Andreas Stathopoulos
Deflation as variance reduction for computing the trace of the inverse

Department of Computer Science
Box 8795
College of William and Mary
Williamsburg
VA 23187-8795
andreas@cs.wm.edu
Arjun Gambhir
Kostas Orginos

A useful statistical measure is the diagonal or the trace of the inverse of a matrix. This has myriad applications from quantum mechanics and Lattice QCD to uncertainty quantification and data mining. Most commonly the matrix A or its Hermitian part $A + A^T$ would be Hermitian positive definite and very large and sparse (large dimension N). We are interested in cases where a direct factorization of the matrix is not possible.

For such large matrices, the current state-of-the-practice is to use a Monte Carlo method known as Hutchinson's which is based on the fact that the trace can be obtained as the expected value of $\operatorname{tr}(A^{-1}) = E(x^T A^{-1} x)$. Therefore, a simple averaging algorithm can be implemented where random vectors are chosen and linear systems $A^{-1}x$ are solved with an iterative method. The method converges as $O(\operatorname{Var}/\sqrt{n})$, where the variance of the estimator, Var, is known to be minimized by choosing x with random Rademacher entries $\{-1, 1\}$ (in the real case). This variance is given by

$$\frac{1}{2}\operatorname{Var}(tr(A)^{-1}) = \|A^{-1}\|_F^2 - \sum_{i=1}^{N}(A_{ii}^{-1})^2.$$

The two computational bottlenecks are (a) for ill conditioned systems iterative methods may take many steps and (b) thousands of linear systems are needed just to get two digits of accuracy in the trace. In earlier work, we showed how approximate eigenvectors and eigenvalues can be used to deflate, and thus remove the ill conditioning in linear systems with multiple right hand sides. We also developed Hierarchical Probing, a method to systematically produce the vectors x in Hutchinsons's method so that variance is reduced significantly. In this talk we focus on the effect of deflation on reducing the variance, and how it can complement Hierarchical Probing to achieve substantial further improvements.

Assume for simplicity that we want to find $\text{tr}(A)$ (not $\text{tr}(A^{-1})$). Consider its singular value decomposition, $A = A_D + A_R = U_1\Sigma_1 V_1^T + U_2\Sigma_2 V_2^T$ where $A_D = U_1\Sigma_1 V_1^T$ corresponds to the largest k singular triplets of A. Clearly, the trace of A_D is trivial to compute direclty. Thus, using the appropriate projector, we can apply Hutchinson's method on the deflated $A_R = U_2\Sigma_2 V_2^T$, hoping this has lower variance. We prove that

$$\frac{1}{2}\text{Var}(tr(A_R)) = \sum_{m=k+1}^{N} \sigma_m^2 - \sum_{m=k+1}^{N}\sum_{l=k+1}^{N} \sigma_m\sigma_l\Delta_{ml},$$

where $\Delta_{ml} = \sum_{i=1}^{N} \bar{u}_{im}v_{im}u_{il}\bar{v}_{il}$, $m,l = 1,\dots,N$. To achieve variance reduction we need $\text{Var}(tr(A_R)) < \text{Var}(tr(A))$. Contrary to low rank matrix approximations, the term subtracting the double sum implies that deflation may not achieve this. In fact, it is easy to come up with such examples. The presence of Δ_{ml} in the formula complicates a precise characterization of when we can expect variance reduction. A crude analysis shows that reduction is guaranteed if the singular spectrum decreases geometrically as $\sigma_{i+1} = 2^{-i}\sigma_1$. However, this is pessimistic.

To bypass the complications caused by Δ_{ml}, we assume that U and V are standard random unitary matrices, i.e., distributed with the Haar probability measure. For large size matrices this is a reasonable assumption, and it is also theoretically supported for matrices in Lattice QCD. Based on an analysis of the moments of the elements of these matrices, we have derived an expression for the expected variance of the deflated estimator. Define the mean and the variance of the $N - k$ singular values of A_R, $\mu_k = \frac{1}{N-k}\sum_{m=k+1}^{N}\sigma_m$, and $V_k = \frac{1}{N-k}\sum_{m=k+1}^{N}(\sigma_m - \mu_k)^2$, respectively. Then, for non-Hermitian matrices it holds

$$\frac{1}{2}E(\text{Var}(t(A_R))) = (N-k)(1 - \frac{1}{N})(V_k + \mu_k^2)$$

and for Hermitian matrices,

$$\frac{1}{2}E(\text{Var}(t(A_R))) = (N-k)\left(V_k\frac{N}{N+1} + \mu_k^2\frac{k}{N+1}\right).$$

This theory facilitates an accurate prediction of the outcome of deflation based solely on the variance and the expectation of the undeflated singular values. For example, we can numerically confirm that it is sufficient that the singular values decrease at a rate faster than linear. In practice, the singular spectrum in Lattice QCD matrices decreases geometrically so deflation is expected to help. Surprisingly, non-Hermitian matrices benefit more than Hermitian matrices with the same singular value distribution. We have also observed that our model is extremely robust on general, non-random matrices.

The second contribution of the paper is that we observed a synergistic action between deflation and hierarchical probing. Specifically, we analyzed the geometric

effect of the two methods and observed that while hierarchical probing removes, by design, the error in neighboring lattice nodes at increasing distances, deflation is not good at that but excels in removing error in long lattice distances. In our experiments on a very large Lattice QCD problem, while deflation alone gave a small improvement, Hierarchical Probing reduced variance by an order of magnitude, and combining the two methods gave us an additional order of magnitude; an overall speedup of more than 200.