Supplementary Information

Self-Directed Online Machine Learning for Topology Optimization

Changyu Deng[1], Yizhou Wang[2], Can Qin[2], and Wei Lu[1,3,*]

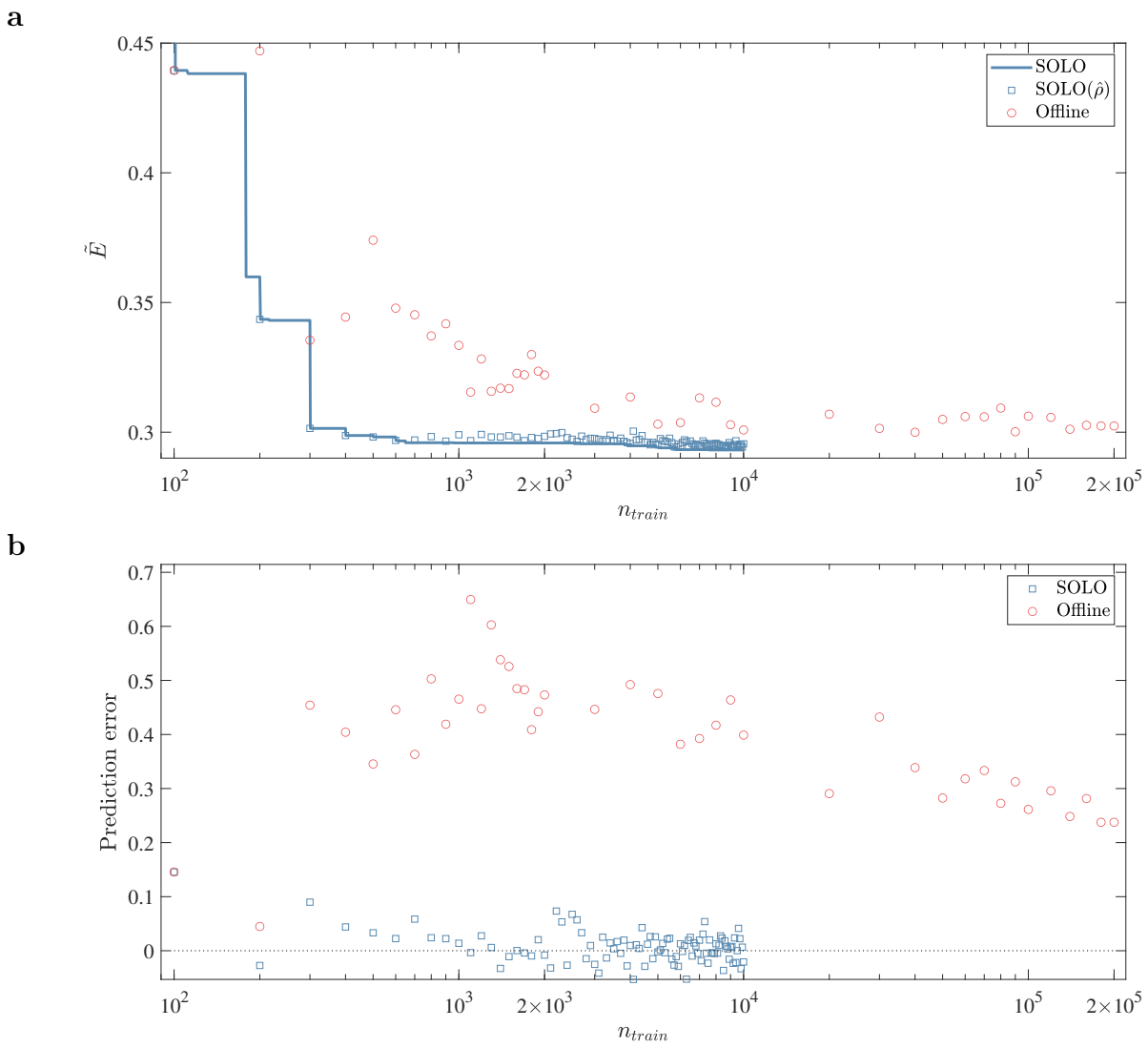
[1]Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109, United States
[2]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, United States
[3]Department of Materials Science and Engineering, University of Michigan, Ann Arbor, MI 48109, United States
[*]Corresponding author: weilu@umich.edu

