Trace and Schatten norm estimation using randomized sampling

1 Preliminaries

Definition. For a matrix $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1(\boldsymbol{B}) \geq \cdots \geq \sigma_r(\boldsymbol{B}) > 0$, we define its Schatten p-norm as:

$$\|\boldsymbol{B}\|_p = \begin{cases} \left(\sum_{i=1}^r \sigma_i(\boldsymbol{B})^p\right)^{\frac{1}{p}}, & p \in [1,\infty) \\ \sigma_1(\boldsymbol{B}), & p = \infty \end{cases}$$

2 Trace estimation

- Given: Non-zero PSD matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ via the matrix-vector product $\boldsymbol{u} \mapsto \boldsymbol{A}\boldsymbol{u}$
- Goal: Estimate $\operatorname{tr}(\boldsymbol{A})$ without computing $\boldsymbol{\delta}_i^\mathsf{T}(\boldsymbol{A}\boldsymbol{\delta}_i)$ for $i = 1,\ldots,n$

2.1 Estimator based on randomized sampling

- For any random test vector $\boldsymbol{\omega} \in \mathbb{R}^n$ that is isotropic ($\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^\mathsf{T}] = \boldsymbol{I}_n$), the random variable $X = \boldsymbol{\omega}^\mathsf{T}(\boldsymbol{A}\boldsymbol{\omega})$ is an unbiased estimator of $\operatorname{tr}(\boldsymbol{A})$.
- Average over k independent copies of X:

  $$\boldsymbol{\omega}_i \sim \boldsymbol{\omega} \text{ i.i.d.}, \quad X_i = \boldsymbol{\omega}_i^\mathsf{T}(\boldsymbol{A}\boldsymbol{\omega}_i), \quad \bar{X}_k = \frac{1}{k}\sum_{i=1}^k X_i$$

  to obtain an unbiased and consistent estimator of $\operatorname{tr}(\boldsymbol{A})$:

  $$\mathbb{E}[\bar{X}_k] = \operatorname{tr}(\boldsymbol{A}), \quad \operatorname{Var}[\bar{X}_k] = \frac{1}{k}\operatorname{Var}[X]$$

- Runtime advantageous for $k \ll n$
- If the coordinates of $\boldsymbol{\omega}$ are independent samples from a standardized random variable Z, then $\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^\mathsf{T}] = \boldsymbol{I}_n$ and $\operatorname{Var}[X] = 2\sum_{i\neq j}(\boldsymbol{A})_{ij}^2 + \left(\mathbb{E}[Z^4]-1\right)\sum_i(\boldsymbol{A})_{ii}^2$.

  Important examples:

  - Girard estimator $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$: $\operatorname{Var}[X] = 2\sum_{i,j}(\boldsymbol{A})_{ij}^2 = 2\|\boldsymbol{A}\|_\mathsf{F}^2 \leq 2\|\boldsymbol{A}\|_\mathsf{s}\operatorname{tr}(\boldsymbol{A})$
  - Hutchinson estimator $\boldsymbol{\omega} \sim \mathcal{U}\{-1,1\}^n$: $\operatorname{Var}[X] = 2\sum_{i\neq j}(\boldsymbol{A})_{ij}^2 < 2\|\boldsymbol{A}\|_\mathsf{F}^2$

2.2 A priori error estimates

We can employ concentration inequalities to derive probability bounds for relative errors. They indicate how many samples k are sufficient to achieve a certain error tolerance. However, such prior bounds depend on information about the input matrix, e.g. $\operatorname{intdim}(\boldsymbol{A})$, which we usually have no access to. Therefore, it is wiser to rely on a posteriori bounds in practice.

