Response to Reviewer 1

Summary Comment

The authors have addressed most all concerns. Just a few remain. If these can be well addressed, I would recommend publication.

Response:

We sincerely appreciate your valuable suggestions, which have greatly enhanced our manuscript. We now believe it is suitable for acceptance. Below, we provide point-by-point responses to your concerns.

Comment 1.1

p. 5: "with l, m, n being the angular momentum, the magnetic momentum, and the multiplicity of projectors, respectively" This would more accurately be stated as "where l and m are the azimuthal and magnetic quantum numbers, respectively, and n is the multiplicity of projectors."

Response 1.1:

Thank you! We've made the changes as requested.

Comment 1.2

p. 8: "To apply to the multiple k points cases, one simply expands the dimension of the pertinent X, by stacking K copies of X, each of the same size." Regarding this, in answer to the question, "is the minimization over the entire set of K copies (unlikely) or over each X_i separately?" the authors' response says (p. 3), "the minimization should be done over the entire set of K copies" However, X_i and X_j, $j \neq i$, are not orthogonal to each other. So, if the minimization is done over all X_l simultaneously, that minimization cannot be on the Stiefel manifold. To eliminate this considerable confusion, the authors should include multiple k points explicitly in Algorithm 1, Fig. 1, and Algorithm 2.

Response 1.2: Thank you for pointing out the possible confusion.

To clarify, the orthogonal constraint is imposed on X_{k_i} for each **k**-point indexed by k_i. Namely,

$$X_{k_i}^\dagger X_{k_i} = I_p,\ k_i = 1, 2, \cdots, K.$$

As discussed in the context of product of manifolds, each **k**-point corresponds to a submanifold. The bold symbol $\mathbf{X}$ is a collection of Xs, of size K, with each column X_i being orthogonal to other X_j. We have moved forward the introduction of the product of manifolds to make the writting more coherent and provided explanations to this. Essentially, a vectorized version of the single k point case is needed. Carrying the index **k** complicates the notation (also conflicts with the iteration index k), but with the explanation, we believe it is no more an issue and readers can extend to the multiple k points case easily.

Comment 1.3

- Table 1 has been added but not referenced in the text.
- Table 1 should include the Euclidean metric as well as the Canonical metric and it should be clarified in Algorithm 1 and throughout the text which metric is being used when the superscript "e" or "c" is omitted.
- Table 1 should have "grad $f(X)$", consistent with Algorithm 1, rather than "grad f."
- Table 1 should have "T_U(V)", consistent with Algorithm 1, rather than "T_U V."
- Table 1 should include "where $Y = R_X(U)$" to clarify the vector transport operation containing Y (if I've interpreted correctly).

Response 1.3: We deeply appreciate your effort on improving the manuscript.

- We in fact referenced it but missed the important notation "in Table" in front of it. We have changed it to "one needs to apply the corresponding linear algebra operations in Table 1".
- We added that "hence all superscripts e is omitted. "
- "grad f" has been changed to "grad $f(X)$.
- ""T_U V" has been changed to "T_U(V)".
- We have made the change to the caption to have "Note, $Y = R_X(U)$ in the vector transport formula."

Comment 1.4

Algorithm 2: How is the initial inverse Hessian approximation H_0 determined? This should be clarified in Algorithm 2 and/or text immediately preceding.

Response 1.4: Thank you for pointing out this omission.

Usually, there are several ways of choosing the initial inverse Hessian approximation H_0. The widely used one is the scaled identity matrix

$$H_0 = \gamma I \tag{1}$$

where γ is a positive scalar and I is the identity matrix. The most naïve choice would be $\gamma = 1$. Often it can also utilize gradient information such as $\gamma = \frac{y_0^T s_0}{y_0^T y_0}$, with initial step $s_0 = x_1 - x_0$ and the initial gradient difference $y_0 = \nabla f(x_1) - \nabla f(x_0)$. This choice can ensures H_0 approximates the inverse Hessian along the initial direction. For a given problem with approximate Hessian available, one can also use problem specific H_0.

- We have explained the common choices for H_0, which now reads " Meanwhile, one has to choose the initial inverse Hessian approximation H_0. The widely used one is the scaled identity matrix

  $$H_0 = \gamma I \tag{2}$$

  where γ is a positive scalar and I is the identity matrix. One can either utilize gradient information or problem specific estimate of Hessian to initialize H_0. In this work, we used the most naïve choice of $\gamma = 1$. "
- We have also made the initial inverse Hessian approximation $H_0 = I$ in the algorithm.