# Matrix-Free Two-to-Infinity and One-to-Two Norms Estimation

**Askar Tsyganov**
HSE University
atsyganov@hse.ru

**Sergey Samsonov**
HSE University

**Maxim Rakhuba**
HSE University

## Abstract

In this paper, we propose new randomized algorithms for estimating the two-to-infinity and one-to-two norms in a matrix-free setting, using only matrix-vector multiplications. Our methods are based on appropriate modifications of Hutchinson's diagonal estimator and its Hutch++ modification. We provide sample complexity bounds for both modifications. We further illustrate the practical utility of our algorithms for Jacobian-based regularization in deep neural network training on image classification tasks.

## 1 Introduction

In recent years, there has been growing interest in randomized linear algebra techniques for estimating matrix functions without explicit access to the matrix entries. This setting, known as *matrix-free*, assumes access only to an oracle that computes matrix-vector products with a matrix $A$ and its transpose $A^\top$. The goal is to approximate important properties or functions of $A$ using only these products. Such a framework is essential in modern machine learning, where matrices such as Jacobian of deep neural networks are prohibitively large to form explicitly but allow efficient computation of matrix-vector products via automatic differentiation (autograd).

A classical problem in this framework is stochastic trace estimation [9], in which the trace of a matrix is approximated using a few matrix-vector products with random vectors. Building on this foundation, a variety of improved estimators have been developed. There are variance-reduced methods such as Hutch++ [15], which exploit low-rank structure to accelerate convergence, and dynamic algorithms [6, 23], which adaptively allocate samples to achieve near-optimal accuracy. Structured estimators based on rank-one vectors [4] also offer computational advantages in certain applications. Closely related is the problem of diagonal estimation [2, 1, 7], for which recent work has provided algorithmic improvements and theoretical guarantees. These approaches form the basis of matrix-free randomized methods for approximating key linear algebraic quantities.

In this work, we focus on matrix-free estimation of two operator norms: $\|\cdot\|_{2\to\infty}$ and $\|\cdot\|_{1\to 2}$. Formally, for a matrix $A \in \mathbb{R}^{d\times n}$, the two-to-infinity norm can be equivalently defined as

$$\|A\|_{2\to\infty} = \max_{i\in[d]} \|A_i\|_2.$$

Given the identity

$$\|A\|_{2\to\infty} = \|A^\top\|_{1\to 2}\ ,$$

it suffices to concentrate on the estimation of the two-to-infinity norm. Compared to classical norms such as the spectral or Frobenius norm, the two-to-infinity norm provides finer control over the row-wise structure of a matrix. This is particularly advantageous when dealing with *tall* matrices, where $d \gg n$. In such cases, each row contains relatively few elements compared to the total number of columns, and bounding the two-to-infinity norm ensures that the norm of each row is tightly

Conference on Learning Theory (COLT 2025) Workshop on the Theory of AI for Scientific Computing.

controlled (see Example 1). This localized control is especially useful in high-dimensional statistical inference [5], perturbation analysis [19], and randomized matrix algorithms [10].

Our main contributions are as follows:

- We introduce a novel algorithm tailored specifically for the $\|\cdot\|_{1\to 2}$ and $\|\cdot\|_{2\to\infty}$ norms under the matrix-free setting. Our method enjoys provable convergence guarantees and empirically demonstrates reliable performance where previous algorithms exhibit instability or divergence. This fills an important gap in the current literature and provides a practical tool for both theoretical analysis and downstream applications.
- We apply our methodology as a regularizer to the problem of image classification using deep neural networks. Our method achieves better generalization performance compared to classical Jacobian regularization techniques [8, 21].

In the following sections, we propose randomized algorithms for estimating the two-to-infinity norm using only access to matrix-vector products, and we analyze their theoretical performance guarantees. We discuss related works in Section 2. Then in Section 3 we present our main algorithm, TwINEst (see Algorithm 1), and analyze its sample complexity. Similarly, in Section 4, we provide an improvement of the TwINEst algorithm based on Hutch++ type modification [15], and study theoretical properties of the modified algorithm. Finally, we provide numerical results in Section 5. Proofs are postponed until Appendix.

**Notations.** For vector $a \in \mathbb{R}^d$, $\|a\|_p = (\sum_{i=1}^d a_i^p)^{1/p}$ denotes the $\ell_p$-norm ($p \geq 1$), $\|a\|_\infty = \max_i |a_i|$ denotes the $\ell_\infty$-norm. For matrix $A \in \mathbb{R}^{n\times d}$, $A_i$ denotes the $i$-th row of $A$. $A^\top$ denotes the transpose of a matrix $A$. $\|A\|_F = (\sum_{i=1}^n \sum_{j=1}^d A_{ij}^2)^{1/2}$ denotes the Frobenius norm. We define the induced norms as $\|A\|_{p\to q} := \sup_{x\neq 0} \|Ax\|_q / \|x\|_p$. For example, $\|A\|_{2\to\infty}$ is equal to the maximum $\ell_2$ norm of the rows and $\|A\|_{1\to 2}$ is equal to the maximum $\ell_2$ norm of the columns. We let $A_k = \arg\min_{B:\mathrm{rk}(B)\leq k} \|A - B\|_F$ denote the best $k$-rank approximation to $A$. For matrices $A, B \in \mathbb{R}^{n\times d}$, $A \odot B$ denotes the Hadamard product (element-wise). We denote by $[d]$ the set $\{1, 2, \ldots, d\}$.

## 2 Related Works

Estimating matrix operator norms induced by different $\ell_p \to \ell_q$ combinations has been an important topic in numerical linear algebra and machine learning, especially in settings where the matrix is not explicitly available, but matrix-vector products with the matrix and its transpose can be performed.

A foundational contribution in this direction is the work [3], which proposes a general iterative algorithm for estimating $\|\cdot\|_{p\to q}$ norm for arbitrary $p$ and $q$, based on a generalization of the classical power method [16]. The approach relies on alternating optimization steps to approximate the optimal input and output vectors. However, this method does not specifically address the cases of $\|\cdot\|_{1\to 2}$ and $\|\cdot\|_{2\to\infty}$ norms, which are of central interest in our work.

