The Journal of Open Source Software

Low scaling BSE implementation in the exciting code

Benedikt Maurer [1*] **and Claudia Draxl** [1*]

1 Department of Physics and CSMB Adlershof, Humboldt-Universit"at zu Berlin, Zum Großen Windkanal 2, D-12489 Berlin, Germany * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Bonan Zhu

Reviewers:

- @jjkas
- @ruiyiQM
- @EderGio

Submitted: 03 February 2025
Published: unpublished

Summary

Solving the Bethe-Salpeter Equation (BSE) is essential for understanding excited-state systems, but often challenging to converge or even computationally prohibitive for large systems. We implement a matrix-free BSE solver leveraging Interpolative Separable Density Fitting (ISDF) to interpolate electron-hole interaction kernels together with the Lanczos algorithm for diagonalization, avoiding full matrix setup. The scaling of our implementation is bounded by $\mathcal{O}(N_o N_u N_\mathbf{k} \log N_\mathbf{k})$ and is thus a massive improvement over methods that set up the whole matrix, scaling with at least $\mathcal{O}((N_o N_u N_\mathbf{k})^3)$, where N_o and N_u are the numbers of occupied and unoccupied states, respectively, and $N_\mathbf{k}$ is the number $\mathbf{k}$-points.

Theoretical background

The Bethe-Salpeter Equation (BSE) within many body perturbation theory (MBPT) provides the state-of-the-art frame work for describing light-matter interaction. In particular, it is used to obtain optical absorption spectra, including the effects of excitons, which are bound electron-hole states. By expanding the electron-hole wavefunctions in the transition basis, solving the BSE can be reduced to a Schr"odinger like equation. Setting up and diagonalizing the Bethe-Salpeter Hamiltonian (BSH) are the computationally expensive tasks(Vorwerk et al., 2019). The BSH is given as

$$H^{BSH} = D + \gamma V - W\,,$$

where $\gamma = 2$ gives the spin-singlet and $\gamma = 0$ spin-triplet channel, respectively. The diagonal term D is given by the differences of the one-particle energies of the occupied (o) and unoccupied (u) states:

$$D_{ou\mathbf{k},o'u'\mathbf{k}'} = (\varepsilon_{u\mathbf{k}} - \varepsilon_{o\mathbf{k}})\delta_{oo'}\delta_{uu'}\delta_{\mathbf{k}\mathbf{k}'}\,.$$

The matrix elements of the repulsive exchange interaction V and the attractive screened Coulomb interaction W are calculated by solving integrals of the form

$$V_{ou\mathbf{k},o'u'\mathbf{k}'} = \int d^3r \int d^3r' \frac{u_{o\mathbf{k}}(\mathbf{r}) u^*_{u\mathbf{k}}(\mathbf{r}) u^*_{o'\mathbf{k}'}(\mathbf{r}') u_{u'\mathbf{k}'}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,,$$

$$W_{ou\mathbf{k},o'u'\mathbf{k}'} = \int d^3r \int d^3r' u^*_{u\mathbf{k}}(\mathbf{r}) u_{u'\mathbf{k}'}(\mathbf{r}) W_{\mathbf{k}-\mathbf{k}'}(\mathbf{r},\mathbf{r}') u_{o\mathbf{k}}(\mathbf{r}') u^*_{o'\mathbf{k}'}(\mathbf{r}')\,,$$

where $W_{\mathbf{k}-\mathbf{k}'}(\mathbf{r},\mathbf{r}')$ is the statically screened Coulomb potential. $u_{i\mathbf{k}}(\mathbf{r})$ is the periodic part of the one-particle wavefunction of state i at the reciprocal lattice point $\mathbf{k}$. In general, we cannot see, which of the matrix elements will be zero, thus we need to compute all of them. Therefore, setting up the full BSH scales with $\mathcal{O}(N_o^2 N_u^2 N_\mathbf{k}^2)$ and diagonalizing it directly with $\mathcal{O}(N_o^3 N_u^3 N_\mathbf{k}^3)$.

