Reply to Report 1a on manuscript LQ17642

Matteo Smerlak[1,1], Anton Zadorin[1], Maseim Kenmoe[1,2], and Ronald Kriemann[1]

[1]Max Planck Institute for Mathematics in the Sciences, Germany
[2]University of Dschang, Cameroon

July 3, 2020

General remarks

We thank the referee for his/her work reviewing our paper.

The report asks us to give further details on (i) numerical stability and (ii) computational performance of our method. We feel that (i) is rather academic, because we prove the convergence of fixed-point iteration under explicit conditions on θ and λ which exclude ill-conditioned eigenvalues—and, numerically, DPT is really just matrix multiplication. Nonetheless, we have added Fig. S4-S5 as well as SI G supporting the claim that DPT is both accurate and numerically stable (within its domain of convergence). Regarding (ii), we have added a timing comparison with LAPACK on large dense matrices (up to size $N = 131,072$), showing a 30-90x speedup (Fig. 4). We also added a link to a public repository containing our code for reproducibility. Edits in the main text are highlighted in blue.

We note that the referee did not offer comments on the physical or theoretical aspects of our method. These are, however, the main focus of our Letter, in line with PRL's scope as a physics journal. We intend to give more details on the implementation of DPT on parallel architectures and other computational aspects in a more suitable venue at a later stage.

Moreover, we observe that some of the referee's comments (on "unphysical" solutions or "subjective [. . .] boundary conditions") appear to refer to algorithms other than DPT. We emphasize that our paper gives explicit conditions under which DPT provably converges to a complete set of eigenvectors, without any ambiguity or user-driven choices. Surely our algorithm cannot be criticized for the shortcomings of its competitors.

Response to specific comments

Comment 1

The proposed dynamic perturbation theory (D-PT) is not general. It is formulated only for matrices with non-degenerate (simple) eigenvalues.

This comment suggests that the exclusion of degenerate eigenvalues is a peculiar deficiency of our method. This is misleading in several ways:

- Degenerate eigenvalues present a challenge for any eigenvalue algorithm because they correspond to singularities in the complex λ plane. (In the language of numerical analysis, degenerate spectra have infinite condition number.) Moreover, many important results in matrix perturbation theory (e.g. the Bauer-Fike theorem) assume simple eigenvalues, as do many physics texts on quantum mechanical perturbation theory.

- Algorithms such as non-degenerate RS-PT (which the simple eigenvalues of the unperturbed matrix to be simple) are not generally seen as defective or "not general". Instead, it is usually understood that degenerate eigenvalues must be treated separately, for instance by diagonalizing the perturbation Δ in the degenerate eigenspace (a procedure called "degenerate perturbation theory" in quantum mechanics). The same is true of our method.
- We prove the convergence of D-PT for any matrix $M = D + \Delta$ such that $\|\theta\|\|\Delta\| < (3-2\sqrt{2})$. This is as "general" a result as one can hope for in the context of perturbation theory. A similar result for RS-PT (see e.g. (Kato, 1966)) is usually considered a "general" theorem of perturbation theory.
- From a strictly mathematical perspective, degenerate spectra are not generic: they form a nowhere dense, zero-measure subset of the space of square matrices. While this comment is moot in physics applications (where symmetries almost always induce degeneracies), it is relevant in other fields where perturbative eigenproblems play an important role, e.g.in molecular evolutionary theory (the application which motivated this work).

The reported illustrations are with highly specialized trivial real symmetric matrices (2 × 2 or 3 × 3) or with a still simple, small size 100 × 100 real non-symmetric matrix with random numbers (as the elements of M) distributed uniformly in the interval [-1,1].

This was already inaccurate in the original submission (Fig. S1 and former Fig. 4 considered sparse random matrices of size up to $10^4 \times 10^4$ and $10^6 \times 10^6$ respectively). The present version includes further timings and accuracy checks for dense, complex matrices without any special structure (other than being near-diagonal) of size up to $10^5 \times 10^5$, see comment 5 below.

No complex matrices are exemplified in the illustrations of the D-PT.