The paper [21] extends the methodology of [3] and applies it to more specialized norms, including the $\|\cdot\|_{1\to 2}$ and $\|\cdot\|_{2\to\infty}$ norm cases. The authors propose using adversarial training techniques as a way to implicitly regularize the operator norm of layers in neural networks. In doing so, they demonstrate that certain adversarial perturbations correspond to directions aligned with large operator norms. While the method provides a practical heuristic, it lacks theoretical convergence guarantees and, as we will demonstrate in our work, often fails to converge in practice for the $\|\cdot\|_{1\to 2}$ and $\|\cdot\|_{2\to\infty}$ norm settings.

The two-to-infinity norm has found increasing utility as a tool for theoretical analysis in various areas of high-dimensional statistics and learning theory. For instance, it plays a central role in understanding the geometry of singular subspaces, particularly in contexts where entrywise control is crucial [5]. In the setting of bandit problems with low-rank structure, the norm has been used to derive tight bounds for subspace recovery, enabling sharper regret guarantees [10]. Moreover, recent advances in spectral perturbation theory have extended classical results such as the Davis–Kahan theorem to the two-to-infinity norm setting, leading to improved guarantees for exact clustering and related tasks [19].

## 3 Main Algorithmic Results

### 3.1 Hutchinson's Diagonal Estimator

Our algorithms build upon a well-known technique for estimating the diagonal of a square matrix using only matrix-vector products. This technique is known as the Hutchinson diagonal estimator [2, 1, 7]. In this section, we briefly introduce the estimator and provide concentration inequality that will be useful for analyzing our algorithms.

**Definition 1** (Hutchinson's Diagonal Estimator)**.** *Let $X^1, \dots, X^m \in \{-1, 1\}^d$ be independent Rademacher random vectors. For a square matrix $A \in \mathbb{R}^{d\times d}$, the Hutchinson's diagonal estimator $D^m(A) \in \mathbb{R}^d$ is defined as*

$$D^m(A) \coloneqq \frac{1}{m}\sum_{i=1}^{m} X^i \odot (AX^i).$$

This estimator is a natural extension of the classical Hutchinson method for trace estimation [9] to the problem of diagonal estimation. It provides an unbiased estimate of the diagonal, satisfying, for each $i \in [d]$:

$$\mathbb{E}[D^m(A)] = \operatorname{diag}(A), \quad \text{and} \quad \operatorname{Var}[D_i^1(A)] = \|A_i\|_2^2 - A_{ii}^2.$$

In addition to being unbiased, the estimator also admits high-probability error bounds that characterize its concentration around the true diagonal. In particular, we rely on the following result from [7], which provides the following bound for the $\ell_2$ norm of the Hutchinson's estimator error:

**Theorem 1** (Theorem 1 in [7])**.** *Let $A \in \mathbb{R}^{d\times d}$, $m \in \mathbb{N}$, $\delta \in (0, 1]$. Then with probability at least $1 - \delta$:*

$$\|D^m(A) - \operatorname{diag}(A)\|_2 \le c\sqrt{\frac{\log(2/\delta)}{m}}\|A - \operatorname{diag}(A)\|_F,$$

*where $c$ is an absolute constant.*

### 3.2 Our method

Now we describe our strategy for estimating $\|\cdot\|_{2\to\infty}$ norm. The main idea is that the diagonal entries of the matrix $AA^\top$ correspond to the squared $\ell_2$ norms of the rows of $A$. Therefore, the $\|\cdot\|_{2\to\infty}$ norm can be equivalently expressed as

$$\|A\|_{2\to\infty}^2 = \max_{i\in[d]} \operatorname{diag}(AA^\top)_i.$$

This identity suggests a natural strategy: instead of computing all row norms explicitly, we can estimate the diagonal of $AA^\top$ using the Hutchinson method, which only requires matrix-vector products with $A$ and $A^\top$. The final estimate of the $\|A\|_{2\to\infty}$ norm is then obtained by taking the maximum of the estimated diagonal.

However, estimating the maximum value through the direct application of Hutchinson's method introduces high variance, leading to a noisy approximation. To mitigate this, we eliminate one source of randomness in the final estimate. Let $D$ be the estimate of the diagonal of $AA^\top$. While the entries of $D$ are typically noisy, we can reduce variance by avoiding direct use of $\max_i D_i$. Instead, we first identify the index of the maximum estimated value, $j = \arg\max_i D_i$, and then explicitly compute the exact $\ell_2$-norm of the $j$-th row using a matrix-vector product with the $j$-th standard basis vector. This approach significantly improves the quality of the estimate, which is supported ablation study carried out in Section 5.1. The detailed procedure is presented in Algorithm 1.

### 3.3 TwINEst Algorithm Analysis

We begin by establishing upper bounds on the sample complexity of our proposed algorithms. In the context of randomized numerical linear algebra, sample complexity typically refers to the number of matrix-vector multiplications required to approximate a matrix quantity within a specified error tolerance and failure probability. However, our analysis does not directly relate sample complexity to the approximation error. Instead, our guarantees are expressed in terms of two key quantities: the

**Algorithm 1** TwINEst: Two-Infinity Norm Estimation

**Input:**
Oracle for matrix-vector multiplication with matrix $A \in \mathbb{R}^{d\times n}$,
Oracle for matrix-vector multiplication with matrix $A^T \in \mathbb{R}^{n\times d}$,
Positive integer $m \in \mathbb{N}$: number of iterations.

**Output:**
An estimate of the $\|A\|_{2\to\infty}$ norm.

Sample $m$ random Rademacher vectors $X^1, X^2, \ldots, X^m$, where each $X^i \in \{-1, 1\}^d$
**for** each $i = 1, 2, \ldots, m$ **do**
  Compute $t_i = X^i \odot AA^\top X^i$
**end for**
Compute $D = \frac{1}{m}\sum_{i=1}^m t_i \in \mathbb{R}^d$ ▷ $D$ - estimate of the $AA^\top$ diagonal
Find $j = \arg\max_i D_i$
Compute $L = \|A^\top e_j\|_2$ ▷ $e_j$ - is the $j$-th standard basis vector
**return** $L$

failure probability $\delta$ and the gap $\Delta$ between the largest $\ell_2$-norm of a row of $A$ and the $\ell_2$-norm of the closest non-maximum row. Formally, let $M = \max_i \|A_i\|_2$. We define

$$\Delta = M - \max_{i\,:\,\|A_i\|_2 < M} \|A_i\|_2. \tag{1}$$

When $\Delta$ is large, the row with the maximum $\ell_2$-norm can be identified more easily, resulting in lower sample complexity. Conversely, a small $\Delta$ indicates that the top norms are close, requiring more samples to reliably identify the maximum.