The primary scaling bottleneck stems from evaluating the matrix elements of the interaction kernels. To mitigate this, we reformulate the wavefunction products using Interpolative Separable Density Fitting (ISDF) (Lu & Ying, 2015). Specifically, we approximate these products on a discrete real-space grid, $\{\mathbf{r}\}$, by expressing them as superpositions of values evaluated on a smaller interpolation grid, $\{\mathbf{r}_\mu\} \subset \{\mathbf{r}\}$:

$$u^*_{i\mathbf{k}}(\mathbf{r})u_{j\mathbf{k}'}(\mathbf{r}) \approx \sum_{\mu=1}^{N_\mu} \zeta_\mu(\mathbf{r})u^*_{i\mathbf{k}}(\mathbf{r}_\mu)u_{j\mathbf{k}}(\mathbf{r}_\mu)\,,$$

where N_μ is the number of interpolation points, and $\zeta_\mu(\mathbf{r})$ are the expansion coefficients. Due to the tensor product structure of $u^*_{i\mathbf{k}}(\mathbf{r})u_{j\mathbf{k}'}(\mathbf{r})$, $\zeta_\mu(\mathbf{r})$ can be computed efficiently, and the scaling is bounded by $\mathcal{O}(N_\mu^3)$ (Hu et al., 2017). We observe that we can always choose $N_\mu \ll N_o N_u N_\mathbf{k}$, thus ISDF is never a bottleneck. The interpolation points are computed efficiently with centroidal Voronoi tessellation within $\mathcal{O}(N_\mu N_r)$ (Dong et al., 2018).

Inserting ISDF in the equations for the interaction kernels yields

$$V_{ou\mathbf{k},o'u'\mathbf{k}'} \approx \frac{1}{N_k^2}\sum_{\mu=1}^{N_\mu^V}\sum_{\nu=1}^{N_\mu^V} u_{o\mathbf{k}}(\mathbf{r}^V_\mu)u^*_{u\mathbf{k}}(\mathbf{r}^V_\mu)\tilde{V}_{\mu\nu}u^*_{o'\mathbf{k}'}(\mathbf{r}^V_\nu)u_{u'\mathbf{k}'}(\mathbf{r}^V_\nu),$$

$$W_{ou\mathbf{k},o'u'\mathbf{k}'} \approx \frac{1}{N_k^2}\sum_{\mu=1}^{N_\mu^{W_u}}\sum_{\nu=1}^{N_\mu^{W_o}} u^*_{u\mathbf{k}}(\mathbf{r}^{W_u}_\mu)u_{u'\mathbf{k}'}(\mathbf{r}^{W_u}_\mu)\tilde{W}_{\mu\nu,\mathbf{k}-\mathbf{k}'}u_{o\mathbf{k}}(\mathbf{r}^{W_o}_\nu)u^*_{o'\mathbf{k}'}(\mathbf{r}^{W_o}_\nu)\,,$$

where we have shifted the integration from the wavefunction pairs to the interpolation coefficients such that

$$\tilde{V}_{\mu\nu} = \int_{\Omega^l\times\Omega^l} dr dr' \zeta^{*V}_\mu(\mathbf{r})V(\mathbf{r},\mathbf{r}')\zeta^V_\nu(\mathbf{r}')\,,$$

$$\tilde{W}_{\mu\nu,\mathbf{k}-\mathbf{k}'} = \int_{\Omega^l\times\Omega^l} dr dr' \zeta^{*W_u}_\mu(\mathbf{r})W_{\mathbf{k}-\mathbf{k}'}(\mathbf{r},\mathbf{r}')\zeta^{W_o}_\nu(\mathbf{r}')\,.$$