The new tests of numerical accuracy of DPT in Fig. S4 are with non-Hermitian complex matrices. Generally speaking, moving from real to complex matrices is completely innocent, as DPT relies only on matrix multiplication.

Stability of the eigensolutions in the D-PT critically depends on symmetry properties of M. This could impact detrimentally on stability of eigensolutions as incorporating good initial conditions becomes a notable problem.

This comment is unclear to us, and appears to conflate the following distinct statements, neither of which is specifically about DPT: *(i)* unlike general matrices, Hermitian matrices are always well-conditioned and are subject to strong stability results (realness of eigenvalues, Weyl's theorem on perturbations of eigenvalues) *(ii)* choosing good initial conditions is important for certain iterative eigenvalue algorithms such as Lanczos or Arnoldi. We answer these distinct comments as follows:

- *(i)* Although important in other contexts, the issue of condition numbers for non-Hermitian matrices is of little relevance to the physics audience of PRL. In physics, M is almost always the Hamiltonian operator of a quantum system, and is therefore Hermitian by definition. Given the space constraint of a Letter, we do not deem it appropriate to address this topic here.
- *(ii)* Unlike other iterative eigenvalue algorithms which leave the starting vector unspecified, our method provides a definite initial condition: the identity matrix $A^{(0)} = I$,

corresponding to approximating the eigenvalues of M by the diagonal elements of D. There is no ambiguity regarding "good initial conditions" in our method other than the partitioning of M itself.

Besides, there is no particular intrinsic dependence of the stability on any symmetry of the matrix. For example, the condition $\|\theta\|\|\Delta\| < (3 - 2\sqrt{2})$, which guarantees the convergence, does not rely on any symmetry considerations.

Comment 2

The Rayleigh-Schrödinger perturbation theory (RS-PT) fails for degenerate spectra. Nevertheless, the perturbation formalism of quantum mechanics has overcome this drawback by resorting to the brilliantly Padé-approximant-based resummation of the Brillouin-Wigner perturbation theory (BW-PT) for operators and/or matrices with degenerate spectra.

- "Brilliantly Padé-approximant-based resummation of the Brillouin-Wigner perturbation theory" is not how degenerate eigenvalues are commonly dealt with in quantum mechanics. Instead, the standard procedure is to diagonalize the perturbation Δ in the degenerate eigenspace and use the corresponding eigenvectors as starting points for RS-PT.
- While it is true that BW-PT does not require D to have distinct diagonal elements (unlike RS-PT), the comparison with D-PT is not pertinent: BW-PT is an *implicit* method which does not by itself provide numerical approximations for the eigenpairs of M. Our method, by contrast, is explicit: it provides formulas approximating the eigenpairs of M to arbitrary order in λ.
- Finally, the introduction recalls that Padé approximants are useful to improve the convergence of (RS- or BW-) PT—as are many other techniques. We are unsure why the referee singles out this particular method here, or what makes it more "brilliant" than other techniques in his/her eyes. However, we have added a reference to Brillouin-Wigner perturbation theory in the introduction.

Comment 3

The D-PT is not practical. The reason is in solving not only one, but many non-linear algebraic equations, say L in total (L = L1 , L2 , . . . , Lmax). Each of these L's gives a set of eigensolutions.

As explained in the main text and SI, D-PT solves as many algebraic equations as eigenpairs are requested. For a complete set of eigenpairs, that is N algebraic equations in $\mathbb{C}^N$. What makes this "not practical"? This is as it must be. Note that although each equation has a form $z = F_n(z)$ and, indeed, generically has N different solutions, D-PT, when converges, nevertheless, picks only one of them for each equation, providing a single solution A.

The non-linear equations necessitate the starting values to initiate the iteration process. The boundary conditions to each of the algebraic equations in the D-PT are unknown. Guessing the initial values is subjective, to say the least.

The initial (not "boundary") conditions are not "unknown": given a partition of M as $D + \lambda\,\Delta$, DPT uses a non-ambiguous set of initial conditions for fixed-point iteration, namely $A^{(0)} = I$. The only "subjective" element in this procedure is the partitioning itself—but this is a feature of any perturbation theory, by definition. The need for starting values has nothing to do with the "non-linear" nature of the equations.