2.3 A posteriori error estimates

- Confidence intervals based on t-distribution:
  - Sample variance: $S_k^2 = \dfrac{1}{k-1}\sum_{i=1}^k (X_i - \bar{X}_k)^2, \quad \mathbb{E}[S_k^2] = \operatorname{Var}[X]$
  - For moderate sample sizes ($k \geq 30$): $\dfrac{\bar{X}_k - \operatorname{tr}(\boldsymbol{A})}{S_k/\sqrt{k}} \sim t(k-1)$

  - $(1-2\alpha)$ confidence interval: $\operatorname{tr}(\boldsymbol{A}) \in \big[\bar{X}_k \pm t_{1-\alpha,k-1}\frac{S_k}{\sqrt{k}}\big]$
- Alternative method: Confidence intervals based on bootstrapping

2.4 Application to Frobenius and Schatten 4-norm

- Given: Matrix $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ accessed via $\boldsymbol{u} \mapsto \boldsymbol{B}\boldsymbol{u}$
- Draw i.i.d. test vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and compute:

$$X_i = \boldsymbol{\omega}_i^\top \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{\omega}_i, \quad \bar{X}_k = \frac{1}{k}\sum\nolimits_{i=1}^{k} X_i$$

- Since $\bar{X}_k$ is the Girard trace estimator for $\boldsymbol{B}^\top\boldsymbol{B}$, we have: $\mathbb{E}[\bar{X}_k] = \operatorname{tr}(\boldsymbol{B}^\top\boldsymbol{B}) = \|\boldsymbol{B}\|_{\mathsf{F}}^2$
- Use the sample variance S_k^2 to estimate $\|\boldsymbol{B}\|_4$: $\mathbb{E}[S_k^2] = \operatorname{Var}[X] = 2\|\boldsymbol{B}^\top\boldsymbol{B}\|_{\mathsf{F}}^2 = 2\|\boldsymbol{B}\|_4^4$

3 Schatten norm estimation

- Goal: Estimate Schatten $2p$-norm $\|\boldsymbol{B}\|_{2p}$ for natural numbers $p \geq 3$
- Idea: Sample i.i.d. isotropic test vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k \in \mathbb{R}^n$ to extract linear information $\boldsymbol{Y}_i = \boldsymbol{B}\boldsymbol{\omega}_i$, form the sample matrix $\boldsymbol{Y} = [\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k]$ and consider $\boldsymbol{X} = \boldsymbol{Y}^\top\boldsymbol{Y} \in \mathbb{R}^{k\times k}$
- For distinct indices $1 \leq i_1, \ldots, i_p \leq k$, we can use the "trace trick" as well as independence and isotropy to conclude:

$$\mathbb{E}\big[(\boldsymbol{X})_{i_1i_2}(\boldsymbol{X})_{i_2i_3}\cdots(\boldsymbol{X})_{i_pi_1}\big] = \operatorname{tr}\big((\boldsymbol{B}^\top\boldsymbol{B})^p\big) = \|\boldsymbol{B}\|_{2p}^{2p}$$

3.1 Approach from classical statistics

- Average over all sequences of distinct indices to obtain an unbiased estimator of $\|\boldsymbol{B}\|_{2p}^{2p}$:

$$U_p = \frac{(k-p)!}{k!}\sum_{1\leq i_1,\ldots,i_p\leq k}^{\circ}(\boldsymbol{X})_{i_1i_2}(\boldsymbol{X})_{i_2i_3}\cdots(\boldsymbol{X})_{i_pi_1}$$

- Computationally demanding (almost k^p summands) and highly variable

  $\Longrightarrow$ Only feasible for small values of p

3.2 Computationally efficient method

- Average only over strictly increasing sequences of indices:

$$V_p = \binom{k}{p}^{-1}\sum_{1\leq i_1<\cdots<i_p\leq k}(\boldsymbol{X})_{i_1i_2}(\boldsymbol{X})_{i_2i_3}\cdots(\boldsymbol{X})_{i_pi_1} = \binom{k}{p}^{-1}\operatorname{tr}\big[\boldsymbol{X}_{\mathsf{U}}^{p-1}\boldsymbol{X}\big]$$

  ($\boldsymbol{X}_{\mathsf{U}}$ denotes the strict upper triangular part of $\boldsymbol{X}$)

- Even higher variance than U_p, but cheaper in computation

  $\Longrightarrow$ Method is feasible for larger values of p and one can deploy a larger sample size k to account for the high variance

Main reference: P. G. Martinsson and J. A. Tropp (2020), 'Randomized numerical linear algebra: Foundations and algorithms', Acta Numerica, Volume 29, pp. 403 – 572, Cambridge University Press