Note that there are three different wavefunction pairings, i.e. $u^*_{o\mathbf{k}}(\mathbf{r})u_{u\mathbf{k}}(\mathbf{r})$ for the exchange kernel (V) and $u^*_{o\mathbf{k}}(\mathbf{r})u_{o'\mathbf{k}'}(\mathbf{r})$, $u^*_{u\mathbf{k}}(\mathbf{r})u_{u'\mathbf{k}'}(\mathbf{r})$ for the screened kernel W. Each pairing requires a separate ISDF calculation, denoted by the superscripts V, W_o, and W_u. Since the number of combinations may vary, the number of interpolation points required may also vary. Reformulating the matrix elements in this way alone does not improve scaling with respect to the system size. To achieve this, we combine it with an iterative solver, here with the Lanczos algorithm. This class of algorithms constructs an approximation to the eigenvalues and eigenvectors by iteratively applying matrix-vector multiplications. Applying the interaction kernels in their interpolated forms to a vector X of dimension $N_o N_u N_\mathbf{k}$ allows efficient computation by rearranging the summations to exploit a separable structure of the kernels. For the exchange kernel we get

$$[V\cdot X]_{ou\mathbf{k}} = \frac{1}{N_\mathbf{k}}\sum_{\mu=1}^{N_\mu^V} u^*_{u\mathbf{k}}(\mathbf{r}^V_\mu)u_{o\mathbf{k}}(\mathbf{r}^V_\mu)\left\{\sum_{\nu=1}^{N_\mu^V}\tilde{V}_{\mu\nu}\left[\sum_{\mathbf{k}'}\left(\sum_{u'}u_{u'\mathbf{k}'}(\mathbf{r}^V_\nu)\left[\sum_{o'}u^*_{o'\mathbf{k}'}(\mathbf{r}^V_\nu)\cdot X_{o'u'\mathbf{k}'}\right]\right)\right]\right\},$$

where we first compute the sums over o', u', and $\mathbf{k}'$ to get a term that depends only on $\mathbf{r}^V_\nu$ with a complexity of $\mathcal{O}(N_\mu^V(N_o N_u N_\mathbf{k} + N_u N_\mathbf{k})$. The remaining sums can be computed with $\mathcal{O}((N_\mu^V)^2 N_\mu^V N_o N_u N_\mathbf{k})$, so the complexity of computing $V\cdot X$ is bounded by $\mathbf{O}((N_\mu^V)^2 + N_\mu N_o N_u N_\mathbf{k})$. Applying the screened kernel to X, after reordering the sums, we get

$$[W\cdot X]_{ou\mathbf{k}} = \frac{1}{N_k}\sum_{\nu=1}^{N_\mu^{W_o}} u_{o\mathbf{k}}(\mathbf{r}^{W_o}_\nu)\left\{\sum_{\mu=1}^{N_\mu^{W_u}} u^*_{u\mathbf{k}}(\mathbf{r}^{W_u}_\mu)\left[\sum_{\mathbf{k}'}\tilde{W}_{\mu\nu,\mathbf{k}-\mathbf{k}'}\left(\sum_{u'}u_{u'\mathbf{k}'}(\mathbf{r}^{W_u}_\mu)\left[\sum_{o'}u^*_{o'\mathbf{k}'}(\mathbf{r}^{W_o}_\nu)X_{o'u}\right.\right.\right.\right.$$

Here we exploit the separable structure of the decomposition so that the terms depending on $\mathbf{k}$ and $\mathbf{k}'$ are on the left and right of $\tilde{W}_{\mu\nu,\mathbf{k}-\mathbf{k}'}$. The evaluation of the two innermost sums over o' and u' to $A^{\mathbf{k}'}_{\mu\nu}$ scales with $\mathcal{O}(N_\mu^{W_u} N_o N_u N_\mathbf{k} + N_\mu^{W_o} N_\mu^{W_u} N_u N_\mathbf{k})$. Then the sum over $\mathbf{k}'$ reads as a discrete convolution

