# E-Companion to *Using Neural Networks To Guide Data-driven Operational Decisions*

Saman Lagzi, Ningyuan Chen, Joseph Milner

**Table EC.1 Table of variables and parameters.**

| Acronym | Meaning |
|---|---|
| $f_*$ | True unknown objective function |
| $Y^i$ | A noisy observation of $f_*$ in epoch $i$ |
| $\boldsymbol{x}$ | Observed covariates |
| $\boldsymbol{z}$ | Decision taken |
| $d_X$ | Dimension of the observed covariates |
| $d_Z$ | Dimension of the decision taken |
| $f_\theta$ | A generic DNN function |
| $f_{\hat{\theta}}$ | A realized fitted DNN function |
| $\sigma(\cdot)$ | The activation function |
| $\sigma'_{h,l}(\boldsymbol{x},\boldsymbol{z})$ | The point-wise first derivative of the activation function at the $h$-th node and $l$-th layer |
| $\sigma''_{h,l}(\boldsymbol{x},\boldsymbol{z})$ | The point-wise second derivative of the activation function at the $h$-th node and $l$-th layer |
| $\boldsymbol{z}^*(\boldsymbol{x_0})$ | The true optimal solution given $\boldsymbol{x_0}$ |
| $\hat{\boldsymbol{z}}(\boldsymbol{x_0})$ | The DNN based optimal solution given $\boldsymbol{x}_0$ |
| $\partial$ | The partial derivative |
| $K_*$ | The Lipschitz constant of $f_*$ |
| $K_{\hat{\theta}}$ | The Lipschitz constant of $f_{\hat{\theta}}$ |
| $K$ | The Lipschitz constant of $\|f_* - f_{\hat{\theta}}\|$ |
| $\mu_{X,Z}$ | The marginal distribution of $(\boldsymbol{X},\boldsymbol{Z})$ |
| $\delta_N$ | The generalization bound with N data points |
| $c_1$ | The lower bound on the density function of $\mu_{X,Z}$ |
| $c_3$ | Min. radius of a ball away from the optimal solution to have probability measure $c_4$ |
| $c_4$ | Probability measure of a ball of radius $c_3$ away from the optimal solution |
| $c_2$ | Lower bound on the increase of $f_*$ for solutions far away from optimal solution |
| $c_5$ | Quadratic increase rate of $f_*$ for solutions near the optimal solution |
| $L_{d_X+d_Z}$ | Constant for volume of a $d_X+d_Z$ dimensional ball |

## EC.1. Using Neural Networks to Fit the Objective Function

As mentioned in Section 3, we propose to find the best fit objective function $\hat{f}(\boldsymbol{x},\boldsymbol{z})$ from the data $\{(\boldsymbol{X}^i,\boldsymbol{Z}^i,Y^i)\}_{i=1}^N$ within the class of DNN functions $\mathscr{F}_{\text{NN}}$. That is, assuming a quadratic loss function, and a clear specification of $\mathscr{F}_{\text{NN}}$, we seek:

$$\hat{f} = \arg\min_{f\in\mathscr{F}_{\text{NN}}} \sum_{i=1}^{N} (f(\boldsymbol{X}^i,\boldsymbol{Z}^i) - Y^i)^2. \tag{EC.1}$$

The class $\mathscr{F}_{\text{NN}}$ is the set of functions that can be represented through a particular DNN. The function class $\mathscr{F}_{\text{NN}}$ is determined by the architecture of the network and the form of interactions between the nodes. We consider *feedforward* DNNs consisting of:

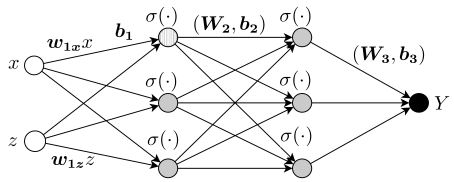


**Figure EC.1 The illustration of a fully connected feedforward neural network with $L=2$ layers and $H=3$ nodes in each layer. The activation function is applied to the input at each of the nodes.**

- $L$ layers (the depth of the network),
- $H$ nodes per layer (the width)[1],
- fully connected layers,
- an activation function $\sigma(\cdot):\mathbb{R}\to\mathbb{R}$.

An example of a network architecture is illustrated in Figure EC.1. The neural network takes as input two scalars $(x,z)$ and outputs a single scalar $(y)$. It has $L=2$ layers and each layer has $H=3$ width (note that we do not count the first and the last layers).

The commonly use activation functions are:

1. **sigmoid**: $\sigma(x)=1/(1+e^{-x})$
2. **Swish**: $\sigma(x)=x/(1+e^{-x})$
3. **ReLU**: $\sigma(x)=\max\{0,x\}$.

Given a network architecture $\mathscr{F}_{\text{NN}}(L,H,\sigma)$, a function withing the class is specified by a parameterization $\theta=(\boldsymbol{W}_{1x},\boldsymbol{W}_{1z},\boldsymbol{b}_1,\boldsymbol{W}_2,\boldsymbol{b}_2,\ldots,\boldsymbol{W}_L,\boldsymbol{b}_L,\boldsymbol{w}_{L+1},b_{L+1})$, where $\boldsymbol{W}_{1x}\in\mathbb{R}^{H\times d_X}$, $\boldsymbol{W}_{1z}\in\mathbb{R}^{H\times d_Z}$, $\boldsymbol{b}_k\in\mathbb{R}^H$ for $k=1,\ldots,L$, $\boldsymbol{W}_k\in\mathbb{R}^{H\times H}$ for $k=2,\ldots,L$, $\boldsymbol{w}_{L+1}\times\mathbb{R}^H$ and $b_{L+1}\in\mathbb{R}$. Given $(\boldsymbol{x},\boldsymbol{z})$ and a parameter $\theta$,

$$f_\theta(\boldsymbol{x},\boldsymbol{z})=\boldsymbol{w}_{\boldsymbol{L+1}}^\top\boldsymbol{\sigma}\Big(\cdots\boldsymbol{\sigma}\Big(\boldsymbol{W_3}\boldsymbol{\sigma}\big(\boldsymbol{W_2}\boldsymbol{\sigma}(\boldsymbol{W_{1x}}\boldsymbol{x}+\boldsymbol{W_{1z}}\boldsymbol{z}+\boldsymbol{b_1})+\boldsymbol{b_2}\big)+\boldsymbol{b_3}\Big)+\cdots\Big)+b_{L+1}.$$

Here, $\boldsymbol{\sigma}(\cdot):\mathbb{R}^H\to\mathbb{R}^H$ represents the vectorized activation function that applies $\sigma(\cdot):\mathbb{R}\to\mathbb{R}$ elementwise.

By this formulation, $\mathscr{F}_{\text{NN}}$ is the set of all functions $f_\theta(\boldsymbol{x},\boldsymbol{z})$. One may use Figure EC.1 to better understand $f_\theta(\boldsymbol{x},\boldsymbol{z})$. Consider the left-most layer (the input layer), the edges leading to the second layer (also referred to as the first hidden layer) and the term $\boldsymbol{w_{1x}}x+\boldsymbol{w_{1z}}z+\boldsymbol{b_1}$. The upper edges coming from node $x$ represent $\boldsymbol{w_{1x}}x$ and the lower edges coming from node $z$ represent $\boldsymbol{w_{1z}}z$. The vector $\boldsymbol{b_1}$ is referred to as the constant term (or bias). Observe that $\boldsymbol{w_{1x}}x:\mathbb{R}^{d_x}\to\mathbb{R}^H$. In this case the scalar $x$ is transferred by the weights $\boldsymbol{w_{1x}}$ into $H$ inputs to the

[1] This is without loss of generality since $H$ can be set equal to the number of nodes in the widest layer, and for other layers, one can consider dummy nodes with weight set to zero and the constant parameter set to be equal to the root point of the activation function (0 for ReLU and Swish and a large negative number such as $-100$ for sigmoid).

second layer, and similarly for $z$. At each node in the second layer the function $\sigma(\cdot):\mathbb{R}^1\to\mathbb{R}^1$ is applied to the $H$-dimensional term $\boldsymbol{w_{1x}}x+\boldsymbol{w_{1z}}z+\boldsymbol{b_1}$ element wise to produce $H$ outputs. Then, recursively, in a feedforward fashion, the results are fed to the edges between the second and third layer, represented by $\boldsymbol{W_2}$ and $\boldsymbol{b_2}$, and so on. In the output layer, an inner product using weight $\boldsymbol{w_{L+1}}$ is applied to the output of the penultimate layer (the last hidden layer). Adding the scalar constant term $b_{L+1}$ gives the scalar output.

Given observations $\{(\boldsymbol{X}^i,\boldsymbol{Z}^i,Y^i)\}_{i=1}^N$, we seek $f_\theta$ that solves (EC.1). That is, we train the DNN by selecting $\boldsymbol{W}$'s and $\boldsymbol{b}$'s that minimize the quadratic loss. A common approach is to use Stochastic Gradient Descent (SGD) which relies on the observation that the gradient of $f_\theta$ with respect to $\theta$ can be given in closed form, when $\sigma$ is differentiable (Rumelhart et al. 1986). SGD has shown good empirical performance (Zhang et al. 2021). See Goodfellow et al. (2016) for an overview of SGD applied to fitting DNNs.

## EC.2. Optimize the Fitted Neural Network

In this section we focus on the case where $\mathcal{P}$ is a box and discuss how to choose $\boldsymbol{z}$ given a covariate $\boldsymbol{X}_0$ so that the output of the fitted DNN is optimized. Let $\hat{\theta}$ denote the fitted DNN (the optimized weights $\boldsymbol{W}$ and constant terms $\boldsymbol{b}$). Then, following (5), we would want to find

$$\hat{\boldsymbol{z}}(\boldsymbol{X}_0)\in\arg\min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0,\boldsymbol{z}). \tag{EC.2}$$

We assume

$$\mathcal{P}=\{\boldsymbol{z}\in R^{d_Z}\,|\,\underline{\boldsymbol{Z}}_i\le\boldsymbol{z}_i\le\bar{\boldsymbol{Z}}_i,\ \forall\ i\le d_Z\},$$

where $\underline{\boldsymbol{Z}}_i$ and $\bar{\boldsymbol{Z}}_i$ indicate the lowest and highest values of the $i^{th}$ coordinate of $\boldsymbol{Z}$ in the data. Consider Figure EC.1. We are suggesting that we have $\boldsymbol{X}_0$ and $f_{\hat{\theta}}$ and want to choose $\boldsymbol{z}$ to minimize $Y$. Solving (EC.2) may be NP-hard, depending on the choice of the activation function (Katz et al. 2017, Anderson et al. 2020).

