Asynchronous Randomized Trace Estimation

Vassilis Kalantzis **Shashanka Ubaru** **Chai Wah Wu** **Georgios Kollias** **Lior Horesh**
IBM Research, Thomas J. Watson Research Center, USA

Abstract

Randomized trace estimation is a popular technique to approximate the trace of an implicitly-defined matrix $\boldsymbol{A}$ by averaging the quadratic form $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ across several samples of a random vector $\boldsymbol{x}$. This paper focuses on the application of randomized trace estimators on asynchronous computing environments where the quadratic form $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ is computed partially by observing only a random row subset of $\boldsymbol{A}$ for each sample of the random vector $\boldsymbol{x}$. Our asynchronous framework treats the number of rows, as well as the row subset observed for each sample, as random variables, and our theoretical analysis establishes the variance of the randomized estimator for Rademacher and Gaussian samples. We also consider an extension where the entries of $\boldsymbol{A}$ are stochastically rounded. We also present error analysis and sampling complexity bounds for the proposed asynchronous randomized trace estimator. Our numerical experiments illustrate that the asynchronous variant can be competitive even when a small number of rows is updated per each sample.

1 Introduction

The problem of computing the trace of an implicitly-defined symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ appears in several applications in engineering, machine learning, and data analysis, e.g., see [38, 37, 36, 41, 23, 24]. Typically, the implicitly-defined matrix has the form $\boldsymbol{A} = f(\boldsymbol{G})$ where $\boldsymbol{G} \in \mathbb{R}^{N\times N}$ is an application-dependent symmetric matrix and $f(\cdot)$ is a certain (real or complex-valued) function. For example, some well-known choices for the function f in graph analytics and machine learning are $f(\beta) = \beta^3$ (triangle counting), $f(\beta) = e^\beta$ (Estrada index), $f(\beta) = \log(\beta)$ (log-determinant), $f(\beta) = \beta^{p/2}$ (Schatten-p norms), and $f(\beta) = \beta\log(\beta)$ (Von-Neumann Entropy).

Assuming that $\boldsymbol{A}$ is accessible only via a Matrix-Vector product routine (MV), i.e., each time we can only observe $\boldsymbol{A}\boldsymbol{v}$ for a specific (but arbitrary) vector $\boldsymbol{v} \in \mathbb{R}^N$, the simplest approach to compute the trace of $\boldsymbol{A}$, $\mathtt{Tr}(\boldsymbol{A})$, is to sum the N individual diagonal entries of $\boldsymbol{A}$ by evaluating N quadratic forms of $\boldsymbol{A}$ (from now on *quadratic forms*) $\boldsymbol{e}_j^\top \boldsymbol{A}\boldsymbol{e}_j$ where $\boldsymbol{e}_j$ denotes the jth column of the $N \times N$ identity matrix. Unfortunately, this requires N MV products with matrix $\boldsymbol{A}$ and quickly becomes impractical for anything but small matrix sizes. Instead, practical efficient algorithms to approximate $\mathtt{Tr}(\boldsymbol{A})$ exploit *randomization*. In a nutshell, randomized trace estimators aim for an approximation of the form $\mathtt{Tr}(\boldsymbol{A}) \approx \left[\sum_{k=1}^{k=M} \boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k\right]/M$, where the vectors $\boldsymbol{x}_k \in \mathbb{R}^N$ are typically samples of a random vector $\boldsymbol{x}$ such that the estimator is unbiased, i.e., $\mathtt{Tr}(\boldsymbol{A}) = \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}]$. For example, when each individual entry of the N-dimensional vectors $\{\boldsymbol{x}_k\}_{k=1}^{k=M}$ is equal to ± 1 with equal probability (i.e. the Rademacher distribution), the estimator is known as *Hutchinson's randomized trace estimator* and has minimum variance over the field of real random vectors [21]. Similarly, when $\{\boldsymbol{x}_k\}_{k=1}^{k=M}$ are independent and their entries are i.i.d standard normal variables, the randomized estimator is known as *Gaussian randomized trace estimator*. An extensive analysis of the Hutchinson and Gaussian trace estimators, including a probabilistic (ϵ, δ) convergence analysis, can be found in the seminal work [2]. Several enhancements and/or variance reduction techniques of randomized estimators can be found in [7, 29, 31, 37, 13, 22, 30].

Contributions. In this paper we consider the problem of randomized trace estimation when the quadratic forms $\boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k,\ k = 1, \ldots, M$, are computed inexactly in the sense that only a random subset of the N entries of the MV product $\boldsymbol{A}\boldsymbol{x}_k$ are observed. As specific contributions of this paper:

- We propose a computational framework to estimate the trace of an implicit matrix $\boldsymbol{A}$ under the constraint that the quadratic forms $\boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k$ are only partially

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

correct and equal to $\boldsymbol{x}_k\boldsymbol{Q}(\mathcal{T}_k)\boldsymbol{x}_k$ where a) the ith row of the matrix $\boldsymbol{Q}(\mathcal{T}_k)$ is identical to the ith row of $\boldsymbol{A}$ if $i \in \mathcal{T}_k \subseteq [N] = \{1, 2, \ldots, N\}$ and zero otherwise, and b) the subset $\mathcal{T}_k$ is an independent random subset of $[N]$ for any value of k. We will refer to this framework as *asynchronous* due to its similarity with the classical asynchronous framework presented in [4].

- We analyze the variance of the asynchronous randomized estimator and derive close form expressions when the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$, are sampled from Rademacher and Gaussian distributions.
- We present error analysis which establishes probabilistic (ϵ, δ)-error bounds with a sampling complexity bound which order-wise (with respect to error tolerance ϵ and probability parameter δ) matches the bounds of classical Hutchinson's trace estimator [2].
- We propose an extension of the asynchronous framework to computing environments equipped with *stochastic rounding* where a real number is approximated by randomly selecting neighboring quantization levels with probability proportional to the distance to the opposite quantization level.

Notation. We use lowercase bold letters to denote vectors and uppercase bold letters to denote matrices. Moreover, we use uppercase Greek letters to denote integers. We denote by $\boldsymbol{e}_i \in \{0, 1\}^N$ the ith column of the $N \times N$ identity matrix we will use the notation A_{ij} to denote the (i, j) entry of the matrix $\boldsymbol{A}$. For any $N \times N$ matrix $\boldsymbol{A}$, we will denote its Frobenius norm by $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{i=N} \sum_{j=1}^{j=N} A_{ij}^2}$. The diagonal matrix holding the N diagonal entries of the matrix $\boldsymbol{A}$ is denoted by $\mathcal{D}(\boldsymbol{A})$. We will use $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$ to denote the probability and expectation, respectively.

2 Related work

Randomized trace estimators. Monte Carlo randomized trace estimators where $\mathtt{Tr}(\boldsymbol{A})$ is approximately estimated as the average of the quadratic transformation $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ were first introduced by Girard in [17]. This approach was later extended and popularized by Hutchinson [21]. The first extensive theoretical error analysis for the method was presented by Avron and Toledo in [2], and the analysis was slightly improved in [31]. The randomized trace estimator has been adapted to solve various matrix computation problems (such as estimating log-determinant, numerical rank, Schatten p-norms, Estrada index, Von-Neumann entropy, spectral density, trace of matrix inverse, diagonal and other spectral approximations), which can all be posed as implicit trace estimation problems, e.g., see [19, 3, 12, 27, 20, 37, 36, 38, 8].

Given the wide range of applications for such implicit trace estimation methods, in recent years, several papers have appeared in the literature [16, 26, 29, 30, 6, 13], which have proposed various variance reduction techniques to improve the randomized trace estimator method. The general idea of these variance reduction techniques is to first compute a low rank approximation of $\boldsymbol{A}$ for which the trace can be computed exactly followed by an estimation of the trace of the residual using the randomised trace estimator. The randomized trace estimator and its variants have also been adapted for dynamic matrix trace estimation in [11, 40].

3 Trace estimates with asynchronous quadratic forms

Asynchronous computations arise naturally in distributed-memory implementations for the computation of stationary points via iterative algorithms in order to reduce idle time between different processing elements via reducing synchronization points. While asynchronous iterations typically lead to slower convergence, the ever-increasing gap between the time required to share a floating-point number between different processing elements and the time needed to perform a single floating-point operation by one of the processing elements, has led to a revived interest in the analysis and application of asynchronous algorithms in numerical linear algebra [39, 32, 33, 1, 18, 15, 4].

Traditionally, asynchronous algorithms are prevalent in the solution of systems of equations of the form $z = G(z),\ G : \mathbb{R}^N \to \mathbb{R}^N$ where the ith entry satisfies $[z]_i = g_i(z),\ i = 1, \ldots, N$. An asynchronous method for computing z can be then defined mathematically as

$$[z]_i^k = \begin{cases} [z]_i^k, \text{ if } i \notin \mathcal{T}_k \\ g_i\left([z]_1^{s_1(k)}, \ldots, [z]_N^{s_N(k)}\right), \text{ if } i \in \mathcal{T}_k \end{cases},$$

where $[z]_i^k$ denotes the ith component of the iterate at time instant k, $\mathcal{T}_k$ is the set of indices updated at instant k, and $s_j(k)$ is the last instant the jth component was updated before being read at instant k [15, 4].

In this work we study the problem of randomized trace estimation from an asynchronous viewpoint. The asynchronous setting considered in this paper does not involve a fixed-point iterative process since each quadratic form $\boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k$ can be computed independently. Instead, we consider the scenario where indices which are not part of the update subset $\mathcal{T}_k$ are simply replaced by zero. In the following we assume that the implicitly-defined matrix $\boldsymbol{A}$ is symmetric and positive-definite, however, with the exception of the high-probability analysis, our theoretical results stand for symmetric non-definite matrices as well.

Definition 1. *Let $\mathcal{T}$ denote a random subset of $T \in \mathbb{N}$ integers (without replacement) from the set*

$\{1, 2, \ldots, N\}$. *We define the asynchronous MV* $\boldsymbol{y} = \boldsymbol{A} \mid_{\mathcal{T}} \boldsymbol{x}$ *between the matrix* $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ *and a vector* $\boldsymbol{x} \in \mathbb{R}^N$ *as a function of* $\mathcal{T}$ *such that:*

$$[\boldsymbol{y}]_i = \begin{cases} [\boldsymbol{A}\boldsymbol{x}]_i & \text{if } i \in \mathcal{T} \\ 0 & \text{if } i \notin \mathcal{T} \end{cases}.$$

In other words, the operator $|_{\mathcal{T}}$ *is equivalent to performing the regular MV* $\boldsymbol{A}\boldsymbol{x}$ *as if the* i*-th row of* $\boldsymbol{A}$ *is replaced by a* N*-length zero row vector for each* $i \notin \mathcal{T}$*.*

Unless mentioned otherwise, throughout this paper we assume that the random integer variable T and the subset $\mathcal{T}$ are independent. Given T, the random subset of $\mathcal{T}$ picks any $T \equiv |\mathcal{T}|$ integers of $\{1, 2, \ldots, N\}$ with equal probability, i.e., any of the $\binom{N}{T}$ possible row sets of $\boldsymbol{A}$ is picked with probability $\binom{N}{T}^{-1}$.

The MV product $\boldsymbol{y} = \boldsymbol{A} \mid_{\mathcal{T}} \boldsymbol{x}$ presented in Definition 1 fits any application of numerical computing where only a random subset of the entries of the exact MV $\boldsymbol{A}\boldsymbol{x}$ are observed, leading to a quadratic form $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ which is only partially complete. One such application is the streaming model of linear algebra where the rows of the matrix $\boldsymbol{A}$ are sampled from some random distribution either due to limits in physical resources, e.g., system memory, or due a stochastic nature of the matrix process [9]. In the first case, only a random row subset of fixed cardinality of $\boldsymbol{A}$ can be accessed at any given time. In the second case, random row subsets of random cardinality become available, and these two sources of randomness are independent.

Another important practical scenario where the framework of Definition 1 is directly applicable is the problem of straggling in large-scale distributed controller-worker architectures with a large number of (heterogeneous) processing elements. Since each processing element might have different computation and network bandwidth, which also vary over time, performing a MV product in parallel can be bottlenecked by unpredictably slow or unresponsive workers which are commonly referred to as stragglers [28, 5]. Even a tiny probability of a processing element becoming a straggler per sample can cause a big increase in the wall-clock time of the trace estimation of the matrix $\boldsymbol{A}$. The analysis presented in this paper gives a very simple solution to the problem of straggling without using any additional hardware. The only requirement is to simply allocate a fixed amount of time per MV product and only utilize the entries of the MV product associated with indices whose computations were completed during the given time-frame. This leads to load balancing and robust computations.

Definition 2. *Let* $k = 1, 2, \ldots, M$, $M \in \mathbb{N}$*, and denote by* $\mathcal{T}_k$ *a random subset of* $|\mathcal{T}_k| \in \mathbb{N}$ *integers (without replacement) from* 1 *to* N*. The deterministic integer* $|\mathcal{T}_k|$ *is an instance of the integer-valued random variable* $T \in \{1, 2 \ldots, N\}$*. Then, for any* N*-length instances* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$*, of a random vector* $\boldsymbol{x}$*, we define the asynchronous randomized trace estimator*

$$\begin{aligned} \Gamma_M &= \frac{1}{M} \sum_{k=1}^{k=M} \boldsymbol{x}_k^\top (\boldsymbol{A} \mid_{\mathcal{T}_k} \boldsymbol{x}_k) \\ &= \frac{1}{M} \sum_{k=1}^{k=M} \sum_{i\in\mathcal{T}_k} [\boldsymbol{x}_k]_i^\top [\boldsymbol{A} \mid_{\mathcal{T}_k} \boldsymbol{x}_k]_i. \end{aligned}$$

The second equality of Γ_M *follows by recalling that the* i*-th entry of the product* $\boldsymbol{A}|_{\mathcal{T}_k}\boldsymbol{x}_k$ *is nonzero if and only if* $i \in \mathcal{T}_k$*.*

Throughout the rest of the paper we assume that the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$, are instances of a random vector $\boldsymbol{x} \in \mathbb{R}^N$ sampled from a distribution such that $\mathbb{E}[\boldsymbol{x}] = 0$ and all N dimensions are statistically independent and have variance equal to one, i.e., $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = I$.

Consider now the diagonal random matrix formed by the summation of T canonical outer products

$$\boldsymbol{D}_{\mathcal{T}} = \sum_{i\in\mathcal{T}} \boldsymbol{e}_i \boldsymbol{e}_i^\top,$$

where both the cardinality T and the row subset $\mathcal{T}$ are random variables. When $T \equiv N$, as in the synchronous case, the matrix $\boldsymbol{D}_{\mathcal{T}}$ is equal to the $N \times N$ identity matrix. The asynchronous randomized trace estimator can be then written equivalently as

$$\begin{aligned} \Gamma_M &= \frac{1}{M} \sum_{k=1}^{k=M} \boldsymbol{x}_k^\top \boldsymbol{D}_{\mathcal{T}_k} \boldsymbol{A}\boldsymbol{x}_k \\ &= \frac{1}{M} \sum_{k=1}^{k=M} \boldsymbol{x}_k^\top \boldsymbol{Q}(\mathcal{T}_k)\boldsymbol{x}_k, \end{aligned} \tag{1}$$

where $\boldsymbol{Q}(\mathcal{T}_k) = \boldsymbol{D}_{\mathcal{T}_k}\boldsymbol{A}$ and $\boldsymbol{D}_{\mathcal{T}_k}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}|_{\mathcal{T}_k}\boldsymbol{x}$.