Some surmised/estimated, insufficiently good starters to nonlinear equations can lead to the wrong results, irrespective of the status of the convergence issue (convergence to the wrong result is not infrequent for implicit non-linear algebraic equations).

Once again, we *prove* convergence to the right results under the condition $\|\theta\|\|\Delta\| < (3-2\sqrt{2})$ with the explicit (not "surmised/estimated") initial condition $A^{(0)} = I$. Here too, the referee appears to be referring to issues of *other* iterative methods for the eigenvalue problems (e.g. the Lanczos or Arnoldi algorithms), and not to our proposed algorithm.

Comment 4

In practice, the elements of M are not ideal entries, i.e. they enter the analysis with their intrinsic inaccuracies. Matrix elements can come from some elaborate computations with finite arithmetics (computational round-off errors) or they can be some empirical data stemming from experimental measurements (inevitably contaminated with noise–systematic, random, etc).

This is true, but since we never require or mention "ideal entries", we do not see the relevance of this comment.

For such matrices routinely encountered in practice, the D-PT would give some non-unique eigensolutions.

This claim is unsubstantiated and incorrect. Once again, we prove convergence of DPT to a unique attracting fixed point containing a complete set of eigenvectors of M whenever $\|\theta\|\|\Delta\| < (3-2\sqrt{2})$. Unless errors in M violate this bound, uniqueness of eigensolutions is guaranteed.

Stated equivalently, there is no guarantee that the eigensolutions would not vary from one to another set of the eigensolutions. Some of the eigensolutions might be found in one or more subsets of all the L sets of the eigensolutions. However, there could also be some spurious eigensolutions that would differ from one to another set in $\{L1,L2,\ldots,Lmax\}$. A key procedure is lacking in the D-PT for separating the unphysical from physical eigensolutions.

Unlike other algorithms, DPT does not generate "spurious" or "unphysical solutions". Our method provably converges to a complete set of eigenpairs under the condition recalled above; other solutions of $A = F(A)$ are not seen by DPT, as explained in SI A. No "key procedure" is lacking.

Comment 5

Some of the existing generic well-known eigenvalue solvers can expediently extract millions of eigen- values from general non-Hermitean complex matrices. Any newly proposed eigenvalue solver (perturbative or nonperturbative alike) should be benchmarked on at least modestly sized non-Hermitean complex matrices (e.g. 1000 × 1000 or so) with elements corrupted by 5-20% random Gaussian-distributed noise (to mimic the experimental data) in a fixed interval.

As already noted, Fig. S1 and 4 examine sparse matrices of size up to $10^4 \times 10^4$ and $10^6 \times 10^6$ respectively. The new figures address performance and numerical stability (two distinct issues) as follows:

- *Performance*: As detailed in SI G, Fig. 4 now compares the performance of DPT on a dual AMD EPYC 7702 with 128 CPU cores using Intel MKL and a NVidia V100 GPU using MAGMA (Dongarra et al., 2014), in single precision. We obtained 30-90x speedups over these (highly optimized) reference routines.

- *Stability*: We added two new supplementary figures. Fig. S4 considers the numerical accuracy of DPT vs. that of LAPACK for non-Hermitean complex matrices of size 2048 with varying $\lambda \in \left[10^{-4}, 0.5\right]$ computed in double precision. We find that DPT is at least one order of magnitude more accurate than Intel MKL's **geevx** (and up to orders of magnitude in single precision, where stability issues are magnified). Fig. S5 illustrates the forward stability of the eigenvalues E of the symmetric matrix defined by Eq. (6) with $N = 1000$ after entries were perturbed by a normal distribution with mean $\mu \in \left[10^{-7}, 10^{-2}\right]$ and deviation 2μ.

References

Perturbation theory in a finite-dimensional space. (1966). In *Perturbation theory for linear operators* (pp. 62–126). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-12678-3_2

Accelerating Numerical Dense Linear Algebra Calculations with GPUs. (2014). In *Numerical Computations with GPUs* (pp. 3–28). Springer International Publishing. https://doi.org/10.1007/978-3-319-06548-9_1