Define $\hat{\boldsymbol{z}}(\boldsymbol{x})$ to be a local minimum if there exists $\delta>0$, such that

$$\forall\boldsymbol{z}\in\mathcal{P}:\|\boldsymbol{z}-\hat{\boldsymbol{z}}(\boldsymbol{x})\|^2\le\delta, f_{\hat{\theta}}(\boldsymbol{x},\hat{\boldsymbol{z}}(\boldsymbol{x}))\le f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z}). \tag{EC.3}$$

We suggest using a gradient-based approach to finding a local minimum to (EC.2). We can use backpropagation to derive the gradient of $f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})$ with respect to $\boldsymbol{z}$ in closed form. Hence, we may hope to efficiently find all the local minima to (EC.2). Moreover, we can use the tools developed for SGD in the training of DNNs to find the $\hat{\boldsymbol{z}}(\boldsymbol{x})$ that minimizes $f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})$.

Similar to the the backpropagation algorithm (Rumelhart et al. 1986), we use the layer by layer structure of the DNN to calculate the derivative of $f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})$ with respect to $\boldsymbol{z}$.

We can find a local minimum of $f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ using a gradient descent algorithm. Let $\hat{g} = \frac{\partial f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{\partial \boldsymbol{z}}$ and let $\hat{K}$ be the Lipschitz constant of $\frac{\partial f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{\partial \boldsymbol{z}}$ and let and $[\boldsymbol{z}]^+$ be the projection operator onto the feasible set $\mathcal{P}$, defined as

$$[\boldsymbol{z}]_i^+ = \begin{cases} \bar{\boldsymbol{Z}}_i & \text{if } \boldsymbol{z}_i \geq \bar{\boldsymbol{Z}}_i, \\ \underline{\boldsymbol{Z}}_i & \text{if } \boldsymbol{z}_i \leq \underline{\boldsymbol{Z}}_i, \\ \boldsymbol{z}_i & \text{otherwise.} \end{cases}$$

**Algorithm 2** Projected gradient update

**Require:** $\hat{K}$, $0 < s < \frac{2}{\hat{K}}$

**Require:** Starting point decision value $\boldsymbol{z}^{(0)} \in \mathcal{P}$

$k \leftarrow 1$

**while** Stopping criterion not met **do**

Compute gradient: $\hat{g} \leftarrow \frac{\partial \hat{f}_{\theta}(\boldsymbol{x}, \boldsymbol{z}^{(k)})}{\partial \boldsymbol{z}}$

Apply update: $\hat{\boldsymbol{z}}^{(k)} \leftarrow \boldsymbol{z}^{(k)} - s\hat{g}$

Project back to feasible set: $\boldsymbol{z}^{(k+1)} \leftarrow [\hat{\boldsymbol{z}}^{(k)}]^+$

$k \leftarrow k + 1$

**end while**

There is a large body of literature on how to choose the step size ($s$) in Algorithm 2 to ensure faster convergence (Bertsekas 1999). However, if $f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ is continuously differentiable, $\mathcal{P}$ is a compact and nonempty convex set, and $\frac{\partial f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{\partial \boldsymbol{z}}$ is Lipschitz continuous with a known Lipschitz constant, Algorithm 2 as stated will converge in the limit to a stationary point of $f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ (Bertsekas (1999), Proposition 2.3.2). If $\frac{\partial f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{\partial \boldsymbol{z}}$ is only locally Lipschitz continuous or its Lipschitz constant is not known, one can alternatively use the celebrated Armijo rule to determine the step size schedule that will guarantee convergence in the limit (Bertsekas (1999), Proposition 2.3.3).

Remark EC.1. The differentiability of $\sigma$ is imperative for Proposition 1. If $\sigma$ is a non-differentiable function such as ReLU, as pointed out by several recent works such as Bolte and Pauwels (2021) and Bianchi et al. (2022), backpropagation for ReLU networks does not necessarily compute any kind of known derivative or even subdifferential of $f_{\hat{\theta}}$ and may not even return the correct derivative of $f_{\hat{\theta}}$ at points where it is differentiable. Moreover, we do not know of any research that shows in such a case, similar to Proposition 1, convergence to a stationary point of $f_{\hat{\theta}}(\boldsymbol{x}, \cdot)$ would result. In particular, the results of Bianchi et al. (2022) which are developed for SGD, can be extended to the deterministic case of Algorithm 2 for ReLU activated DNNs, with backpropagation used to calculate $\hat{g}$. However their results only guarantee convergence to the set

of zeros of the conservative field of $f_{\hat{\theta}}(\boldsymbol{x}, \cdot)$ and this set is not necessarily identical to the set of its stationary points. (See Bolte and Pauwels (2021) for a comprehensive review). □

## EC.3. The Choice of Activation Functions

The choice of activation functions has a substantial impact on the performance of DNNs, as documented in a number of empirical studies (for example, see Ramachandran et al. 2017). Although the field of deep learning is evolving rapidly and new activation functions are constantly proposed, ReLU remains one of the most popular choices (Ramachandran et al. 2017, Agarap 2018, Nwankpa et al. 2018). In addition to the empirical success, DNNs with ReLU activation functions possess benign theoretical properties, achieving the minimax convergence rate (Schmidt-Hieber 2020).

In our framework, since the quality of the data-driven decision is closely related to how well $f_*$ can be learned from the data and ReLU has had tremendous empirical success in deep learning, it seems a natural choice to use ReLU. However, we point out a surprising observation that is specific to our framework using DNNs for data-driven decision-making: using ReLU may lead to *worse* data-driven decisions compared to smooth activation functions such as sigmoid and Swish. We illustrate this observation using the following toy example.

Example EC.1 (Using ReLU leads to worse outcomes). Consider a scenario without covariates and a one-dimensional objective function $f_*(z) = z - z^2$, $z \in [0, 1]$. The function is shown by the dashed lines in Figure EC.2. We compare the ability of two DNNs to fit this function. The DNNs have one hidden layer and 29 nodes. They differ only in their activation function, one with ReLU and one with Swish. We then sample a large number of observations of $f_*(z)$ and fit the DNNs. The ReLU network has a smaller out of sample prediction error (MSE $7.24 \times 10^{-6}$) compared to that of the Swish network (MSE $1.86 \times 10^{-5}$). After fitting the DNNs, we solve for the optimal $\hat{z}$ for the DNNs. In Figure EC.2 we observe the optimal $\hat{z}$ for the Swish network is much closer to $z^*$ than that of the ReLU network.

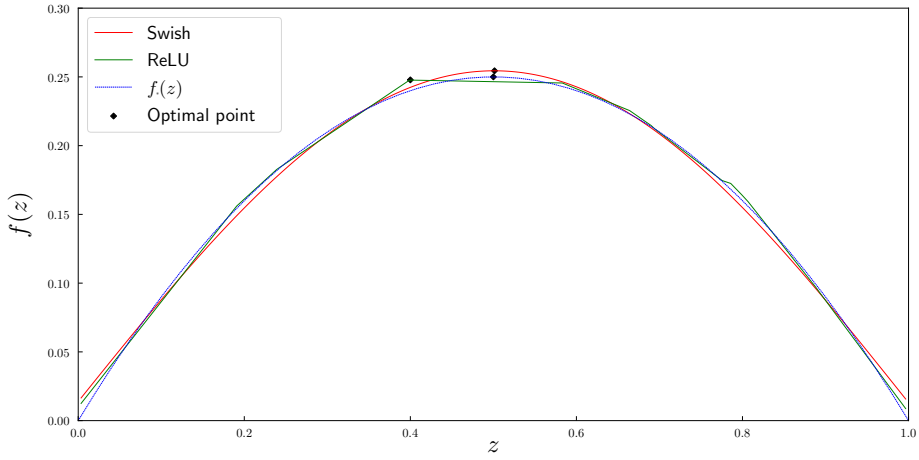


**Figure EC.2 The illustration of the prediction value for DNNs using ReLU and Swish activation functions.**

## EC.4. Proofs

*Proof of Proposition 1* As $\sigma$ is differentiable everywhere $f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})$ is differentiable. Thus, we can apply the chain rule layer by layer, inductively.

Then, we observe

$$\frac{\partial \sigma_{h,1}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \sigma'(\boldsymbol{w}_{\boldsymbol{h,1x}}^{\top}\boldsymbol{x} + \boldsymbol{w}_{\boldsymbol{h,1z}}^{\top}\boldsymbol{z} + b_{h,1}).\boldsymbol{w}_{\boldsymbol{h,1z}}.$$

So

$$\frac{\partial \boldsymbol{\sigma_1}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \operatorname{diag}\big(\sigma'_{1,1}(\boldsymbol{x},\boldsymbol{z}),\ldots,\sigma'_{H,1}(\boldsymbol{x},\boldsymbol{z})\big) \times \boldsymbol{W_{1z}}.$$

Furthermore,

$$\frac{\partial \boldsymbol{\sigma_2}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \frac{\partial \boldsymbol{\sigma_2}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{\sigma_1}(\boldsymbol{x},\boldsymbol{z})} \times \frac{\partial \boldsymbol{\sigma_1}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \operatorname{diag}\big(\sigma'_{1,2}(\boldsymbol{x},\boldsymbol{z}),\ldots,\sigma'_{H,2}(\boldsymbol{x},\boldsymbol{z})\big) \times \boldsymbol{W_2} \times \frac{\partial \boldsymbol{\sigma_1}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}}.$$

Assume for some $l \geq 2$, that

$$\frac{\partial \boldsymbol{\sigma_l}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \operatorname{diag}\big(\sigma'_{1,l}(\boldsymbol{x},\boldsymbol{z}),\ldots,\sigma'_{H,l}(\boldsymbol{x},\boldsymbol{z})\big) \times \boldsymbol{W_l} \times \frac{\partial \boldsymbol{\sigma_{l-1}}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}}.$$