4 Analysis of the asynchronous randomized trace estimator

Lemma 1. *Let* $\boldsymbol{Q}$ *denote a random matrix and* $\boldsymbol{x}$ *denote an independent random vector of the same length as* $\boldsymbol{Q}$ *such that* $\mathbb{E}[\boldsymbol{x}] = 0$ *and* $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}$*. Then,*

$$\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}] = \mathtt{Tr}(\mathbb{E}[\boldsymbol{Q}]).$$

Lemma 1 states that we can apply randomized trace estimation to approximate the trace of a random matrix $\boldsymbol{Q}$ in a similar fashion as for a deterministic matrix $\boldsymbol{A}$. If the sample space of the random matrix $\boldsymbol{Q}$ is formed by all possible matrices $\boldsymbol{Q}(\mathcal{T}) = \boldsymbol{D}_{\mathcal{T}}\boldsymbol{A}$ such

that, for a given sample integer value of a uniform T in the interval $[1, N]$, the random subset of $\mathcal{T}$ picks any $T \equiv |\mathcal{T}|$ integers of $\{1, 2, \ldots, N\}$ with equal probability, then Γ_M (1) is an unbiased estimator of $\mathtt{Tr}(\mathbb{E}[\boldsymbol{Q}])$.

The main question now becomes whether we can exploit Γ_M to approximate the trace of the matrix $\boldsymbol{A}$. As we show in the following Proposition, the answer is affirmative.

Proposition 1. *Let $\mu_T = \mathbb{E}[T]$ denote the expectation of the random variable T. Then,*

$$\mathbb{E}[\boldsymbol{Q}] = \frac{\mu_T}{N}\boldsymbol{A}, \quad \text{and} \quad \mathbb{E}\left[\Gamma_M\right] = \frac{\mu_T}{N}\mathtt{Tr}(\boldsymbol{A}),$$

i.e., the randomized estimator $\dfrac{N}{\mu_T}\Gamma_M$ is an unbiased estimator of $\mathtt{Tr}(\boldsymbol{A})$.

Proposition 1 tells us that the asynchronous randomized estimator Γ_M is an unbiased estimator of $\mathtt{Tr}(\boldsymbol{A})$ up to multiplication with the factor μ_T/N. Note here that when $T \equiv N$, we have $\mu_T = N$, $\mathcal{T} = \{1, 2, \ldots, N\}$ and $\boldsymbol{Q}(\mathcal{T}) = \boldsymbol{A}$, i.e., Γ_M becomes a synchronous randomized trace estimator. Next, we consider the variance $\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x})$ of a single sample of the asynchronous randomized trace estimator Γ_M.

Theorem 1. *Let σ_T^2 denote the variance of the random variable T, and define the scalars*

$$K_1 = \frac{(\sigma_T^2 + \frac{1}{N}\mu_T^2 - \mu_T)}{N(N-1)}, \qquad K_2 = \frac{(N\mu_T - \sigma_T^2 - \mu_T^2)}{N(N-1)},$$
$$\text{and} \qquad K_3 = \frac{\left((N-2)\mu_T + \sigma_T^2 + \mu_T^2\right)}{N(N-1)}.$$

The variance of a single sample of the asynchronous randomized trace estimator Γ_M is equal to

$$\frac{2\mu_T}{N}\|\boldsymbol{A}\|_F^2 + K_1\mathtt{Tr}\left(\boldsymbol{A}\right)^2 + K_2\mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right),$$

when $\boldsymbol{x} \in \mathcal{N}(0, \boldsymbol{I})$, and equal to

$$\frac{2\mu_T}{N}\|\boldsymbol{A}\|_F^2 + K_1\mathtt{Tr}\left(\boldsymbol{A}\right)^2 - K_3\mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right),$$

when $\boldsymbol{x}$ is a Rademacher random vector.

Theorem 1 tells us that the variance of the estimator Γ_m depends on scalar multiples of the three following terms involving the matrix $\boldsymbol{A}$: $\|\boldsymbol{A}\|_F^2$, $\mathtt{Tr}\left(\boldsymbol{A}\right)^2$, and $\mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right)$. Notice that when $T \equiv N$ we have $\sigma_T^2 = 0$ and $\mu_T = N$. Plugging these values in Theorem 1 gives us $K_1 = K_2 = 0$, $K_3 = 2$, and $\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}) = 2\|\boldsymbol{A}\|_F^2$ when $\boldsymbol{x} \in \mathcal{N}(0, \boldsymbol{I})$, and $\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}) = 2(\|\boldsymbol{A}\|_F^2 - \sum_{i=1}^N A_{ii}^2)$ when $\boldsymbol{x}$ is a Rademacher, which are identical to those in the synchronous case [2].

In the general asynchronous case, T can be less than N, and we can distinguish three important cases:

1. T is a fixed integer (deterministic) in the range $1 \leq T \leq N$,
2. T takes an integer values in $[1, N]$ with equal probability, and
3. $\mathcal{T}$ is obtained by choosing each element in $[1, N]$ with probability p. Note that for a fixed T, each subset $\mathcal{T}$ such that $T \equiv |\mathcal{T}|$ occurs with the same probability.

The variance of the trace estimator Γ_M for these three cases is shown in the following three corollaries.

Corollary 1. *Let T take some fixed integer value between 1 and N. The variance of the asynchronous randomized trace estimator Γ_M is then equal to*

$$\frac{T}{N}\left[2\|\boldsymbol{A}\|_F^2 + \frac{N-T}{N-1}\left(\mathrm{Tr}(\mathcal{D}(\boldsymbol{A})^2) - \frac{1}{N}\mathrm{Tr}(\boldsymbol{A})^2\right)\right],$$

when $\boldsymbol{x} \in \mathcal{N}(0, \boldsymbol{I})$, and is equal to

$$\frac{T}{N}\left[2\|\boldsymbol{A}\|_F^2 - \frac{T+N-2}{N-1}\mathrm{Tr}(\mathcal{D}(\boldsymbol{A})^2) - \frac{N-T}{N^2-N}\mathrm{Tr}(\boldsymbol{A})^2\right]$$

when $\boldsymbol{x}$ is a Rademacher random vector.

The variance reported in Corollary 1 concerns the biased estimator Γ_M. In practice we exploit the unbiased estimator $\dfrac{N}{\mu_T}\Gamma_M$ for which $\dfrac{N}{\mu_T}\mathbb{E}\left[\Gamma_M\right] \equiv \mathtt{Tr}(\boldsymbol{A})$. The variances for the unbiased estimator are listed in the following remark. Note that now the variance no longer approaches zero as N increases.

Remark 1. *The variance of the unbiased estimator $\dfrac{N}{\mu_T}\Gamma_M$ is equal to*

$$\frac{N}{T}\left[2\|\boldsymbol{A}\|_F^2 + \frac{N-T}{N-1}\left(\mathrm{Tr}(\mathcal{D}(\boldsymbol{A})^2) - \frac{1}{N}\mathrm{Tr}(\boldsymbol{A})^2\right)\right],$$

when $\boldsymbol{x} \in \mathcal{N}(0, \boldsymbol{I})$, and

$$\frac{N}{T}\left[2\|\boldsymbol{A}\|_F^2 - \frac{T+N-2}{N-1}\mathrm{Tr}(\mathcal{D}(\boldsymbol{A})^2) - \frac{N-T}{N^2-N}\mathrm{Tr}(\boldsymbol{A})^2\right]$$

when $\boldsymbol{x}$ is a Rademacher random vector.

Figure 1 plots the magnitude of the variance coefficients associated with the terms $\mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right)$, $\|\boldsymbol{A}\|_F^2$, and $\mathtt{Tr}\left(\boldsymbol{A}\right)^2$, for the unbiased Gaussian estimator shown in Remark 1 where for simplicity we pick $N = 200$. As expected, when $T = N$, the coefficients of the terms $\mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right)$ and $\mathtt{Tr}\left(\boldsymbol{A}\right)^2$ become zero, and the variance of a single sample becomes $2\|\boldsymbol{A}\|_F^2$ as in the synchronous case. On the other hand, when $T = 1$, the variance of the unbiased estimator is approximately equal to $N(\|\boldsymbol{A}\|_F^2 + \mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right))$ and can be quite high for large-scale problems. Nonetheless, as $T \to N$, the variance coefficients of the terms $\mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right)$ and $\mathtt{Tr}\left(\boldsymbol{A}\right)^2$ approach zero. In particular, the variance contribution of the terms $\mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right)$ and $\mathtt{Tr}\left(\boldsymbol{A}\right)^2$ plateaus

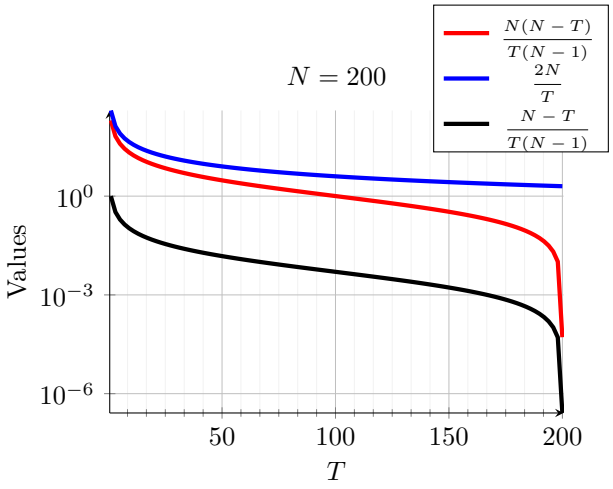

Figure 1: Magnitude of the coefficients multiplying the terms $\mathtt{Tr}\left(\mathcal{D}(\boldsymbol{A})^2\right)$, $\|\boldsymbol{A}\|_F^2$, and $\mathtt{Tr}\left(\boldsymbol{A}\right)^2$, for the unbiased Gaussian estimator (Remark 1).

around $T \approx \dfrac{N}{2}$ but varies faster around the endpoints 1 and N.

The following corollaries list the variance of a single sample of the (biased) estimator Γ_M for the cases where T either takes any integer value in $[1, N]$ with equal probability (Corollary 2) or the row subset $\mathcal{T}$ is obtained by choosing each element in $[1, N]$ with some fixed probability p (Corollary 3).

Corollary 2. *Let the random variable T be any of the values $1, 2, \ldots, N$, with equal probability, i.e., $\mathbb{P}[T = t] = p_t = 1/N$, $t = 1, 2, \ldots, N$. The variance of the asynchronous randomized trace estimator Γ_M is then equal to*

$$\frac{N+1}{N}\left[\|\boldsymbol{A}\|_F^2 + \frac{1}{12}\left((1-\frac{3}{N})\operatorname{Tr}(\boldsymbol{A})^2 + 2\operatorname{Tr}(\mathcal{D}(\boldsymbol{A})^2)\right)\right]$$

when $\boldsymbol{x} \in \mathcal{N}(0, \boldsymbol{I})$, and is equal to

$$\frac{N+1}{N}\left[\|\boldsymbol{A}\|_F^2 + \frac{1}{12}\left((1-\frac{3}{N})\operatorname{Tr}(\boldsymbol{A})^2 - 10\operatorname{Tr}(\mathcal{D}(\boldsymbol{A})^2)\right)\right]$$

when $\boldsymbol{x}$ is a Rademacher random vector.

Proof. Follows from Theorem 1 and the fact that $\mu_T = \frac{N+1}{2}$ and $\operatorname{Var}(T) = \frac{N^2-1}{12}$. □

Corollary 3. *For $0 < p \leq 1$, let $\mathcal{T}$ be random subsets of $\{1, \cdots, N\}$ with probability equal to $p^{|\mathcal{T}|}(1-p)^{N-|\mathcal{T}|}$, i.e. each element in $\{1, \cdots, N\}$ is chosen independently with probability p, then the variance of the asynchronous randomized trace estimator Γ_M is then equal to*

$$2p\|\boldsymbol{A}\|_F^2 + p(1-p)\operatorname{Tr}(\mathcal{D}(\boldsymbol{A})^2),$$

when $\boldsymbol{x} \in \mathcal{N}(0, \boldsymbol{I})$, and is equal to

$$2p\|\boldsymbol{A}\|_F^2 - p(1+p)\operatorname{Tr}(\mathcal{D}(\boldsymbol{A})^2),$$

when $\boldsymbol{x}$ is a Rademacher random vector.

Proof. Follows by noticing that $\mu_T = Np$ and $\operatorname{Var}(T) = Np(1-p)$. □

To compare the sampling schemes in Corollary 2 and Corollary 3, let $p = \frac{N+1}{2N}$, i.e., the sampling rate μ_T is the same for both schemes. Then, for Corollary 3, $\operatorname{Var}(T) = \frac{2N^2-1}{4N}$ which grows on the order of N, in contrast to $\operatorname{Var}(T)$ for Corollary 2 which grows on the order of N^2. We also note that the variance of the mean in the (unbiased) estimator $\frac{N}{\mu_T}\Gamma_M$ is $\frac{N\operatorname{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x})}{M\mu_T}$, i.e., the standard error of the mean is $\frac{\sqrt{N}\sigma(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x})}{\sqrt{M\mu_T}}$.

4.1 Sampling complexity

Finally, when $\boldsymbol{A}$ is symmetric and semi-definite, we can establish relative error bounds for the asynchronous trace estimator, i.e., (ϵ, δ)-approximation error bounds. These results yield us *a*) lower bounds on the number random samples required to achieve a desired $\epsilon \in \mathbb{R}$ error guarantee; *b*) a convergence rate for the trace estimation with respect to sampling complexity; and *c*) a computational complexity of the algorithm to achieve a desired ϵ error. We then have the following result.

Theorem 2. *Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be a symmetric positive semi-definite matrix, and $\delta \in (0, \frac{1}{2}]$, $\epsilon \in (0, 1]$. The asynchronous randomized trace estimator $\frac{1}{p}\Gamma_M$ with row selection probability p, is an (ϵ, δ)-approximator of* $\operatorname{Tr}(\boldsymbol{A})$,

$$\Pr\left(\left|\frac{1}{p}\Gamma_M - \operatorname{Tr}(\boldsymbol{A})\right| \leq \epsilon \operatorname{Tr}(\boldsymbol{A})\right) \geq 1 - \delta,$$

for sampling complexity with a fixed constant C, which only depends on the sub-Gaussianity of the random vectors $\boldsymbol{x}$:

- $M > \dfrac{C\log(1/\delta)}{p\epsilon^2}$, *for the case where rows are chosen with probability p (i.e., Corollary 3), and*
- $M > \dfrac{CN\log(1/\delta)}{\mu_T\epsilon^2}$, *for the i.i.d. observation variable T with mean μ_T, since $p = \frac{\mu_T}{N}$.*

The proof of Theorem 2 (the details of which are in the supplement) depends on a sparse variant of the Hanson-Wright inequality, similar to the ones in [42].