Table of Contents

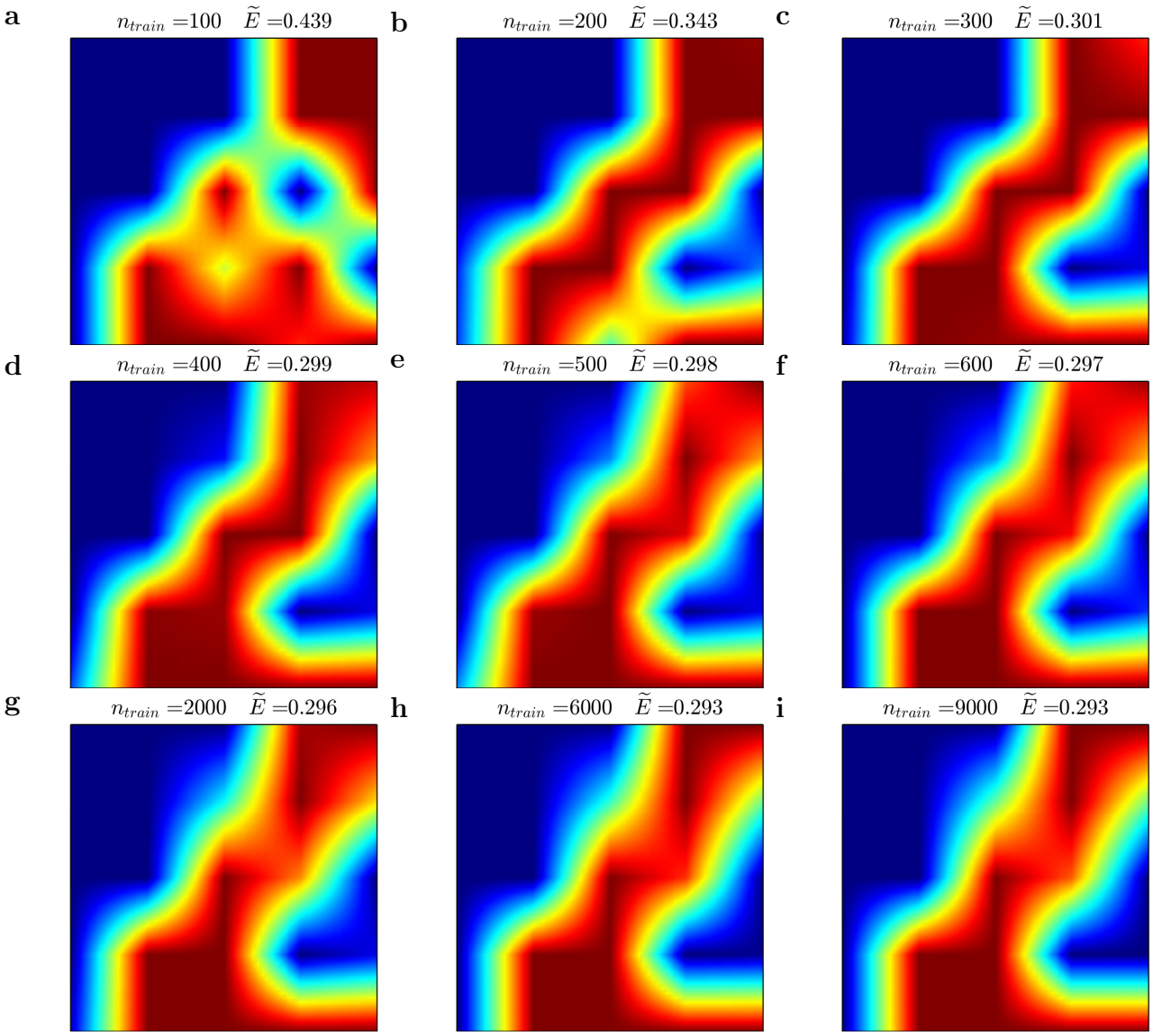
1 Supplementary table and figures

Supplementary Table 1: Average wall time within a loop. There are three major steps in each loop: FEM calculation to obtain corresponding objective function values, DNN training, and optimization which searches for the optimum based on DNN's prediction. We give a very rough estimate on our personal computer (CPU: Intel i7-8086K, GPU: NVidia RTX 2080 Super). *Italic numbers* indicate GPU computing and the others are computed entirely on CPU. Actual running time is sensitive to hardware environment, software packages, parameter setting and so forth. Further, FEM calculation is approximately proportional to the number of additional samples per loop; training time depends on existing training data obtained from previous loops; optimization depends on the number of function evaluations. Similar to other SMBO methods, our surrogate model introduces overhead computation. Although the overhead is comparable with FEM calculation time, it is almost negligible considering the huge benefit of reducing FEM calculations from $10^5 \sim 10^8$ (see the table) to $10^2 \sim 10^4$. Besides, we chose relatively simple problems and thus each calculation only cost < 0.5s for compliance problems and < 6s for fluid problems; smaller portion of the overhead is expected for more complicated problems with higher FEM computation time.

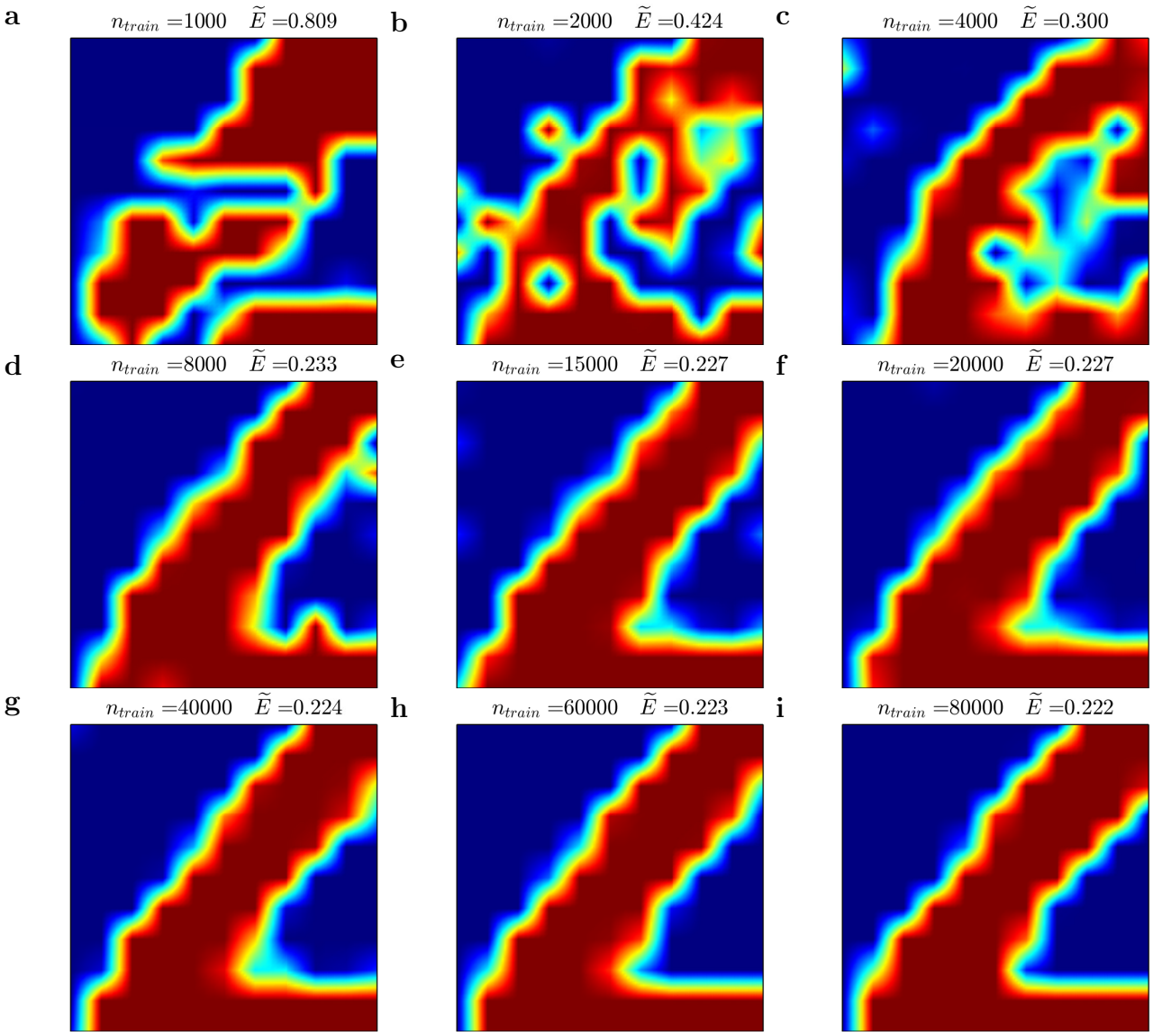
Problem	Number of additional samples	Wall time /s		
		FEM	Training	Optimization (evaluations)
Compliance 5x5	100	40	35	70 (2×10^5)
Compliance 11x11	1000	500	150	1000 (4×10^6)
Fluid 20x8 (G)	10	35	*10*	*35* (1×10^8)
Fluid 20x8 (R)	100	350	*20*	*35* (1×10^8)
Fluid 40x16 (G)	10	60	*25*	*140* (2×10^8)