$$\sum_{\mathbf{k}'} W_{\mathbf{k}-\mathbf{k}'} A^{\mathbf{k}'}_{\mu\nu} \,,$$

which can be efficiently evaluated with fast Fourier transforms simultaneously within the $\mathcal{O}(N\mathbf{k} \log N\mathbf{k})$ scaling for each $\mu\nu$ pair. The remaining summations scale with $\mathcal{O}(N_\mu^{W_o} N_\mu^{W_u} N_u N_\mathbf{k})$. So the complexity for the computation of $W \cdot X$ is bounded by $\mathcal{O}(N_\mu^{W_u} N_o N_u N_\mathbf{k} + N_\mu^{W_o} N_\mu^{W_u} N_u N_\mathbf{k} + N_\mu^{W_o} N_\mu^{W_u} N_\mathbf{k} \log N_\mathbf{k})$.

Statement of need

Due to the unfavorable scaling of solving the BSE directly, many interesting problems such as complex materials with large unit cells or systems requiring a dense Brioullin-zone sampling are not feasible. Even though Henneke and coworkers (Henneke et al., 2020) have already described the new algorithm and demonstrated the scaling improvement, an easy-to-use and scalable implementation was still missing. We have implemented and fully integrated this approach in the existing BSE infrastructure of the all-electron, full-potential package exciting (Gulans et al., 2014). Users can now easily choose which algorithm they prefer to use and have the full suite of exciton analysis implemented in exciting at hand.

Results

To demonstrate that our implementation does indeed scale as proposed, we run BSE calculations for increasing $N_\mathbf{k}$ using the direct and the new implementation. In Fig.(Figure 1) we show the wall times for the example of diamond. The new algorithm massively outperforms the direct, and the speed up increases more than linearly with $N_\mathbf{k}$.

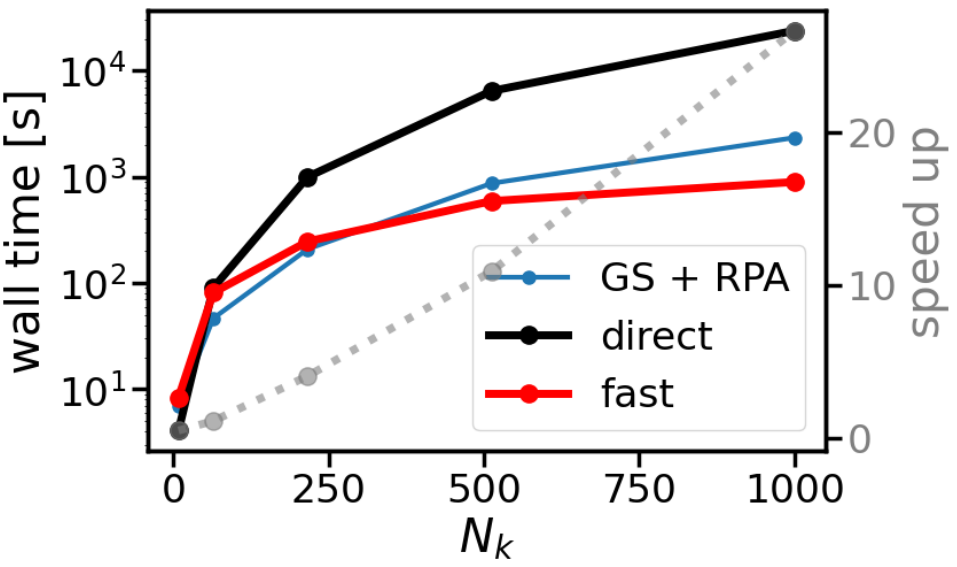

Figure 1: Runtimes of the RPA screening (blue) and the direct (black) and new (red) BSE implementations as a function of the number of $\mathbf{k}$-points, $N_\mathbf{k}$. The speedup of the algorithm is shown by the gray line.

