.2 (Estimator for the Signal). *Fix $h \in [n]$, a noise support set $M \subset [n] \setminus \{h\}$, and a noise scale factor c . Consider the distributions $F_{h,M,c}[w]$ parametrized by w .*

If the column $A_{\bullet,h}$ is nonzero, then there exists an unbiased, complete, and sufficient statistic $T_{h,M} : \mathbb{R}^k \rightarrow \mathbb{R}$ for the signal w based on the sketch $A\mathbf{v}$.

Furthermore, the deviation of this estimator from its mean, defined as

$$\Delta_{h,M}(cA\mathbf{u}) := T_{h,M}(A\mathbf{v}) - w,$$

depends only on the sketch of the noise $\mathbf{u} \sim \mathcal{N}_{n,M}$, and is distributed as a Gaussian random variable $\mathcal{N}(0, c^2\sigma_T^2(h, M))$.

Proof. If the column $A_{\bullet,h}$ is zero, then the sketch contains no information on the signal w and unbiased estimation is not possible. Otherwise, our goal is to express the unbiased sufficient statistic $T_{h,M}(A\mathbf{v})$ for the unknown scalar w .

Because row operations preserve the information in a sketch, there exists an invertible matrix $G \in \mathbb{R}^{k \times k}$ such that for $A' = GA$,

- the transformed column h has value 1 in the first row and 0 elsewhere: $A'_{1h} = 1$ and $A'_{ih} = 0$ for $i > 1$; and
- every row A'_{iM} for $i > 1$ of the submatrix restricted to the noise coordinates, is either orthogonal to the first row A'_{1M} or is the zero vector $\mathbf{0}$.

Since G is invertible, A' is equivalent to the original sketching matrix A in that we can obtain the sketch $A'\mathbf{v}$ from $A\mathbf{v}$ and vice versa.

We specify a sufficient statistics T in terms of $A'\mathbf{v}$ (so that $T_{h,M}(A\mathbf{v}) \equiv T(GA\mathbf{v})$). Consider the distribution of the sketch $\mathbf{y} = A'\mathbf{v} = GA\mathbf{v}$ for $\mathbf{v} = w\mathbf{e}_h + c\mathbf{u}$ where $\mathbf{u} \sim \mathcal{N}_{n,M}$. We have $y_1 = w + c \sum_{j \in M} A'_{ij} u_j$, hence it has distribution

$$c\mathcal{N}(w, \sigma_T^2), \text{ where } \sigma_T^2 := \|A'_{1M}\|_2^2 \sigma^2 = \frac{1}{m} \|A'_{1M}\|_2^2. \quad (7)$$

From orthogonality of the rows A'_{iM} restricted to M , the random variables y_i for $i > 1$ are independent of $A'_{1\bullet}\mathbf{u}$ and hence convey no information on w . The information on the signal w in the sketch, and the unbiased sufficient statistic is therefore $y_1 = (GA\mathbf{v})_1$, which has distribution (7) and is an unbiased estimator of w .

To establish the claim for the additive error, note that

$$\Delta_{h,M}(cA\mathbf{u}) = T_{h,M}(A\mathbf{v}) - w = (GA\mathbf{v})_1 - w = (cGA\mathbf{u})_1.$$

□

C Proof of Lower Bound on Error Lemma

This section contains a restatement and proof of Lemma 4.5.

Lemma 4.5 (Lower Bound on Error). *Let $A \in \mathbb{R}^{k \times (m+1)}$ be a matrix with $m \geq 20k \log^2 k$. Then, for at least 0.9 fraction of columns $h \in [m+1]$, it holds that either $A_{\bullet,h}$ is all zeros or $\sigma_T^2(h, M = [m+1] \setminus \{h\}) = \Omega\left(\frac{1}{k \log k}\right)$.*

We will use the following technical claims

Definition C.1 (Fragile Columns). Let $A \in \mathbb{R}^{k \times m}$. For each column $h \in [m]$, define

$$K_h = \{i \in [k] : A_{ih} \neq 0\}, \quad b_{ih} = |\{j \in [m] : A_{ij}^2 \geq A_{ih}^2\}|$$

the set K_h of active rows and the dominated number b_{ih} of each active row. Let $i \in K_h$

$$c_1^{(h)} \leq c_2^{(h)} \leq \dots \leq c_{|K_h|}^{(h)}$$

be the nondecreasing rearrangement of $\{b_{ih} : i \in K_h\}$. We say that a column h is *fragile*, if:

$$\forall 1 \leq i \leq |K_h| : c_i^{(h)} \geq \frac{i m}{10k \log_2 k}.$$

Note that zero columns are fragile by definition.

