Hutchinson's Estimator is bad at Kronecker Trace Estimation

Raphael A. Meyer,
NYU

Haim Avron
Tel Aviv University

Matrix-Vector Complexity

Many matvec-optimal algorithms proven recently

Great for applications where matvecs are:

1. Efficiently Computable
2. Computational Bottleneck

E.g. $\underline{x} \rightarrow f(A)\underline{x}$ via Lanczos Iteration

But what if 1 does not hold?

We can only compute $A\underline{x}$ for some $\underline{x}$!

Kronecker Product

Let $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times q}$. Then $A \otimes B \in \mathbb{R}^{mp \times nq}$

$$\mathbf{A} \otimes \mathbf{B} := \begin{bmatrix} [\mathbf{A}]_{11}\mathbf{B} & [\mathbf{A}]_{12}\mathbf{B} & \dots & [\mathbf{A}]_{1m}\mathbf{B} \\ [\mathbf{A}]_{21}\mathbf{B} & [\mathbf{A}]_{22}\mathbf{B} & \dots & [\mathbf{A}]_{2m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ [\mathbf{A}]_{n1}\mathbf{B} & [\mathbf{A}]_{n2}\mathbf{B} & \dots & [\mathbf{A}]_{nm}\mathbf{B} \end{bmatrix}$$

For vectors $\underline{x} \in \mathbb{R}^m$, $\underline{y} \in \mathbb{R}^p$, $\underline{x} \otimes \underline{y} \in \mathbb{R}^{mp}$

vectors are underlined

$$\underline{x} \otimes \underline{y} = \begin{bmatrix} x_1 \underline{y} \\ x_2 \underline{y} \\ \vdots \\ x_m \underline{y} \end{bmatrix}$$

Kronecker-Matrix-Vector Oracle Model

Before: $A \in \mathbb{R}^{d \times d}$. We can compute $A\underline{x}$ for any $\underline{x} \in \mathbb{R}^d$

Now: $A \in \mathbb{R}^{d^K \times d^K}$. We can compute $A\underline{x}$ for any $\underline{x} = \underline{x}_1 \otimes \underline{x}_2 \otimes \cdots \otimes \underline{x}_K$

$\underline{x}_1, \cdots, \underline{x}_K \in \mathbb{R}^d$

Can we still solve linear algebra problems efficiently?

poly$(d, K, \frac{1}{\varepsilon})$?

Core Issue: d^K vs dK parameters

Trace Estimation

Estimate $tr(A)$ from few matvecs

Find $\tilde{t}$ such that

$$(1-\varepsilon)\, tr(A) \le \tilde{t} \le (1+\varepsilon)\, tr(A)$$

Classically,

Hutchinson's Estimator uses $\Theta(1/\varepsilon^2)$ matvecs

Hutch++ uses $\Theta(1/\varepsilon)$ matvecs

Variance Reduction

Our Contribution: Analyze Kronecker-Hutchinson

Hutchinson's Estimator can easily be made Kronecker

How many matvecs are needed for std dev $\leq \varepsilon \, tr(A)$?

Answer: $\ell = \Theta(\frac{3^k}{\varepsilon^2})$ are needed [Ahle et al. '20]

Further: Exact variance,

$O(\frac{2^k}{\varepsilon^2})$ for random rank-one matrices

$\Theta(\frac{2^k}{\varepsilon^2})$ needed for complex matvecs

The matrices where $exp(K)$ matvecs are need are either:

(1) Low Rank or (2) $A = A_1 \otimes A_2 \otimes \cdots \otimes A_k$

We can compute $tr(A)$ exactly efficiently in both cases

Hutchinson's Estimator is bad at Kronecker Trace Estimation

Hutchinson's Estimator

$H_\ell(A) := \frac{1}{\ell}\sum_{i=1}^{\ell} g_i^T A g_i$ where $g_i \sim N(\underline{0}, I)$