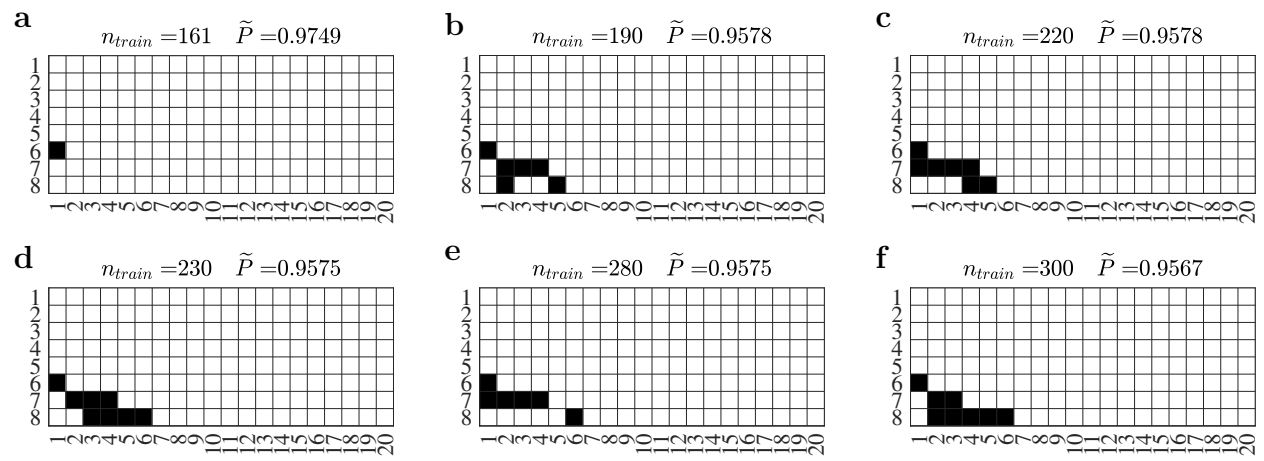
Supplementary Fig. 1: Objective (energy) and prediction error of the compliance minimization problem with 5×5 variables. a, Dimensionless energy as a function of n_{train}. For SOLO, the solid line denotes the best objective values and the squares denote $\widetilde{E}(\widehat{\rho})$. **b**, Energy prediction error of $\widehat{\rho}$.



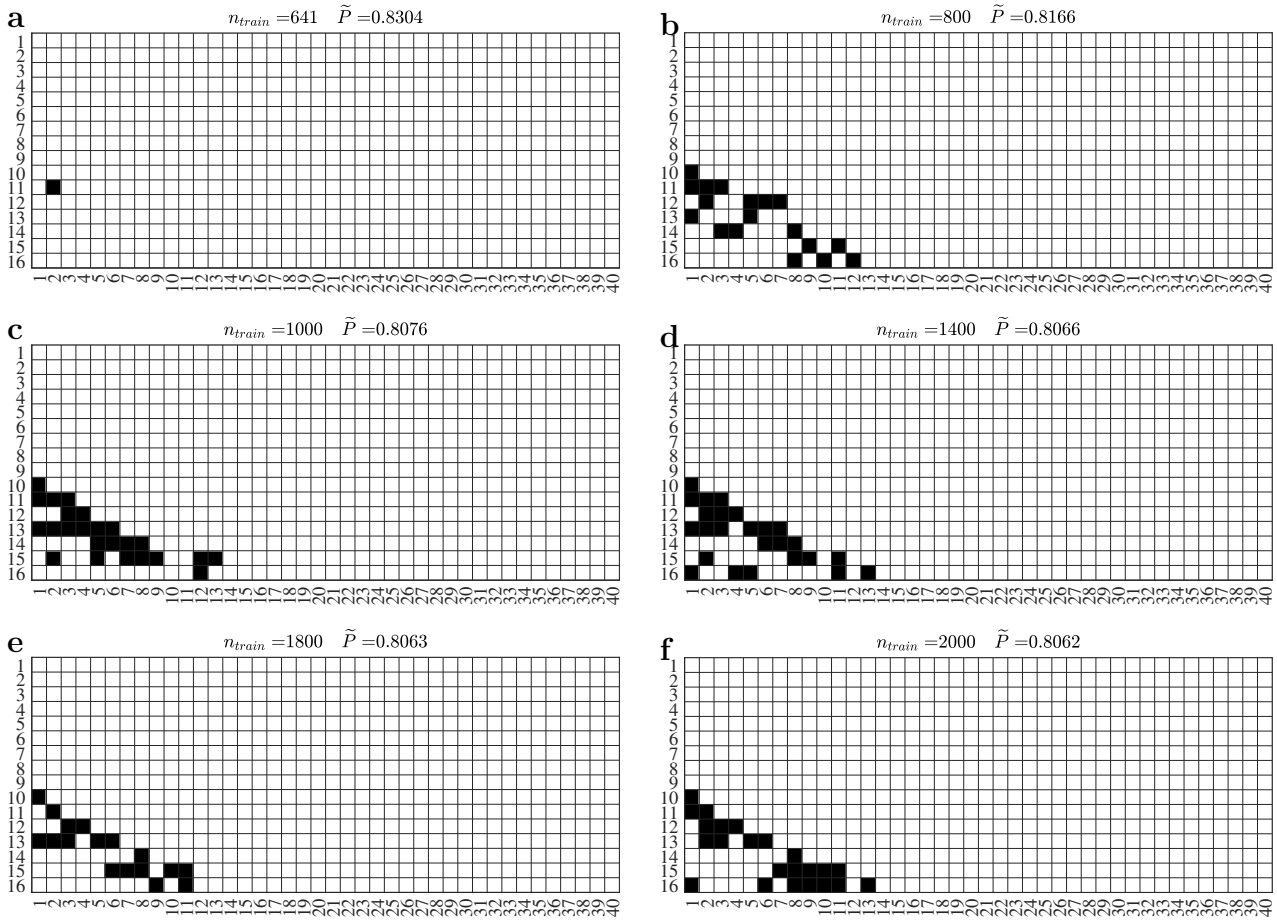
Supplementary Fig. 2: Evolution of the solution from SOLO for the compliance minimization problem with 5×5 variables. Each plot is the best among n_{train} accumulated training data and the corresponding energy $\widetilde{E}$ is marked. There is no obvious change after hundreds of training samples.