Then, we have

$$\frac{\partial \boldsymbol{\sigma_{l+1}}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \frac{\partial \boldsymbol{\sigma_{l+1}}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{\sigma_l}(\boldsymbol{x},\boldsymbol{z})} \times \frac{\partial \boldsymbol{\sigma_l}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \operatorname{diag}\big(\sigma'_{1,l+1}(\boldsymbol{x},\boldsymbol{z}),\ldots,\sigma'_{H,l+1}(\boldsymbol{x},\boldsymbol{z})\big) \times \boldsymbol{W_{l+1}} \times \frac{\partial \boldsymbol{\sigma_l}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}}.$$

Then, finally,

$$\frac{\partial f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \frac{\partial f_{\hat{\theta}}}{\partial \boldsymbol{\sigma_L}(\boldsymbol{x},\boldsymbol{z})} \times \frac{\partial \boldsymbol{\sigma_L}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}} = \boldsymbol{w}_{L+1}^{\top} \times \frac{\partial \boldsymbol{\sigma_L}(\boldsymbol{x},\boldsymbol{z})}{\partial \boldsymbol{z}},$$

which completes the proof. □

*Proof of Proposition 2* We note that by assumption $\sigma(\cdot)$ is twice continuously differentiable and hence $f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})$ is twice continuously differentiable. Moreover, we can apply the chain rule to the nodes in each layer, inductively. Thus, the proof follows from applying the chain rule to $\sigma(\cdot)$ and its derivative. To elaborate, it is trivial to show that

$$\frac{\partial \sigma_{h,1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j} = w_{h1}^{j}\sigma'_{h,1}(\boldsymbol{x},\boldsymbol{z}),$$

$$\frac{\partial^2 \sigma_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j \partial z_k} = w_{h1}^{j} w_{h1}^{k}\sigma''_{h,l}(\boldsymbol{x},\boldsymbol{z}).$$

Now, assume $l \geq 2$, from the chain rule we have

$$\frac{\partial \sigma_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j} = \frac{\partial \sigma_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\sigma_{h',l-1}(\boldsymbol{x},\boldsymbol{z})+b_{h,l}\big)}\frac{\partial\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\sigma_{h',l-1}(\boldsymbol{x},\boldsymbol{z})+b_{h,l}\big)}{\partial z_j} = \sigma'_{h,l}(\boldsymbol{x},\boldsymbol{z})\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\frac{\partial \sigma_{h,l-1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j}\big),$$

while

$$\frac{\partial^2 \sigma_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j \partial z_k} = \frac{\partial \sigma'_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial z_k}\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\frac{\partial \sigma_{h,l-1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j}\big) + \frac{\partial\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\frac{\partial \sigma_{h,l-1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j}\big)}{\partial z_k}\sigma'_{h,l}(\boldsymbol{x},\boldsymbol{z})$$

by applying the chain rule to the term $\frac{\partial \sigma'_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial z_k}$, we get

$$\begin{aligned}\frac{\partial^2 \sigma_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j \partial z_k} =& \frac{\partial \sigma'_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\sigma_{h',l-1}(\boldsymbol{x},\boldsymbol{z})+b_{h,l}\big)}\frac{\partial\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\sigma_{h',l-1}(\boldsymbol{x},\boldsymbol{z})+b_{h,l}\big)}{\partial z_k}\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\frac{\partial \sigma_{h,l-1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j}\big)\\ &+ \sigma'_{h,l}(\boldsymbol{x},\boldsymbol{z})\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\frac{\partial^2 \sigma_{h',l-1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j \partial z_k}\big),\end{aligned}$$

which leads to

$$\frac{\partial^2 \sigma_{h,l}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j \partial z_k} = \sigma''_{h,l}(\boldsymbol{x},\boldsymbol{z})\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\frac{\partial \sigma_{h',l-1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_k}\big)\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\frac{\partial \sigma_{h',l-1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j}\big) + \sigma'_{h,l}(\boldsymbol{x},\boldsymbol{z})\big(\sum_{h'=1}^{H} w_{h,l}^{h'}\frac{\partial^2 \sigma_{h',l-1}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j \partial z_k}\big).$$

The last part of the Proposition follows trivially from noticing that $f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z}) = \sum_{h'=1}^{H} w_{L+1}^{h'}\sigma_{h,L}(\boldsymbol{x},\boldsymbol{z})$, and hence:

$$\frac{\partial^2 f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j \partial z_k} = \sum_{h=1}^{H} w_{L+1}^{h}\frac{\partial^2 \sigma_{h',L}(\boldsymbol{x},\boldsymbol{z})}{\partial z_j \partial z_k}.$$

□

*Proof of Lemma 1.* From Algorithm 1 we have:

$$\hat{\boldsymbol{z}}(\boldsymbol{X}_0) \in \arg\min_{\boldsymbol{z}\in\mathcal{P}} \|\boldsymbol{z} - \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)\|.$$

Thus we must have:

$$\|\hat{\boldsymbol{z}}(\boldsymbol{X}_0) - \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)\| \leq \max_{\boldsymbol{z}\in\bar{\mathcal{P}}}\min_{\boldsymbol{z}'\in\mathcal{P}} \|\boldsymbol{z} - \boldsymbol{z}'\|.$$

Moreover, $f_{\hat{\theta}}$ is Lipschitz continuous, as it is a composition of Lipschitz continuous functions. Hence, if $K_{\hat{\theta}}$ is its Lipschitz constant, then we have:

$$|f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}(\boldsymbol{X}_0)) - f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0))| \leq K_{\hat{\theta}}(\max_{\boldsymbol{z}\in\bar{\mathcal{P}}}\min_{\boldsymbol{z}'\in\mathcal{P}} \|\boldsymbol{z} - \boldsymbol{z}'\|).$$

Now, if $f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)) \leq \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z})$ then by definition we have

$$f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)) \leq \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z}) \leq f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}(\boldsymbol{X}_0))$$

which implies

$$|f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}(\boldsymbol{X}_0)) - \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z})| \leq K_{\hat{\theta}}(\max_{\boldsymbol{z}\in\bar{\mathcal{P}}} \min_{\boldsymbol{z}'\in\mathcal{P}} \|\boldsymbol{z} - \boldsymbol{z}'\|).$$

Hence, assume $f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)) \geq \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z})$. We have

$$\begin{aligned}
&|f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}(\boldsymbol{X}_0)) - \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z})| = \\
&|f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}(\boldsymbol{X}_0)) - f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)) + f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)) - \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z})| \leq \\
&|f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}(\boldsymbol{X}_0)) - f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0))| + |f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)) - \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z})| \leq \\
&K_{\hat{\theta}}(\max_{\boldsymbol{z}\in\bar{\mathcal{P}}} \min_{\boldsymbol{z}'\in\mathcal{P}} \|\boldsymbol{z} - \boldsymbol{z}'\|) + |f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)) - \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z})| \leq \\
&K_{\hat{\theta}}(\max_{\boldsymbol{z}\in\bar{\mathcal{P}}} \min_{\boldsymbol{z}'\in\mathcal{P}} \|\boldsymbol{z} - \boldsymbol{z}'\|) + |f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)) - \min_{\boldsymbol{z}\in\bar{\mathcal{P}}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z})|
\end{aligned}$$

where the last inequality follows from the fact that by definition $\bar{\mathcal{P}}$ is a superset of $\mathcal{P}$ and hence

$$\min_{\boldsymbol{z}\in\bar{\mathcal{P}}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z}) \leq \min_{\boldsymbol{z}\in\mathcal{P}} f_{\hat{\theta}}(\boldsymbol{X}_0, \boldsymbol{z}) \leq f_{\hat{\theta}}(\boldsymbol{X}_0, \hat{\boldsymbol{z}}_{\bar{\mathcal{P}}}(\boldsymbol{X}_0)),$$

which completes the proof. □

*Proof of Proposition 3.* We first express $f_*(\boldsymbol{x}, z)$ in a closed form and then its weak derivatives with respect to its input. Note that the following expressions are all in the almost-everywhere sense, which is how weak derivatives are defined. Because

$$f_*(\boldsymbol{x}, z) = \int_{-\infty}^{z-g(\boldsymbol{x})} h(z - g(\boldsymbol{x}) - u) f_\mu(u) du + \int_{z-g(\boldsymbol{x})}^{\infty} b(g(\boldsymbol{x}) + u - z) f_\mu(u) du,$$

we have

$$\frac{\partial f_*(\boldsymbol{x}, z)}{\partial z} = \int_{-\infty}^{z-g(\boldsymbol{x})} h f_\mu(u) du - \int_{z-g(\boldsymbol{x})}^{\infty} b f_\mu(u) du \leq \max\{h, b\} M_\epsilon \|\bar{\epsilon} - \underline{\epsilon}\|$$

This suggests that

$$\frac{\partial^2 f_*}{\partial z^2} = (h - b) f_\mu(z - g(\boldsymbol{x}))$$

and according to the assumptions on $g$ and$f_\mu$, we have

$$\frac{\partial^2 f_*}{\partial z^2} \in \mathcal{W}^{\min\{m,p\},\infty}([-1, 1]^{d_X+1}, (h - b) M_\epsilon).$$

Similarly, we have

$$\frac{\partial^2 f_*}{\partial z \, \partial x_i} = (b - h) \frac{\partial g}{\partial x_i} f_\mu(z - g(\boldsymbol{x})).$$

By the assumption on $g$ and $f_\mu$,

$$\frac{\partial^2 f_*}{\partial z\,\partial x_i} \in \mathcal{W}^{\min\{m,p-1\},\infty}([-1,1]^{d_X+1}, (b-h)M_\epsilon M_g^2).$$

Furthermore, we have

$$\frac{\partial f_*}{\partial x_i} = \int_{z-g(\boldsymbol{x})}^{\infty} b\frac{\partial g}{\partial x_i} f_\mu(u)du - \int_{-\infty}^{z-g(\boldsymbol{x})} h\frac{\partial g}{\partial x_i} f_\mu(u)du \le \max\{h,b\}M_\epsilon M_g \|\bar{\epsilon}-\underline{\epsilon}\|.$$