We now establish the sample complexity required for our algorithm, TwINEst, to converge to the exact value of the matrix norm $\|\cdot\|_{2\to\infty}$ with high probability.

**Theorem 2** (TwINEst Sample Complexity)**.** *Let $A \in \mathbb{R}^{d\times n}$, $m \in \mathbb{N}$, and $\Delta$ be defined in* (1)*. Let $T^m(A)$ be the result of Algorithm 1 based on $m$ random vectors. Then, it suffices to take*

$$m > \frac{8\log(2d/\delta)}{\Delta^2}\|AA^\top - \mathrm{diag}(AA^\top)\|_{2\to\infty}^2$$

*to ensure $T^m(A) = \|A\|_{2\to\infty}$ with probability at least $1-\delta$.*

**Discussion.** Proof of Theorem 2 is given in Appendix B.1. Results in a similar vein to Theorem 2 were previously obtained for the Hutchinson estimator [20, 11], Hutch++ [15], and some other methods. Importantly, our analysis goes beyond simply bounding the probability of deviation from the true value; instead, we directly bound the probability that our algorithm returns the exact value. To the best of our knowledge, our bound is the first one on the sample complexity of randomized estimation of two-to-infinity norm.

Notably, our method offers practical advantages over power-iteration based algorithms [21], partly due to its straightforward parallelization.

## 4 Improved Algorithm

In this section, we improve the sample complexity bounds by modifying our algorithm using the variance reduction technique introduced in [15]. The key insight is that a low-dimensional random sketch suffices to capture the dominant eigenspace of $AA^\top$. Once this component is computed exactly, applying a Hutchinson estimator to the residual yields reduced variance.

More precisely, we divide the budget of $m$ matrix-vector multiplications into three parts. First, we compute a random sketch $AA^\top S$, where $S$ is a Rademacher matrix (with i.i.d. $\{\pm 1\}$ entries), and obtain its orthonormal basis $Q$ (e.g., via QR decomposition). This allows us to decompose the matrix $AA^\top$ as follows:

$$AA^\top = AA^\top QQ^\top + AA^\top(I - QQ^\top),$$

where the diagonal of the low-rank component $AA^\top QQ^\top$ can be computed exactly using the second portion of the budget. The diagonal of the residual term $AA^\top(I - QQ^\top)$ is then estimated using

**Algorithm 2** TwINEst++

**Input:**
Oracle for matrix-vector multiplication with matrix $A \in \mathbb{R}^{d\times n}$,
Oracle for matrix-vector multiplication with matrix $A^T \in \mathbb{R}^{n\times d}$,
Positive integer $m \in \mathbb{N}$: number of iterations.

**Output:**
An estimate of the $\|A\|_{2\to\infty}$ norm.

Sample $\frac{m}{3}$ random Rademacher vectors $X^1, X^2, \ldots, X^{\frac{m}{3}}$, where each $X^i \in \{-1, 1\}^d$
Sample random Rademacher matrix $S \in \mathbb{R}^{d\times \frac{m}{3}}$, with i.i.d. $\{-1, 1\}$ entries
Compute an orthonormal basis $Q$ for $AA^\top S$ ▷ via QR decomposition
**for** each $i = 1, 2, \ldots, \frac{m}{3}$ **do**
    Compute $t_i = X^i \odot AA^\top (I - QQ^\top)X^i$
**end for**
Compute $\hat{D} = \frac{1}{m}\sum_{i=1}^{m} t_i \in \mathbb{R}^d$
Compute $D = \hat{D} + \mathrm{diag}(AA^\top QQ^\top)$ ▷ $D$ - estimate of the $AA^\top$ diagonal
Find $j = \arg\max_i D_i$
Compute $L = \|A^\top e_j\|_2$ ▷ $e_j$ - is the $j$-th standard basis vector
**return** $L$

Hutchinson's method with the remaining budget. Summing these two components yields an estimate of $\mathrm{diag}(AA^\top)$ with lower variance than that produced by the TwINEst algorithm.

This modification results to improved sample complexity bounds, as formalized in Theorem 3. A detailed description of the algorithm is provided in Algorithm 2.

### 4.1 TwINEst++ Algorithm Theoretical Analysis

Here, we analyze an improved version of our algorithm, TwINEst++, which leverages low-rank approximations to enhance estimation efficiency. The structure of the proof follows the same reasoning as in the case of Theorem 2 for the base TwINEst algorithm. In particular, we again rely on a concentration inequality for the diagonal estimator. However, for TwINEst++, we employ a refined concentration result provided in Theorem 1, which yields a tighter control over the estimation error and enables the improved complexity result stated below.

**Theorem 3** (TwINEst++ Sample Complexity)**.** *Let $A \in \mathbb{R}^{d\times n}$, $m \in \mathbb{N}$, and $\Delta$ be defined as in Equation 1. Let $T^m_{++}(A)$ be the output of Algorithm 2 based on $m$ matrix-vector multiplications. Then, it suffices to choose*

$$m = O\left(\frac{\sqrt{\log(2/\delta)}}{\Delta}\|A\|_F^2 + \log(1/\delta)\right)$$

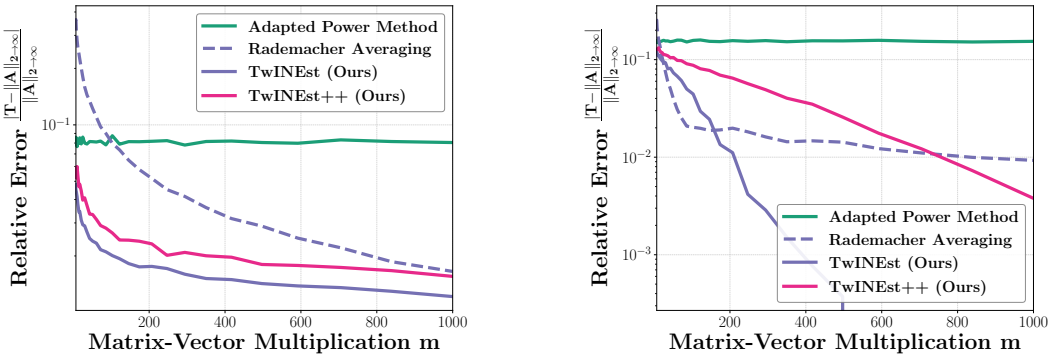
*to ensure that $T^m_{++}(A) = \|A\|_{2\to\infty}$ with probability at least $1 - \delta$.*

*Proof.* See Appendix B.2 of the Appendix. □

Theorem 3 shows that TwINEst++ achieves an improved query complexity compared to the original algorithm, particularly in challenging scenarios when $\Delta \to 0$, making identification of the correct row difficult. Specifically, the sample complexity is reduced from $O(1/\Delta^2)$ in the original TwINEst algorithm to $O(1/\Delta)$ in TwINEst++.

