A.4. Discussion on convergence analysis

To understand how neural networks enhance ResKoopNet and build up the theoretical framework of convergence, it is important to first introduce Barron space (Pinkus, 1999; Cybenko, 1989; Haykin, 2009; Barron, 1993). Barron space characterizes functions efficiently approximated by two-layer neural networks, which is central to ResKoopNet. By leveraging networks that approximate functions within this space, ResKoopNet can flexibly optimize the dictionary functions for Koopman operator approximation, making it highly effective for complex, high-dimensional systems.

A function f belongs to Barron space $\mathcal{B}$ if it can be represented as:

$$f(x) = \int_{\Omega} a\sigma(w^T x)\rho(da, dw),$$

where σ is the activation function, w is a weight vector, a is a coefficient, and ρ is a probability distribution. The complexity of f is measured by the Barron norm $\|f\|_{\mathcal{B}}$:

$$\|f\|_{\mathcal{B}} = \inf_{\rho \in P_f} \left(\int_{\Omega} |a| \|w\|_1 \rho(da, dw) \right),$$

where P_f is the set of distributions for which f can be represented. This framework provides a foundation for analyzing approximation errors in neural networks.

The following theorem (E et al., 2020) discusses the approximation capabilities of two-layer neural networks within this context, establishing a foundation for the subsequent analysis.

Theorem A.2 (Direct Approximation Theorem, L^2-version)**.** *For any $f \in \mathcal{B}$ and $r \in \mathbb{N}$, there exists a two-layer neural network f_r with r neurons $\{(a_i, \mathbf{w}_i)\}$ such that*

$$\|f - f_r\|_{\mu} \lesssim \frac{\|f\|_{\mathcal{B}}}{\sqrt{r}}.$$

This implies an approximation error decreasing at a rate of $O(1/\sqrt{r})$, where r is the number of neurons. In ResKoopNet, the dictionary $\Psi(x;\theta) = \{\psi_i(x;\theta)\}_{i=1}^{N_K}$ is parameterized by a neural network with parameters θ. Assuming the true dictionary functions $\psi_i \in \mathcal{B}$, Theorem A.2 ensures that $\Psi(x;\theta)$ can approximate the optimal dictionary spanning the Koopman invariant subspace $\mathcal{B}_{N_K} \subset \mathcal{F}$ with error $\epsilon > 0$, provided r is sufficiently large.

Remark A.3. Notice that, the "two-layer neural network" in the Theorem A.2 statement refers to a hidder layer + an output layer, which is the most standard and general setting. Our implementation uses three hidden layers. This does not invalidate the result, as deeper networks can achieve at least the same approximation power (Pinkus, 1999).

We want to show two convergence results here: (1) $\Psi(x;\theta)$ approaches the true invariant subspace of $\mathcal{K}$; (2) The eigenpairs (λ_i, ϕ_i) and pseudospectrum approximate $\mathcal{K}$'s true spectrum as $J(\theta) \to 0$.

Assumption A.4. To formalize convergence, we make the following assumptions:

(a) The optimal dictionary functions $\{\psi_i^*\}_{i=1}^{N_K}$ spanning $\mathcal{K}$'s invariant subspace lie in $\mathcal{B}$.

(b) The loss $J(\theta)$ is Lipschitz continuous in θ, and the neural network $\Psi(x;\theta)$ has bounded gradients, facilitating gradient-based optimization.

Now, consider a Barron space $\mathcal{B}$ which is dense in $\mathcal{F}$. Given N_K fixed, let $\mathcal{B}_{N_K} = \text{span}\{\psi_i^*\}_{i=1}^{N_K}$ be the true invariant subspace. By Theorem A.2, for each ψ_i^*, there exists a neural network approximation $\psi_i(x;\theta_r)$ with r neurons such that:

$$\|\psi_i^* - \psi_i(\cdot;\theta_r)\|_{\mu} \leq \frac{\|\psi_i^*\|_{\mathcal{B}}}{\sqrt{r}}.$$

The total dictionary error is:

$$\|\Psi^* - \Psi(\cdot;\theta_r)\|_F^2 = \sum_{i=1}^{N_K} \|\psi_i^* - \psi_i(\cdot;\theta_r)\|_{\mu}^2 \leq \frac{1}{r} \sum_{i=1}^{N_K} \|\psi_i^*\|_{\mathcal{B}}^2 = \frac{C_{N_K}}{r},$$

where $C_{N_K} = \sum_{i-1}^{N_K} \|\psi_i^\star\|_{\mathcal{B}}^2$ is finite under Assumption A.4(a). As $r \to \infty$, $\Psi(x;\theta_r) \to \Psi^\star$ in the Frobenius norm, which ensures the dictionary approximated by neural network can represent $\mathcal{B}_{N_K}$.

Algorithm 1 updates θ via stochastic gradient descent (SGD) with step size δ and computes $\tilde{K}(\theta_n)$ iteratively until $J(\theta_n) < \epsilon$ where θ_n represents n-th iteration of parameters θ. For a Lipschitz continuous $J(\theta)$ with constant L (by Assumption A.4(b)) and a strongly convex region near the optimum θ^* (assumed locally for simplicity), SGD converges at a rate of $O(1/n)$ in expectation (Bottou et al., 2018):

$$\mathbb{E}[J(\theta_n) - J(\theta^*)] \leq \frac{L}{2\eta n},$$

where η is the strong convexity constant and n is the iteration number. In practice, $J(\theta)$ is non-convex due to the neural network, but empirical convergence is observed (Section 4), and the alternating update with $\tilde{K}(\theta)$ stabilizes the process. As iteration step $n \to \infty$ and amount of data points $m \to \infty$, $J(\theta_n) \to 0$, which implies $\widehat{\text{res}}(\lambda_i, \phi_i) \to 0$ for all i.

With $J(\theta) \to 0$, we now assess the spectral error: if $\widehat{\text{res}}(\lambda_i, \phi_i) < \epsilon$, then $\|\mathcal{K}\phi_i - \lambda_i\phi_i\|_\mu / \|\phi_i\|_\mu < \sqrt{\epsilon}$, indicating λ_i and ϕ_i are approximate eigenpairs of $\mathcal{K}$ converges to $\mathcal{K}$'s true spectrum as $N_K \to \infty$ and $\epsilon \to 0$, as pointed out in (Colbrook & Townsend, 2024, Theorem B.1). Uniform convergence on compact subsets of $\mathbb{C}$ follows from the density of $\mathcal{B}_{N_K}$ in $\mathcal{F}$ and Dini's theorem, as pointed out in Colbrook & Townsend (2024, Lemma B.1).

The non-convexity and system complexity may slow the spectral approximation in practice. High-order convergence (e.g., polynomial) could arise for smooth dynamics (See Colbrook & Townsend (2024, Theorem 3.1) for more details), which is useful for further study.

Remark A.5. The computational cost (Appendix A.6) scales with r, t, and m, which trades off with accuracy. Adaptive selection of r and early stopping when $J(\theta) < \epsilon$ can optimize this balance.