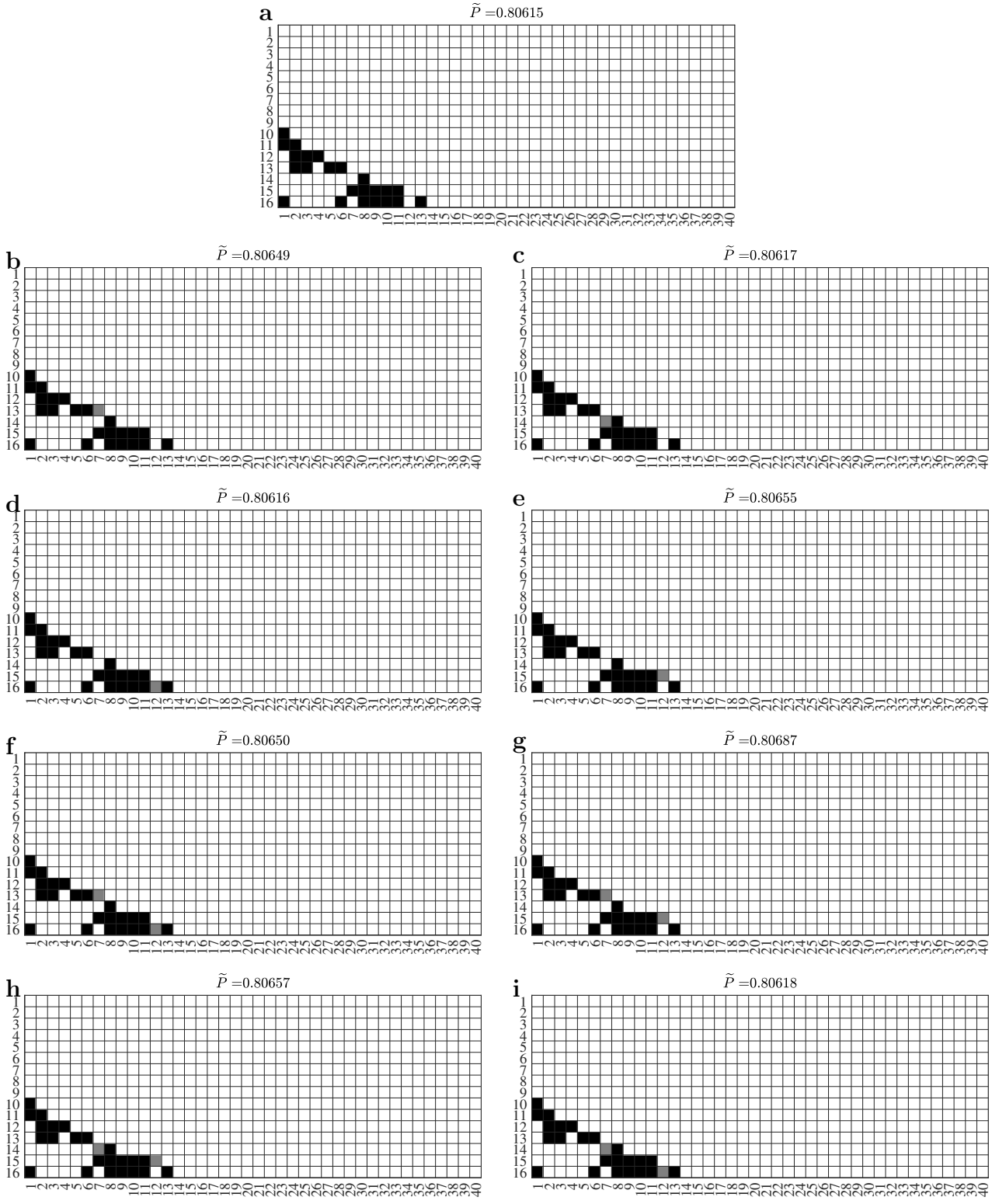
Supplementary Fig. 3: Evolution of the solution from SOLO for the compliance minimization problem 11×11 variables. Each plot is the best among n_{train} accumulated training data and the corresponding energy $\widetilde{E}$ is marked. There is no obvious change after ten thousand training samples.



Supplementary Fig. 4: Evolution of the solution from SOLO-G for the fluid-structure optimization problem with 20×8 mesh. Each plot is the best among n_{train} samples.



Supplementary Fig. 5: Evolution of the solution from SOLO-G for the fluid-structure optimization problem with 40×16 mesh. Each plot is the best among n_{train} samples.



Supplementary Fig. 6: Perturbation of the optimum from SOLO-G for the fluid-structure optimization problem with 40×16 mesh. Intuitively the ramp should be smooth, yet we observe two gaps in the optimum given by SOLO-G. We try filling the gaps. **a**, the optimum from SOLO-G. **b-i**, one or two blocks (gray) are added to fill the gap, with higher $\widetilde{P}$.

2 Theory on convergence

In the main text, we presented a simplified version of convergence (Eq. (4)). In this section, we give a detailed description of our theoretical result. We first present the main result (Theorem 1). Then, we introduce some preliminary definitions and knowledge used in the proof. At the end, we approach the proof.

2.1 Formulation and theorem

The unknown object function is denoted as $F(\rho)$, where $\rho \in \mathbb{R}^N$.We denote the domain of $\{\rho \mid 0 \leq \rho_i \leq 1, 1 \leq i \leq N\}$ as K. We suppose the global minimizer $\rho^* = \text{argmin}_\rho F(\rho)$.

We consider the total iteration number to be T. At iteration $t(1 \leq t \leq T)$, the DNN is denoted as $f_t(\cdot)$ and we denote the empirical minimizer of this DNN function to be $\widehat{\rho}^{(t)}$, i.e.

$$\widehat{\rho}^{(t)} = \underset{\rho}{\text{argmin}}\, f_t(\rho). \tag{S1}$$

Besides, we denote our DNN as a D-layer neural network which is formulated as follows:

$$f_t(\rho) = W_D^\top \sigma(W_{D-1}\sigma(...\sigma(W_1\rho))),$$

where $\mathcal{W} = \{W_k \mid W_k \in \mathbb{R}^{d_{k-1}\times d_k},\ k = 1, ..., D-1,\ W_D \in \mathbb{R}^{d_{D-1}}\}, d_0 = N$(number of input dimentions) and $\sigma(v) = [\max\{v_1, 0\}, ..., \max\{v_d, 0\}]^\top$ is the ReLU[1] activation function for $v \in \mathbb{R}^d$. We further denote $d = \max\{d_i\}$ and the function class of such neural networks as $\mathcal{H}_f$.

At time step t, given the empirical optimal point $\widehat{\rho}^{(t-1)}$, the additional m training points is generated through the following process:

$$\rho^{(j_t)} = \widehat{\rho}^{(t-1)} + \xi^{(j_t)}, j_t = mt - m + 1, mt - m + 2, \cdots, mt.$$

Here $\xi^{(j)}$ denotes random noise for perturbation. Hence throught the iterating process, the sampled points are random variables. Since $\widehat{\rho}^{(t)}$ is the minimizer of f_t and f_t is neural network trained on random variables, we also view $\widehat{\rho}^{(t)}$ as random variables in our theoretical analysis. At time step t, we denote all the realizations of random training data points set as $K_t = \{\rho^{(i)} \mid i = 1, \cdots, mt\}$.

Now before we proceed, we need to impose some mild assumptions on the problem.

Assumption 1. We suppose that

1) the spectral norm of the matrices in DNNs are uniformly bounded, i.e., there exists $B_W > 0$ s.t. $\|W_k\|_2 \leq B_W, \forall k = 1, \cdots, D$.
2) the target function is bounded, i.e., there exists $B_F > 0$ s.t. $\|F\|_\infty \leq B_F$.

1) of Assumption 1 is a commmonly studied assumption in existing generalization theory literature on deep neural networks[2–4]. 2) of Assumption 1 assumes F is bounded, which is standard and intuitive since F has a physical meaning.

Assumption 2. We assume that for any iteration t, $\xi^{(j_t)}(j_t = mt - m + 1, \cdots, mt)$ are i.i.d. (independent and identically distributed) perturbation noise

The assumption of the i.i.d. properties of noise in Assumption 2 is common in optimization literature[5–8]. The difference is that in traditional optimization literature noise refers to the difference between the true gradient and the stochastic gradient while the noise here denotes perturbations to generate new samples in each iteration. Note that our Assumption 2 only needs the i.i.d. property of noise, which is weaker than the standard assumptions for stochastic gradient methods which require unbiased property and bounded variance[6–8].

Assumption 3. We suppose that for any iteration t, $\{\widehat{\rho}^{(t)}|t = 1 \cdots T\}$ are mutually independent.

Since our fitting DNN f_ts are continuously changing throughout iterations, it is reasonable for us to assume their empirical minimizers $\widehat{\rho}^{(t)}$ to be independent for the ease of analysis.

We denote the distribution of samples $\{\rho^{(j_t)}|j_t = mt-m+1, mt-m+2, \cdots, mt\}$ as $D_t(1 \leq t \leq T)$, with which we can introduce the following definition.

