SOAP	0.51 (0.17)	0.19 (0.09)	0.00	0.46 (0.18)	0.21 (0.10)	0.00	0.62 (0.15)	0.14 (0.07)	0.00
SRT	0.00 (0.00)	0.01 (0.02)	0.35	0.00 (0.00)	0.01 (0.01)	0.45	0.00 (0.00)	0.00 (0.01)	0.65
CSSP	0.49 (0.18)	0.83 (0.06)	0.00	0.46 (0.16)	0.84 (0.06)	0.00	0.51 (0.16)	0.83 (0.07)	0.00
Go	**0.99 (0.04)**	**0.00 (0.00)**	**0.90**	**0.99 (0.04)**	**0.00 (0.00)**	**1.00**	**0.97 (0.06)**	**0.00 (0.00)**	**0.90**
IPU	0.91 (0.11)	0.01 (0.01)	0.50	**0.99 (0.04)**	**0.00 (0.00)**	**1.00**	**0.97 (0.06)**	**0.00 (0.00)**	**0.90**

The authors thank the reviewers for their careful readings and insightful and constructive comments. We will improve the manuscript in the revised version. Below please find our responses to the major points raised.

To R1: Thanks for your valuable comments. **On experiments**, with the spike model following [45] with $\mathbf{x}_i = \mathbf{V}\mathbf{z}_i + \sigma\mathbf{w}_i$, where $\mathbf{V} \in \mathbb{R}^{d\times m}, \sigma = 0.3$, and Gaussian $\mathbf{z}_i \in \mathbb{R}^m$, $\mathbf{w}_i \in \mathbb{R}^d$. The partial (due to space limit) results are reported above. More will be included. **On statistical properties,** it could be an easy corollary from Prop. 1.1 of [54].

To R2: Thanks for appreciating the contributions of our work!

To R3: We appreciate the very detailed and thoughtful comments from the reviewer. We'd like to do some clarifications here. We proved two types of theoretical results. One is on approximation which controls the absolute error of the output objective value, that is, accuracy. The other is on convergence which ensures the convergence, which is exact, and finite time termination of IPU and allows for bounding the approximation error of IPU with Theorem 5.1.

On tightness when $\kappa = \lambda_1\lambda_d^{-1} \to 1$. Thanks! We prove following improved Theorem 5.1, in which $\varepsilon \to 0$ when $\kappa \to 1$, or $k \to d$, or $\mathbf{A} \to \mathbf{A}_m$. Besides, there is no polynomial algorithm that has small ε for all $\mathbf{A}$ [8].

Claim R.1. *Let the condition number of* $\mathbf{A}$ *be* $\kappa = \lambda_1\lambda_d^{-1} \geq 1$. *In Theorem 5.1 and following corollaries, it holds*

$$\varepsilon \leq \min\left\{\tfrac{dG_1}{k}, \tfrac{dG_2}{m}, 1-\kappa^{-1}, 1-\tfrac{k}{d}\right\}.$$

Proof. Using the Poincaré separation theorem in Lemma D.1, we have $\varepsilon \leq 1-\kappa^{-1}$ by $\mathrm{Tr}(\mathbf{W}_m^\top\mathbf{A}\mathbf{W}_m) \geq \sum_{i=d-m+1}^{d}\lambda_i \geq m\cdot\lambda_d$, and $\mathrm{Tr}(\mathbf{W}_*^\top\mathbf{A}\mathbf{W}_*) \leq \sum_{i=1}^{m}\lambda_i \leq m\cdot\lambda_1 = m\cdot\kappa\lambda_d$. Meanwhile, $\varepsilon \leq 1-kd^{-1}$ holds by using $\mathrm{Tr}(\mathbf{W}_m^\top\mathbf{A}\mathbf{W}_m) \geq \mathrm{Tr}(\mathbf{W}_m^\top\mathbf{A}_m\mathbf{W}_m) \geq \frac{k}{d}\mathrm{Tr}(\mathbf{A}_m) \geq \frac{k}{d}\sum_{i=1}^{m}\lambda_i$, and $\mathrm{Tr}(\mathbf{W}_*^\top\mathbf{A}\mathbf{W}_*) \leq \sum_{i=1}^{m}\lambda_i$. □

On converging to fixed-point. In the continuous non-convex optimization literature, it is very common to show the algorithm converges to a stationary/critical point [52] as the general non-convex optimization are NP-hard even for computing a local minimizer [50]. However, a stationary point might still be a local maximum/minimum, or saddle point and far from the global one. Indeed, to our knowledge, it is very difficult (if not impossible) to show any global convergence guarantee to global optima in general non-convex setting, unless the interested problem has very special properties, e.g., benign landscape, robust bistability. Moreover, we emphasize that our problem is not a convex one, and has no known good properties. We are trying to *maximize* a convex objective function with non-convex combinatorial constraints. So both the objective and the feasible domain bring difficulty. Actually, for our problem, assuming SSE-hard and NP≠P, it is impossible [8] to have any polynomial running time algorithm that provably returns $\mathbf{W}$ such that $c\cdot\mathrm{Tr}(\mathbf{W}^\top\mathbf{A}\mathbf{W}) \geq \mathrm{Tr}(\mathbf{W}_*^\top\mathbf{A}\mathbf{W}_*)$ for arbitrary large but finite $c > 1$ and general $\mathbf{A}$. Thus, we think it is reasonable to have a fixed-point convergence result, especially when the approximation error (accuracy) of the fixed-point is controlled by Corollary 5.3. Besides, a local analysis near the optima is possible but meaningless, to us, as it is still NP-hard to ensure the initialization to be in the basin of attraction.

On related work. To the best knowledge of us, there is no directly comparable work in the literature. The most related work to ours are the very new [54, 53]. [53] proposed algorithm for FSPCA problem with computational complexity exponential in m and $\mathrm{rank}(\mathbf{A})$. [53] clearly said (page 4) that their algorithm is of theoretical nature and may not be practically implementable. In [54], they proposed a heuristic algorithm and a relaxed problem whose optimal value is upper bound by $(1+\sqrt{m})^2$ times the FSPCA optimal value. However, [54] provided no rounding procedure for extracting feasible solution from their relaxation and no approximation guarantee for their heuristic algorithm. In contrast, our approximation guaranteed algorithm runs in polynomial time and highly implementable in practice. Besides, [51] proposed an algorithm that runs exponential in the $\mathrm{rank}(\mathbf{A})$ and m for the *disjoint*-FSPCA problem that requires the support of different eigenvectors to be *disjoint*, which is clearly different from our setting. Finally, we note that there are many work [41, 45, 26, 5, 19, 27, 15, 43] that prove results assuming statistical models. They are not directly comparable to ours as our results are model-free, i.e., same as [54, 53], and applicable for *any* model and might have applications in other machine learning problems as noted by R4.

To R4: Thanks for your insightful and positive comments! We will highlight that in the revised paper and move the suggested parts to the main paper. The typos have been fixed.

[50] Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 1987.

[51] Sparse pca via bipartite matchings. *NeurIPS*, 2015.

[52] Non-convex optimization for machine learning. *arXiv preprint arXiv:1712.07897*, 2017.

[53] Sparse pca on fixed-rank matrices. *submitted manuscript*, 2019.

[54] Upper bounds for model-free row-sparse principal component analysis. *ICML*, 2020.