Claim C.2 (Most Columns are Fragile). If $m > 2k \log_2 k$ then at least $0.9m$ of the columns are fragile.

Proof. We look at $\lfloor \log_2 k \rfloor$ ranges of b_{ih} values where $R_1 = [1, \dots, \frac{m}{10k \log_2 k})$ and $R_t = [2^{t-2}, 2^t) \frac{m}{10k \log_2 k}$ for $t \geq 2$.

We say a column h is *strong* for range t if it has 2^{t-1} or more b_{ih} values in R_t . That is, $|\{i : b_{ih} \in R_t\}| \geq 2^{t-1}$. Clearly if a column h is not fragile then it must be strong for some range t , but the converse may not hold.

We bound the number of distinct columns that are strong for at least one range t . This bounds the total number of non-fragile columns.

Consider range $t = 1$. Each of the k rows contributes the columns of its $\frac{m}{10k \log_2 k}$ largest entries. So the total number of columns that can be strong for range $t = 1$ is at most $\frac{m}{10 \log_2 k}$.

Now consider $t > 1$. Each row has at most $2^{t-1} \frac{m}{10k \log_2 k}$ distinct columns h with $b_{ih} \in R_t$. But for a column to be strong for t it has to participate in 2^{t-1} such rows. So in total, the range contributes at most $\frac{m}{10 \log_2 k}$ strong columns.

Summing over all ranges, we obtain a bound of $m/10$ on the total number of columns that are strong for at least one range. This also bounds the number of non-fragile columns. This concludes the proof. \square

Claim C.3 (Nearly to fully orthogonal rows). Let $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(k)} \in \mathbb{R}^n$ for $k \geq 1$ be linearly independent and satisfy

$$\langle \mathbf{v}^{(i)}, \mathbf{v}^{(j)} \rangle = 1 \quad (i \neq j), \quad \|\mathbf{v}^{(i)}\|^2 > i(2 + \ln k) \quad (i = 1, \dots, k).$$

Then there are orthogonal $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(k)}$ such that $\mathbf{u}^{(1)} = \mathbf{v}^{(1)}$, and

$$\begin{aligned} \|\mathbf{u}^{(i)}\|_2^2 &> \|\mathbf{v}^{(i)}\|_2^2 - 1 \geq i(1 + \ln k) \quad (i = 2, \dots, k) \\ \mathbf{u}^{(i)} &\text{ is an affine combination of } \mathbf{v}^{(1)}, \dots, \mathbf{v}^{(i-1)} \quad (i = 2, \dots, k) \end{aligned}$$

Proof. We construct vectors $\mathbf{u}^{(2)}, \dots, \mathbf{u}^{(k)}$ in order using the following operations:

$$\begin{aligned} \tilde{\mathbf{u}}^{(i)} &= \mathbf{v}^{(i)} - \sum_{j=1}^{i-1} \frac{\langle \mathbf{v}^{(i)}, \mathbf{u}^{(j)} \rangle}{\|\mathbf{u}^{(j)}\|^2} \mathbf{u}^{(j)} \\ \mathbf{u}^{(i)} &= \frac{\tilde{\mathbf{u}}^{(j)}}{1 - \sum_{j=1}^{i-1} \frac{\langle \mathbf{v}^{(i)}, \mathbf{u}^{(j)} \rangle}{\|\mathbf{u}^{(j)}\|^2}}. \end{aligned}$$

We establish the claim by induction on i . The claim clearly holds for $i = 1$.

Assume it holds for i . Then $\mathbf{u}^{(i)} = \sum_{j=1}^i \gamma_{ij} \mathbf{v}^{(j)}$, where $\sum_{j=1}^i \gamma_{ij} = 1$.

Therefore for all $h > i$,

$$\langle \mathbf{v}^{(h)}, \mathbf{u}^{(i)} \rangle = \sum_{j=1}^i \gamma_{ij} \langle \mathbf{v}^{(h)}, \mathbf{v}^{(j)} \rangle = \sum_{j=1}^i \gamma_{ij} = 1.$$

Therefore

$$\tilde{\mathbf{u}}^{(i+1)} = \mathbf{v}^{(i+1)} - \sum_{j=1}^i \frac{1}{\|\mathbf{u}^{(j)}\|^2} \mathbf{u}^{(j)}$$

Using orthogonality of $(\mathbf{u}^{(j)})_{j \leq i}$ we obtain:

$$\|\tilde{\mathbf{u}}^{(i)}\|_2^2 = \|\mathbf{v}^{(i)}\|_2^2 - \sum_{j=1}^{i-1} \frac{1}{\|\mathbf{u}^{(j)}\|_2^2} \geq \|\mathbf{v}^{(i)}\|_2^2 - \sum_{j=1}^{i-1} \frac{1}{i(1 + \ln k)} \geq \|\mathbf{v}^{(i)}\|_2^2 - 1.$$

(using an upper bound on the Harmonic sum).