Definition 1. For a measurable function f, we denote

$$\mathbb{E}_{D_{1:T}} f(\rho) = \frac{\sum_{t=1}^{T} \mathbb{E}_{\rho \sim D_t} f(\rho)}{T}, \tag{S2}$$

where $\mathbb{E}$ denotes expectation.

Assumption 4. For any t and $f_t \in \mathcal{H}_f$,

$$\|F - f_t\|_\infty^2 = C(t)\mathbb{E}_{\rho \sim D_{1:t}}(F - f_t)^2,$$

where $C(t)$ is a monotonely decreasing function w.r.t. iteration number t.

Assumption 4 basically describes that the Chebyshev distance of our DNN at time t and F is bounded by a constant number(w.r.t. t) times the average true loss of $(F - f_t)^2$ till time t. This assumption is reasonable in that the the average true loss can be seen as a variant of Euclidean distance between our DNN at time t and F.

Eventually we arrive at our main result.

Theorem 1. Under Assumptions 1, 2, 3 and 4, given iteration number T and any $\delta > 0$, for any trained DNN $f_T \in \mathcal{H}_f$ with empirical MSE training error ϵ at iteration T, we have that with probability at least $1 - \delta$ over the joint distribution of $\rho^{(1)}, \rho^{(2)}, \cdots, \rho^{(mT)}$,

$$(F(\widehat{\rho}^{(T)}) - F(\rho^*))^2 \leq 4C(T)\left(\frac{96B^2}{\sqrt{mT}}\sqrt{d^2 D \log(1 + 8BB_W^D D\sqrt{mTd})} + 12B^2\sqrt{\frac{2\log\frac{2}{\delta}}{mT}} + \frac{8}{mT} + \epsilon\right),$$

where $B = \max\{B_F, B_W^D\}$.

2.2 Preliminaries

Before showing the proof, we introduce some definitions and lemmas.

Lemma 1 (McDiarmid's Inequality[9])**.** Let $X_1, \cdots, X_m \in \mathcal{X}^m$ be a set of $m \geq 1$ independent random variables and assume that there exist $c_1, \cdots, c_m > 0$ such that $f : \mathcal{X}^m \to \mathbb{R}$ satisfies the following conditions:

$$|f(x_1, \cdots, x_i, \cdots, x_m) - f(x_1, \cdots, x_i', \cdots, x_m)| \leq c_i,$$

for all $i \in [m]$ and any points $x_1, \cdots, x_m, x_i' \in \mathcal{X}$. Let $f(S)$ denote $f(X_1, \cdots, X_m)$, then, for all $s > 0$, the following inequality hold:

$$\mathbb{P}\big\{f(S) - \mathbb{E}[f(S)] \geq s\big\} \leq \exp\Big(\frac{-2s^2}{\sum_{i=1}^m c_i^2}\Big), \tag{S3}$$

$$\mathbb{P}\big\{f(S) - \mathbb{E}[f(S)] \leq -s\big\} \leq \exp\Big(\frac{-2s^2}{\sum_{i=1}^m c_i^2}\Big), \tag{S4}$$

where $\mathbb{P}$ denotes probability and $\mathbb{E}$ denotes expectation.

Definition 2 (Covering Number[10])**.** Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. Vector set $\{V_i \in V | i = 1, \cdots, N\}$ is an ϵ-covering of Θ if $\Theta \subset \cup_{i=1}^N B(V_i, \epsilon)$ where $B(V_i, \epsilon)$ denotes the ball with center V_i and radius ϵ, equivalently, $\forall \theta \in \Theta, \exists i$ such that $\|\theta - V_i\| \leq \epsilon$. The covering number is defined as :

$$\mathcal{N}(\Theta, \|\cdot\|, \epsilon) := \min\{n : \exists \epsilon\text{-covering over } \Theta \text{ of size } n\}.$$

Definition 3 (Rademacher Complexity & Empirical Rademacher Complexity[10,11])**.** Given a sample $S = \{x_1, x_2, \cdots, x_n\}$ and a set of real-valued function $\mathcal{H}$, the *Empirical Rademacher Complexity* is defined as

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) = \mathfrak{R}_n(\mathcal{H}_{|S}) := \frac{1}{n}\mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \sum_{i=1}^n \sigma_i h(x_i),$$

where sup denotes supremum and the expectation is over the Rademacher random variables $(\sigma_1, \sigma_2, \cdots, \sigma_n)$, which are i.i.d. (independent and identically distributed) with $\mathbb{P}(\sigma_1 = 1) = \mathbb{P}(\sigma_1 = -1) = \frac{1}{2}$. The *Rademacher Complexity* is defined as

$$\mathfrak{R}_n(\mathcal{H}) := \mathbb{E}_S \mathfrak{R}_n(\mathcal{H}_{|S}) = \frac{1}{n}\mathbb{E}_{S,\sigma} \sup_{h \in \mathcal{H}} \sum_{i=1}^n \sigma_i h(x_i),$$

which is the expectation of the Empirical Rademacher Complexity over sample S.

Lemma 2 (Dudley's Entropy Integral Bound[4])**.** Given a sample $S = \{x_1, x_2, \cdots, x_n\}$, let $\mathcal{H}$ be a real-valued function class taking values in $[0, r]$ for some constant r, and assume that zero function $\mathbf{0} \in \mathcal{H}$. Then we have

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) \leq \inf_{\alpha > 0}\Big(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n}\int_\alpha^{r\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{H}, \epsilon, \|\cdot\|_\infty)}d\epsilon\Big),$$

where inf denotes infimum.

Lemma 3. (Covering number bound using volume ratio[4]) Let $\mathcal{W} = \{W \in \mathbb{R}^{a\times b} : \|W\|_2 \leq \lambda\}$ be the set of matrices with bounded spectral norm and ϵ be given. The covering number $\mathcal{N}(\mathcal{W}, \epsilon, \|\cdot\|_F)$ is upper bounded by

$$\mathcal{N}(\mathcal{W}, \epsilon, \|\cdot\|_{\mathrm{F}}) \leq \left(1 + 2\frac{\min\{\sqrt{a}, \sqrt{b}\}\lambda}{\epsilon}\right)^{ab}.$$

2.3 Proof

This subsection presents the complete proof of Theorem 1. We first introduce an auxilliary lemma here.

Lemma 4. Under Assumptions 2 and 3, we have

1) the whole generated data points $\{\rho^{(i)} \mid i = 1, 2, \cdots, mT\}$ are mutually independent.
2) for any t, $\{\rho^{(j_t)} | j_t = mt - m + 1, \cdots, mt\}$ are i.i.d..

Lemma 4 is a straightforward result employing the assumptions.

Now we can approach the proof of Theorem 1.