5 Extension to stochastic rounding

We can consider the asynchronous setting as a case of approximate and inaccurate computing where only a random approximation $\boldsymbol{Q}$ of the matrix $\boldsymbol{A}$ is used each time, but requiring that the expectation of the

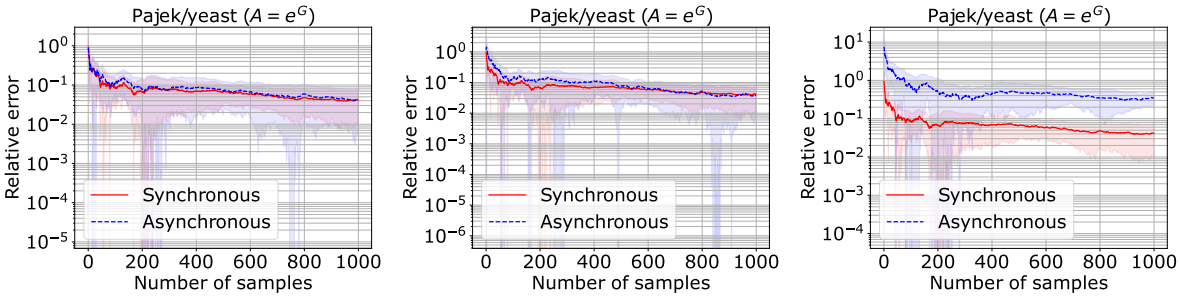

Figure 2: $A = e^G$. *Left to right: fixed* $T = \lceil Np \rceil$*, uniform* T*, fixed* p*.*

random variable is proportional to $\boldsymbol{A}$ (as expressed by $\mathbb{E}[\boldsymbol{Q}] = \frac{\mu_T}{N}\boldsymbol{A}$). Another important method of random approximation is *stochastic rounding* [10], where a real number is approximated by neighboring quantization levels with probability proportional to the distance to the opposite quantization level. More precisely, if $q_1 \le x \le q_2$ lies between quantization levels q_1 and q_2, the stochastic rounding of x is defined as $\mathrm{sr}(x) = q_1$ with probability $\frac{q_2 - x}{q_2 - q_1}$ and $\mathrm{sr}(x) = q_2$ otherwise. We define $\mathrm{sr}(x)$ when x is a vector or matrix by applying stochastic rounding to each element of x independently. It is easy to see that $\mathbb{E}[\mathrm{sr}(x)] = x$ and $\mathrm{Var}(\mathrm{sr}(x)) = x(q_1 + q_2 - x) - q_1 q_2$. Let us denote $r(x) = x - q_1$ and $\Delta(x) = q_2 - q_1$ in which case we can write $\mathrm{Var}(\mathrm{sr}(x)) = r(x)(\Delta(x) - r(x))$, and $\mathbb{E}[\mathrm{sr}(x)^2] = q_2 r(x) + x q_1$. In the sequel, we will assume that Δ does not depend on x, i.e. all the quantization levels are equally spaced. Let $\widetilde{\boldsymbol{A}}$ be the random matrix where each entry $\widetilde{A}_{ij} = \mathrm{sr}(A_{ij})$ independently. Then $\mathbb{E}[\widetilde{\boldsymbol{A}}] = \boldsymbol{A}$ and $\boldsymbol{Q}(\mathcal{T}_k) = \boldsymbol{D}_{\mathcal{T}_k}\widetilde{\boldsymbol{A}}$ with $\mathbb{E}[\boldsymbol{Q}(\mathcal{T}_k)] = \mathbb{E}[\boldsymbol{D}_{\mathcal{T}_k}]\boldsymbol{A}$ as before. Similarly, $\mathbb{E}[\mathtt{Tr}(\widetilde{\boldsymbol{A}})] = \mathtt{Tr}(\boldsymbol{A})$, $\mathrm{Var}(\mathtt{Tr}(\widetilde{\boldsymbol{A}})) = \sum_i r(A_{ii})(\Delta - r(A_{ii}))$ and $\mathbb{E}[\mathtt{Tr}(\widetilde{\boldsymbol{A}})^2] = \mathrm{Var}(\mathtt{Tr}(\widetilde{\boldsymbol{A}})) + \mathbb{E}[\mathrm{diag}(\widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{A}}]$.

Definition 3. *Let* $\mathcal{T}$ *denote a random subset of* $T \in \mathbb{N}$ *integers (without replacement) from the set* $\{1, 2, \ldots, N\}$*. We define the stochastically rounded asynchronous matrix-vector product (SRAMVP)* $\boldsymbol{y} = \boldsymbol{A}\,|_{\mathcal{T}}\,\boldsymbol{x}$ *between* $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ *and a vector* $\boldsymbol{x} \in \mathbb{R}^N$ *as a function of* $\mathcal{T}$ *such that:*

$$[\boldsymbol{y}]_i = \begin{cases} \left[\widetilde{\boldsymbol{A}}\boldsymbol{x}\right]_i & \text{if } i \in \mathcal{T} \\ 0 & \text{if } i \notin \mathcal{T}. \end{cases}$$

In other words, the operator $|_{\mathcal{T}}$ *is equivalent to the regular matrix-vector multiplication* $\boldsymbol{A}\boldsymbol{x}$ *with the difference that the matrix entries are replaced with a stochastic rounding representation and the* i*th row of* $\boldsymbol{A}$ *is replaced by an* N*-length zero row vector unless* $i \in \mathcal{T}$*. We assume that the stochastic rounding is independent from the random subset* $\mathcal{T}$*.*

As for the random vectors $\boldsymbol{x}$, note that by symmetry the Rademacher vectors can be considered a stochastic rounding of Gaussian vectors with two quantization levels when the stochastic rounding is independent from the Gaussian random variable. More generally, we replace $\boldsymbol{x}$ with $\mathrm{sr}(\boldsymbol{x})$ and obtain

$$\widetilde{\Gamma}_M = \frac{1}{M}\sum_{k=1}^{k=M} \mathrm{sr}(\boldsymbol{x}_k)^\top \boldsymbol{Q}(\mathcal{T}_k)\mathrm{sr}(\boldsymbol{x}_k). \tag{2}$$

Assuming the quantization levels are symmetric around 0, then for $\boldsymbol{x}$ symmetric around 0 (e.g., Gaussian) we have $\mathbb{E}\left[\mathrm{sr}(\boldsymbol{x})\mathrm{sr}(\boldsymbol{x})^\top\right] \propto \boldsymbol{I}$ and Eq. (2) after scaling is an unbiased estimator of $\mathtt{Tr}(\boldsymbol{A})$.

Table 1: *Matrices used in this section. The variables* N*,* $\mathtt{nnz}(\boldsymbol{A})$*, and* $\mathtt{Tr}(\boldsymbol{A})$ *denote the size, number of non-zero entries, and trace of matrix* $\boldsymbol{A}$*, respectively.*

Id	Matrix name	N	$\mathtt{nnz}(\boldsymbol{A})$	$\mathtt{Tr}(\boldsymbol{A})$
1	`Pajek/yeast`	2361	13828	536
2	`SNAP/ca-HepTh`	9877	51971	25
3	`Botonakis/thermomech_TC`	102158	711558	585.871
4	`SNAP/web-Stanford`	281903	2312497	0
5	`LAW/cnr-2000`	325557	3216152	87442

Theorem 3. *The variance of the stochastically rounded asynchronous randomized trace estimator* $\widetilde{\Gamma}_M$ *is equal to*

$$\frac{2\mu_T}{N}\mathbb{E}\left[\|\widetilde{\boldsymbol{A}}\|_F^2\right] + K_1\mathbb{E}\left[\mathtt{Tr}(\widetilde{\boldsymbol{A}})^2\right] + K_2\mathtt{Tr}\left(\mathbb{E}\left[\mathcal{D}(\widetilde{\boldsymbol{A}})^2\right]\right),$$

when $\boldsymbol{x} \in \mathcal{N}(0, \boldsymbol{I})$*, and equal to*

$$\frac{2\mu_T}{N}\mathbb{E}\left[\|\widetilde{\boldsymbol{A}}\|_F^2\right] + K_1\mathbb{E}\left[\mathtt{Tr}(\widetilde{\boldsymbol{A}})^2\right] - K_3\mathtt{Tr}\left(\mathbb{E}\left[\mathcal{D}(\widetilde{\boldsymbol{A}})^2\right]\right),$$

when $\boldsymbol{x}$ *is a Rademacher random vector, where* K_1*,* K_2*, and* K_3 *are defined in Theorem 1.*

6 Numerical experiments

In this section we illustrate the numerical performance of the asynchronous randomized trace estimator applied

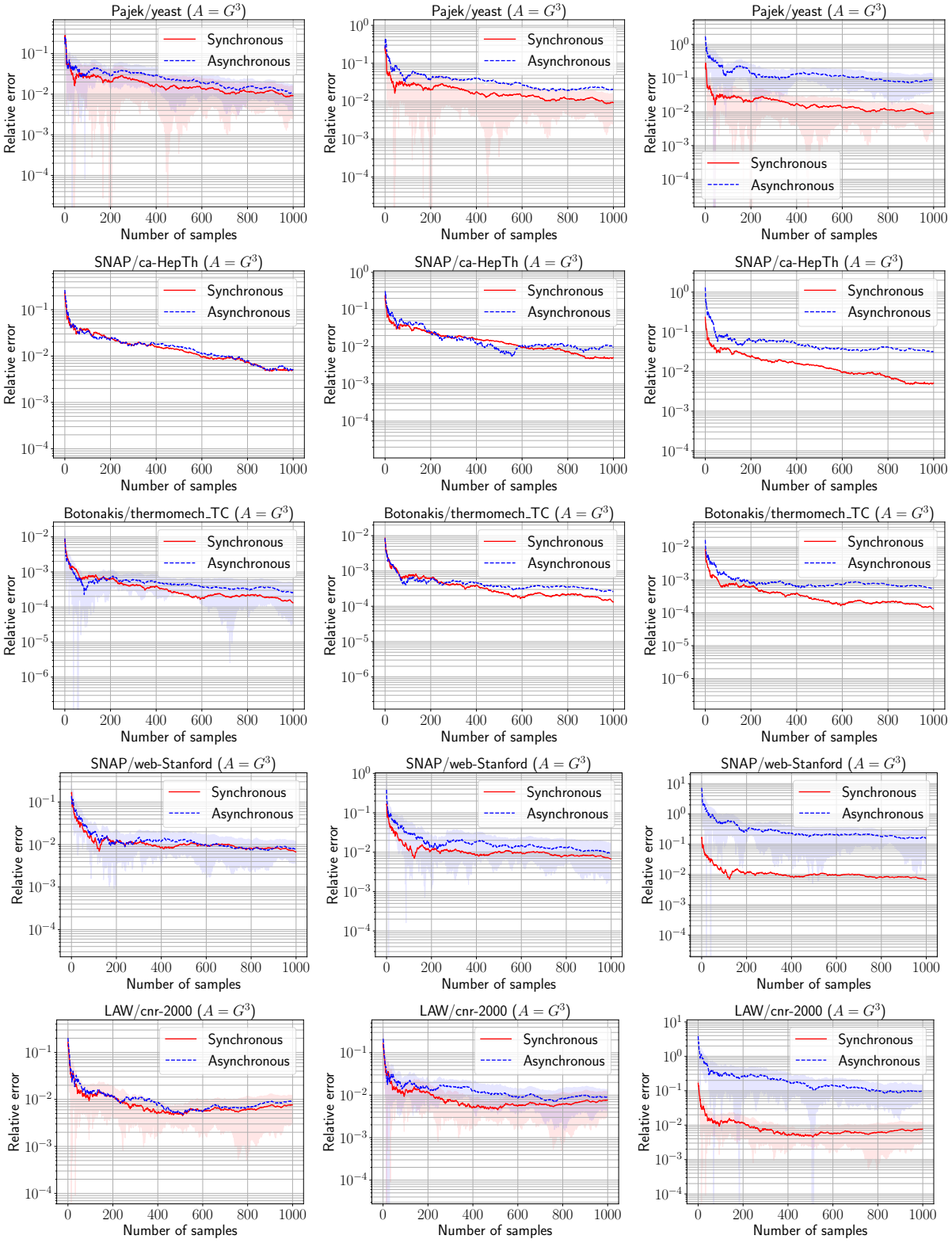

Figure 3: $A = G^3$. *Left to right: fixed* $T = \lceil Np \rceil$, *uniform* T, *fixed* p.

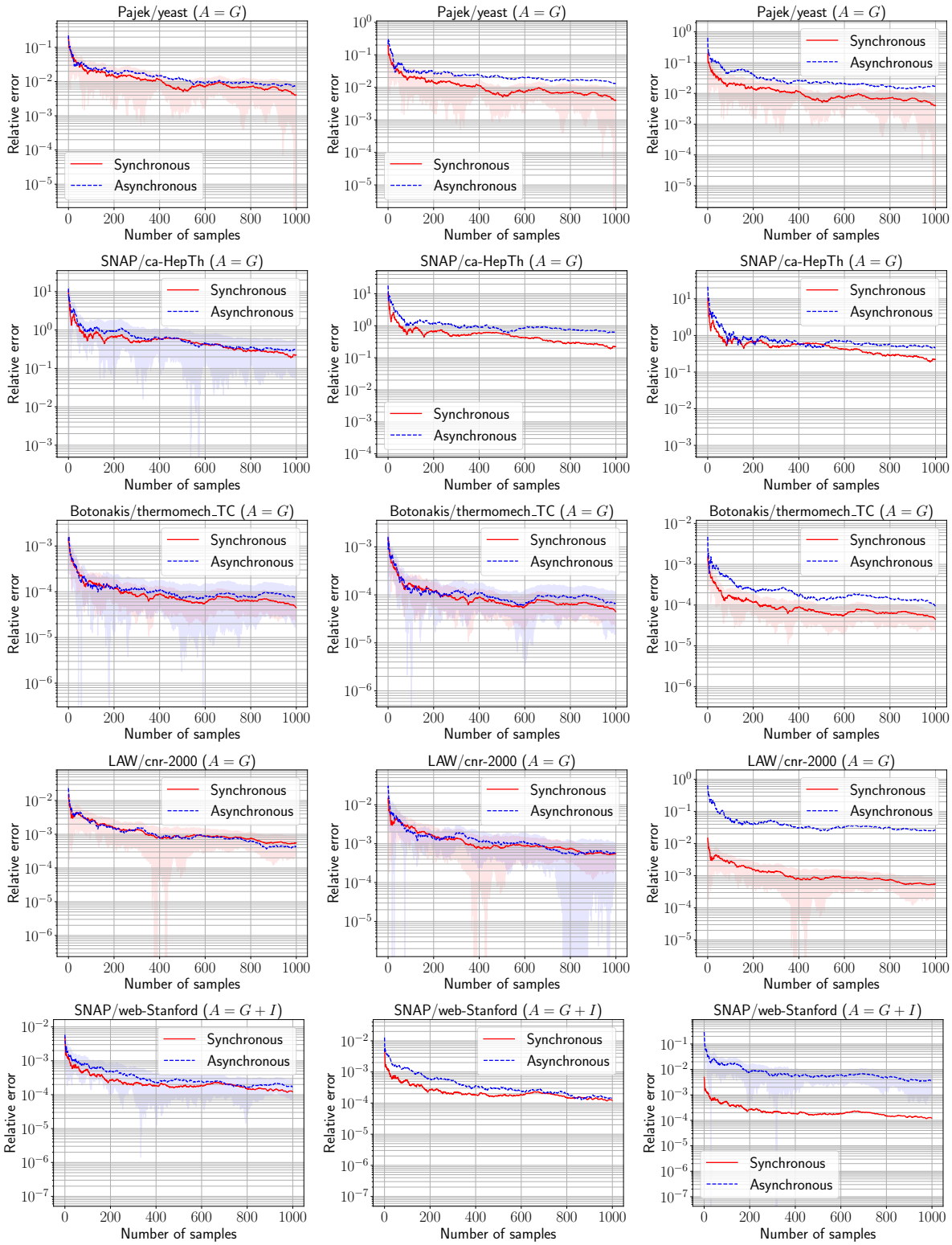

Figure 4: *Left to right: fixed $T = \lceil Np \rceil$, uniform T, fixed p.*

to the collection of sparse matrices listed in Table 1. Additional matrices and numerical results, including experiments with stochastic rounding, can be found in our supplement.