## 5 Experiments

We present an empirical evaluation of the proposed algorithms. Our experiments cover two settings: synthetic and real-world matrices (see Section 5.1 and Section 5.2), and applications to deep learning tasks (see Section 5.3). The results indicate that the algorithms yield accurate estimates of the two-to-infinity norm and exhibit improved convergence behavior compared to existing methods. The source code is available at:https://anonymous.4open.science/r/jacobian-image-classification-BB91.




(a) Synthetic data with $\Delta = 10^{-2}$.

(b) Synthetic data with $\Delta = 10^{-1}$.

Figure 1: Comparison of methods for estimating the two-to-infinity norm on random square matrices. Shown is the relative error versus the number of matrix-vector multiplications, averaged over 500 trials.

### 5.1 Synthetic Data

In the following two sections, we compare the following methods:

- **Adapted Power Method.** A modification of the power iteration method for estimating the two-to-infinity norm, introduced by [21]. This approach incorporates a projection operator onto the $\ell_\infty$ unit ball within the power method. Further details are provided in Algorithm 3.
- **Rademacher Averaging.** Similar to our TwINEst method (Algorithm 1), but the final estimate is obtained by taking the maximum of diagonal estimates instead of computing the argmax row.
- **TwINEst.** Our algorithm introduced in Algorithm 1.
- **TwINEst++.** An enhanced version of TwINEst incorporating random projections, described in Algorithm 2.

The Adapted Power Method lacks theoretical guarantees and may diverge on certain matrices (see Example 2). Therefore, we hypothesize that our algorithms will outperform it. In contrast, the TwINEst and TwINEst++ methods are supported by theoretical convergence guarantees. We also expect the TwINEst algorithm to consistently outperform Rademacher Averaging.

For our synthetic experiments, we generate random Gaussian matrices $A \in \mathbb{R}^{5000\times 5000}$. Specifically, we fix a parameter $\Delta < 1$ and sample values $c_1, \ldots, c_{5000} \sim \mathcal{U}[0,1]$, setting $c_1 = 1 + \Delta$, $c_2 = 1$. Each row of the matrix $A$ is normalized such that its $\ell_2$-norm is equal to $c_i$, ensuring a gap of magnitude $\Delta$ between the largest and second-largest row norms. The singular value densities of these matrices are quite similar for different values of $\Delta$ (see Figure 3a). Therefore, the parameter $\Delta$ may have a significant impact on the convergence rate.

The results of synthetic experiments for different values of $\Delta$ are illustrated in Figure 1. As anticipated, TwINEst consistently outperforms Rademacher Averaging. The Adapted Power Method from [21] fails to converge, as evidenced by its flat performance line. For a relatively large gap $\Delta = 10^{-1}$, the TwINEst algorithm rapidly converges, achieving accurate results consistently within approximately 400 iterations.

### 5.2 Real World Data

To validate our methods on real-world data, we evaluate them using the Jacobian matrix $J \in \mathbb{R}^{3\cdot 32\cdot 32\times 100}$ of a WideResNet-16-10 [24] pre-trained on CIFAR-100 [13]. Given the low-rank structure of $J$, we expect TwINEst++ to outperform all other methods.

Figure 2 confirms our hypothesis: the TwINEst++ algorithm achieves rapid convergence consistent with the approximate rank of matrix $J$, significantly outperforming other algorithms. Again, the

Adapted Power Method from [21] fails to converge, highlighting the practical efficacy of our algorithms. For a comprehensive ablation, we provide a comparison between the relative error of the methods and their floating point operation counts (FLOPs) in Appendix D.3.

### 5.3 Deep Learning Applications

We study whether penalizing the $\|\cdot\|_{2\to\infty}$ norm of the input-output Jacobian can improve the generalization ability of neural networks in image classification. The Jacobian of a standard image classifier is typically a tall matrix: the number of output classes is much smaller than the number of input features (pixels). In this setting, the $\|\cdot\|_{2\to\infty}$ norm provides finer control over the worst-case directional response of individual output units, which is not captured by global norms like the Frobenius or spectral norm. As illustrated in Example 1, the $\|\cdot\|_{2\to\infty}$ norm remains bounded even when the number of rows grows, unlike spectral and Frobenius norms, which scale with dimensionality. This makes it particularly well-suited for regularizing tall Jacobians in high-dimensional input spaces.

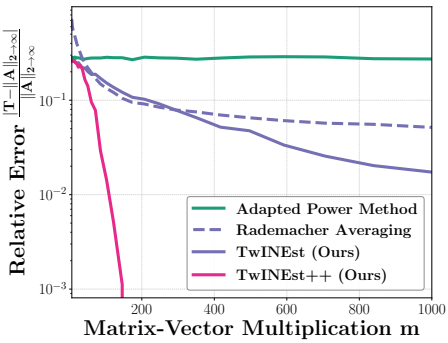



Figure 2: Comparison of methods for estimating the two-to-infinity norm of the Jacobian matrix of WideResNet-16-10 pre-trained on CIFAR-100. The plot shows the relative error versus the number of matrix-vector multiplications, averaged over 500 trials.

We compare our regularizer to established Jacobian-based penalties: Frobenius norm [8], spectral norm [21], and $\ell_\infty$ norm [21]. The $\|\cdot\|_{2\to\infty}$ norm is estimated using TwINEst Algorithm 1. We minimize the following objective function:

$$\mathcal{L}(x, y) = \mathcal{L}_{\text{CE}}(f(x), y) + \lambda \cdot \|J_f(x)\|^2,$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, $f$ is the network, $J_f(x)$ is the Jacobian of the logits with respect to the input, and $\|\cdot\|$ is one of the following norms: Frobenius, spectral, $\ell_\infty$, or $\|\cdot\|_{2\to\infty}$.