The scaling of $\tilde{\mathbf{u}}^{(i)}$ ensures that $\mathbf{u}^{(i)}$ is an affine combination of $\mathbf{v}^{(i)}$ and $(\mathbf{u}^{(j)})_{j < i}$. Since by induction each $\mathbf{u}^{(j)}$ for $j < i$ is an affine combination of $(\mathbf{v}^{(h)})_{h \leq j}$, combining we obtain that so is $\mathbf{u}^{(i)}$.

The scale factor is $s_i = 1 - \sum_{j=1}^i \frac{1}{j(1 + \ln k)} \in (0, 1]$. Therefore,

$$\|\mathbf{u}^{(i)}\|_2^2 = \frac{1}{s_i^2} \|\tilde{\mathbf{u}}^{(i)}\|_2^2 \geq \|\tilde{\mathbf{u}}^{(i)}\|_2^2 \geq \|\mathbf{v}^{(i)}\|_2^2 - 1.$$

\square

Proof of Lemma 4.5. We first give a characterization of $\sigma_T^2(h, [m+1] \setminus \{h\})$ that we will use, and follows from a similar argument to the proof of Lemma 4.2. Let $G \in \mathbb{R}^{k \times k}$ be invertible so that the column $GA_{\bullet, h}$ has only values in $\{0, 1\}$ and the rows of the submatrix $GA_{\bullet, [m+1] \setminus \{h\}}$ with column h removed are orthogonal. Then

$$\frac{1}{\sigma_T^2(h, [m+1] \setminus \{h\})} = \sum_{i: GA_{ih}=1} \frac{m}{\|GA_{i, [m+1] \setminus \{h\}}\|_2^2}. \quad (8)$$

We assume, without loss of generality, that the rows of A are either orthogonal or $\mathbf{0}$. From Claim C.2, it follows that most columns of A are fragile.

We now fix a nonzero fragile column h . We scale the rows so that $A_{\bullet,h}$ are in $\{0, 1\}$. Note that the fragility of columns is invariant to rescaling of rows. From fragility, and with the rescaling, the squared norms of the rows in increasing order are at least $\frac{im}{10k \log k}$.

We now consider the submatrix $B = A_{L_h, [m+1] \setminus \{h\}} \in \mathbb{R}^{k \times m}$ of A with column h removed and all rows in which column h was not active are removed. Let the rows of B be $\mathbf{v}^{(i)}$ and observe that from orthogonality of the rows of A and from the fragility of h , the vectors satisfy the conditions in the statement of Claim C.3. Therefore by applying the claim we obtain a matrix B' (with row vectors \mathbf{u}^i that are orthogonal, are affine transformations of the rows of B , and that

$$\sum_i \frac{1}{\|B'_{i\bullet}\|_2^2} \leq \sum_i \frac{10k \log k}{im} = O\left(\frac{k \log k}{m}\right).$$

It follows that if we applied the same transformations with the column h , the column would be invariant. The matrix A' with column A with B' substituting the matrix B could be obtained using the same transformation. This matrix satisfies the conditions of the characterization and by applying (8) we obtain $\sigma_T^2(h, [m+1] \setminus \{h\}) = \Omega\left(\frac{1}{k \log k}\right)$ and this concludes the proof. \square

D Proof of the Gain Lemma

This section contains a restatement and proof of Lemma 4.9.

Because $T_{h,M}$ is a sufficient statistic for w , the distribution of the sketch $A\mathbf{v}$, conditioned on $T_{h,M}(A\mathbf{v}) = (G^{(h,M)} A\mathbf{v})_1 = \tau$, does not depend on w . Let $f_\tau : \mathbb{R}^k$ be the density function of this distribution.