While our framework targets implicitly-defined matrices, for simplicity in our experiments we form each matrix explicitly and sample T and the corresponding row subsets $\mathcal{T}_k$ using a random generator. As our accuracy metric we use the relative error achieved by the estimator, defined as $|\mathtt{Tr}_M(\boldsymbol{A}) - \mathtt{Tr}(\boldsymbol{A})|/|\mathtt{Tr}(\boldsymbol{A})|$ for any approximation $\mathtt{Tr}_M(\boldsymbol{A})$ of the trace $\mathtt{Tr}(\boldsymbol{A})$. If the trace of a matrix is zero (e.g., `SNAP/web-Stanford`) we pre-process it by shifting with the identity matrix so as to avoid division by zero. Our experiments are conducted in Python 3.9 with NumPy and SciPy libraries for matrix computations in 64-bit arithmetic on a system equipped with an Apple M1 Max chip and 64 GB of LPDDR5 memory. Due to space limitations, the results shown in this section assume $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ are sampled from the Rademacher distribution.

Our experiments evaluate three separate scenarios for T as described in Corollaries 1, 2, and 3, respectively: *a*) a fixed value of T in the interval $[1, 2, \ldots, N]$, *b*) random T sampled uniformly from the set $\{1, 2, \ldots, N\}$, and *c*) a fixed probability p of choosing an element from $\{1, 2, \ldots, N\}$ in the subset $\mathcal{T}$. For "*a*)" we set $T = \lceil Np \rceil$ where $p = 0.6$; this value of p is also used throughout "*c*)". This implies that options "*a*)" and "*c*)" exploit about 60% of the rows of A when forming the quadratic forms $\boldsymbol{x}_k^\top \boldsymbol{A}\boldsymbol{x}_k$. Comparisons using various values of p are deferred to our supplement. For each test matrix and scenario we use $M = 1000$ Rademacher samples and perform each run ten times using a different random seed. We then accumulate all results and compute the mean and standard deviation of the relative error. The performance of the asynchronous (dashed lines) and synchronous (solid lines) trace estimators is demonstrated in Figure 4, where the shaded areas correspond to the standard deviation of each sample. In summary, the accuracy of the asynchronous randomized trace estimator is generally inferior to that of its its synchronous counterpart. This behavior is generally expected due to the higher variance of the asynchronous estimator. Nonetheless, in practice the accuracy achieved by the asynchronous estimator can be very close to that of the synchronous estimator even when only half of the rows are retained for each Rademacher sample.

Our last set of experiments considers the application of trace estimation in two important graph analytics tasks. The first task is that of counting the number of triangles of a graph $\mathcal{G}$, an important summarization feature in the analysis of patterns in networks [34, 35, 25]. This quantity is given by $\mathrm{Tr}(\boldsymbol{A} \equiv \boldsymbol{G}^3/3!)$ where $\boldsymbol{G}$ is the adjacency matrix of $\mathcal{G}$. Our results for the triangle counting problem are listed in Figure 3. The second task considers the determination of the Estrada index, a topological index of protein folding suggested by Ernesto Estrada as a measure of the degree of folding of a protein [14]. This quantity is given by $\mathrm{Tr}(\boldsymbol{A} \equiv e^{\boldsymbol{G}})$ where $\boldsymbol{G}$ represents the adjacency matrix of the protein network. Due to space limitations, we only plot the performance of the estimators for the `Pajek/yeast` matrix in Figure 2, deferring additional results to our supplement.

7 Conclusion

This paper considered the problem of randomized trace estimation following an asynchronous setting under which quadratic forms $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ are computed partially and is equivalent to observing only a random row subset of the matrix $\boldsymbol{A}$. Our theoretical results indicate that, up to scaling, the asynchronous randomized trace estimator is an unbiased estimator of $\mathrm{Tr}(\boldsymbol{A})$. Both Gaussian and Rademacher vector sampling was discussed, while extensions to environments with stochastic rounding were analyzed. Our numerical experiments, including problems from graph analytics, suggest that asynchronous randomized estimators generally exhibit higher variance but can achieve an accuracy that is on par with that in the synchronous case.

Several possible directions are left as future work. For example, while our analysis assumed that each row subset is picked with equal probability, in practice it might be beneficial to pick rows with higher norm more often, which is akin to an importance sampling scheme. Another important direction is to extend the asynchronous estimator to compute the main diagonal of an implicit matrix, i.e., compute all N individual diagonal entries of $\boldsymbol{A}$. This problem is of great interest in several applications in physics and statistics. Finally, a limitation that needs to be overcome in order to make asynchronous estimators practical and apply them in real production codes is the efficient formation of the asynchronous quadratic $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions.

Reproducibility

Our computer implementation, datasets, and interactive notebooks are available at `https://github.com/gidiko/async-trace`.

References

[1] H. Avron, A. Druinsky, and A. Gupta. Revisiting asynchronous linear solvers: Provable convergence rate through randomization. *Journal of the ACM (JACM)*, 62(6):1–27, 2015.

[2] H. Avron and S. Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011.

[3] Z. Bai, G. Fahey, and G. Golub. Some large-scale matrix computation problems. *Journal of Computational and Applied Mathematics*, 74(1):71–89, 1996.

[4] D. Bertsekas and J. Tsitsiklis. *Parallel and distributed computation: numerical methods*. Athena Scientific, 2015.

[5] R. Bitar, Y. Xing, Y. Keshtkarjahromi, V. Dasari, S. El Rouayheb, and H. Seferoglu. Private and rateless adaptive coded matrix-vector multiplication. *EURASIP Journal on Wireless Communications and Networking*, 2021:1–25, 2021.

[6] T. Chen and E. Hallman. Krylov-aware stochastic trace estimation. *arXiv preprint arXiv:2205.01736*, 2022.

[7] T. Chen, T. Trogdon, and S. Ubaru. Analysis of stochastic lanczos quadrature for spectrum approximation. In *International Conference on Machine Learning*, pages 1728–1739. PMLR, 2021.

[8] T. Chen, T. Trogdon, and S. Ubaru. Randomized matrix-free quadrature for spectrum and spectral sum approximation. *arXiv preprint arXiv:2204.01941*, 2022.

[9] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214, 2009.

[10] M. P. Connolly, N. J. Higham, and T. Mary. Stochastic rounding and its probabilistic backward error analysis. *SIAM Journal on Scientific Computing*, 43(1):A566–A585, 2021.

[11] P. Dharangutte and C. Musco. Dynamic trace estimation. *Advances in Neural Information Processing Systems*, 34:30088–30099, 2021.

[12] E. Di Napoli, E. Polizzi, and Y. Saad. Efficient estimation of eigenvalue counts in an interval. *Numerical Linear Algebra with Applications*, 23(4):674–692, 2016.

[13] E. N. Epperly, J. A. Tropp, and R. J. Webber. Xtrace: Making the most of every sample in stochastic trace estimation. *arXiv preprint arXiv:2301.07825*, 2023.

[14] E. Estrada. Characterization of 3d molecular structure. *Chemical Physics Letters*, 319(5-6):713–718, 2000.

[15] A. Frommer and D. B. Szyld. On asynchronous iterations. *Journal of computational and applied mathematics*, 123(1-2):201–216, 2000.

[16] A. S. Gambhir, A. Stathopoulos, and K. Orginos. Deflation as a method of variance reduction for estimating the trace of a matrix inverse. *SIAM Journal on Scientific Computing*, 39(2):A532–A558, 2017.

[17] A. Girard. A fast 'monte-carlo cross-validation'procedure for large least squares problems with noisy data. *Numerische Mathematik*, 56:1–23, 1989.

[18] C. Glusa, E. G. Boman, E. Chow, S. Rajamanickam, and D. B. Szyld. Scalable asynchronous domain decomposition solvers. *SIAM Journal on Scientific Computing*, 42(6):C384–C409, 2020.

[19] G. H. Golub and G. Meurant. Matrices, moments, and quadrature. In D. F. Griffiths and G. A. Watson, editors, *Numerical Analysis 1993*, volume 303, pages 105–1–6. Pitman, Research Notes in Mathematics, 1994.

[20] I. Han, D. Malioutov, H. Avron, and J. Shin. Approximating spectral sums of large-scale matrices using stochastic chebyshev approximations. *SIAM Journal on Scientific Computing*, 39(4):A1558–A1585, 2017.

[21] M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.

[22] S. Jiang, H. Pham, D. Woodruff, and R. Zhang. Optimal sketching for trace estimation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23741–23753. Curran Associates, Inc., 2021.

[23] V. Kalantzis, C. Bekas, A. Curioni, and E. Gallopoulos. Accelerating data uncertainty quantification by solving linear systems with multiple right-hand sides. *Numerical Algorithms*, 62(4):637–653, 2013.

[24] V. Kalantzis, A. C. I. Malossi, C. Bekas, A. Curioni, E. Gallopoulos, and Y. Saad. A scalable iterative dense linear system solver for multiple right-hand sides in data analytics. *Parallel Computing*, 74:136–153, 2018.

[25] M. N. Kolountzakis, G. L. Miller, R. Peng, and C. E. Tsourakakis. Efficient triangle counting in large graphs via degree-based vertex partitioning. *Internet Mathematics*, 8(1-2):161–185, 2012.

[26] L. Lin. Randomized estimation of spectral densities of large matrices made accurate. *Numerische Mathematik*, 136:183–213, 2017.

[27] L. Lin, Y. Saad, and C. Yang. Approximating spectral densities of large matrices. *SIAM Review*, 58(1):34–65, 2016.

[28] A. Mallick, M. Chaudhari, U. Sheth, G. Palanikumar, and G. Joshi. Rateless codes for near-perfect load balancing in distributed matrix-vector multiplication. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–40, 2019.

[29] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 142–155. SIAM, 2021.

[30] D. Persson, A. Cortinovis, and D. Kressner. Improved variants of the Hutch++ algorithm for trace estimation. *SIAM Journal on Matrix Analysis and Applications*, 43(3):1162–1185, 2022.

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

[32] O. Teke and P. P. Vaidyanathan. The asynchronous power iteration: A graph signal perspective. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4059–4063. IEEE, 2018.

[33] O. Teke and P. P. Vaidyanathan. Random node-asynchronous updates on graphs. *IEEE Transactions on Signal Processing*, 67(11):2794–2809, 2019.

[34] C. E. Tsourakakis. Fast counting of triangles in large real networks without counting: Algorithms and laws. In *2008 Eighth IEEE International Conference on Data Mining*, pages 608–617. IEEE, 2008.

[35] C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos. Doulion: counting triangles in massive graphs with a coin. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 837–846, 2009.

[36] S. Ubaru, J. Chen, and Y. Saad. Fast estimation of tr(f(a)) via stochastic Lanczos quadrature. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1075–1099, 2017.

[37] S. Ubaru and Y. Saad. Fast methods for estimating the numerical rank of large matrices. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2016.

[38] S. Ubaru and Y. Saad. Applications of trace estimation techniques. In *High Performance Computing in Science and Engineering: Third International Conference, HPCSE 2017, Karolinka, Czech Republic, May 22–25, 2017, Revised Selected Papers*, pages 19–33. Springer, 2018.

[39] J. Wolfson-Pou and E. Chow. Asynchronous multigrid methods. In *2019 IEEE international parallel and distributed processing symposium (IPDPS)*, pages 101–110. IEEE, 2019.

[40] D. Woodruff, F. Zhang, and R. Zhang. Optimal query complexities for dynamic trace estimation. *Advances in Neural Information Processing Systems*, 35:35049–35060, 2022.

[41] L. Wu, J. Laeuchli, V. Kalantzis, A. Stathopoulos, and E. Gallopoulos. Estimating the trace of the matrix inverse by interpolating from the diagonal of an approximate inverse. *Journal of Computational Physics*, 326:828–844, 2016.

[42] S. Zhou. Sparse Hanson–Wright inequalities for subgaussian quadratic forms. *Bernoulli*, 25(3):1603–1639, 2019.

Supplementary material

A Theoretical Analysis

A.1 Proof of Lemma 1

Proof. The proof follows directly by applying the conditional independence of the random variables $\boldsymbol{Q}$ and $\boldsymbol{x}$ in tandem with the cyclic property of the $\texttt{Tr}(.)$ linear operator:

$$\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}] = \mathbb{E}[\texttt{Tr}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x})] = \mathbb{E}[\texttt{Tr}(\boldsymbol{Q}\boldsymbol{x}\boldsymbol{x}^\top)] = \texttt{Tr}(\mathbb{E}[\boldsymbol{Q}\boldsymbol{x}\boldsymbol{x}^\top]) = \texttt{Tr}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\mathbb{E}[\boldsymbol{Q}]) = \texttt{Tr}(\mathbb{E}[\boldsymbol{Q}]).$$

□

A.2 Proof of Proposition 1

Proof. Following Lemma 1, the expectation $\mathbb{E}\left[\Gamma_M\right]$ is equal to $\texttt{Tr}(\mathbb{E}\left[\boldsymbol{Q}\right])$, where $\boldsymbol{Q}$ is a random matrix whose samples are of the form $\boldsymbol{Q}(\mathcal{T}_k)$. Thus, it suffices to show that

$$\mathbb{E}[\boldsymbol{Q}] = \frac{\mu_T}{N}\boldsymbol{A}.$$

To this end, we apply the same approach as in [5, Lemma 1]. By the Law of Total Expectation [1], the expectation $\mathbb{E}[\boldsymbol{Q}]$ can be written as $\mathbb{E}_T[\mathbb{E}_{\mathcal{T}}[\boldsymbol{Q}|T]]$ where the outer expectation is with respect to the cardinality T of the random integer set $\mathcal{T}$ and the inner expectation is with respect to the content of $\mathcal{T}$. Denoting by $\mathbb{P}[\boldsymbol{Q} = \boldsymbol{Q}(\mathcal{T})|T]$ the probability that $\boldsymbol{Q}(\mathcal{T})$ is realized for a random row subset $\mathcal{T}$ of cardinality T, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}}[\boldsymbol{Q}|T] &= \sum_{\mathcal{T}} \mathbb{P}[\boldsymbol{Q} = \boldsymbol{Q}(\mathcal{T})|T]\boldsymbol{Q}(\mathcal{T}) \\
&= \sum_{\mathcal{T}} \binom{N}{T}^{-1} \boldsymbol{D}_{\mathcal{T}}\boldsymbol{A} \\
&= \binom{N}{T}^{-1} \binom{N-1}{T-1}\boldsymbol{A} \\
&= \frac{T}{N}\boldsymbol{A}.
\end{aligned}$$

The proof follows by noticing that $\mathbb{E}_T[\mathbb{E}_{\mathcal{T}}[\boldsymbol{Q}|T]] = \mathbb{E}_T\left[\frac{T}{N}\boldsymbol{A}\right] = \frac{\mathbb{E}[T]}{N}\boldsymbol{A}$. □

A.3 Proof of Theorem 1

The proof of this Theorem is derived by combining the following theoretical results.