Experiments are conducted on CIFAR-100 [13] and TinyImageNet [14] using the WideResNet-16-10 architecture [24], implemented in PyTorch [18] and trained on a single NVIDIA Tesla V100 GPU. The hyperparameters are detailed in Appendix D.2. We evaluate each method by reporting the final test accuracy, the stable rank of the Jacobian (computed as $\|J\|_F^2/\|J\|_2^2$), total training time (in wall-clock hours), and the value of the regularization weight $\lambda$ used during training.

| | | | **CIFAR-100** | | **TinyImageNet** | |
|---|---|---|---|---|---|---|
| **Regularizer** | $\lambda$ | Time | Acc. ↑ | S. Rank ↓ | Acc. ↑ | S. Rank ↓ |
| No regularization | – | **4h** | $75.5 \pm 0.2$ | $32.0 \pm 1.1$ | $57.8 \pm 1.3$ | $30.9 \pm 4.3$ |
| Frobenius | $10^{-7}$ | 12h | $75.7 \pm 0.5$ | $31.6 \pm 0.2$ | $58.6 \pm 0.3$ | $27.8 \pm 0.9$ |
| Spectral | $10^{-6}$ | 10h | $75.7 \pm 0.3$ | $32.0 \pm 1.0$ | $57.4 \pm 0.8$ | $28.2 \pm 0.3$ |
| Infinity | $10^{-6}$ | 10h | $75.8 \pm 0.4$ | $30.7 \pm 1.2$ | $57.1 \pm 0.7$ | $28.8 \pm 0.9$ |
| TwINEst (ours) | $10^{-8}$ | 13h | $\mathbf{77.3 \pm 0.1}$ | $\mathbf{18.3 \pm 0.8}$ | $\mathbf{59.6 \pm 0.9}$ | $\mathbf{24.9 \pm 0.3}$ |

Table 1: Comparison of Jacobian regularization methods on CIFAR-100 and TinyImageNet datasets using WideResNet-16-10. Metrics averaged for 3 trials.

As shown in Section 5.3, our method improves the generalization ability of WideResNet-16-10 on the CIFAR-100 and TinyImageNet datasets. In contrast, while other methods require slightly less training time, they do not yield significant improvements over the baseline.

## 6 Conclusion

In this paper, we proposed two novel matrix-free stochastic algorithms for estimating the two-to-infinity and one-to-two norms, and provided theoretical analysis of their behavior. Our empirical results demonstrate that the proposed methods outperform existing approaches in terms of both

accuracy and computational efficiency. Furthermore, we showed that our algorithms can be easily integrated into deep learning pipelines. In particular, we illustrated their utility in improving the generalization performance of neural networks on image classification tasks.

## Acknowledgement

This work was supported by the Ministry of Economic Development of the Russian Federation (code 25-139-66879-1-0003). This research was supported in part through computational resources of HPC facilities at HSE University [12].

## References

[1] Robert A Baston and Yuji Nakatsukasa. Stochastic diagonal estimation: probabilistic bounds and an improved algorithm. *arXiv preprint arXiv:2201.10684*, 2022.

[2] Costas Bekas, Effrosyni Kokiopoulou, and Yousef Saad. An estimator for the diagonal of a matrix. *Applied numerical mathematics*, 57(11-12):1214–1229, 2007.

[3] David W Boyd. The power method for lp norms. *Linear Algebra and its Applications*, 9:95–101, 1974.

[4] Zvonimir Bujanovic and Daniel Kressner. Norm and trace estimation with random rank-one vectors. *SIAM Journal on Matrix Analysis and Applications*, 42(1):202–223, 2021.

[5] Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439, 2019.

[6] Prathamesh Dharangutte and Christopher Musco. Dynamic trace estimation. *Advances in Neural Information Processing Systems*, 34:30088–30099, 2021.

[7] Prathamesh Dharangutte and Christopher Musco. A tight analysis of hutchinson's diagonal estimator. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 353–364. SIAM, 2023.

[8] Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

[9] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[10] Yassir Jedra, William Réveillard, Stefan Stojanovic, and Alexandre Proutiere. Low-rank bandits via tight two-to-infinity singular subspace recovery. *arXiv preprint arXiv:2402.15739*, 2024.

[11] Shuli Jiang, Hai Pham, David Woodruff, and Richard Zhang. Optimal sketching for trace estimation. *Advances in Neural Information Processing Systems*, 34:23741–23753, 2021.

[12] PS Kostenetskiy, RA Chulkevich, and VI Kozyrev. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, page 012050. IOP Publishing, 2021.

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[14] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[15] Raphael A Meyer, Cameron Musco, Christopher Musco, and David P Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 142–155. SIAM, 2021.

[16] RV Mises and Hilda Pollaczek-Geiringer. Praktische verfahren der gleichungsauflösung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 9(1):58–77, 1929.

[17] Cameron Musco and Christopher Musco. Projection-cost-preserving sketches: Proof strategies and constructions. *arXiv preprint arXiv:2004.08434*, 2020.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. 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.

[19] Marianna Pensky. Davis-kahan theorem in the two-to-infinity norm and its application to perfect clustering. *arXiv preprint arXiv:2411.11728*, 2024.

[20] Farbod Roosta-Khorasani and Uri Ascher. Improved bounds on sample size for implicit matrix trace estimators. *Foundations of Computational Mathematics*, 15(5):1187–1212, 2015.

[21] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. Adversarial training is a form of data-dependent operator norm regularization. *Advances in Neural Information Processing Systems*, 33:14973–14985, 2020.

[22] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[23] David Woodruff, Fred Zhang, and Richard Zhang. Optimal query complexities for dynamic trace estimation. *Advances in Neural Information Processing Systems*, 35:35049–35060, 2022.