In Fig.(Figure 2) we show the differences in exciton binding energies obtained from the direct implementation, which sets up and diagonalizes the full BSH, and the new implementation with fixed numbers of interpolation points: $N_\mu^V = 202$, $N_\mu^{W_o} = 322$, and $N_\mu^{W_u} = 360$ for the ISDF. Additionally, we present the spectral similarity, defined here as the area between the two curves. Even with these fixed numbers of interpolation points, the results of the new algorithm converge to those of the direct algorithm as N_k increases. This demonstrates that the new implementation yields results equivalent to the direct solution of the BSE but with significantly reduced computational time. Consequently, it enables more precise calculations and facilitates the study of more complex problems.

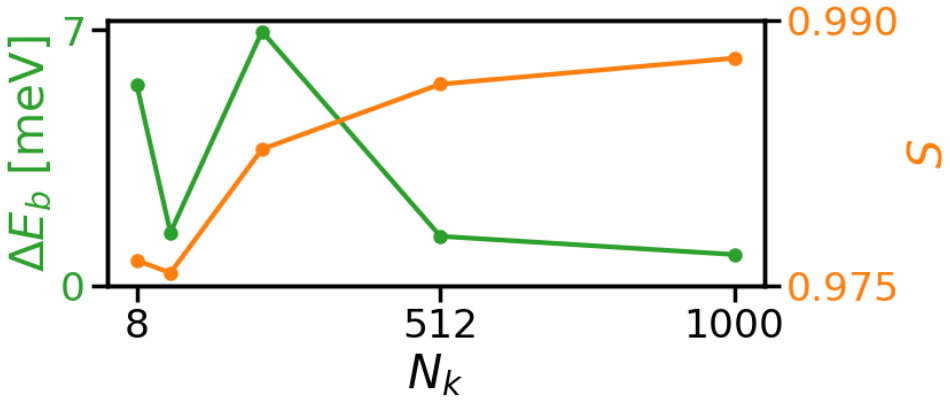

Figure 2: Difference of the exciton binding energy and spectral similarity between the new method and the direct method as functions of N_k for the case of diamond.

References

Dong, K., Hu, W., & Lin, L. (2018). Interpolative separable density fitting through centroidal voronoi tessellation with applications to hybrid functional electronic structure calculations. *Journal of Chemical Theory and Computation*, *14*, 1311–1320. https://doi.org/10.1021/acs.jctc.7b01113

Gulans, A., Kontur, S., Meisenbichler, C., Nabok, D., Pavone, P., Rigamonti, S., Sagmeister, S., Werner, U., & Draxl, C. (2014). Exciting: A full-potential all-electron package implementing density-functional theory and many-body perturbation theory. *Journal of Physics: Condensed Matter*, *26*, 363202. https://doi.org/10.1088/0953-8984/26/36/363202

Henneke, F., Lin, L., Vorwerk, C., Draxl, C., Klein, R., & Yang, C. (2020). Fast optical absorption spectra calculations for periodic solid state systems. *Communications in Applied Mathematics and Computational Science*, *15*, 89–113. https://doi.org/10.2140/camcos.2020.15.89

Hu, W., Lin, L., & Yang, C. (2017). Interpolative separable density fitting decomposition for accelerating hybrid density functional calculations with applications to defects in silicon. *Journal of Chemical Theory and Computation*, *13*, 5420–5431. https://doi.org/10.1021/acs.jctc.7b00807

Lu, J., & Ying, L. (2015). Compression of the electron repulsion integral tensor in tensor hypercontraction format with cubic scaling cost. *Journal of Computational Physics*, *302*, 329–335. https://doi.org/10.1016/j.jcp.2015.09.014

Vorwerk, C., Aurich, B., Cocchi, C., & Draxl, C. (2019). Bethe–salpeter equation for absorption

and scattering spectroscopy: Implementation in the exciting code. *Electronic Structure*, *1*, 037001–037001. https://doi.org/10.1088/2516-1075/ab3123