Lemma A.1. *The variance of a single sample of the asynchronous randomized trace estimator Γ_M is equal to*

$$\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}) = \texttt{Tr}(2\mathbb{E}\left[\boldsymbol{Q}^2\right]) + \mathrm{Var}(\texttt{Tr}(\boldsymbol{Q})),$$

when $\boldsymbol{x} \in \mathcal{N}(0, \boldsymbol{I})$, and

$$\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}) = \texttt{Tr}(2\mathbb{E}\left[\boldsymbol{Q}^2 - \mathrm{diag}(\boldsymbol{Q})^2\right]) + \mathrm{Var}(\texttt{Tr}(\boldsymbol{Q})),$$

when $\boldsymbol{x}$ is a Rademacher random vector.

Proof. Recall that the variance of the randomized trace estimator for a constant $N \times N$ matrix $\boldsymbol{K}$ is equal to $\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{K}\boldsymbol{x}) = 2\texttt{Tr}(\boldsymbol{K}^2)$ when the samples of $\boldsymbol{x}$ are drawn from the standard normal distribution and

$\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{K}\boldsymbol{x}) = 2\mathtt{Tr}(\boldsymbol{K}^2 - \mathrm{diag}(\boldsymbol{K}^2))$ when the samples of $\boldsymbol{x}$ are drawn from the Rademacher distribution. Returning to the asynchronous setting, notice that by assumption the variance of the random vector $\boldsymbol{x}$ is finite. Thus, we can apply Eve's law (also known as "law of total variance") [6]. More specifically, focusing on random $\boldsymbol{Q}$, we can write

$$\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{Q}}[\mathrm{Var}_{\boldsymbol{x}}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x})|\boldsymbol{Q}] + \mathrm{Var}_{\boldsymbol{Q}}(\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}|\boldsymbol{Q}]).$$

The first term is equal to the expectation of the variance of the quadratic form $\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$ when $\boldsymbol{Q}$ is a random matrix. Thus, $\mathbb{E}_{\boldsymbol{Q}}[\mathrm{Var}_{\boldsymbol{x}}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x})|\boldsymbol{Q}] = \mathbb{E}[2\mathtt{Tr}(\boldsymbol{Q}^2)]$ for the Rademacher case and $\mathbb{E}_{\boldsymbol{Q}}[\mathrm{Var}_{\boldsymbol{x}}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x})|\boldsymbol{Q}] = \mathtt{Tr}(2\mathbb{E}\left[\boldsymbol{Q}^2 - \mathrm{diag}(\boldsymbol{Q})^2\right])$ for the Gaussian case. Likewise, the second term is equal to $\mathrm{Var}(\mathtt{Tr}(\boldsymbol{Q}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]))$. Recalling once again that $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}$ concludes the proof. □

The variance of the asynchronous estimator depends mainly on the terms $\mathtt{Tr}(2\mathbb{E}\left[\boldsymbol{Q}^2\right])$ and $\mathrm{Var}(\mathtt{Tr}(\boldsymbol{Q}))$. In the following, we analyze each term separately.

Lemma A.2. *The term* $\mathrm{Var}(\mathtt{Tr}(\boldsymbol{Q}))$ *satisfies*

$$\mathrm{Var}(\mathtt{Tr}(\boldsymbol{Q})) = \mathbb{E}[\mathtt{Tr}(\boldsymbol{Q})^2] - \frac{\mu_T^2}{N^2}\mathtt{Tr}(\boldsymbol{A})^2. \tag{1}$$

Proof. The expectation $\mathbb{E}[\mathtt{Tr}(\boldsymbol{Q})]$ of $\mathtt{Tr}(\boldsymbol{Q})$ is equal to $\frac{\mu_T}{N}\mathtt{Tr}(\boldsymbol{A})$. By definition, we have

$$\mathrm{Var}(\mathtt{Tr}(\boldsymbol{Q})) = \mathbb{E}[(\mathtt{Tr}(\boldsymbol{Q}) - \mathbb{E}[\mathtt{Tr}(\boldsymbol{Q})])^2] = \mathbb{E}[\mathtt{Tr}(\boldsymbol{Q})^2] - \mathbb{E}[\mathtt{Tr}(\boldsymbol{Q})]^2 = \mathbb{E}[\mathtt{Tr}(\boldsymbol{Q})^2] - \frac{\mu_T^2}{N^2}\mathtt{Tr}(\boldsymbol{A})^2.$$

□

According to Lemma A.2, the computation of $\mathrm{Var}(\mathtt{Tr}(\boldsymbol{Q}))$ requires that of $\mathbb{E}\left[\mathtt{Tr}(\boldsymbol{Q})^2\right]$. The latter quantity is listed in the following lemma.

Lemma A.3. *The term* $\mathbb{E}\left[\mathtt{Tr}(\boldsymbol{Q})^2\right]$ *satisfies*

$$\mathbb{E}\left[\mathtt{Tr}(\boldsymbol{Q})^2\right] = \frac{1}{N(N-1)}\left((N\mu_T - \sigma_T^2 - \mu_T^2)\,\mathrm{Tr}(\mathcal{D}(\boldsymbol{A})^2) + (\sigma_T^2 + \mu_T^2 - \mu_T)\,\mathrm{Tr}(\boldsymbol{A})^2\right).$$

Proof. Starting from $\mathbb{E}_{\boldsymbol{Q}}\left[\mathrm{Tr}(\boldsymbol{Q})^2\right]$, we can write:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Q}}\left[\mathrm{Tr}(\boldsymbol{Q})^2\right] &= \mathbb{E}_T\left[\mathbb{E}_{\mathcal{T}}\left(\sum_{i\in\mathcal{T}} A_{ii}\right)^2\right] \\
&= \mathbb{E}_T\left[\sum_{\mathcal{T}}\binom{N}{T}^{-1}\left(\sum_{i\in\mathcal{T}} A_{ii}\right)^2\right] \\
&= \mathbb{E}_T\left[\sum_{\mathcal{T}}\binom{N}{T}^{-1}\left(\sum_{i\in\mathcal{T}} A_{ii}^2 + \sum_{i\in\mathcal{T}}\sum_{j\neq i} A_{ii}A_{jj}\right)\right] \\
&= \mathbb{E}_T\left[\binom{N}{T}^{-1}\left(\binom{N-2}{T-1}\sum_{i=1}^{i=N} A_{ii}^2 + \binom{N-2}{T-2}\left(\sum_{i=1}^{N} A_{ii}\right)^2\right)\right] \\
&= \mathbb{E}_T\left[\frac{T(N-T)}{N(N-1)}\,\mathrm{Tr}(\mathcal{D}(\boldsymbol{A})^2) + \frac{T(T-1)}{N(N-1)}\,\mathrm{Tr}(\boldsymbol{A})^2\right].
\end{aligned}$$

The first three equalities follow from the definition of trace and the expansion of the square of sum. The fourth equality follows by counting the number of times each term appears in all possible sets of size T. Indeed, the cross terms $A_{ii}A_{jj}$ appear $\binom{N-2}{T-2}$ times while the square terms A_{ii}^2 appear $\binom{N-1}{T-1}$ times. The final equality then follows immediately.

□

Finally, the derivation of $\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x})$ requires an expression for the quantity $\mathbb{E}[2\mathtt{Tr}(\boldsymbol{Q}^2)]$.

Lemma A.4. *The term* $2\mathbb{E}[\mathtt{Tr}(\boldsymbol{Q}^2)]$ *satisfies*

$$2\mathbb{E}[\mathtt{Tr}(\boldsymbol{Q}^2)] = \frac{2\mu_T}{N}\mathtt{Tr}\left(\boldsymbol{A}^2\right). \tag{2}$$

Proof. First, notice that $\mathbb{E}[\mathtt{Tr}(\boldsymbol{Q}^2)] = \mathbb{E}[\|\boldsymbol{Q}\|_F^2])$, and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Q}}\left[\|\boldsymbol{Q}\|_F^2\right] &= \mathbb{E}_T\left[\mathbb{E}_{\mathcal{T}}\sum_{i\in\mathcal{T}}\|\boldsymbol{A}_i\|_2^2\right] \\
&= \mathbb{E}_T\left[\sum_{\mathcal{T}}\binom{N}{T}^{-1}\sum_{i\in\mathcal{T}}\|\boldsymbol{A}_i\|_2^2\right] \\
&= \mathbb{E}_T\left[\binom{N}{T}^{-1}\binom{N-1}{T-1}\sum_{i=1}^{i=N}\|\boldsymbol{A}_i\|_2^2\right] \\
&= \mathbb{E}_T\left[\frac{T}{N}\|\boldsymbol{A}\|_F^2\right] = \frac{\mu_T}{N}\|\boldsymbol{A}\|_F^2.
\end{aligned}$$

□

Similarly, for the Radamacher vectors case, we have

$$\mathtt{Tr}(2\mathbb{E}\left[\boldsymbol{Q}^2 - \mathrm{diag}(\boldsymbol{Q})^2\right]) = \frac{2\mu_T}{N}\left(\mathtt{Tr}(\boldsymbol{A}^2) - \mathtt{Tr}(\mathcal{D}(\boldsymbol{A})^2\right)$$

Combining Lemmas A.1, A.2, A.3, and A.4, for Gaussian vectors, we get

$$\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}) = \frac{2\mu_T}{N}\mathtt{Tr}\left(\boldsymbol{A}^2\right) + \frac{1}{N(N-1)}\left((\sigma_T^2 + \frac{1}{N}\mu_T^2 - \mu_T)\,\mathrm{Tr}(\boldsymbol{A})^2 + (N\mu_T - \sigma_T^2 - \mu_T^2)\,\mathrm{Tr}(\mathcal{D}(\boldsymbol{A})^2)\right)$$

and for Radamacher vectors, we have

$$\mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}) = \frac{2\mu_T}{N}\mathtt{Tr}\left(\boldsymbol{A}^2\right) + \frac{1}{N(N-1)}\left((\sigma_T^2 + \frac{1}{N}\mu_T^2 - \mu_T)\,\mathrm{Tr}(\boldsymbol{A})^2 - ((N-2)\mu_T + \sigma_T^2 + \mu_T^2)\,\mathrm{Tr}(\mathcal{D}(\boldsymbol{A})^2)\right)$$

A.4 Proof of Theorem 2

For the standard stochastic trace estimator, we can obtain the (ϵ, δ) error bounds and the bound on the sampling complexity using the Hanson-Wright inequality [4], see [2]. In the asynchronous setting, in order to prove Theorem 2 and obtain the bound on sampling complexity, we consider a sparse variant of the Hanson-Wright inequality, see [7, 3]. We have the following Lemma, which is a modification of the main results (Theorem 1.1 and particularly Corollary 2.3) in [7].

Lemma A.5 (Sparse Hanson-Wright)**.** *Let* $\boldsymbol{x} \in \mathbb{R}^N$ *be a random vector of mean zero, i.i.d. sub-Gaussian random entries with constant sub-Gaussian parameter* C, $\boldsymbol{\xi} \in \{0,1\}^N$ *be a random vector independent of* $\boldsymbol{x}$*, with independent Bernoulli random variables* ξ_i *such that* $P[\xi_i = 1] = p$, $\boldsymbol{D}_\xi = \mathcal{D}(\boldsymbol{\xi})$ *and* $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ *be a given matrix. Then, there exists a constant* c *only depending on* C *such that for every* $t \geq 0$*,*

$$\Pr\left(\left|\boldsymbol{x}^\top \boldsymbol{D}_\xi \boldsymbol{A}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{D}_\xi \boldsymbol{A}\boldsymbol{x}]\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{p\|\boldsymbol{A}\|_F^2}, \frac{t}{\|\boldsymbol{A}\|_2}\right)\right), \tag{3}$$

Proof. The proof of the lemma follows the proof of Theorem 1.1 in [7]. We can write:

$$\boldsymbol{x}^\top \boldsymbol{D}_\xi \boldsymbol{A}\boldsymbol{x} = \sum_i \sum_j x_i \xi_i A_{ij} x_j = \sum_i x_i^2 \xi_i A_{ii} + \sum_{i\neq j} x_i x_j \xi_i A_{ij}.$$

Note that for the diagonal sum, we have $\mathbb{E}[\sum_i x_i^2 \xi_i A_{ii}] = p \sum_i \mathbb{E}[x_i^2] A_{ii}$, and for the off-diagonal sums, $\mathbb{E}[\sum_{i \neq j} x_i x_j \xi_i A_{ij}] = 0$. For these two terms, we obtain the same tail bounds as in [7]. For the diagonal sum, we can use Lemma 3.1 in [7] directly, and we have

$$\Pr\left(\left|\sum_i x_i^2 \xi_i A_{ii} - p \sum_i \mathbb{E}[x_i^2] A_{ii}\right| \geq t\right) \leq 2 \exp\left(-c_1 \min\left(\frac{t^2}{p \sum_i A_{ii}^2}, \frac{t}{\max_i(A_{ii})}\right)\right).$$

For the off-diagonal sum, we can use similar arguments as in [7, Eqn. 8] and obtain the following by setting $p_i = p, p_j = 1$ in that equation,

$$\Pr\left(\left|\sum_{i \neq j} x_i x_j \xi_i A_{ij}\right| \geq t\right) \leq 2 \exp\left(-c_2 \min\left(\frac{t^2}{p \sum_i A_{ij}^2}, \frac{t}{\|\boldsymbol{A}\|_2}\right)\right).$$

Combining these two, we get the sought result. □

Next, we use the above Lemma to get the result:

Lemma A.6. *Let $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ be a given PSD matrix, $\delta \in (0, 1/2], \epsilon \in (0, 1]$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$, be M random vectors of mean zero, i.i.d. sub-Gaussian random entries with constant sub-Gaussian parameter, $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M$ be M random vectors independent of $\boldsymbol{x}_i$'s with independent Bernoulli random variables ξ_i such that $P[\xi_i = 1] = p$, and $\boldsymbol{D}_{\xi_i} = \mathcal{D}(\boldsymbol{\xi}_i)$. For fixed constants c, C, we have*

$$\Pr\left(\left|\frac{1}{pM} \sum_{k=1}^{M} \boldsymbol{x}_k^\top \boldsymbol{D}_{\xi_k} \boldsymbol{A} \boldsymbol{x}_k - \operatorname{Tr}(\boldsymbol{A})\right| \leq \epsilon \operatorname{Tr}(\boldsymbol{A})\right) \geq 1 - \delta, \tag{4}$$

if $M > \frac{C \log(1/\delta)}{p \epsilon^2}$.