[24] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

# A Technical Lemmas

**Lemma 4.** *Let $A \in \mathbb{R}^{d\times d}$, $m \in \mathbb{N}$, $i \in [d]$, and $\varepsilon \geq 0$. Then*

$$\mathbb{P}\left(|D_i^m(A) - A_{ii}| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2 m}{2(\|A_i\|_2^2 - A_{ii}^2)}\right).$$

*Proof.* This statement follows from Theorem 2 in [1], but we provide an independent argument. Since $(X_i^k)^2 = 1$ for Rademacher random variables,

$$\begin{aligned} D_i^m(A) - A_{ii} = \frac{1}{m}\sum_{k=1}^m (X^k \odot AX^k)_i - A_{ii} = \frac{1}{m}\sum_{k=1}^m \left(A_{ii}(X_i^k)^2 + \sum_{j\neq i} X_i^k A_{ij} X_j^k\right) - A_{ii} \\ = \frac{1}{m}\sum_{k=1}^m \sum_{j\neq i} X_i^k A_{ij} X_j^k. \end{aligned}$$

Define $Y_j^k = X_i^k X_j^k$. Since the product of two independent Rademacher variables is again Rademacher, and they remain mutually independent,

$$\mathbb{P}\left(|D_i^m(A) - A_{ii}| \geq \varepsilon\right) = \mathbb{P}\left(\left|\sum_{k=1}^m \sum_{j\neq i} \frac{A_{ij}}{m} Y_j^k\right| \geq \varepsilon\right).$$

Applying Hoeffding's inequality (see [22]) yields the desired result. □

**Theorem 5.** *Let $A \in \mathbb{R}^{d\times d}$, $m \in \mathbb{N}$, $\varepsilon \geq 0$, and let $\bar{A}$ be the matrix $A$ with diagonal entries set to zero. Then*

$$\mathbb{P}\left(\|D^m(A) - \operatorname{diag}(A)\|_\infty \geq \varepsilon\right) \leq 2d\exp\left(-\frac{\varepsilon^2 m}{2\|\bar{A}\|_{2\to\infty}^2}\right).$$

*Proof.*

$$\begin{aligned} \mathbb{P}\left(\|D^m(A) - \operatorname{diag}(A)\|_\infty \geq \varepsilon\right) &= \mathbb{P}\left(\max_i |D_i^m(A) - \operatorname{diag}(A)_i| \geq \varepsilon\right) \\ &\leq \sum_{i=1}^d \mathbb{P}\left(|D_i^m(A) - \operatorname{diag}(A)_i| \geq \varepsilon\right) \qquad \text{(By the union bound)} \\ &\leq \sum_{i=1}^d 2\exp\left(-\frac{\varepsilon^2 m}{2\|\bar{A}_i\|_2^2}\right) \qquad \text{(By Lemma 4)} \\ &\leq 2d\exp\left(-\frac{\varepsilon^2 m}{2\|\bar{A}\|_{2\to\infty}^2}\right). \end{aligned}$$

□

**Lemma 6.** *Let $k \in \mathbb{N}$ and let $A \in \mathbb{R}^{d\times d}$ be a positive semidefinite (PSD) matrix. Denote by $A_k$ the best rank-$k$ approximation of $A$ in the Frobenius norm. Then,*

$$\|A - A_k\|_F \leq \frac{1}{\sqrt{k}}\operatorname{tr}(A).$$

*Proof.* Since $A$ is PSD, it admits an eigenvalue decomposition $A = U\Lambda U^\top$ with non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. The best rank-$k$ approximation $A_k$ is obtained by keeping the top $k$ eigenvalues. Therefore,

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^d \lambda_i^2.$$

Applying the inequality $\lambda_i \le \lambda_{k+1}$ for all $i > k$ and using Cauchy–Schwarz, we obtain

$$\sum_{i=k+1}^{d} \lambda_i^2 \le \lambda_{k+1} \sum_{i=k+1}^{d} \lambda_i \le \frac{1}{k} \left( \sum_{i=1}^{d} \lambda_i \right)^2 = \frac{1}{k} \operatorname{tr}(A)^2.$$

Taking the square root gives the desired bound. □

# B High Dimensional Proofs

## B.1 Proof of Theorem 2

For simplicity of notation, let $B := AA^\top$ and $\bar{B}$ denote the matrix $B$ with its diagonal entries set to zero. Let $D := D^m(B)$ be the diagonal estimate of $B$.

Recall that, as discussed in Section 3, the goal of the algorithm is to find an index corresponding to a row of maximal $\ell_2$-norm. The key observation is that for any $\gamma \in \arg\max_i B_{ii}$ (there might be multiple rows with maximal norm), we need to show that its estimate $D_\gamma$ dominates all other estimates $D_j$ for $j \in S$, where $S := \{i \mid i \notin \arg\max_i B_{ii}\}$ be the set of non-maximal rows.

By Theorem 5, with probability at least $1 - \delta$:

$$\|D - \operatorname{diag}(B)\|_\infty \le \varepsilon, \quad \text{where} \quad \varepsilon = \sqrt{\frac{2\log(2d/\delta)}{m}} \|\bar{B}\|_{2\to\infty}.$$

This bound implies that for each $i$, $D_i \in [B_{ii} - \varepsilon, B_{ii} + \varepsilon]$ . Moreover, by definition of $\Delta$ for any $j$, $B_{\gamma\gamma} \ge B_{jj} + \Delta$. Combining these facts, we conclude that for any $\gamma \in \arg\max_i B_{ii}$,

$$D_\gamma - D_j \ge (B_{\gamma\gamma} - \varepsilon) - (B_{jj} + \varepsilon) = \underbrace{(B_{\gamma\gamma} - B_{jj})}_{\ge \Delta} - 2\varepsilon > \Delta - 2\,(\Delta/2) = 0,$$

where the last inequality holds when $\varepsilon < \Delta/2$.

This shows that for any maximal row $\gamma$ and any non-maximal row $j$, $D_\gamma > D_j$ with probability at least $1 - \delta$. Therefore, the algorithm correctly identifies a maximal row, meaning that $T^m(A) = \|A\|_{2\to\infty}$.

Finally, the condition $\varepsilon < \Delta/2$ is equivalent to

$$m > \frac{8\log(2d/\delta)}{\Delta^2} \|\bar{B}\|_{2\to\infty}^2 = \frac{8\log(2d/\delta)}{\Delta^2} \|AA^\top - \operatorname{diag}(AA^\top)\|_{2\to\infty}^2.$$

## B.2 Proof of Theorem 3

Define $B := AA^\top$. Let $S \in \mathbb{R}^{d\times l}$ be a random Rademacher matrix, and let $Q$ be an orthonormal basis for the range of $BS$. We decompose $B$ as

$$B = BQQ^\top + B(I - QQ^\top),$$

where $BQQ^\top$ can be computed exactly using $l$ matrix-vector products with $B$, and the challenge is to estimate $\operatorname{diag}(B(I - QQ^\top))$.