Finally,

$$\frac{\partial^2 f_*}{\partial x_i\,\partial x_j} = \frac{\partial g}{\partial x_j}(h-b)(\frac{\partial g}{\partial x_i}) f_\mu(z-g(\boldsymbol{x})) + \int_{z-g(\boldsymbol{x})}^{\infty} b\frac{\partial^2 g}{\partial x_i\,\partial x_j} f_\mu(u)du - \int_{-\infty}^{z-g(\boldsymbol{x})} h\frac{\partial^2 g}{\partial x_i\,\partial x_j} f_\mu(u)du.$$

By the assumption that $f_\mu(u)=0,\ \forall\ u\notin[\underline{\epsilon},\bar{\epsilon}]$ and given $g$ is bounded and in $\mathcal{W}^{p,\infty}([-1,1]^{d_X}, M_g)$, there must exist $\hat{M}<\infty$ such that

$$\frac{\partial^2 f_*}{\partial x_i\,\partial x_j} \in \mathcal{W}^{\min\{m,(p-2)^+\},\infty}([-1,1]^{d_X}, \hat{M}).$$

Given the fact that $\underline{\epsilon}\le\epsilon\le\bar{\epsilon}$, while $g$ is bounded and in $\mathcal{W}^{p,\infty}([-1,1]^{d_X}, M_g)$, we must have $f_*(\boldsymbol{x},z)\le\bar{M}$ for some $\bar{M}<\infty$. Moreover, remember we assumed the marginal probability measure of $(\boldsymbol{X},Z)$ is absolutely continuous with respect to the Lebesgue measure, while $[-1,1]^{d_X+1}$, which is a compact set, is the support of $(\boldsymbol{x},z)$. As such, for $\beta=\min\{m+1,p\}$, there exists $M<\infty$ such that

$$f_*(\boldsymbol{x},z) \in \mathcal{W}^{\beta,\infty}([-1,1]^{d_X+1}, M).$$

□

*Proof of Proposition 4.* By Assumption 2, we have that $\mathbb{E}[(f_*(\boldsymbol{X},\boldsymbol{Z})-f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2]\le\delta_N$. Recall that $\underline{u}>0$ is the lower bound for the conditional PMF $\mu_{\cdot|X}$ for all $\boldsymbol{x}$ and $\boldsymbol{z}$. It is easy to show that the marginal PMF $\mu_Z\ge\underline{u}$ as well:

$$\mu_{\boldsymbol{Z}}(\boldsymbol{z}) = \int_{\boldsymbol{x}\in[-1,1]^{d_X}} \mu_{\cdot|\boldsymbol{X}}(\boldsymbol{z})d\mu_{\boldsymbol{X}} \ge \int_{\boldsymbol{x}\in[-1,1]^{d_X}} \underline{u}\,d\mu_{\boldsymbol{X}} = \underline{u}.$$

Then we have for all $\boldsymbol{z}'\in\mathcal{P}$:

$$\mathbb{E}[(f_*(\boldsymbol{X},\boldsymbol{Z})-f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2|\boldsymbol{Z}=\boldsymbol{z}']\times\underline{u} \le \mathbb{E}[(f_*(\boldsymbol{X},\boldsymbol{Z})-f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2|\boldsymbol{Z}=\boldsymbol{z}']\times\Pr(\boldsymbol{Z}=\boldsymbol{z}').$$

Moreover, by the law of total probability, we have

$$\begin{aligned}\mathbb{E}[(f_*(\boldsymbol{X},\boldsymbol{Z})-f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2|\boldsymbol{Z}=\boldsymbol{z}']\times\Pr(\boldsymbol{Z}=\boldsymbol{z}') &\le \sum_{\boldsymbol{z}\in\mathcal{P}} \mathbb{E}[(f_*(\boldsymbol{X},\boldsymbol{Z})-f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2|\boldsymbol{Z}=\boldsymbol{z}]\times\Pr(\boldsymbol{Z}=\boldsymbol{z}) \\ &= \mathbb{E}[(f_*(\boldsymbol{X},\boldsymbol{Z})-f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2].\end{aligned}$$

As a result, we have

$$\mathbb{E}[(f_*(\boldsymbol{X},\boldsymbol{Z})-f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2|\boldsymbol{Z}=\boldsymbol{z}']\leq\frac{\delta_N}{\underline{u}}.$$

For any $(\boldsymbol{x}_0,\boldsymbol{z}')$, where $\boldsymbol{x_0}$ is not an extreme point, consider a ball $B\subset\mathbb{R}^{d_X}$ of radius $b$ that contains $\boldsymbol{x}_0$, and the ball itself is fully contained in $[-1,1]^{d_X}$ (later we will show that $b$ will shrink in $\delta_N$ and hence if $\delta_N$ is small enough, this is always possible). Let $\boldsymbol{x}^{min}\in\arg\min_{\boldsymbol{x}\in B}|f_*(\boldsymbol{x},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z'})|$ and $\boldsymbol{x}^{max}\in\arg\max_{\boldsymbol{x}\in B}|f_*(\boldsymbol{x},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z'})|$. Because $d\mu_{\boldsymbol{X}}\geq\underline{u}dx$ and $\mu_{\boldsymbol{z}}\leq 1$:

$$d\mu_{\boldsymbol{X}|\boldsymbol{z}}=\frac{d\mu_{\boldsymbol{X}}\times\mu_{\cdot|\boldsymbol{X}}(\boldsymbol{z})}{\mu_{\boldsymbol{z}}}\geq\underline{u}^2dx$$

while we can recall that a ball of radius $b$ in the $d_X$-dimensional space has volume $L_{d_X}b^{d_X}$ (see Definition 1), we have

$$\begin{aligned}\underline{u}^2L_{d_X}b^{d_X}(f_*(\boldsymbol{x}^{min},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x}^{min},\boldsymbol{z'}))^2&\leq\int_{\boldsymbol{x}\in B}(f_*(\boldsymbol{x}^{min},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x}^{min},\boldsymbol{z'}))^2d\mu_{X|\boldsymbol{z}'}\\&\leq\int_{\boldsymbol{x}\in B}(f_*(\boldsymbol{x},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z'}))^2d\mu_{X|\boldsymbol{z}'}\leq\frac{\delta_N}{\underline{u}}\end{aligned}$$

Hence, we can upper bound $|f_*(\boldsymbol{x}^{min},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x}^{min},\boldsymbol{z'})|$ as:

$$|f_*(\boldsymbol{x}^{min},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x}^{min},\boldsymbol{z'})|\leq\sqrt{\frac{\delta_N}{\underline{u}^3L_{d_X}b^{d_X}}}.$$

By Assumption 3, the function $|f_*(\boldsymbol{x},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z'})|$ is Lipschitz continuous with a constant $K$. Moreover, because $\|\boldsymbol{x}^{max}-\boldsymbol{x}^{min}\|_2\leq 2b$, we have

$$|f_*(\boldsymbol{x_0},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x_0},\boldsymbol{z'})|\leq|f_*(\boldsymbol{x}^{max},\boldsymbol{z'})-f_{\hat{\theta}}(\boldsymbol{x}^{max},\boldsymbol{z'})|\leq\sqrt{\frac{\delta_N}{\underline{u}^3L_{d_X}b^{d_X}}}+2bK:=\epsilon.\tag{EC.4}$$

We note that this inequality holds for all $\boldsymbol{z}'\in\mathcal{P}$.

Now consider $\hat{\boldsymbol{z}}(\boldsymbol{x_0})\in\arg\min_{\boldsymbol{z}\in\mathcal{P}}f_{\hat{\theta}}(\boldsymbol{x_0},\boldsymbol{z})$, the optimal DNN solution. Furthermore, with a slight abuse of notation let $\boldsymbol{z}^*\in\arg\max_{\boldsymbol{z}\in\mathcal{P}}f_*(\boldsymbol{x}_0,\boldsymbol{z})$, an optimal solution of $f_*$ in $\mathcal{P}$.

Applying the inequality (EC.4) first for $\boldsymbol{z}'=\hat{\boldsymbol{z}}(\boldsymbol{x_0})$ and then for $\boldsymbol{z}'=\boldsymbol{z}^*(\boldsymbol{x_0})$, we have

$$f_*(\boldsymbol{x_0},\hat{\boldsymbol{z}}(\boldsymbol{x_0}))\leq f_{\hat{\theta}}(\boldsymbol{x_0},\hat{\boldsymbol{z}}(\boldsymbol{x_0}))+\epsilon\leq f_{\hat{\theta}}(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))+\epsilon\leq f_*(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))+2\epsilon.$$

To complete the proof, we note that by choosing a suitable $b$ for diminishing $\delta_N$, we can shrink $\epsilon$. In particular, let

$$b=\left(\frac{(d_X)^2\delta_N}{16K^2\underline{u}^3L_{d_X}}\right)^{\frac{1}{d_X+2}}.$$

At this point note that unless $\boldsymbol{x_0}$ is an extreme point, if $\delta_N$ is small enough, then $b$ will also be small for the ball $B$ to be fully contained in $[-1,1]^{d_X}$. By plugging the obtained $b$ into

$$\epsilon=\sqrt{\frac{\delta_N}{\underline{u}^3L_{d_X}b^{d_X}}}+2bK$$

we obtain the bound

$$2(\frac{\delta_N d_X}{4K\underline{u}^3 L_{d_X}} + 2\delta_N^{\frac{1}{d_X+2}}(\frac{K^{d_X} d_X^2}{16\underline{u}^3 L_{d_X}})^{\frac{1}{d_X+2}}).$$