Proof. Let $\boldsymbol{D}_\xi \bar{\boldsymbol{A}} \in \mathbb{R}^{MN \times MN}$ be a block diagonal matrix with $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M]^\top$ and $\boldsymbol{D}_\xi = \mathcal{D}(\boldsymbol{\xi})$ given by

$$\boldsymbol{D}_\xi \bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{D}_{\xi_1} \boldsymbol{A} & 0 & \cdots & 0 \\ 0 & \boldsymbol{D}_{\xi_2} \boldsymbol{A} & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{D}_{\xi_m} \boldsymbol{A} \end{bmatrix}$$

Let $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M]^\top \in \mathbb{R}^{MN}$, then using Lemma A.5, we have

$$\Pr\left(\left|\boldsymbol{x}^\top \boldsymbol{D}_\xi \bar{\boldsymbol{A}} \boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{D}_\xi \bar{\boldsymbol{A}} \boldsymbol{x}]\right| \geq t\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{p \|\bar{\boldsymbol{A}}\|_F^2}, \frac{t}{\|\bar{\boldsymbol{A}}\|_2}\right)\right),$$

Next, we have $\boldsymbol{x}^\top \boldsymbol{D}_\xi \bar{\boldsymbol{A}} \boldsymbol{x} = \sum_{k=1}^{M} \boldsymbol{x}_k^\top \boldsymbol{D}_{\xi_k} \boldsymbol{A} \boldsymbol{x}_k$, $\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{D}_\xi \bar{\boldsymbol{A}} \boldsymbol{x}] = p \operatorname{Tr}(\bar{\boldsymbol{A}}) = pM \operatorname{Tr}(\boldsymbol{A})$, and $\|\bar{\boldsymbol{A}}\|_F^2 = M \|\boldsymbol{A}\|_F^2, \|\bar{\boldsymbol{A}}\|_2 = \|\boldsymbol{A}\|_2$, setting $t' = t/(pM)$, we get

$$\Pr\left(\left|\frac{1}{pM} \sum_{k=1}^{M} \boldsymbol{x}_k^\top \boldsymbol{D}_{\xi_k} \boldsymbol{A} \boldsymbol{x}_k - \operatorname{Tr}(\boldsymbol{A})\right| \geq t'\right) \leq 2 \exp\left(-c \min\left(\frac{pMt'^2}{\|\boldsymbol{A}\|_F^2}, \frac{pMt'}{\|\boldsymbol{A}\|_2}\right)\right).$$

Setting $t' = \epsilon \operatorname{Tr}(\boldsymbol{A})$, choosing $\delta \geq 2 \exp\left(-c \frac{pMt'^2}{\|\boldsymbol{A}\|_F^2}\right)$, noting for SPD matrices $\operatorname{Tr}^2(\boldsymbol{A}) \geq \|\boldsymbol{A}\|_F^2$ and that for $\delta < 1/2$, we have $\log(2/\delta) \leq 2 \log(1/\delta)$, we obtain the result for a constant C when

$$M > \frac{C \log(1/\delta)}{p \epsilon^2}.$$

□

For the different cases in Theorem 2, we get the appropriate sampling complexity by setting $p = \mu_T / N$.

Table 1: *Additional set of test matrices. The variables N, $\texttt{nnz}(\boldsymbol{A})$, and $\texttt{Tr}(\boldsymbol{A})$ denote the size, number of non-zero entries, and trace of matrix $\boldsymbol{A}$, respectively.*

Id	Matrix name	N	$\texttt{nnz}(\boldsymbol{A})$	$\texttt{Tr}(\boldsymbol{A})$
1	`Pajek/Roget`	1022	5075	1
2	`Arenas/email`	1133	10902	0
3	`TKK/plbuckle`	1282	30644	3.208e+08
4	`SNAP/wiki-Vote`	8297	103689	0
5	`SNAP/ca-CondMat`	23133	186936	58

A.5 Proof of Theorem 3

The proof relies on the previous Lemmas and the following Lemmas whose proofs are similar to those of Lemmas A.4 and A.3, respectively.

Lemma A.7. *The term* $\mathbb{E}[2\texttt{Tr}(\boldsymbol{Q}^2)] = \texttt{Tr}(2\mathbb{E}[\boldsymbol{Q}^2])$ *satisfies*

$$\mathbb{E}[2\texttt{Tr}(\boldsymbol{Q}^2)] = \frac{2\mu_T}{N}\texttt{Tr}\left(\mathbb{E}\left[\widetilde{\boldsymbol{A}}^2\right]\right).$$

Lemma A.8. *The term* $\mathbb{E}\left[\texttt{Tr}(\boldsymbol{Q})^2\right]$ *satisfies*

$$\mathbb{E}\left[\texttt{Tr}(\boldsymbol{Q})^2\right] = \frac{(N\mu_T - \mathrm{Var}(T) - \mu_T^2)\,\mathrm{Tr}(\mathbb{E}[\mathcal{D}(\widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{A}}]) + (\mathrm{Var}(T) + \mu_T^2 - \mu_T)\mathbb{E}[\mathrm{Tr}(\widetilde{\boldsymbol{A}})^2]}{N(N-1)}.$$

B Additional numerical experiments

In this section we present additional numerical experiments to accompany the main paper. More specifically, in addition to presenting results on asynchronous randomized trace estimation with Gaussian vectors, as well as stochastic rounding, we include five additional test matrices reported in Table 1.

Figure 1 presents the accuracy of the Rademacher asynchronous and synchronous randomized trace estimators when applied to the matrices in Table 1.

Figures 2 and 3 present the accuracy of the Gaussian asynchronous and synchronous randomized trace estimators when applied to all reported matrices (including the main paper).

Figures 4 and 5 present the accuracy of the Rademacher/Gaussian asynchronous and synchronous randomized trace estimators with stochastic rounding when applied to a i.i.d. dense matrix of size $N = 1000$.

Finally, Figures 6-10 present a comparison[1] of the asynchronous and synchronous randomized trace estimators as $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, for the five largest matrices listed in the main paper and our supplement. Notice that the synchronous randomized trace estimator does not depend on p but we list its accuracy on each figure for comparison purposes. The same is true for the asynchronous randomized estimator when T is uniform (no dependence on p) but we plot the same figure across each different row of subfigures for the sake of completeness. In summary, the accuracy of the asynchronous randomized trace estimator improves as p increases, and becomes more similar to that obtained by the synchronous randomized trace estimator. The latter is expected since increasing p forces the asynchronous estimator to sample more rows per sample (i.e., if we were to use $p = 1$ we would retrieve the classical synchronous trace estimator). Moreover, the two different sampling mechanisms outlined in Corolarries 1 and 3 of our main paper lead to similar accuracy when p is small. On the other hand, as p approaches one, the sampling mechanism outlined in Corollary 1 leads to a better matching of the accuracy achieved by the synchronous trace estimator.

[1]For economy, we only list the case where the random vector $\boldsymbol{x}$ is sampled from the Rademacher distribution.

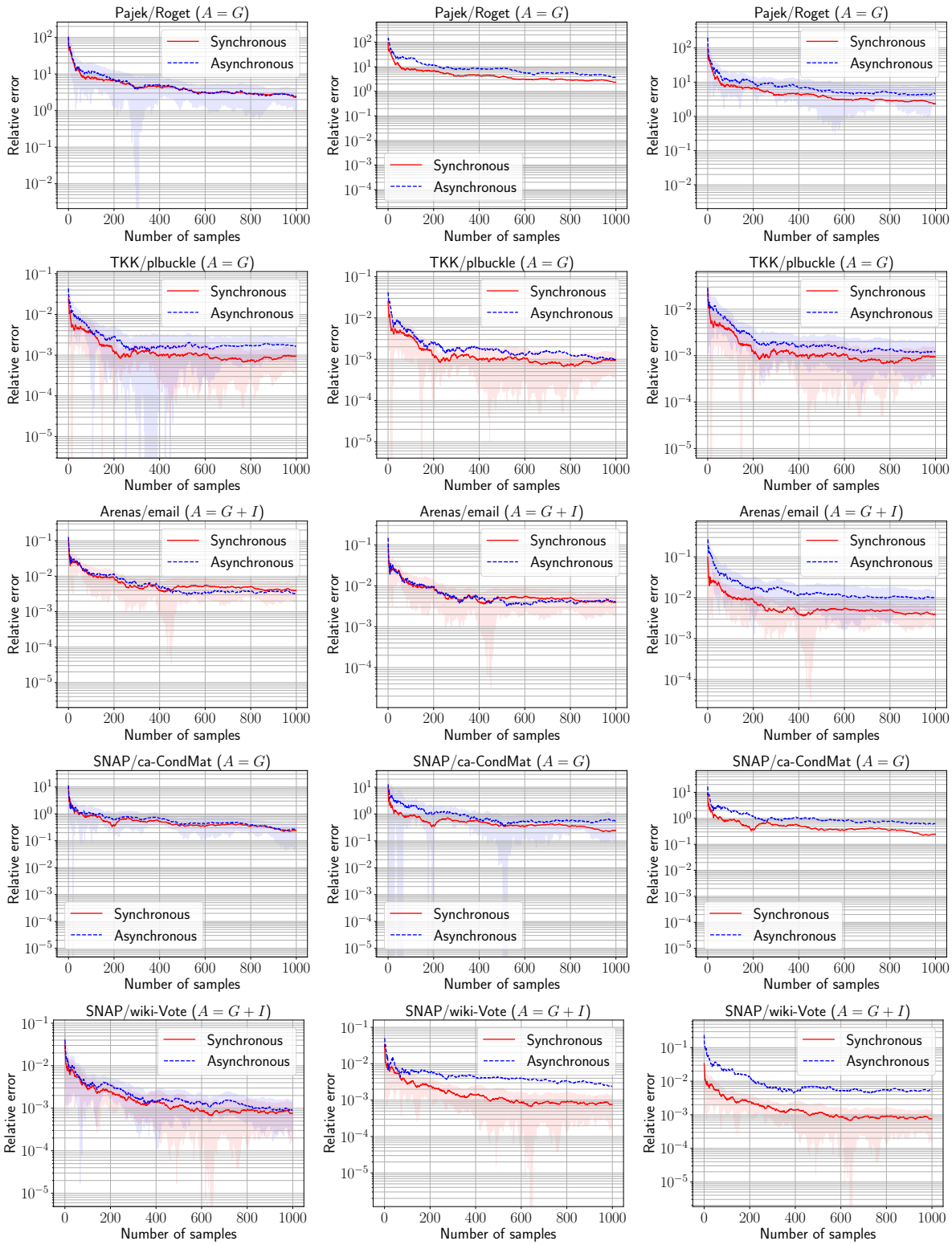

Figure 1: *[Rademacher samples]* Left to right: fixed T, uniform T, fixed p; $p = 0.6$, $T = \lceil Np \rceil$.

References

[1] T. Chen, T. Trogdon, and S. Ubaru. Analysis of stochastic lanczos quadrature for spectrum approximation. In *International Conference on Machine Learning*, pages 1728–1739. PMLR, 2021.

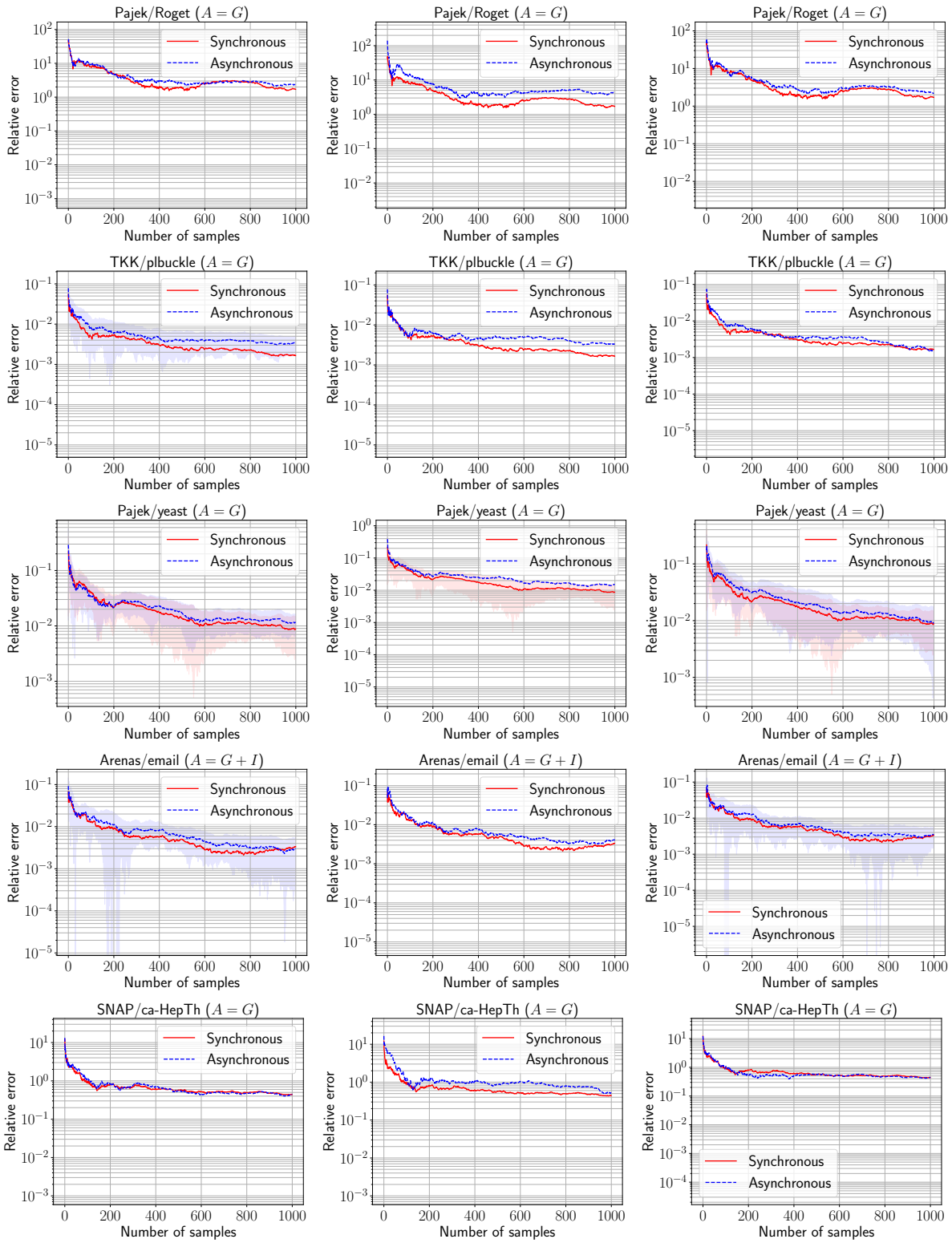

Figure 2: *[Gaussian samples]* Left to right: fixed T, uniform T, fixed p; $p = 0.6$, $T = \lceil Np \rceil$.

[2] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 142–155. SIAM, 2021.

[3] S. Park, X. Wang, and J. Lim. Sparse Hanson–Wright inequality for a bilinear form of sub-gaussian variables.