Define $\hat{D} := D^k(B(I - QQ^\top))$. Let $k \in \mathbb{N}$ and $l = O(k + \log(1/\delta))$. Then we have with probability at least $1 - \delta$:

$$
\begin{aligned}
\|\hat{D} - \operatorname{diag}(B(I - QQ^\top))\|_\infty &\leq \|\hat{D} - \operatorname{diag}(B(I - QQ^\top))\|_2 \\
&\leq c\sqrt{\frac{\log(2/\delta)}{k}}\|B(I - QQ^T)\|_F && \text{(By Theorem 1)} \\
&\leq 2c\sqrt{\frac{\log(2/\delta)}{k}}\|B - B_k\|_F \\
& && \text{(By Corollary 7 and Claim 1 from [17])} \\
&\leq 2c\sqrt{\frac{\log(2/\delta)}{k^2}}\operatorname{tr}(B) && \text{(By Lemma 6)} \\
&= 2c\sqrt{\frac{\log(2/\delta)}{k^2}}\|A\|_F^2 && \text{(since } B = AA^\top\text{)}
\end{aligned}
$$

Setting $k = O\left(\frac{\sqrt{\log(2/\delta)}}{\Delta}\|A\|_F^2\right)$ ensures that

$$\|\hat{D} - \operatorname{diag}(B(I - QQ^\top))\|_\infty < \Delta/2.$$

Finally, following the same reasoning as in the proof of Theorem 2, we conclude that $T_{++}^m(A) = \|A\|_{2\to\infty}$ with probability at least $1 - \delta$, when

$$m = O\left(\frac{\sqrt{\log(2/\delta)}}{\Delta}\|A\|_F^2 + \log(1/\delta)\right).$$

# C Two-To-Infinity Norm Properties

**Example 1.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix with entries*

$$A_{ij} = 1/\sqrt{n}.$$

*Then*

$$\|A\|_{2\to\infty} = 1, \quad \|A\|_2 = \|A\|_F = \sqrt{d}.$$

*This example highlights that while spectral and Frobenius norms grow with the dimension $d$, the two-to-infinity norm remains bounded, emphasizing its effectiveness in controlling row-wise behavior independently of $d$.*

**Lemma 7.** *For any matrix $A \in \mathbb{R}^{n\times d}$*

$$\|A\|_{2\to\infty} = \max_i \|A_i\|_2$$

*Proof.* Using the fact that the vector norm $\|\cdot\|_2$ is dual to itself,

$$
\begin{aligned}
\|A\|_{2\to\infty} = \sup_{\|x\|_2\leq 1}\max_i |(Ax)_i| = \sup_{\|x\|_2\leq 1}\max_i |\langle A_{i\cdot}, x\rangle| &= \max_i \overbrace{\sup_{\|x\|_2\leq 1}|\langle A_{i\cdot}, x\rangle|}^{\text{dual norm}} \\
&= \max_i \|A_{i\cdot}\|_2
\end{aligned}
$$

We can swap supremum and maximum because any vector norm is a continuous function and the closed unit ball is a compact set. □

**Lemma 8** (Right Unitarily Invariant)**.** *Let $U \in \mathbb{R}^{n\times n}$ be a unitary matrix and $A \in \mathbb{R}^{d\times n}$, then*

$$\|AU\|_{2\to\infty} = \|A\|_{2\to\infty}$$

*Proof.* Using the definition of two-to-infinity norm and the commonly known fact that the $\|\cdot\|_2$ vector norm is unitarily invariant,

$$\|AU\|_{2\to\infty} = \sup_{x\neq 0} \frac{\|AUx\|_\infty}{\|x\|_2} = \sup_{x\neq 0} \frac{\|A(Ux)\|_\infty}{\|Ux\|_2}$$

Let us denote $y = Ux$. Since $U$ is unitary, we have $\text{rank}(U) = n$, so $x = U^{-1}y$ and we can take the supremum over the vector $y$.

$$= \sup_{y\neq 0} \frac{\|Ay\|_\infty}{\|y\|_2} = \|A\|_{2\to\infty}$$

□

**Lemma 9.** *For any matrices $A \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{d\times m}$, and $C \in \mathbb{R}^{k\times n}$*

$$\|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty}\|B\|_2$$

$$\|CA\|_{2\to\infty} \leq \|C\|_\infty\|A\|_{2\to\infty}$$

*Proof.* Using the fact that the $\|\cdot\|_2$ vector norm is sub-multiplicative ($\|Bx\|_2 \leq \|B\|_2\|x\|_2$),

$$\begin{aligned}\|AB\|_{2\to\infty} = \sup_{x\neq 0} \frac{\|ABx\|_\infty}{\|x\|_2} = \sup_{x\neq 0} \frac{\|ABx\|_\infty\|B\|_2}{\|B\|_2\|x\|_2} &\leq \sup_{x\neq 0} \frac{\|ABx\|_\infty}{\|Bx\|_2}\|B\|_2 \\ &\leq \sup_{y\neq 0} \frac{\|Ay\|_\infty}{\|y\|_2}\|B\|_2 = \|A\|_{2\to\infty}\|B\|_2\end{aligned}$$

The last expression follows from Hölder's inequality:

$$\begin{aligned}\|CA\|_{2\to\infty} = \sup_{\|x\|_2=1} \|CAx\|_\infty = \sup_{\|x\|_2=1} \max_i |\langle C_i, Ax\rangle| &\leq \sup_{\|x\|_2=1} \max_i \|C_i\|_1\|Ax\|_\infty \\ &= \max_i \|C_i\|_1 \sup_{\|x\|_2=1} \|Ax\|_\infty = \|C\|_\infty\|A\|_{2\to\infty}\end{aligned}$$

□

**Remark 10.** *The norm $\|\cdot\|_{2\to\infty}$ is not sub-multiplicative. For example, $\|AB\|_{2\to\infty} = \sqrt{8} > 2 = \|A\|_{2\to\infty}\|B\|_{2\to\infty}$ when*

$$A = B = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, \quad AB = \begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix}.$$

# D Details for Experiments

## D.1 Adapted Power Method

**Definition 2** (Projection Operator for $\ell_p$)**.** *Let $x \in \mathbb{R}^d$. The projection operator $\psi_p(x)$ onto the $\ell_p$ unit sphere is defined as*

$$\psi_p(x) = \begin{cases}\text{sign}(x) \odot |x|^{p-1}/\|x\|_p^{p-1}, & \text{if } p < \infty,\\ |\mathcal{I}|^{-1}\,\text{sign}(x) \odot \mathbf{1}_\mathcal{I}, & \text{if } p = \infty,\end{cases}$$