We can thus express the expected value of ψ , conditioned on the value of the statistic $T_{h,M}(\mathbf{y}) = \tau$, (as it does not depend on the signal value w):

$$\Psi(\tau) := \int_{\mathbb{R}^k} \psi(\mathbf{y}) f_\tau(\mathbf{y}) d\mathbf{y}. \quad (9)$$

We express the error rate of ψ (see Definition 4.8) in terms of Ψ and σ_T^2 :

$$\begin{aligned} \text{err}(\psi) &= \int_a^1 \int_{\mathbb{R}} \frac{\Psi(w+x) + 1}{2} \varphi_{0, c^2 \sigma_T^2}(x) dx \nu(w) dw \\ &\quad + \int_{1+\alpha}^b \int_{\mathbb{R}} \frac{1 - \Psi(w+x)}{2} \varphi_{0, c^2 \sigma_T^2}(x) dx \nu(w) dw, \end{aligned} \quad (10)$$

where $\nu(w)$ be the density function of the distribution W on the signal w (see Definition 3.3) and $\varphi_{0, c^2 \sigma_T^2}$ is the distribution of the deviation (see Lemma 4.2).

Lemma 4.9 (Gain Lemma). *If $\text{err}[\psi] < \delta$ and attack parameters are as in Definition 3.3, and $\sigma_T < \alpha / (c\sqrt{\log(1/\delta)})$ then*

$$\mathbb{E}_{\mathbf{v}} [\psi(A\mathbf{v})(T_{h,M}(A\mathbf{v}) - w)] = \Omega(\alpha^{-1} c^3 \sigma_T(h, M)^2)$$

Proof. We bound from below the expected value, over our query distribution, of

$$\psi(A\mathbf{v})(T_{h,M}(A\mathbf{v}) - w) = c\psi(A\mathbf{v})\Delta_{h,M}(A\mathbf{u}).$$

We express the expected value of ψ over our query distribution, conditioned on the deviation $\Delta_{h,M}(cA\mathbf{u}) = x$:

$$\begin{aligned}\mathbf{E}[\psi(A\mathbf{v}) \mid \Delta_{h,M}(cA\mathbf{u}) = x] &= \frac{1}{\varphi_{0,c^2\sigma_T^2}(x)} \cdot \int_a^b \Psi(w+x) \varphi_{0,c^2\sigma_T^2}(x) \nu(w) dw \\ &= \int_a^b \Psi(w+x) \nu(w) dw \\ &= \int_{a+x}^{b+x} \Psi(w) \nu(w-x) dw.\end{aligned}\tag{11}$$

We express the expected value of $\psi(A\mathbf{v})\Delta_{h,M}(cA\mathbf{u})$ over our query distribution.

$$\begin{aligned}\mathbf{E}_{\mathbf{v}}[\psi(A\mathbf{v})\Delta_{h,M}(cA\mathbf{u})] &= \mathbf{E}_{x \sim \mathcal{N}(0,c^2\sigma_T^2)} x \mathbf{E}_{\mathbf{v}}[\psi(A\mathbf{v}) \mid \Delta_{h,M}(cA\mathbf{u}) = x] \\ &= \int_{-\infty}^{\infty} x \varphi_{0,c^2\sigma_T^2}(x) \left(\int_{a+x}^{b+x} \Psi(w) \nu(w-x) dw \right) dx \quad ; \text{ using (11)} \\ &= \int_0^{\infty} x \varphi_{0,c^2\sigma_T^2}(x) \left(\int_{a+x}^{b+x} \Psi(w) \nu(w-x) dw - \int_{a-x}^{b-x} \Psi(w) \nu(w+x) dw \right) dx\end{aligned}\tag{12}$$

(using that φ is a symmetric function).

We separately bound the parts of the integral in (12) due to $x \in [\alpha, \infty]$ and $x \in [0, \alpha]$.

We first consider $x \in [\alpha, \infty]$. Since ν is a density function and $|\Psi| \leq 1$, the absolute value of the difference expression is bounded by 2. Hence,

$$|(12) \text{ due to } x \in [\alpha, \infty]| \leq 2 \int_{\alpha}^{\infty} x \varphi_{0,c^2\sigma_T^2}(x) dx = \frac{2}{\sqrt{2\pi}} \exp(-(\alpha/(c\sigma_T))^2/2).$$

We now bound the contribution to (12) due to $x \in [0, \alpha]$. We use the following:

Claim D.1. For $x \in [0, \alpha]$ and assuming $\text{err}(\psi) \leq 0.05$ and our setting of a, b ,

$$\int_{a+x}^{b+x} \Psi(w) \nu(w-x) dw - \int_{a-x}^{b-x} \Psi(w) \nu(w+x) dw = \Omega(c\alpha^{-1}x).\tag{13}$$

Proof. We break the range of integration into parts according to the density function ν (see Definition 3.3) and bound each part.