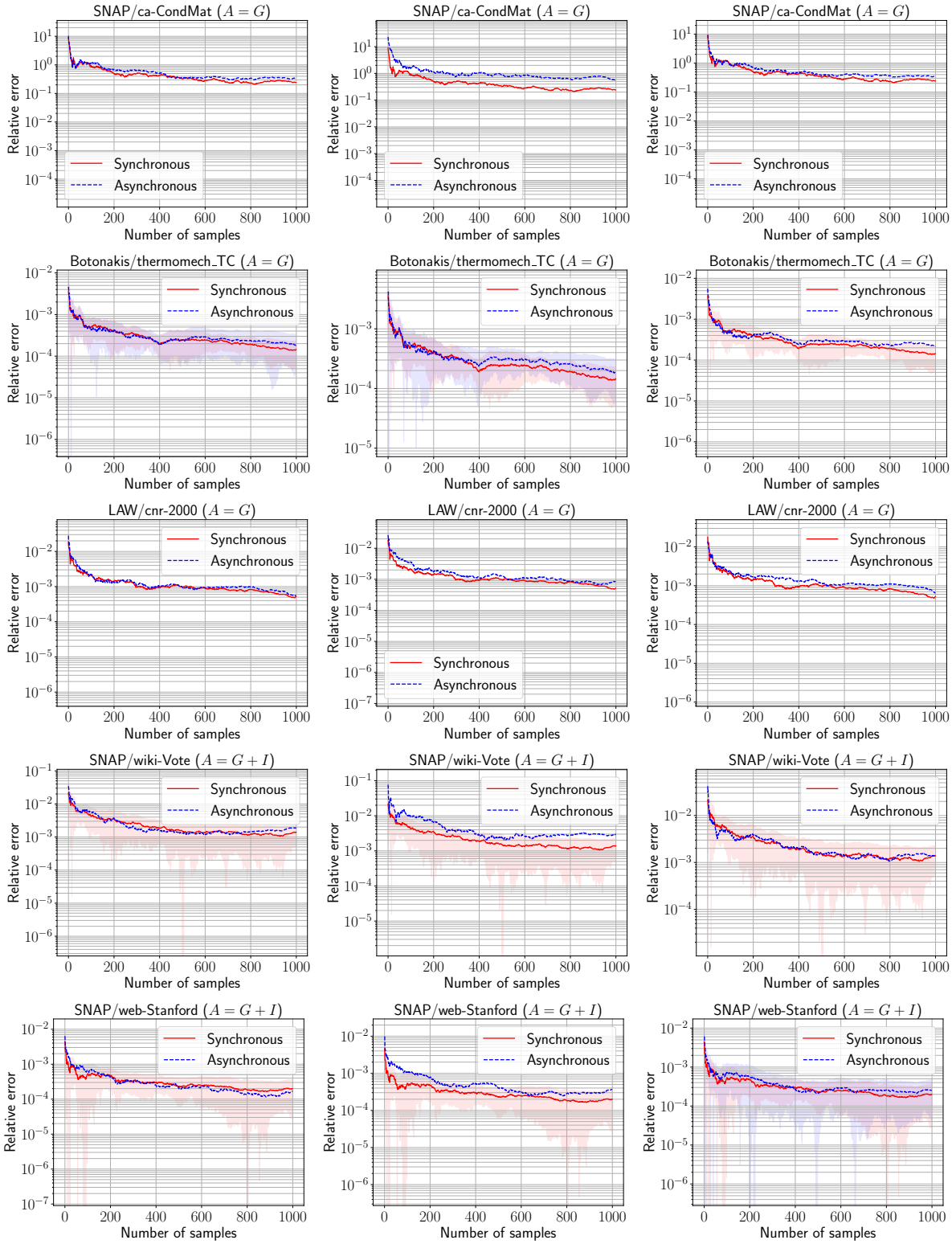

Figure 3: *[Gaussian samples]* Left to right: fixed T, uniform T, fixed p; $p = 0.6$, $T = \lceil Np \rceil$.

Stat, 12(1):e539, 2023.

[4] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *ELECTRONIC COMMUNICATIONS IN PROBABILITY*, 18:1–9, 2013.

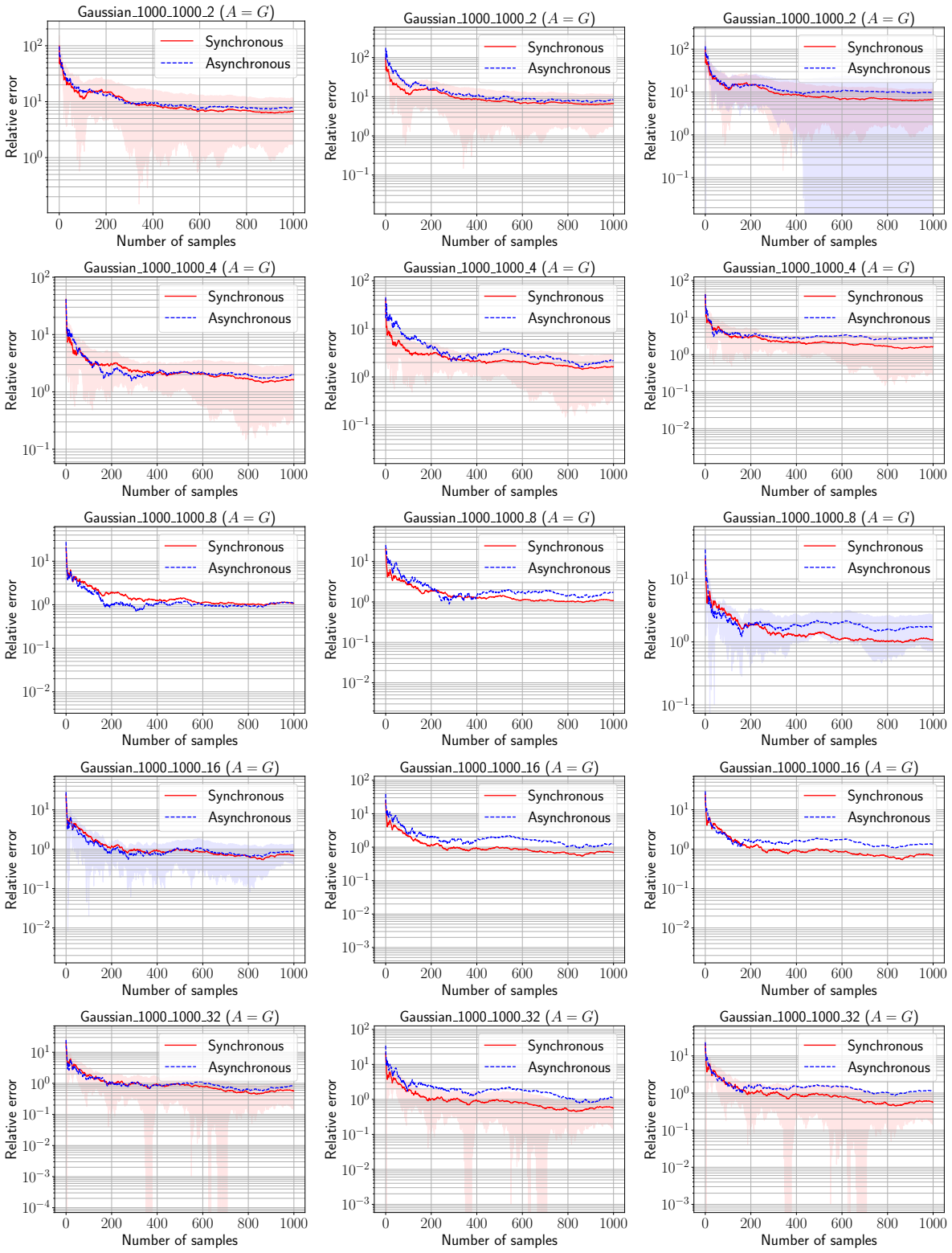

Figure 4: *[Rademacher samples]* Stochastic rounding. Matrix of size $N = 1000$ with entries sampled from standard normal distribution scaled by 1000. Left to right: fixed T, uniform T, fixed p; $p = 0.6$, $T = \lceil Np \rceil$. Top to bottom: Different numbers of quantization levels: $2, 4, 8, 16, 32$.

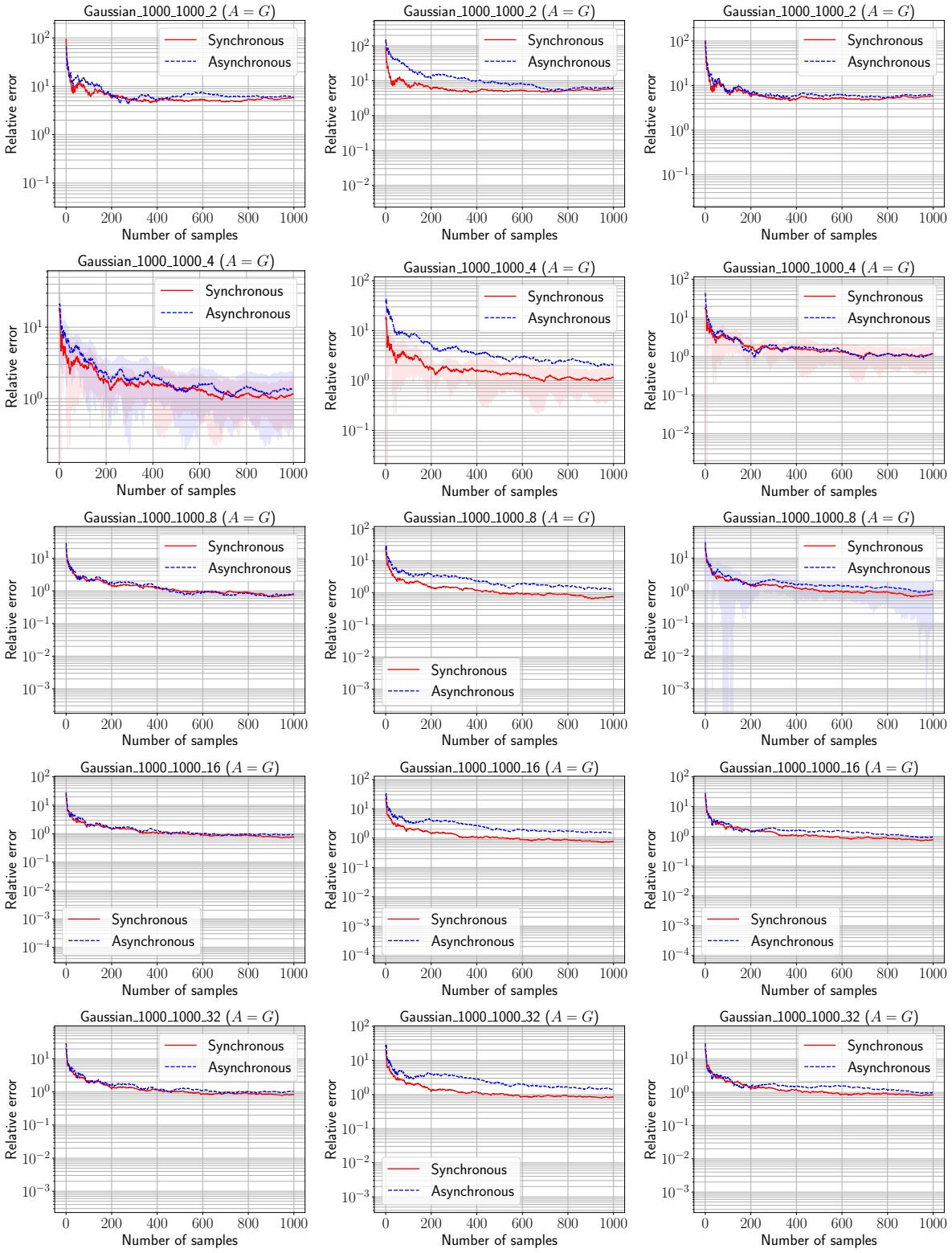

Figure 5: *[Gaussian samples]* Stochastic rounding. Matrix of size $N = 1000$ with entries sampled from standard normal distribution scaled by 1000. Left to right: fixed T, uniform T, fixed p; $p = 0.6$, $T = \lceil Np \rceil$. Top to bottom: Different numbers of quantization levels: $2, 4, 8, 16, 32$.

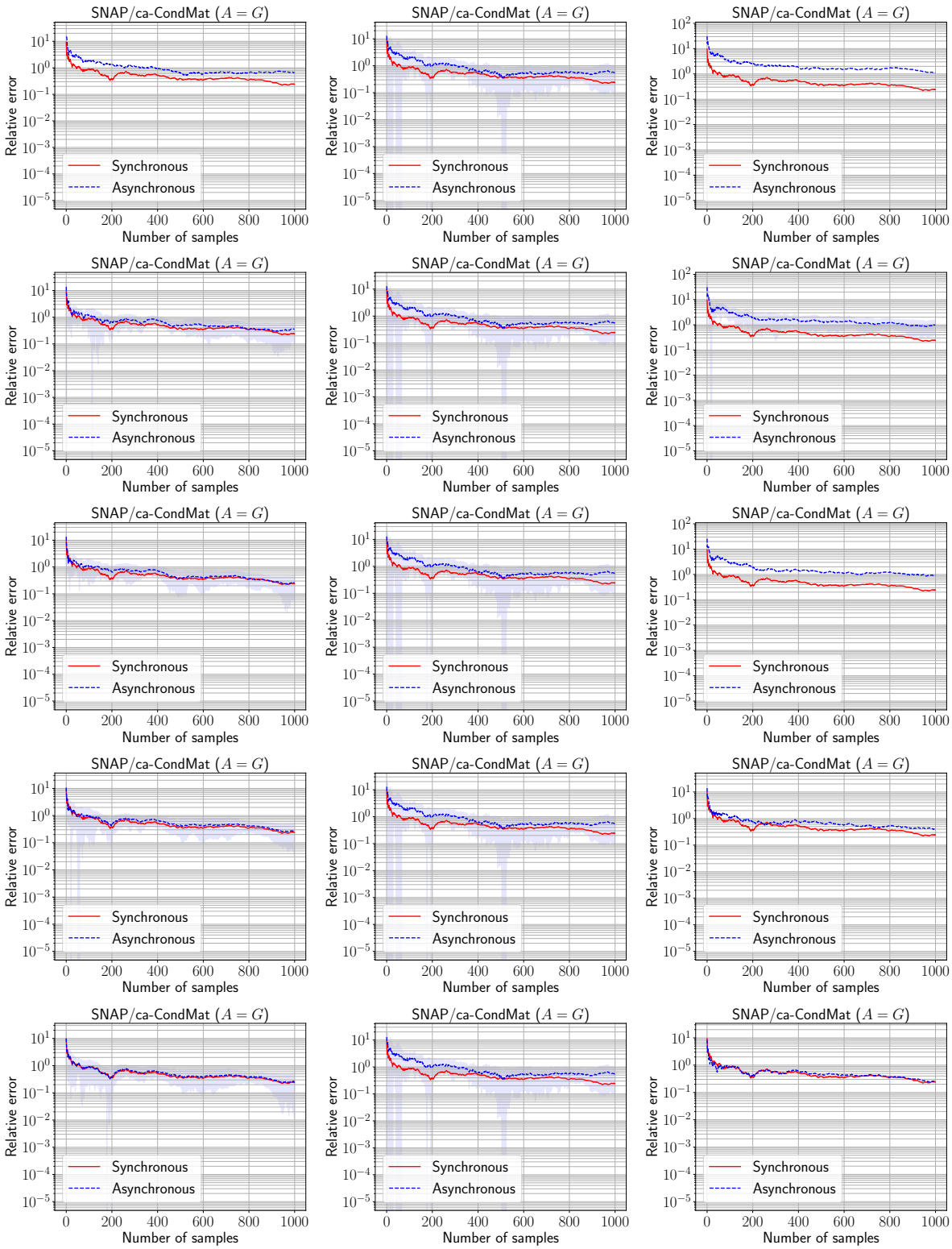

Figure 6: *[Rademacher samples]* Comparing the asynchronous and synchronous randomized trace estimators for various values of p (matrix: `SNAP/ca-CondMat`). Left to right: fixed T, uniform T, fixed p; $T = \lceil Np \rceil$. Top to bottom: $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