*where $\mathcal{I} := \{i \in [d] : |x_i| = \|x\|_\infty\}$, and $\mathbf{1}_\mathcal{I} := \sum_{i\in\mathcal{I}} e_i$ is the indicator vector over $\mathcal{I}$.*

**Example 2** (Divergence of the Adapted Power Method)**.** *Let $A \in \mathbb{R}^{2\times 2}$ be given by*

$$A = \begin{pmatrix}2 & 0\\ 0 & 1\end{pmatrix}.$$

*Then, with probability $\frac{1}{2}$, the Adapted Power Method diverges when applied to the matrix $A$.*

**Algorithm 3** Adapted Power Method for Two-to-Infinity Norm from [21]

**Input:**
Oracle for matrix-vector multiplication with matrix $A \in \mathbb{R}^{d\times n}$,
Oracle for matrix-vector multiplication with matrix $A^T \in \mathbb{R}^{n\times d}$,
Positive integer $m \in \mathbb{N}$: number of iterations.
**Output:**
An estimate of the $\|A\|_{2\to\infty}$ norm.
1: Sample random vector $X^0 \in \mathbb{R}^n$ from $\mathcal{N}(0, I_n)$
2: **for** each $i = 1, 2, \ldots, m$ **do**
3: Compute $Y^i = \psi_\infty(AX^{i-1})$
4: Compute $X^i = \psi_2(A^\top Y^i)$
5: **end for**
6: Compute $L = (Y^m)^\top AX^m$
7: **return** $L$

*Explanation.* We follow the notation from Algorithm 3. Let $X' = AX^0$, where the components of the initial vector $X^0$ are independent. Then,

$$\mathbb{P}(X'_1 < X'_2) = \mathbb{P}(2X_1^0 < X_2^0) = \mathbb{P}(2X_1^0 - X_2^0 < 0) = \mathbb{P}(\mathcal{N}(0, 5) < 0) = \frac{1}{2}.$$

By the definition of $\psi_\infty$, we then have $Y^1 = (0, 1)^\top$ with probability $\frac{1}{2}$. It follows that

$$X^1 = \frac{A^\top Y^1}{\|A^\top Y^1\|_2} = (0, 1)^\top.$$

It is easy to verify that this condition is preserved in all subsequent iterations, i.e., $X^i = Y^i = (0, 1)^\top$ for all $i$. Consequently, the final estimate is

$$L = (Y^m)^\top AX^m = 1,$$

which is incorrect, as the two-to-infinity norm of $A$ is 2. □

### D.2 Hyperparameters for Image Classification

Each model is trained for 200 epochs using stochastic gradient descent (SGD) with Nesterov momentum of $0.9$ and weight decay of $5 \cdot 10^{-5}$. The initial learning rate is set to $0.1$, decayed by a factor of $0.1$ at epochs $60$, $120$, and $160$. We use a batch size of $128$ and apply the data augmentations listed below.

#### D.2.1 CIFAR-100

Table 2: Data augmentation used for CIFAR-100.

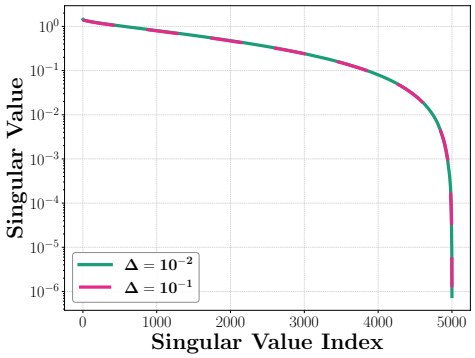
| **Transform** | **Parameters** |
|---|---|
| RandomHorizontalFlip | — |
| Pad | padding = 4, padding_mode = "symmetric" |
| RandomCrop | size = 32 |
| Normalize | mean = [0.5, 0.5, 0.5], std = [0.5, 0.5, 0.5] |

### D.2.2 TinyImageNet

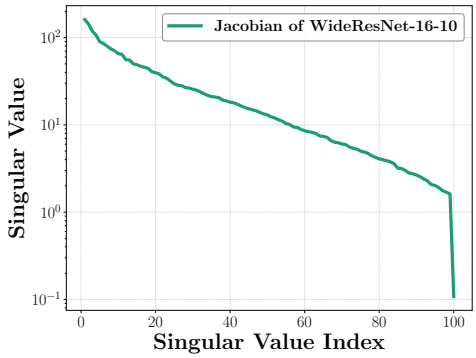
Table 3: Data augmentation used for TinyImageNet.

| Transform | Parameters |
|---|---|
| RandomHorizontalFlip | — |
| Pad | padding = 4, padding_mode = "symmetric" |
| RandomCrop | size = 64 |
| ColorJitter | brightness = 0.2, contrast = 0.2, saturation = 0.2, hue = 0.1 |
| Normalize | mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225] |

## D.3 Supplementary Figures

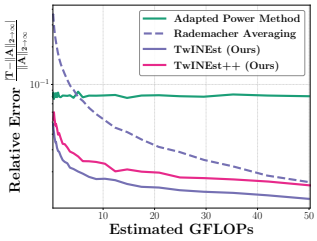



(a) Singular values of synthetic matrices for different $\Delta$.

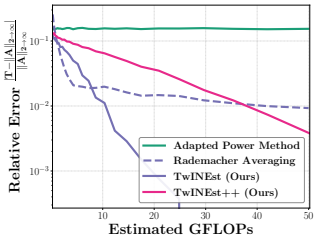



(b) Singular values of the Jacobian matrix of WideResNet-16-10.

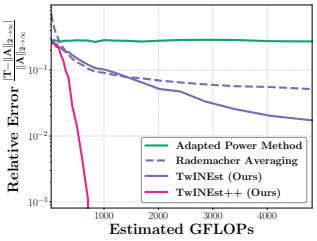
Figure 3: Singular values of synthetic and real world matrices.




(a) Synthetic data with $\Delta = 10^{-2}$.




(b) Synthetic data with $\Delta = 10^{-1}$.




(c) Jacobian matrix of pre-trained WideResNet-16-10.

Figure 4: Comparison of methods for estimating the two-to-infinity matrix norm. The plot shows the relative error versus GFLOPs, averaged over 500 trials. For the Jacobian matrix, matrix-vector multiplications were computed using JVP and VJP via autograd, whereas for synthetic data, explicit matrix-vector multiplications were used.