If $\boldsymbol{x_0}$ is an extreme point, then let the ball $B$ be centered around $\boldsymbol{x_0}$. We have

$$\frac{1}{2^{d_X}}\underline{u}^2 L_{d_X} b^{d_X}(f_*(\boldsymbol{x}^{min}, \boldsymbol{z'}) - f_{\hat{\theta}}(\boldsymbol{x}^{min}, \boldsymbol{z'}))^2 \leq \frac{\delta_N}{\underline{u}}$$

giving raise to the following alternative bound:

$$2(\frac{\delta_N d_X}{4K\underline{u}^3 2^{d_X} L_{d_X}} + 2\delta_N^{\frac{1}{d_X+2}}(\frac{K^{d_X} d_X^2}{16\underline{u}^3 2^{d_X} L_{d_X}})^{\frac{1}{d_X+2}}),$$

which completes the proof. □

*Proof of Lemma 2* By Assumption 2, we have that $\mathbb{E}\,\|f_*(\boldsymbol{X}, \boldsymbol{Z}) - f_{\hat{\theta}}(\boldsymbol{X}, \boldsymbol{Z})\|^2 \leq \delta_N$. Now let $(\boldsymbol{x}^{min}, \boldsymbol{z}^{min}) \in \arg\min_{(\boldsymbol{x},\boldsymbol{z})\in B\cap[-1,1]^{d_X}\times\mathcal{P}} \|f_*(\boldsymbol{x}, \boldsymbol{z}) - f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})\|$ and $(\boldsymbol{x}^{max}, \boldsymbol{z}^{max}) \in \arg\max_{(\boldsymbol{x},\boldsymbol{z})\in B\cap[-1,1]^{d_X}\times\mathcal{P}} \|f_*(\boldsymbol{x}, \boldsymbol{z}) - f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})\|$.
We have:

$$\int_{(\boldsymbol{x},\boldsymbol{z})\in B\cap[-1,1]^{d_X}\times\mathcal{P}} \|f_*(\boldsymbol{x}^{min}, \boldsymbol{z}^{min}) - f_{\hat{\theta}}(\boldsymbol{x}^{min}, \boldsymbol{z}^{min})\|^2 d\mu_{X,Z} \leq$$
$$\int_{(\boldsymbol{x},\boldsymbol{z})\in B\cap[-1,1]^{d_X}\times\mathcal{P}} \|f_*(\boldsymbol{x}, \boldsymbol{z}) - f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})\|^2 d\mu_{X,Z},$$

while we know :

$$\int_{(\boldsymbol{x},\boldsymbol{z})\in B\cap[-1,1]^{d_X}\times\mathcal{P}} \|f_*(\boldsymbol{x}, \boldsymbol{z}) - f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})\|^2 d\mu_{X,Z} \leq \int_{(\boldsymbol{x},\boldsymbol{z})\in[-1,1]^{d_X}\times\mathcal{P}} \|f_*(\boldsymbol{x}, \boldsymbol{z}) - f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})\|^2 d\mu_{X,Z} \leq \delta_N.$$

$$\implies S\|f_*(\boldsymbol{x}^{min}, \boldsymbol{z}^{min}) - f_{\hat{\theta}}(\boldsymbol{x}^{min}, \boldsymbol{z}^{min})\|^2 \leq \delta_N$$

Thus we can see that:

$$|f_*(\boldsymbol{x}^{min}, \boldsymbol{z}^{min}) - f_{\hat{\theta}}(\boldsymbol{x}^{min}, \boldsymbol{z}^{min})| \leq \sqrt{\frac{\delta_N}{S}}. \tag{EC.5}$$

At this point, we need to relate $|f_*(\boldsymbol{x}^{max}, \boldsymbol{z}^{max}) - f_{\hat{\theta}}(\boldsymbol{x}^{max}, \boldsymbol{z}^{max})|$ to $|f_*(\boldsymbol{x}^{min}, \boldsymbol{z}^{min}) - f_{\hat{\theta}}(\boldsymbol{x}^{min}, \boldsymbol{z}^{min})|$. We have

$$|f_*(\boldsymbol{x}^{max}, \boldsymbol{z}^{max}) - f_{\hat{\theta}}(\boldsymbol{x}^{max}, \boldsymbol{z}^{max})| - |f_*(\boldsymbol{x}^{min}, \boldsymbol{z}^{min}) - f_{\hat{\theta}}(\boldsymbol{x}^{min}, \boldsymbol{z}^{min})| \leq$$
$$||f_*(\boldsymbol{x}^{max}, \boldsymbol{z}^{max}) - f_{\hat{\theta}}(\boldsymbol{x}^{max}, \boldsymbol{z}^{max})| - |f_*(\boldsymbol{x}^{min}, \boldsymbol{z}^{min}) - f_{\hat{\theta}}(\boldsymbol{x}^{min}, \boldsymbol{z}^{min})||.$$

Moreover, since we know $|f_*(\boldsymbol{x}, \boldsymbol{z}) - f_{\hat{\theta}}(\boldsymbol{x}, \boldsymbol{z})|$ is $K$-Lipschitz continuous, and by definition $\|(\boldsymbol{x}^{max}, \boldsymbol{z}^{max}) - (\boldsymbol{x}^{min}, \boldsymbol{z}^{min})\| \leq 2b$ (both points lie in a ball with radius $b$), we have

$$\big|\,|f_*(\boldsymbol{x}^{max},\boldsymbol{z}^{max})-f_{\hat{\theta}}(\boldsymbol{x}^{max},\boldsymbol{z}^{max})|-|f_*(\boldsymbol{x}^{min},\boldsymbol{z}^{min})-f_{\hat{\theta}}(\boldsymbol{x}^{min},\boldsymbol{z}^{min})|\,\big|\le 2bK.$$

Therefore, it follows that

$$|f_*(\boldsymbol{x}^{max},\boldsymbol{z}^{max})-f_{\hat{\theta}}(\boldsymbol{x}^{max},\boldsymbol{z}^{max})|-|f_*(\boldsymbol{x}^{min},\boldsymbol{z}^{min})-f_{\hat{\theta}}(\boldsymbol{x}^{min},\boldsymbol{z}^{min})|\le 2bK.$$

Hence, we have

$$|f_*(\boldsymbol{x}^{max},\boldsymbol{z}^{max})-f_{\hat{\theta}}(\boldsymbol{x}^{max},\boldsymbol{z}^{max})|\le|f_*(\boldsymbol{x}^{min},\boldsymbol{z}^{min})-f_{\hat{\theta}}(\boldsymbol{x}^{min},\boldsymbol{z}^{min})|+2bK\le\sqrt{\frac{\delta_N}{S}}+2bK,$$

where the last inequality follows from (EC.5). □

*Proof of Lemma 3.* We proceed with this proof by constructing an alternative target function $\bar{f}_*$ that agrees with $f_*$ everywhere except within an area of zero measure. To that avail let Assumptions 1 (continuous version) to 4 hold. Consider $\boldsymbol{x_0}$ for which condition 1 in the lemma holds. In other words, for all $(\boldsymbol{x},\boldsymbol{z})\in[-1,1]^{d_X}\times\mathcal{P}$ and $\|(\boldsymbol{x},\boldsymbol{z})-(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))\|\le r_1$, the PDF of $\mu_{X,Z}$ is 0. Consider an alternative $\bar{f}_*$ that has the exact same value as $f_*$ everywhere, except for $(\boldsymbol{x},\boldsymbol{z})\in[-1,1]^{d_X}\times\mathcal{P}$ where $\|(\boldsymbol{x},\boldsymbol{z})-(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))\|\le r_1$. For $M$ defined in Assumption 1, let

$$\bar{f}_*(\boldsymbol{x},\boldsymbol{z})=\begin{cases}\frac{(4+\epsilon_M)(r_1-\|(\boldsymbol{x},\boldsymbol{z})-(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))\|)}{r_1}M+f_*(\boldsymbol{x},\boldsymbol{z}), & \text{if } (\boldsymbol{x},\boldsymbol{z})\in[-1,1]^{d_X}\times\mathcal{P}\ \&\ \|(\boldsymbol{x},\boldsymbol{z})-(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))\|\le r_1\\ f_*(\boldsymbol{x},\boldsymbol{z}), & \text{otherwise.}\end{cases}$$

$\bar{f}_*$ is continuous and hence, by the universal approximation result for neural networks (e.g. Leshno et al. (1993)), for any $\epsilon>0$ there exists $f_{\hat{\theta}}$ that for all $(\boldsymbol{x},\boldsymbol{z})\in[-1,1]^{d_X}\times\mathcal{P}$, $|\bar{f}_*(\boldsymbol{x},\boldsymbol{z})-f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})|\le\epsilon$. Meanwhile, it is trivial to check that for any $\delta_N$, if $\epsilon$ small enough, we have

$$\mathbb{E}_{\mu_{X,Z}}[(f_*(\boldsymbol{X},\boldsymbol{Z})-f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2]\le\delta_N.$$

Let $\hat{\boldsymbol{z}}(\boldsymbol{X}_0)$ be the optimizer of $f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})$. Then, it is easy to see that given

$$\bar{f}_*(\boldsymbol{x},\boldsymbol{z})\ge M+M\epsilon_M/2,\ \forall\ (\boldsymbol{x},\boldsymbol{z}):\|(\boldsymbol{x},\boldsymbol{z})-(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))\|\le r_1/2,$$

it must be that for some $\epsilon<M\epsilon_M/4$,

$$\|\hat{\boldsymbol{z}}(\boldsymbol{x_0})-\boldsymbol{z}^*(\boldsymbol{x_0})\|\ge r_1/2,$$

which completes the proof for condition 1.

The proof for the case under condition 2 is similar. Assume for some $\boldsymbol{z'}\in\mathcal{P}$, there exists $r_2>0$ such that $\|(\boldsymbol{x_0},\boldsymbol{z'})-(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))\|\ge r_1/2+r_2$, and we have $\int_{(\boldsymbol{x},\boldsymbol{z})\in[-1,1]^{d_X}\times\mathcal{P}:\|(\boldsymbol{x},\boldsymbol{z})-(\boldsymbol{x_0},\boldsymbol{z'})\|\le r_2}d\mu_{X,Z}=0$.