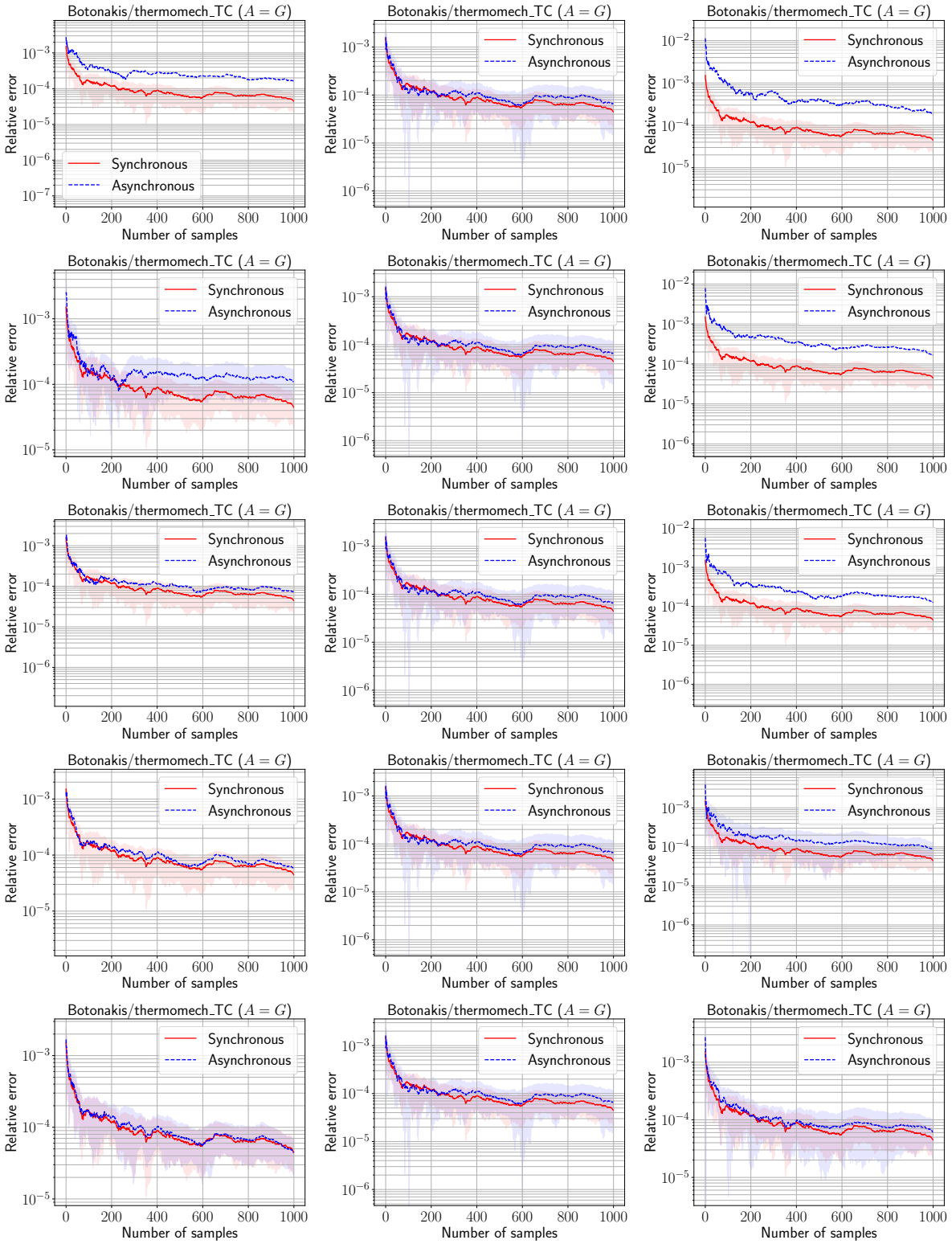

Figure 7: *[Rademacher samples]* Comparing the asynchronous and synchronous randomized trace estimators for various values of p (matrix: `Botonakis/thermotech`). Left to right: fixed T, uniform T, fixed p; $T = \lceil Np \rceil$. Top to bottom: $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

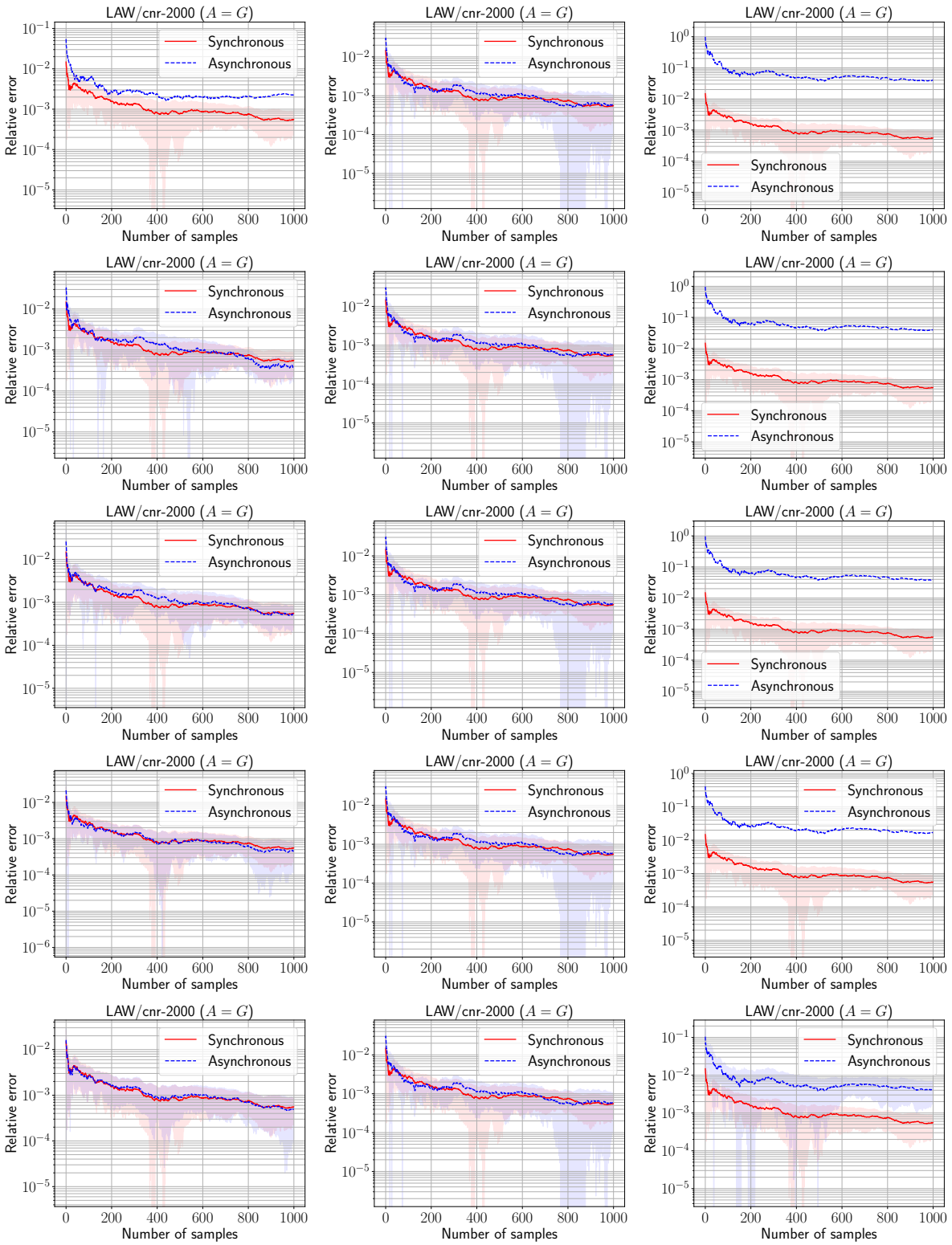

Figure 8: *[Rademacher samples]* Comparing the asynchronous and synchronous randomized trace estimators for various values of p (matrix: `LAW/cnr-2000`). Left to right: fixed T, uniform T, fixed p; $T = \lceil Np \rceil$. Top to bottom: $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

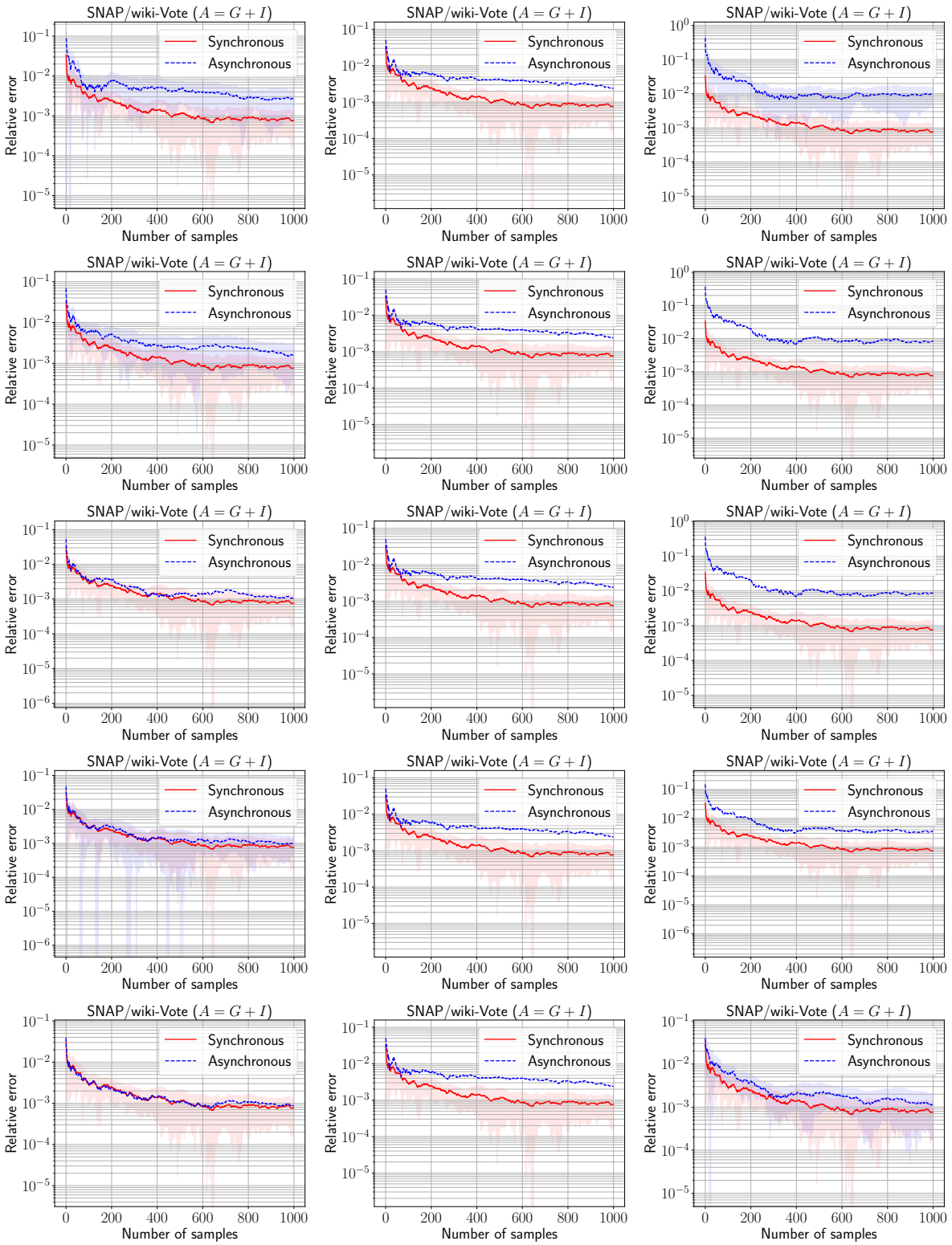

Figure 9: *[Rademacher samples]* Comparing the asynchronous and synchronous randomized trace estimators for various values of p (matrix: `SNAP/wiki-Vote`). Left to right: fixed T, uniform T, fixed p; $T = \lceil Np \rceil$. Top to bottom: $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

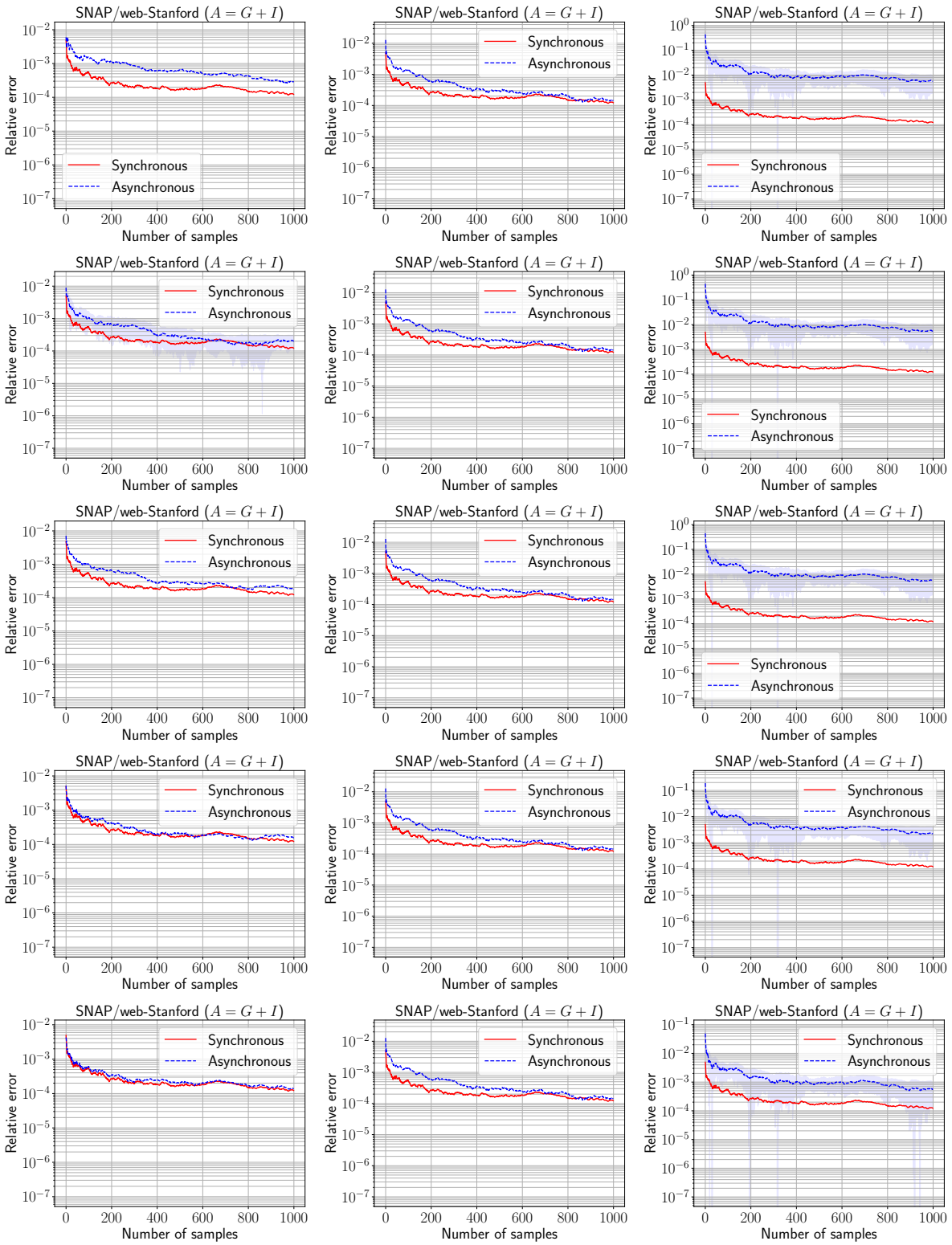

Figure 10: *[Rademacher samples]* Comparing the asynchronous and synchronous randomized trace estimators for various values of p (matrix: `SNAP/web-Stanford`). Left to right: fixed T, uniform T, fixed p; $T = \lceil Np \rceil$. Top to bottom: $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

[5] O. Teke and P. P. Vaidyanathan. Random node-asynchronous updates on graphs. *IEEE Transactions on Signal Processing*, 67(11):2794–2809, 2019.

[6] N. A. Weiss, P. T. Holmes, and M. Hardy. *A course in probability*. Pearson Addison Wesley Boston, Massachusetts, USA, 2006.

[7] S. Zhou. Sparse Hanson–Wright inequalities for subgaussian quadratic forms. *Bernoulli*, 25(3):1603–1639, 2019.