Let $\hat{A} = \frac{1}{2}(A + A^T)$ be the symmetrized A

$$= U \underbrace{\Lambda}_{\text{eigenvalues}} U^T$$

Then $g^T A g = g^T \hat{A} g \overset{dist}{=} g^T \Lambda g = \sum_i \lambda_i g_i^2$

So $\mathbb{E}[g^T A g] = tr(\hat{A})$

$= tr(A)$

$Var[g^T A g] = 2\|\hat{A}\|_F^2$

$\leq 2\|A\|_F^2$

$\leq 2(tr(A))^2$ for PSD A

Kronecker Hutchinson

Let $\underline{x} = \underline{x}_1 \otimes \cdots \otimes \underline{x}_k$ for $\underline{x}_i \overset{iid}{\sim} N(\underline{0}, I)$

Sample $\underline{x}^T A \underline{x}$

What is $\mathbb{E}[\underline{x}^T A \underline{x}]$? $Var[\underline{x}^T A \underline{x}]$?

Problem: $\underline{x}$ is not rotationally inveriant!

Solution: "Extract" one $\underline{x}_i$ at a time

Shapes Ink to Shape Ink to Text Ruler Math

Extraction

$$\mathbf{x} \otimes \mathbf{y} = (\mathbf{I}_n \otimes \mathbf{y})\mathbf{x} = (\mathbf{x} \otimes \mathbf{I}_m)\mathbf{y}$$

$K=2$

$$(\underline{x}_1 \otimes \underline{x}_2)^T A \, (\underline{x}_1 \otimes \underline{x}_2)$$

$$\underline{x}_1^T \underbrace{(\underline{x}_1 \otimes \underline{x}_2)^T A \, (I \otimes \underline{x}_2)}_{} \, \underline{x}_1$$

$$\underline{x}_1^T M \underline{x}_1$$

$$\mathop{E}_{\underline{x}_1}[\underline{x}^T A \underline{x}] = tr(M) = \underline{x}_2^T tr_1(A)\, \underline{x}_2$$

Partial Trace of A

$$\mathop{Var}_{\underline{x}_1}[\underline{x}^T A \underline{x}] = 2\left\|\tfrac{1}{2}(M + M^T)\right\|_F^2$$

Partially Symmetrize A

Core Theorem

Let $\bar{A} = \frac{1}{2^K} \sum_{\nu \subseteq \{1,\cdots,K\}} A^{T_\nu}$ be the average of all partial symmetrizations of A.

Then
$$\mathrm{Var}[\underline{x}^T A \underline{x}] = \sum_{S \subseteq \{1,\cdots,K\}} 2^{K-|S|} \|\mathrm{tr}_S(A)\|_F^2$$
$$\leq \sum_{S \subseteq \{1,\cdots,K\}} 2^{K-|S|} \|\mathrm{tr}_S(A)\|_F^2$$
$$\leq 3^K (\mathrm{tr}(A))^2 \qquad \text{for PSD } A$$

Proof: Lots of induction

Conclusions

- Introduce Kron-Mat-Vec Complexity
- Variance of Kron-Hutchinson Algo
- Surprising connection to Partial Trace, Partial Transpose
- $\Omega\left(\frac{3^K}{\varepsilon^2}\right)$ lower bound when A is the all-ones matrix
- $\Omega\left(\frac{\sqrt{K}}{\varepsilon^2}\right)$ lower bound against all Kron-Mat-Vec Algos

Double-Sparse Model

Suppose $A \in \mathbb{R}^{d \times d}$ on hard drive, but d is huge

You cannot store $\underline{x} \in \mathbb{R}^d$ in memory (RAM)

But cols of A are c-sparse

E.g. banded matrices

If $\underline{x}$ is s-sparse, then $A\underline{x}$ is cs-sparse

Allower $O(d)$ time but $o(d)$ memory

[Jonathan Weare, Robert Webber]