We construct $\bar{f}_*$ such that it agrees with $f_*$ everywhere except for $(\boldsymbol{x},\boldsymbol{z}) \in [-1,1]^{d_X} \times \mathcal{P} : \|(\boldsymbol{x},\boldsymbol{z}) - (\boldsymbol{x_0},\boldsymbol{z'})\| \leq r_2$ as follows:

$$\bar{f}_*(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} -\frac{(4+\epsilon_M)(r_2 - \|(\boldsymbol{x},\boldsymbol{z})-(\boldsymbol{x_0},\boldsymbol{z'})\|)}{r_2} M + f_*(\boldsymbol{x},\boldsymbol{z}), & \text{if } (\boldsymbol{x},\boldsymbol{z}) \in [-1,1]^{d_X} \times \mathcal{P} \ \&\ \|(\boldsymbol{x},\boldsymbol{z}) - (\boldsymbol{x_0},\boldsymbol{z'})\| \leq r_2 \\ f_*(\boldsymbol{x},\boldsymbol{z}), & \text{otherwise.} \end{cases}$$

$\bar{f}_*$ is continuous and hence, by the universal approximation result for neural networks, for any $\epsilon > 0$ there exists $f_{\hat{\theta}}$ that for all $(\boldsymbol{x},\boldsymbol{z}) \in [-1,1]^{d_X} \times \mathcal{P}$, $|\bar{f}_*(\boldsymbol{x},\boldsymbol{z}) - f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})| \leq \epsilon$. Meanwhile, it is trivial to check that for any $\delta_N$, if $\epsilon$ small enough, we have

$$\mathbb{E}_{\mu_{X,Z}}[(f_*(\boldsymbol{X},\boldsymbol{Z}) - f_{\hat{\theta}}(\boldsymbol{X},\boldsymbol{Z}))^2] \leq \delta_N.$$

Let $\hat{\boldsymbol{z}}(\boldsymbol{X}_0)$ be the optimizer of $f_{\hat{\theta}}(\boldsymbol{x},\boldsymbol{z})$. Then, it is easy to see that given

$$\bar{f}_*((\boldsymbol{x},\boldsymbol{z})) \leq -M - M\epsilon_M/2, \ \forall\ (\boldsymbol{x},\boldsymbol{z}) : \|(\boldsymbol{x},\boldsymbol{z}) - (\boldsymbol{x_0},\boldsymbol{z'}))\| \leq r_2/2,$$

it must be that for some $\epsilon < M\epsilon_M/4$,

$$\|\hat{\boldsymbol{z}}(\boldsymbol{x_0}) - \boldsymbol{z'}\| \leq r_2,$$

which suggests we have

$$\|\hat{\boldsymbol{z}}(\boldsymbol{x_0}) - \boldsymbol{z}^*(\boldsymbol{x_0})\| \geq r_1/2.$$

This completes the proof for condition 2. □

*Proof of Proposition 5.* We first will show that $\|\hat{\boldsymbol{z}}(\boldsymbol{x_0}) - \boldsymbol{z}^*(\boldsymbol{x_0})\| \leq r$ where $r$ is defined in Assumption 5. By Assumption 6 we have $c_2 - 2c_3K > 0$. Thus $c_2 - 2(c_3 + r')K > 0$, for some $0 < r' < r$. Let $\delta_N$ be small enough such that $c_2 - 2(c_3 + r')K > \sqrt{\frac{\delta_N}{c_4}} + \sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} c_1 r'^{d_X+d_Z}}}$. Define $S_{r'} = \int_{(\boldsymbol{x},\boldsymbol{z}) \in B_{r'}} d\mu_{X,Z}$, where $B_{r'}$ is a ball of radius $r'$ centered at $(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0}))$. Then we have

$$c_2 - 2(c_3 + r')K > \sqrt{\frac{\delta_N}{c_4}} + \sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} c_1 r'^{d_X+d_Z}}} \geq \sqrt{\frac{\delta_N}{c_4}} + \sqrt{\frac{\delta_N}{S_{r'}}}. \tag{EC.6}$$

The last inequality follows from Assumption 5 and the fact that

$$\int_{(\boldsymbol{x},\boldsymbol{z}) \in B_{r'}} d\mu_{X,Z} \geq c_1 L_{d_X+d_Z} r'^{d_X+d_Z}.$$

Now, consider some $\boldsymbol{z} \in \mathcal{P}$ such that $\|\boldsymbol{z} - \boldsymbol{z}^*(\boldsymbol{x_0})\| > r$. Then $\|(\boldsymbol{x_0},\boldsymbol{z}) - (\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0}))\| > r$. It directly follows from the definition of $\hat{\boldsymbol{z}}(\boldsymbol{x_0})$ and Lemma 2 by considering a ball of radius $r'$ surrounding $(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0}))$ that:

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \hat{\boldsymbol{z}}(\boldsymbol{x_0})) \leq f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) \leq f_*(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) + \sqrt{\frac{\delta_N}{S_{r'}}} + 2r'K. \tag{EC.7}$$

Moreover, from Assumption 6, we have that $f_*(\boldsymbol{x_0}, \boldsymbol{z}) - f_*(\boldsymbol{x_0}, \boldsymbol{z}^*(x_0)) \geq c_2$. Thus

$$f_*(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) + \sqrt{\frac{\delta_N}{S_{r'}}} + 2r'K \leq f_*(\boldsymbol{x_0}, \boldsymbol{z}) - c_2 + \sqrt{\frac{\delta_N}{S_{r'}}} + 2r'K. \tag{EC.8}$$

Furthermore, given (EC.6), from Inequalities (EC.7) and (EC.8) we have:

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \hat{\boldsymbol{z}}(\boldsymbol{x_0})) \leq f_*(\boldsymbol{x_0}, \boldsymbol{z}) - c_2 + \sqrt{\frac{\delta_N}{S_{r'}}} + 2r'K < f_*(\boldsymbol{x_0}, \boldsymbol{z}) - \sqrt{\frac{\delta_N}{c_4}} - 2c_3K. \tag{EC.9}$$

Now, we need to bound the error of $f_{\hat{\theta}}$ at $(\boldsymbol{x_0}, \boldsymbol{z})$. By Assumption 6 we know that a $d_X + d_Z$ dimensional ball of radius $c_3$ around $(\boldsymbol{x_0}, \boldsymbol{z})$ has a probability measure of at least $c_4$. Hence we can use Lemma 2, suggesting:

$$|f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}) - f_*(\boldsymbol{x_0}, \boldsymbol{z})| \leq \sqrt{\frac{\delta_N}{c_4}} + 2c_3K.$$

This implies:

$$f_*(\boldsymbol{x_0}, \boldsymbol{z}) - \sqrt{\frac{\delta_N}{c_4}} - 2c_3K \leq f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z})$$

From Inequality (EC.9) we finally have:

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \hat{\boldsymbol{z}}(\boldsymbol{x_0})) \leq f_*(\boldsymbol{x_0}, \boldsymbol{z}) - c_2 + \sqrt{\frac{\delta_N}{S_{r'}}} + 2r'K < f_*(\boldsymbol{x_0}, \boldsymbol{z}) - \sqrt{\frac{\delta_N}{c_4}} - 2c_3K \leq f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}).$$

Thus, we have proven that for any $\boldsymbol{z}$ such that $\|\boldsymbol{z} - \boldsymbol{z}^*(\boldsymbol{x_0})\| > r$, we have:

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \hat{\boldsymbol{z}}(\boldsymbol{x_0})) < f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}),$$

which proves:

$$\|\hat{\boldsymbol{z}}(\boldsymbol{x_0}) - \boldsymbol{z}^*(\boldsymbol{x_0})\| \leq r.$$

In other words, the optimal decision output by the fitted NN is within a neighborhood of the actual optimal decision. Next we will find the smallest such neighborhood and show that they are getting smaller as $\delta_N$ decreases.

For some $r_1 < r$, we want to show that for any $\boldsymbol{z}$ such that $\|\boldsymbol{z} - \boldsymbol{z}^*(\boldsymbol{x_0})\| > r_1$, we have

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) < f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}).$$

Consider $\boldsymbol{z'} \in \mathcal{P}$ such that $r_1 < \|\boldsymbol{z'} - \boldsymbol{z}^*(\boldsymbol{x_0})\| \leq r$. We know from Assumption 7 that:

$$f_*(\boldsymbol{x_0}, \boldsymbol{z'}) > f_*(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) + c_5 r_1^2.$$

At this point, we would like to lower bound $f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z'})$ using $f_*(\boldsymbol{x_0}, \boldsymbol{z'})$. To that avail, consider a $(d_X + d_Z)$ dimensional ball $B_{r_2}$ of radius $r_2 \leq r_1/2$ centered at $(\boldsymbol{x}_0, \boldsymbol{z''})$. The value $\boldsymbol{z''}$ is chosen such that $\boldsymbol{z''}$ is a convex combination of $\boldsymbol{z'}$ and $\boldsymbol{z}^*(\boldsymbol{x}_0)$, and such that $\|\boldsymbol{z'} - \boldsymbol{z''}\| = r_2$. Because $\mathcal{P}$ is

convex according to Assumption 1, we can always find such a ball with $(\boldsymbol{x}_0, \boldsymbol{z}'') \in \mathcal{P}$. Note that by construction, $(\boldsymbol{x_0}, \boldsymbol{z'}) \in B_{r_2}$. From Lemma 2 we have:

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z'}) \geq f_*(\boldsymbol{x_0}, \boldsymbol{z'}) - \sqrt{\frac{\delta_N}{S}} - 2r_2K \geq f_*(\boldsymbol{x_0}, \boldsymbol{z'}) - \sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} r_2^{d_X+d_Z}}} - 2r_2K, \tag{EC.10}$$

where $S = \int_{(\boldsymbol{x},\boldsymbol{z}) \in B_{r_2} \cap [-1,1]^{d_X} \times \mathcal{P}} d\mu_{X,Z}$. The last inequality follows from Assumption 5 as we have

$$\int_{(\boldsymbol{x},\boldsymbol{z}) \in B_{r_2} \cap [-1,1]^{d_X} \times \mathcal{P}} d\mu_{X,Z} \geq c_1 L_{d_X+d_Z} r_2^{d_X+d_Z}.$$

Next, we would like also to develop an upper bound on $f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0}))$. Thus, let us consider another Euclidean ball around $(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0}))$ with radius $r_2$. Then, by Lemma 2 and Assumption 5, we know that

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) \leq f_*(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) + \sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} r_2^{d_X+d_Z}}} + 2r_2K. \tag{EC.11}$$

Recall that by construction we have $\|\boldsymbol{z'} - \boldsymbol{z}^*(\boldsymbol{x_0})\| > r_1$. From Assumption 7, (EC.11) implies that:

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) \leq f_*(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) + \sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} r_2^{d_X+d_Z}}} + 2r_2K < f_*(\boldsymbol{x_0}, \boldsymbol{z'}) - c_5 r_1^2 + \sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} r_2^{d_X+d_Z}}} + 2r_2K. \tag{EC.12}$$