Proof.

$$\begin{aligned}
&\sup_{f_T\in H_f} (F(\widehat{\rho}^{(T)}) - F(\rho^*))^2 \\
&\qquad \overset{\text{(i)}}{\leq} \sup_{f_T\in H_f} (F(\widehat{\rho}^{(T)}) - f_T(\widehat{\rho}^{(T)}) + f_T(\rho^*) - F(\rho^*))^2 \\
&\qquad \overset{\text{(ii)}}{\leq} \sup_{f_T\in H_f} 2\big\{[F(\widehat{\rho}^{(T)}) - f_T(\widehat{\rho}^{(T)})]^2 + [f_T(\rho^*) - F(\rho^*)]^2\big\} \\
&\qquad \leq 4 \sup_{f_T\in H_f} \|F - f_T\|_\infty^2 \\
&\qquad \overset{\text{(iii)}}{=} 4C(T) \sup_{f_T\in H_f} \frac{\sum_{t=1}^T \mathbb{E}_{\rho\sim D_t}(F(\rho) - f_T(\rho))^2}{T}.
\end{aligned} \tag{S5}$$

Here (i) comes from Eq. (S1), (ii) uses the fact that for any real number x and y, we have $(x + y)^2 \leq 2(x^2 + y^2)$. (iii) arises from Assumption 4.

We further denote

$$\Phi(K_T) = \sup_{f_T\in\mathcal{H}_f} \left[\mathbb{E}_{D_{1:T}}(F - f_T)^2 - \widehat{\mathbb{E}}_{K_T}(F - f_T)^2\right], \tag{S6}$$

where $\widehat{\mathbb{E}}_{K_T}(F - f_T)^2 = \frac{1}{mT}\sum_{i=1}^{mT}(F(\rho^{(i)}) - f_T(\rho^{(i)}))$ corresponds to the empirical MSE loss when training our neural network.

Suppose K_T' and K_T are two samples only different in the k-th point, namely $K_T = \{\rho^{(1)}, ..., \rho^{(k)}, ..., \rho^{(mT)}\}$ and $K_T' = \{\rho^{(1)}, ..., \rho^{(k)\prime}, ..., \rho^{(mT)}\}$, we have

$$\begin{aligned}
|\Phi(K_T') - \Phi(K_T)| &\leq \sup_{f_T \in \mathcal{H}_f} |\widehat{\mathbb{E}}_{K_T}(F - f_T)^2 - \widehat{\mathbb{E}}_{K_T'}(F - f_T)^2| \\
&= \sup_{f_T \in \mathcal{H}_f} \left| \frac{(F(\rho_k) - f_T(\rho_k))^2}{mT} - \frac{(F(\rho^{(k)\prime}) - f_T(\rho^{(k)\prime}))^2}{mT} \right| \\
&\leq \frac{8B^2}{mT},
\end{aligned}$$

then by Mcdiarmid's Inequality (Eq.(S3) in Lemma 1), we get

$$\mathbb{P}(\Phi(K_T) - \mathbb{E}_{K_T}\Phi(K_T) \geq s) \leq \exp\left(\frac{-2s^2}{mT \cdot (\frac{8B^2}{mT})^2}\right). \tag{S7}$$

Given any $\delta > 0$, by setting the right handside of (S7) to be $\frac{\delta}{2}$, we have with probability at least $1 - \frac{\delta}{2}$,

$$\Phi(K_T) \leq \mathbb{E}_{K_T}\Phi(K_T) + 4B^2\sqrt{\frac{2\log\frac{2}{\delta}}{mT}}. \tag{S8}$$

Notice that

$$\begin{aligned}
\mathbb{E}_{K_T}\Phi(K_T) &= \mathbb{E}_{K_T}\left\{\sup_{f_T \in \mathcal{H}_f}\left[\mathbb{E}_{D_{1:T}}(F - f_T)^2 - \widehat{\mathbb{E}}_{K_T}(F - f_T)^2\right]\right\} \\
&= \mathbb{E}_{K_T}\left\{\sup_{f_T \in \mathcal{H}_f} \mathbb{E}_{K_T'}[\widehat{\mathbb{E}}_{K_T'}(F - f_T)^2 - \widehat{\mathbb{E}}_{K_T}(F - f_T)^2]\right\}.
\end{aligned} \tag{S9}$$

Here the second equality in Eq. (S9) is because:

$$\begin{aligned}
\mathbb{E}_{K_T'}[\widehat{\mathbb{E}}_{K_T'}(F - f_T)^2] =& \frac{1}{mT}\sum_{i=1}^{mT}\mathbb{E}_{K_T'}[F(\rho^{(i)}) - f_T(\rho^{(i)})]^2 \\
\overset{\text{(i)}}{=}& \frac{1}{mT}\left\{\sum_{i=1}^{m}\mathbb{E}_{\rho^{(i)}\sim D_1}[F(\rho^{(i)}) - f_T(\rho^{(i)})]^2 + \sum_{i=m+1}^{2m}\mathbb{E}_{\rho^{(i)}\sim D_2}[F(\rho^{(i)}) - f_T(\rho^{(i)})]^2 + \cdots \right. \\
&\left. + \sum_{i=mT-T+1}^{mT}\mathbb{E}_{\rho^{(i)}\sim D_T}[F(\rho^{(i)}) - f_T(\rho^{(i)})]^2\right\} \\
\overset{\text{(ii)}}{=}& \frac{1}{mT}[m\mathbb{E}_{D_1}(F - f_T)^2 + m\mathbb{E}_{D_2}(F - f_T)^2 + \cdots + m\mathbb{E}_{D_T}(F - f_T)^2] \\
=& \mathbb{E}_{D_{1:T}}(F - f_T)^2.
\end{aligned}$$

Here (i) results from 1) of Lemma 4 and (ii) comes from 2) of Lemma 4.

Further we have

$$
\begin{aligned}
&\mathbb{E}_{K_T}\left\{\sup_{f_T\in\mathcal{H}_f}\mathbb{E}_{K_T'}[\widehat{\mathbb{E}}_{K_T'}(F-f_T)^2-\widehat{\mathbb{E}}_{K_T}(F-f_T)^2]\right\}\\
&\overset{\text{(i)}}{\leq}\mathbb{E}_{K_T,K_T'}\sup_{f_T\in\mathcal{H}_f}[\widehat{\mathbb{E}}_{K_T'}(F-f_T)^2-\widehat{\mathbb{E}}_{K_T}(F-f_T)^2]\\
&=\mathbb{E}_{K_T,K_T'}\sup_{f_T\in\mathcal{H}_f}\frac{1}{mT}\sum_{i=1}^{mT}[(F(\rho^{(i)\prime})-f_T(\rho^{(i)\prime}))^2-(F(\rho^{(i)})-f_T(\rho^{(i)}))^2]\\
&\overset{\text{(ii)}}{=}\mathbb{E}_{\sigma,K_T,K_T'}\sup_{f_T\in\mathcal{H}_f}\frac{1}{mT}\sum_{i=1}^{mT}\sigma_i[(F(\rho^{(i)\prime})-f_T(\rho^{(i)\prime}))^2-(F(\rho^{(i)})-f_T(\rho^{(i)}))^2]\\
&\overset{\text{(iii)}}{\leq}\mathbb{E}_{\sigma,K_T'}\sup_{f_T\in\mathcal{H}_f}\frac{1}{mT}\sum_{i=1}^{mT}[\sigma_i(F(\rho^{(i)\prime})-f_T(\rho^{(i)\prime}))^2]+\mathbb{E}_{\sigma,K_T}\sup_{f_T\in\mathcal{H}_f}\frac{1}{mT}\sum_{i=1}^{mT}[-\sigma_i(F(\rho^{(i)})-f_T(\rho^{(i)}))^2]\\
&=2\mathbb{E}_{\sigma,K_T}\sup_{f_T\in\mathcal{H}_f}\frac{1}{mT}\sum_{i=1}^{mT}[\sigma_i(F(\rho^{(i)})-f_T(\rho^{(i)}))^2],
\end{aligned}
\tag{S10}
$$

where σ_i are Rademacher variables (Definition 3), which are uniformly distributed independent random variables taking values in $\{-1,+1\}$. Here (i) and (iii) hold due to the sub-additivity of the supremum function (considering the convexity of supremum function, by Jensen's Inequality, we have for any function f, $\sup\int_x f(x)\leq\int_x\sup f(x)$ holds). (ii) combines the definition of Rademacher variable σ_i and the fact that the expectation is taken over both K_T and $K_{T'}$.