$$\int_{a+x}^{b+x} \Psi(w) \nu(w-x) dw - \int_{a-x}^{b-x} \Psi(w) \nu(w+x) dw \quad (14)$$

$$= - \int_{a-x}^{a+x} \Psi(w) \nu(w+x) dw \quad (15)$$

$$+ \int_{a+x}^{1-x} \Psi(w) (\nu(w-x) - \nu(w+x)) dw \quad (16)$$

$$+ \int_{1-x}^{1+x} \Psi(w) (\nu(w-x) - \nu(w+x)) dw \quad (17)$$

$$+ \int_{1+x}^{1+\alpha-x} \Psi(w) (\nu(w-x) - \nu(w+x)) dw \quad (18)$$

$$+ \int_{1+\alpha-x}^{1+\alpha+x} \Psi(w) (\nu(w-x) - \nu(w+x)) dw \quad (19)$$

$$+ \int_{1+\alpha+x}^{b-x} \Psi(w) (\nu(w-x) - \nu(w+x)) dw \quad (20)$$

$$+ \int_{b-x}^{b+x} \Psi(w) \nu(w-x) dw \quad (21)$$

We now bound each part. We use that $C = \Theta(c/\alpha)$ and $(1-a)b - (1+\alpha) = \Theta(\alpha/c)$.

$$|(15)| \leq C \frac{4x^2}{1-a} = O(c^2 \alpha^{-2} x^2)$$

$$(16) = \int_{a+x}^{1-x} \Psi(w) (-2x \frac{C}{1-a}) dw = -2x \frac{C}{1-a} \int_{a+x}^{1-x} \Psi(w) dw \geq 0.1 c \alpha^{-1} x$$

$$|(17)|, |(19)| \leq 2Cx^2 = O(c\alpha^{-1}x^2)$$

$$(18) = 0$$

$$(20) = 2x \frac{C}{b-(1+\alpha)} \int_{1+\alpha+x}^{b-x} \Psi(w) dw \geq 0.1 x c \alpha^{-1}$$

$$|(21)| \leq C \frac{4x^2}{b-(1+\alpha)} = O(c^2 \alpha^{-2} x^2)$$

Our bounds for (16) and (20) use that for Ψ with error at most $\delta < 0.05$ and $x \in [0, \alpha]$, it holds (with our choices for a, b) that $\int_{a+x}^{1-x} \Psi(w) dw < -0.1(1-a)$, $\int_{1+\alpha+x}^{b-x} \Psi(w) dw > 0.1(b-(1+\alpha))$. \square

$$(12) \text{ due to } x \in [0, \alpha] = \int_0^\alpha x \varphi_{0, c^2 \sigma_T^2}(x) \Omega(Cx) dx = \Omega(c^3 \alpha^{-1} \sigma_T^2) . \quad (22)$$

Combining (22) and (13) we establish the claim in the statement of the Lemma: (12) = $\Omega(\alpha^{-1} c^3 \sigma_T^2)$. \square

E Signed sum of Gaussian vectors

Lemma E.1 (Upper bound for the signed sum of Gaussian vectors). *Let $X_1, \dots, X_r \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}_m(0, 1)$ in \mathbb{R}^m . Define*

$$M := \max_{s \in \{\pm 1\}^r} \left\| \sum_{i=1}^r s_i X_i \right\|_2 .$$

Then for every $t > 0$

$$\Pr[M \leq (\sqrt{r} + \sqrt{m} + t) \sqrt{r}] \geq 1 - 2e^{-t^2/2}.$$

In particular,

$$M \leq 2r + \sqrt{rm} \quad \text{with probability } 1 - e^{-r/2}.$$

Proof. Collect the random vectors into the Gaussian matrix $X = [X_1 \dots X_r] \in \mathbb{R}^{m \times r}$. For any sign vector $s \in \{\pm 1\}^r$ we have

$$\sum_{i=1}^r s_i X_i = Xs, \quad \text{hence} \quad \|Xs\|_2 \leq \sigma_{\max}(X) \|s\|_2 = \sigma_{\max}(X) \sqrt{r},$$

where $\sigma_{\max}(X)$ is the largest singular value of X . Since this holds for all s , we get

$$M \leq \sigma_{\max}(X) \sqrt{r}. \quad (23)$$

For an $m \times r$ matrix with i.i.d. $\mathcal{N}(0, 1)$ entries, applying the standard concentration bound (see, e.g., Theorem 4.4.5 in Vershynin [38]) we obtain

$$\Pr[\sigma_{\max}(X) \geq \sqrt{r} + \sqrt{m} + t] \leq 2e^{-t^2/2} \quad (t > 0).$$

Combining this bound with (23) yields the first stated probability bound. For the simplified bound, we plug-in $t = \sqrt{r}$ in the above. \square