Therefore, if the following inequality holds

$$\sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} r_2^{d_X+d_Z}}} + 2r_2K \leq \frac{c_5 r_1^2}{2}, \tag{EC.13}$$

then for the right-hand side of Inequality (EC.11) and from Inequality (EC.10) we will have:

$$f_*(\boldsymbol{x_0}, \boldsymbol{z'}) - c_5 r_1^2 + \sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} r_2^{d_X+d_Z}}} + 2r_2K \leq f_*(\boldsymbol{x_0}, \boldsymbol{z'}) - \sqrt{\frac{\delta_N}{c_1 L_{d_X+d_Z} r_2^{d_X+d_Z}}} - 2r_2K \leq f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z'}).$$

This, along with Inequality (EC.12), proves that

$$f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z}^*(\boldsymbol{x_0})) < f_{\hat{\theta}}(\boldsymbol{x_0}, \boldsymbol{z'}).$$

To find the smallest $r_1$ for which Inequality (EC.13) holds, we note that $r_1$ will achieve its lowest value (that satisfies Inequality (EC.13)) when $r_2$ is set to be the minimizer of the left-hand side of Inequality (EC.13), which happens to be convex with respect to $r_2$. To prove the convexity, we notice that:

$$\frac{d(\sqrt{\frac{\delta_N}{L_{d_X+d_Z} c_1 r_2^{d_X+d_Z}}} + 2r_2K)}{dr_2} = -\frac{d_X+d_Z}{2}\sqrt{\frac{\delta_N}{L_{d_X+d_Z} c_1}} r_2^{-\frac{d_X+d_Z+2}{2}} + 2K,$$

while:

$$\frac{-\frac{d_X+d_Z}{2}\sqrt{\frac{\delta_N}{L_{d_X+d_Z}c_1}}r_2^{-\frac{d_X+d_Z+2}{2}}+2K}{dr_2}=\frac{(d_X+d_Z)(d_X+d_Z+2)}{4}\sqrt{\frac{\delta_N}{c_1L_{d_X+d_Z}}}r_2^{-\frac{d_X+d_Z+4}{2}}\geq 0.$$

Hence, the best $r_2$ would satisfy the first-order condition of the left hand side of Inequality (EC.13) with respect to $r_2$. Therefore, we have:

$$r_2=(\frac{(d_X+d_Z)^2\delta_N}{16K^2c_1L_{d_X+d_Z}})^{\frac{1}{d_X+d_Z+2}}.$$

Then by plugging $r_2$ into the left hand side of Inequality (EC.13), we will have:

$$\sqrt{\frac{\delta_N}{c_1L_{d_X+d_Z}r_2^{d_X+d_Z}}}+2r_2K=\frac{\delta_N(d_X+d_Z)}{4Kc_1L_{d_X+d_Z}}+2K(\frac{(d_X+d_Z)^2\delta_N}{16K^2c_1L_{d_X+d_Z}})^{\frac{1}{d_X+d_Z+2}}.$$

Hence, setting

$$r_1=\sqrt{2\frac{\frac{\delta_N(d_X+d_Z)}{4Kc_1L_{d_X+d_Z}}+2K(\frac{(d_X+d_Z)^2\delta_N}{16K^2c_1L_{d_X+d_Z}})^{\frac{1}{d_X+d_Z+2}}}{c_5}},$$

completes the proof. We note that as long as $\delta_N$ is small enough we have:

$$r_2=(\frac{(d_X+d_Z)^2\delta_N}{16K^2c_1L_{d_X+d_Z}})^{\frac{1}{d_X+d_Z+2}}\leq\sqrt{\frac{\frac{\delta_N(d_X+d_Z)}{4Kc_1L_{d_X+d_Z}}+2K(\frac{(d_X+d_Z)^2\delta_N}{16K^2c_1L_{d_X+d_Z}})^{\frac{1}{d_X+d_Z+2}}}{2c_5}}=\frac{r_1}{2}<\frac{r}{2},$$

where the inequality holds since its left hand side diminishes much quicker in $\delta_N$.

To finalize the proof, we note that since $\|\hat{\boldsymbol{z}}(\boldsymbol{x_0})-\boldsymbol{z}^*(\boldsymbol{x_0})\|\leq r_1$, it follows directly from Assumption 3 that:

$$|f_*(\boldsymbol{x_0},\boldsymbol{z}^*(\boldsymbol{x_0}))-f_*(\boldsymbol{x_0},\hat{\boldsymbol{z}}(\boldsymbol{x_0}))|\leq K_*\sqrt{2\frac{\frac{\delta_N(d_X+d_Z)}{4Kc_1L_{d_X+d_Z}}+2K(\frac{(d_X+d_Z)^2\delta_N}{16K^2c_1L_{d_X+d_Z}})^{\frac{1}{d_X+d_Z+2}}}{c_5}}.$$

□

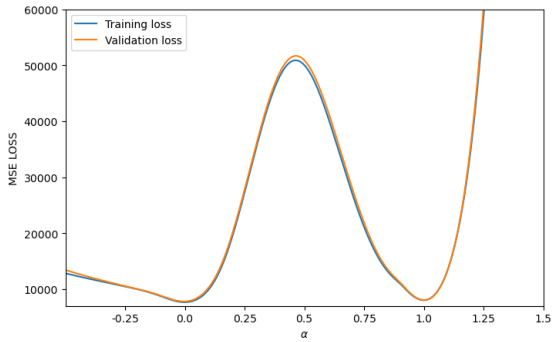


**Figure EC.3 MSE loss of the linear interpolation of small and large batch DNNs.**

## EC.5. Additional Investigations and Figures

In some applications of neural networks such as image classification, researchers have pointed out that neural networks trained with SGD with large batch sizes may not generalize well (Keskar et al. 2016, Li et al. 2018). The existing research suggests this could be because large batch methods tend to converge to sharp minimizers of the training and testing loss functions (Keskar et al. 2016). Given the importance of generalization bounds in our paper, we look into such potential sharpness in Equation (6). Similar to Li et al. (2018), to visualize such sharpness, we consider two batch sizes of 50 (small batch) and 1000 (large batch) in the newsvendor experiment. On a data set comprised of 20,000 training data points and 5,000 test data points, we look at the converged DNNs of the small batch SGD and that of the large batch one. Let $\hat{\theta}^S$ denote the parameters of the fitted DNN with small batch, and $\hat{\theta}^L$ that of the large batch. Following Li et al. (2018), let $f_{\hat{\theta}^\alpha}$ be the neural network with parameters $\alpha \times \hat{\theta}^L + (1-\alpha) \times \hat{\theta}^S$. Figure EC.3 portrays the training and validation MSE loss of $f_{\hat{\theta}^\alpha}$ as a function of $\alpha$ varying for $-.5$ to 1.5. The result suggests that in the application, such sharpness of loss function for neural networks trained with large batch size SGD may not be present.

Figure EC.4 depicts the training and validation error as a function of training epochs for one of the DNNs in Example 1 when training with 20,000 data points and validating with 5,000. It portrays a robust effect across the experiments. The training error does not shrink to zero as observed in other applications of DNNs (most probably because we use relatively small DNNs), and further, the validation error tends to be slightly larger than the training error. In other words, we do not observe benign overfitting documented in, e.g., Bartlett et al. (2020). This is because we focus on the classic statistical learning regime in which the data size is not huge and the number of parameters is significantly less than the number of data points. Moreover, the figure provides evidence that in the newsvendor experiment, the SGD method used to train the DNNs does not suffer from unstable convergence, as reported elsewhere in the literature (Ahn et al. 2022).

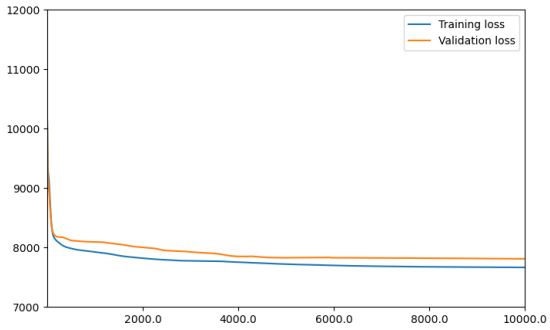


**Figure EC.4 Training versus Validation Loss.**

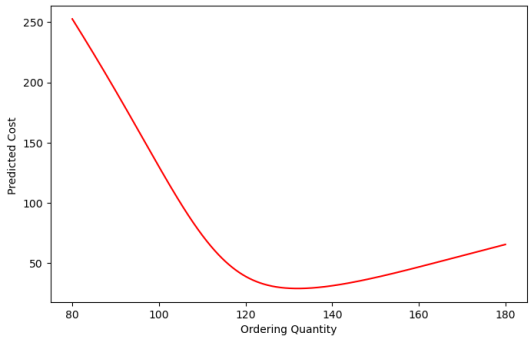


**Figure EC.5 The illustration of the effect of order quantity on predicted cost.**

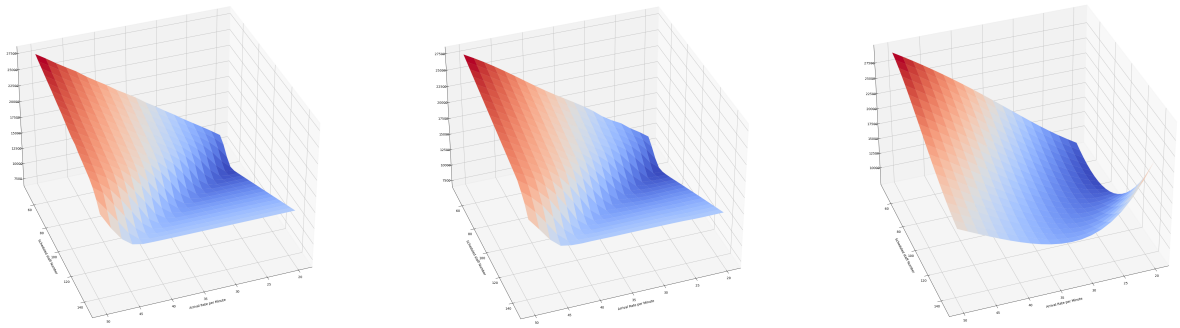

**Figure EC.6 $G/GI/z+G$ actual and predicted costs. Left is the actual cost, middle is the DONN prediction and right is the PR prediction.**

Figure EC.5 depicts the predicted newsvendor cost as a function of order quantity when all other covariates are fixed (after training with 250 days of data for one of the DNNs). The figure suggests a smooth and well-behaved objective function that strongly resembles the actual newsvendor objective function in Equation 13. Hence, if the true objective function does not suffer from local optima, we expect the fitted DNNs prediction value to exhibit a similar trend. In Example 1 we optimize the DNNs from multiple starting points and all of the results are nearly identical suggesting a lack of local optima issues.