For notational simplity, given any function $f_T\in\mathcal{H}_f$, we define the non-negative loss function $M(f_T):\rho\to(f_T(\rho)-F(\rho))^2$ and its function class $\mathcal{H}_M=\{M(f_T):f_T\in\mathcal{H}_f\}$.

Then combining (S9) and (S10) we obtain

$$
\mathbb{E}_{K_T}\Phi(K_T)\leq 2\mathfrak{R}_{mT}(\mathcal{H}_M),
\tag{S11}
$$

where $\mathfrak{R}_{mT}(\mathcal{H}_M)=\mathbb{E}_{\sigma,K_T}\sup_{f_T\in\mathcal{H}_f}\frac{1}{mT}\sum_{i=1}^{mT}\sigma_i(F(\rho^{(i)})-f_T(\rho^{(i)}))^2$ is the Rademacher Complexity (Definition 3) of $\mathcal{H}_M$.

Now, we define the Empirical Rademacher Complexity of $\mathcal{H}_M$ as

$$
\widehat{\mathfrak{R}}_{K_T}(\mathcal{H}_M):=\mathbb{E}_{\sigma}\sup_{f_T\in\mathcal{H}_f}\frac{1}{mT}\sum_{i=1}^{mT}\sigma_i(F(\rho^{(i)})-f_T(\rho^{(i)}))^2.
$$

Again, suppose K_T' and K_T are two samples only different in the k-th point, namely $K_T=$

$\{\rho^{(1)}, ..., \rho^{(k)}, ..., \rho^{(mT)}\}$ and $K_T' = \{\rho^{(1)}, ..., \rho^{(k)\prime}, ..., \rho^{(mT)}\}$, we have

$$
\begin{aligned}
&|\widehat{\mathfrak{R}}_{K_T}(\mathcal{H}_M) - \widehat{\mathfrak{R}}_{K_T'}(\mathcal{H}_M)| \\
&= \left| \mathbb{E}_\sigma \sup_{f_T \in \mathcal{H}_f} \frac{1}{mT} \sum_{i=1}^{mT} \sigma_i (F(\rho^{(i)}) - f_T(\rho^{(i)}))^2 - \mathbb{E}_\sigma \sup_{f_T \in \mathcal{H}_f} \frac{1}{mT} \sum_{i=1}^{mT} \sigma_i (F(\rho^{(i)\prime}) - f_T(\rho^{(i)\prime}))^2 \right| \\
&\leq \sup_{f_T \in \mathcal{H}_f} \left| \frac{(F(\rho^{(k)}) - f_T(\rho^{(k)}))^2}{mT} - \frac{(F(\rho^{(k)\prime}) - f_T(\rho^{(k)\prime}))^2}{mT} \right| \\
&\leq \frac{8B^2}{mT},
\end{aligned}
$$

then by Mcdiarmid's Inequality (Eq.(S4) in Lemma 1), we get

$$
\mathbb{P}(\widehat{\mathfrak{R}}_{K_T}(\mathcal{H}_M) - \mathfrak{R}_{mT}(\mathcal{H}_M) \leq -s) \leq \exp\left(\frac{-2s^2}{mT \cdot (\frac{8B^2}{mT})^2} \right). \tag{S12}
$$

Given any $\delta > 0$, by setting the right handside of Eq.(S12) to be $\frac{\delta}{2}$, we have with probability at least $1 - \frac{\delta}{2}$,

$$
\mathfrak{R}_{mT}(\mathcal{H}_M) \leq \widehat{\mathfrak{R}}_{K_T}(\mathcal{H}_M) + 4B^2 \sqrt{\frac{2 \log \frac{2}{\delta}}{mT}}. \tag{S13}
$$

Now combining (S8), (S11) and (S13), we get with probability at least $1 - \delta$,

$$
\Phi(K_T) \leq 2\widehat{\mathfrak{R}}_{K_T}(\mathcal{H}_M) + 12B^2 \sqrt{\frac{2 \log \frac{2}{\delta}}{mT}}, \tag{S14}
$$

here we use the fact that (S8) and (S13) hold with probability $1 - \frac{\delta}{2}$ respectively and that $(1 - \frac{\delta}{2})^2 > 1 - \delta$.

It is straightforward that $\|M(f_T)\|_\infty \leq 4B^2$, then Dudley's Entropy (Lemma 2) gives us

$$
\begin{aligned}
\widehat{\mathfrak{R}}_{K_T}(\mathcal{H}_M) &\leq \frac{4\alpha}{\sqrt{mT}} + \frac{12}{mT} \int_\alpha^{4B^2\sqrt{mT}} \sqrt{\log \mathcal{N}(\mathcal{H}_M, \epsilon, \|\cdot\|_\infty)} d\epsilon \\
&\leq \frac{4\alpha}{\sqrt{mT}} + \frac{48B^2}{\sqrt{mT}} \sqrt{\log \mathcal{N}(\mathcal{H}_M, \alpha, \|\cdot\|_\infty)},
\end{aligned} \tag{S15}
$$

where $\mathcal{N}$ denotes the covering number. We pick $\alpha = \frac{1}{\sqrt{mT}}$, and combine (S5), (S14) and (S15) to get

$$
\begin{aligned}
&\sup_{f_T \in H_f} (F(\widehat{\rho}^{(T)}) - F(\rho^*))^2 \\
&\leq 4C(T) \left(\frac{96B^2}{\sqrt{mT}} \sqrt{\log \mathcal{N}(\mathcal{H}_M, \frac{1}{\sqrt{mT}}, \|\cdot\|_\infty)} + 12B^2 \sqrt{\frac{2 \log \frac{2}{\delta}}{mT}} + \frac{8}{mT} + \widehat{\mathbb{E}}_{K_T}(F - f_T)^2 \right).
\end{aligned} \tag{S16}
$$

Next we need to compute the covering number $\mathcal{N}(\mathcal{H}_M, \frac{1}{\sqrt{mT}}, \|\cdot\|_\infty)$.