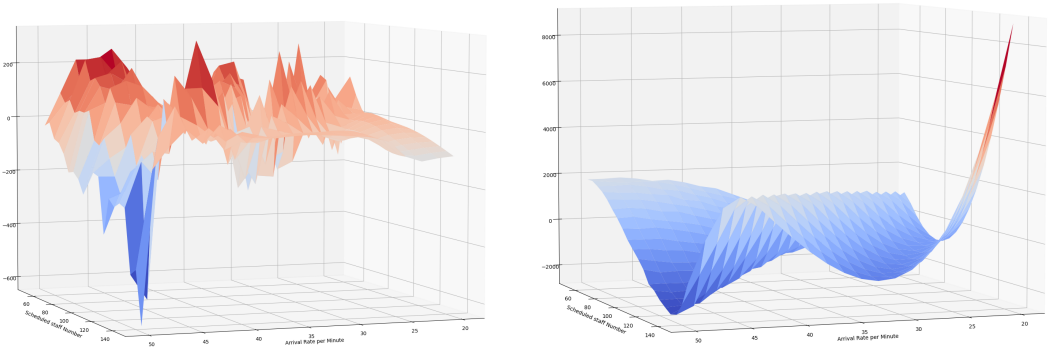

**Figure EC.7 Prediction error of DNN and Polynomial Regression in the $G/GI/z+G$ case.**

Figure EC.6 presents the actual cost (left figure) and predicted cost (DONN center figure, PR right figure) of the $G/GI/z+G$ queue for various arrival rates and staffing levels. Figure EC.7 compliments Figure EC.6. The left-hand side figure suggests that when predicting the daily cost of the call center staffing problem (Example 3) for the $G/GI/z+G$ case, the error by the DNN is quite small (generally less than 1%) and centered around zero throughout the tested ranges of staffing levels and arrival rates. The error is never larger than 5% of the actual objective function (portrayed in the left figure in Figure EC.6). The right figure in Figure EC.7 suggests the same cannot be said about Polynomial Regression method that we tested and it may suffer from very large errors in certain regions.

Figures EC.8a, EC.8b and EC.8c pertain to the predicted optimal cost using each methodology, in the linear staffing case, increasing convex staffing and $G/GI/z+G$ case, respectively. We emphasize that these results are not simulated and are the values obtained by plugging each method's optimal staffing level into its approximation of Equation (22).

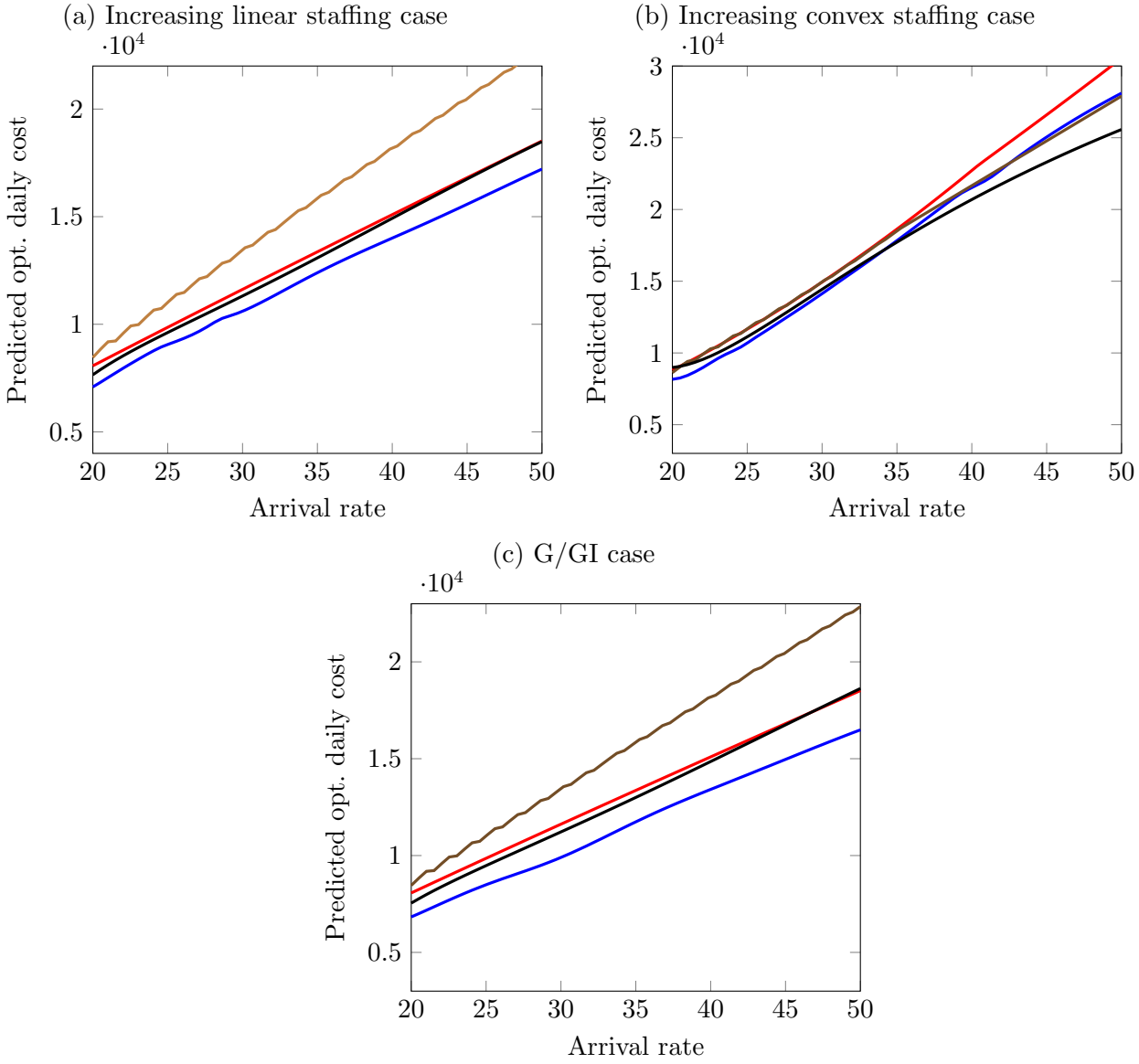


**Figure EC.8 The illustration of the predicted costs of each methodology in the queuing setting.**

## EC.6. Example

In this section, we look at an example on how to apply our framework in details. [2]. Consider an example with one scalar covariate $x$ and one scalar decision variable $z$ where $f_*(x,z) = (x-z)^2$. We simulate a large data set with $50,000$ data points where $x$ and $z$ are uniformly and independently drawn from $[-10,10]$. We train a DNN with $L=3$, $H=3$ and Swish activation function. The fitted DNN has the following weight matrices and constant terms.

[2] See https://github.com/saman-lagzi/Data-driven-Optimization-with-Neural-Networks for a detailed Python implementation of Example EC.6 along with the generated dataset.

$$\boldsymbol{W_1}=\begin{bmatrix} -1.7124 & 1.7100 \\ 0.0081 & -0.0080 \\ 1.7235 & -1.7209 \end{bmatrix}, \boldsymbol{W}_{1z}=\begin{bmatrix} 1.7100 \\ -0.0080 \\ -1.7209 \end{bmatrix}, \boldsymbol{b}_1=\begin{bmatrix} -1.6824 \\ 1.3022 \\ -1.7184 \end{bmatrix},$$

$$\boldsymbol{W_2}=\begin{bmatrix} 4.6518 & 1.9831 & 4.5889 \\ 0.2762 & -1.0031 & -0.2200 \\ -1.4228 & 3.8292 & -1.4879 \end{bmatrix}, \boldsymbol{b}_2=\begin{bmatrix} 2.3721 \\ -2.2213 \\ 3.6026 \end{bmatrix},$$

$$\boldsymbol{W_3}=\begin{bmatrix} 1.4862 & -1.1508 & -5.0347 \\ -0.1134 & -0.1684 & -0.1325 \\ 4.0344 & -0.6634 & -1.0397 \end{bmatrix}, \boldsymbol{b}_3=\begin{bmatrix} 2.4532 \\ 0.1484 \\ 2.0276 \end{bmatrix},$$

$$\boldsymbol{W_4}=\begin{bmatrix} 2.1921 \\ 0.8658 \\ 3.2429 \end{bmatrix}, b_4 = 0.9145.$$

First, we note that the derivative of the fitted DNN with respect to the decision variable $z$ at any given point, calculated through Proposition 1 is identical to the value calculated using torch.autograd function in PyTorch. For example, at input point $[1,1]$ both methods return a gradient of $\frac{\partial f_{\hat{\theta}}}{\partial z}(1,1) = 0.183$. Moreover, the point $x = 0$ translates into 0.0014 when standardized for input to the fitted DNN, and using both Proposition 1 and the torch.autograd in PyTorch we have $\hat{z}(0.0014) = -0.0055$ where $\frac{\partial f_{\hat{\theta}}}{\partial z}(0.0014, -0.0055) = 0$.

To find $\hat{z}(0.0014) = \arg\min_z f_{\hat{\theta}}(0.0014, z)$, we use the scipy.optimize package from Python with a starting point of $z_0 = -1.96$. We provide the gradient $\frac{\partial f_{\hat{\theta}}(0.0014,z)}{\partial z}$, using both our approach highlighted in Proposition 1 and the torch.autograd function in PyTorch. While both methods lead to $\hat{z}(0.0014) = -0.0055$, the execution time is 0.10 second for the first method and 0.12 second for the latter. The faster execution time based on Proposition 1 is well within expectation as it provides the closed form formula for $\frac{\partial f_{\hat{\theta}}(0.0014,z)}{\partial z}$ based on the weight matrices and constant terms of $f_{\hat{\theta}}$, however, torch.autograd is a generalized method and needs to apply the chain rule, layer by layer to $f_{\hat{\theta}}$ to find $\frac{\partial f_{\hat{\theta}}(0.0014,z)}{\partial z}$.