Consider $f_T(\rho) = W_D^\top \sigma(W_{D-1}\sigma(...\sigma(W_1\rho)))$ and $f'_T(\rho) = W_D'^\top \sigma(W'_{D-1}\sigma(...\sigma(W'_1\rho)))$ with different sets of weight matrices, we first notice that

$$\begin{aligned}\|M(f_T) - M(f'_T)\|_\infty &= \sup_\rho |(f_T(\rho) - F(\rho))^2 - (f'_T(\rho) - F(\rho))^2| \\ &= \sup_\rho |(f_T(\rho) + f'_T(\rho) - 2F(\rho))(f_T(\rho) - f'_T(\rho))| \\ &\le 4B \|f_T - f'_T\|_\infty .\end{aligned}$$

Next we get the bound based on weight matrices. Specifically, given two different sets of matrices $W_1, \cdots, W_D$ and $W'_1, \cdots, W'_D$, we have

$$\begin{aligned}&\|f_T - f'_T\|_\infty \\ \le& \left\|W_D^\top \sigma(W_{D-1}\sigma(...\sigma(W_1\rho)...)) - (W'_D)^\top \sigma(W'_{D-1}\sigma(...\sigma(W'_1\rho)...))\right\|_2 \\ \le& \left\|W_D^\top \sigma(W_{D-1}\sigma(...\sigma(W_1\rho)...)) - (W'_D)^\top \sigma(W_{D-1}\sigma(...\sigma(W_1\rho)...))\right\|_2 \\ &+ \left\|(W'_D)^\top \sigma(W_{D-1}\sigma(...\sigma(W_1\rho)...)) - (W'_D)^\top \sigma(W'_{D-1}\sigma(...\sigma(W'_1\rho)...))\right\|_2 \\ \le& \|W_D - W'_D\|_2 \|\sigma(W_{D-1}\sigma(...\sigma(W_1\rho)...))\|_2 \\ &+ \|W'_D\|_2 \left\|\sigma(W_{D-1}\sigma(...\sigma(W_1\rho)...)) - \sigma(W'_{D-1}\sigma(...\sigma(W'_1\rho)...))\right\|_2 .\end{aligned}$$

Note that we have

$$\begin{aligned}\|\sigma(W_{D-1}\sigma(...\sigma(W_1\rho)...))\|_2 &\overset{\text{(i)}}{\le} \|W_{D-1}\sigma(...\sigma(W_1\rho)...)\|_2 \\ &\le \|W_{D-1}\|_2 \|\sigma(...\sigma(W_1\rho)...)\|_2 \overset{\text{(ii)}}{\le} B_W^{D-1} \|\rho\|_2 \overset{\text{(iii)}}{\le} B_W^{D-1},\end{aligned}$$

where (i) comes from the definition of the ReLU activation, (ii) comes from $\|W_i\|_2 \le B_W$ and recursion, and (iii) comes from the boundedness of ρ. Accordingly, we have

$$\begin{aligned}\|M(f_T) - M(f'_T)\|_\infty &\le 4B \|f_T(\rho) - f'_T(\rho)\|_\infty \\ &\le 4B(B_W^{D-1} \|W_D - W'_D\|_2 + \|W'_D\|_2 \left\|\sigma(W_{D-1}\sigma(...)) - \sigma(W'_{D-1}\sigma(...))\right\|_2) \\ &\overset{\text{(i)}}{\le} 4B(B_W^{D-1} \|W_D - W'_D\|_2 + B_W \left\|W_{D-1}\sigma(...) - W'_{D-1}\sigma(...)\right\|_2) \\ &\overset{\text{(ii)}}{\le} 4BB_W^{D-1} \sum_{i=1}^D \|W_i - W'_i\|_2 ,\end{aligned} \tag{S17}$$

where (i) comes from the fact that $\forall A_1, A_2 \in \mathbb{R}^{a\times b}, \|\sigma(A_1) - \sigma(A_2)\|_2 \le \|A_1 - A_2\|_2$, and (ii) comes from the recursion. We then derive the covering number of $\mathcal{H}_M$ by the Cartesian product

of the matrix covering of $W_1, ..., W_D$:

$$\begin{aligned}\mathcal{N}(\mathcal{H}_M, \epsilon, \|\cdot\|_\infty) &\overset{\text{(i)}}{\leq} \prod_{i=1}^{D} \mathcal{N}\left(W_i, \frac{\epsilon}{4BB_W^{D-1}D}, \|\cdot\|_2\right) \\ &\overset{\text{(ii)}}{\leq} \prod_{i=1}^{D} \mathcal{N}\left(W_i, \frac{\epsilon}{4BB_W^{D-1}D}, \|\cdot\|_{\mathrm{F}}\right) \\ &\overset{\text{(iii)}}{\leq} \left(1 + \frac{8BB_W^D D\sqrt{d}}{\epsilon}\right)^{d^2 D}. \end{aligned} \tag{S18}$$

Here (i) utilizes the fact that if $\forall i = 1, 2, \cdots, D$, matrix set

$$\left\{ V_{i,j_i} \middle| j_i = 1, 2, \cdots, \mathcal{N}\left(W_i, \frac{\epsilon}{4BB_W^{D-1}D}, \|\cdot\|_2\right) \right\}$$

is a $\frac{\epsilon}{4BB_W^{D-1}D}-$covering of set $\{W_i | \|W_i\|_2 \leq B_W\}$, then by (S17) we have function set

$$\left\{ V_{D,j_D}^\top \sigma(V_{D-1,j_{D-1}} \sigma(...\sigma(V_{1,j_1}\rho)...)) \middle| 1 \leq j_i \leq \mathcal{N}\left(W_i, \frac{\epsilon}{4BB_W^{D-1}D}, \|\cdot\|_2\right), \forall 1 \leq i \leq D \right\}$$

is an $\epsilon-$covering of $\mathcal{H}_M$. (ii) comes from the fact that for any matrix W we have $\|W\|_2 \leq \|W\|_{\mathrm{F}}$, and (iii) employs Lemma 3. Plugging (S18) into (S16), we get

$$\begin{aligned}&\sup_{f_T \in H_f} (F(\widehat{\rho}^{(T)}) - F(\rho^*))^2 \\ &\leq 4C(T)\left(\frac{96B^2}{\sqrt{mT}}\sqrt{d^2 D \log(1 + 8BDB_W^D\sqrt{mTd})} + 12B^2\sqrt{\frac{2\log\frac{2}{\delta}}{mT}} + \frac{8}{mT} + \widehat{\mathbb{E}}_{K_T}(F - f_T)^2\right).\end{aligned} \tag{S19}$$

Since we consider the empirical MSE training loss to be less than ϵ, i.e.,

$$\widehat{\mathbb{E}}_{K_T}(F - f_T)^2 \leq \epsilon, \tag{S20}$$

so by plugging (S20) into (S19), we get the desired result. □

References

[1] Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML* (2010).

[2] Bartlett, P. L., Foster, D. J. & Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 6240–6249 (2017).

[3] Neyshabur, B., Bhojanapalli, S. & Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564* (2017).

[4] Chen, M., Li, X. & Zhao, T. On generalization bounds of a family of recurrent neural networks. *arXiv preprint arXiv:1910.12947* (2019).

[5] Blair, C. Problem complexity and method efficiency in optimization (as nemirovsky and db yudin). *SIAM Review* **27**, 264 (1985).

[6] Nemirovski, A., Juditsky, A., Lan, G. & Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* **19**, 1574–1609 (2009).

[7] Lin, T., Jin, C. & Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331* (2019).

[8] Chen, M. *et al.* On computation and generalization of generative adversarial imitation learning. *arXiv preprint arXiv:2001.02792* (2020).

[9] McDiarmid, C. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, 195–248 (Springer, 1998).

[10] Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of machine learning* (MIT press, 2018).

[11] Bartlett, P. L. & Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3**, 463–482 (2002).