Informed Learning by Wide Neural Networks: Convergence, Generalization and Sampling Complexity

Jianyi Yang [1] **Shaolei Ren** [1]

Abstract

By integrating domain knowledge with labeled samples, informed machine learning has been emerging to improve the learning performance for a wide range of applications. Nonetheless, rigorous understanding of the role of injected domain knowledge has been under-explored. In this paper, we consider an informed deep neural network (DNN) with over-parameterization and domain knowledge integrated into its training objective function, and study how and why domain knowledge benefits the performance. Concretely, we quantitatively demonstrate the two benefits of domain knowledge in informed learning — regularizing the label-based supervision and supplementing the labeled samples — and reveal the trade-off between label and knowledge imperfectness in the bound of the population risk. Based on the theoretical analysis, we propose a generalized informed training objective to better exploit the benefits of knowledge and balance the label and knowledge imperfectness, which is validated by the population risk bound. Our analysis on sampling complexity sheds lights on how to choose the hyper-parameters for informed learning, and further justifies the advantages of knowledge informed learning.

1. Introduction

The remarkable success of deep neural networks (DNNs), or more generally machine learning, largely relies on the proliferation of data samples with ground-truth labels for supervised learning. Nonetheless, labeled data of high quality can often be very limited and/or extremely expensive to collect in real application domains, including medical sciences, security-related fields, and specialized engineering areas (von Rueden et al., 2021).

In parallel with the data-driven learning paradigm, domain *knowledge* (which we simply refer to as knowledge) has been utilized to assist with decision making and system designs, with a long history of success. As its name would suggest, domain knowledge is naturally domain-specific and can come from various sources in multiple forms, such as subjective experiences (e.g., medical prognosis), external sources, and scientific laws. For example, partial differential equations are used to govern many flow dynamics in physics, and the Shannon channel capacity is the fundamental principle to guide the design of modern communications systems (Goldsmith, 2005; Willard et al., 2020).

Importantly, domain knowledge has already been, sometimes implicitly, integrated into every stage of the machine learning pipeline, including training data augmentation, hypothesis set selection, model training and hypothesis finalization (more details in Appendix E). For example, differential equations and logic rules from physical sciences and/or common knowledge provide additional constraints or new functional regularization terms for model training (Battaglia et al., 2016; Borghesi et al., 2020; Silvestri et al., 2021; Muralidhar et al., 2018; Xu et al., 2018).

Despite the numerous successful examples (von Rueden et al., 2021; Deng et al., 2020), there still lacks a rigorous understanding of the role of domain knowledge in informed learning. In this paper, we focus on informed DNNs — DNNs with domain knowledge explicitly integrated into the training risk/loss function. Concretely, we consider an over-parameterized DNN with a sufficiently large network width (Neyshabur et al., 2018), and study how domain knowledge affects the DNN from three complementary aspects: convergence, generalization, and sampling complexity.

Convergence (Theorem 4.1): We show the convergence of training an informed risk function under milder technical assumptions than the prior works (Section 4.1). More specifically, we show that for inputs within a smooth set (Definition 1), the network outputs converge to the optimal solution jointly determined by all the samples in the set.

Generalization (Theorems 4.2 and 5.1): We show in Theo-

[1]Department of Electrical and Computer Engineering, University of California, Riverside, CA 92521, United States. Correspondence to: Shaolei Ren <sren@ece.ucr.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

rem 4.2 that the population risk relies on the knowledge imperfectness (Definition 3) as well as knowledge-regularized label imperfectness (Definition 4). Specifically, knowledge has two benefits: regularization for noisy labels and supplementing labels. We propose a generalized informed risk function which disentangles the two effects by introducing another hyper-weight β, followed by the population risk bounds in Theorem 5.1 and Corollary 5.2.

Sampling Complexity (Corollay 5.3): By establishing a quantitative equivalence between domain knowledge and labeled samples, we show that domain knowledge (with a reasonable quality) can effectively reduce the number of labeled samples while achieving the same generalization performance, compared to the no-knowledge case.

2. Related Work

Informed Machine Learning. The broad paradigm of informed machine learning (von Rueden et al., 2021) includes several existing learning frameworks, such as learning using privileged information (LUPI) (Vapnik & Vashist, 2009) where side knowledge is available for labeled samples (Vapnik & Vashist, 2009; Motiian et al., 2016; Sharmanska et al., 2013). Likewise, knowledge distillation (Rahbar et al., 2020; Hinton et al., 2014; Gou et al., 2021; Cho & Hariharan, 2019) transfers prior knowledge from teacher networks to a student network. Some recent studies have also focused on understanding knowledge distillation (Allen-Zhu & Li, 2020). In (Phuong & Lampert, 2019), a generalization bound is derived for knowledge distillation based on linear classifiers and deep linear classifiers, providing insights towards the mechanism of knowledge distillation. The subsequent analysis (Ji & Zhu, 2020; Rahbar et al., 2020) extends to neural networks, showing that the student network may generalize better by exploiting soft labels from the teacher model. Teacher imperfectness is investigated in (Dao et al., 2021), which bounds the learning error and proposes enhanced methods to address imperfect teachers.

Physics-informed neural networks (PINNs) have been recently proposed to solve partial differential equations (PDEs) (Yin et al., 2021; Institute, 2020; Raissi et al., 2017; Baker et al., 2019; Deng et al., 2020; Willard et al., 2020). Besides empirical studies, (Shin et al., 2020) bounds the expected PINN loss, showing that the minimizer of the regularized loss converges to the PDE solution.

More broadly, informed machine learning also includes weakly-supervised learning (Zhou, 2018; Robinson et al., 2020) and few-shot learning (Wang et al., 2020), where knowledge provides weak supervision. Domain-specific constraints (Muralidhar et al., 2018) and semantic information (Xu et al., 2018; Diligenti et al., 2017a) can also be viewed as knowledge injected into training. Our work complements these empirical studies and provides a rigorous understanding of knowledge in a unified framework.

Over-parameterized neural networks. Several recent studies (Bahri et al., 2021; Song et al., 2021; Gao et al., 2021; Khanduri et al., 2021; Jacot et al., 2018; Lee et al., 2019; Yang, 2019; Allen-Zhu et al., 2019b; Arora et al., 2019b;a; Cao & Gu, 2019; Allen-Zhu et al., 2019a; Neyshabur et al., 2018) show that over-parameterized neural networks have good convergence and generalization performance. In addition to assuming data separability in a strong sense, another crucial assumption often made in the existing studies is that the network widths increase polynomially with the total number of training samples. In informed DNNs, however, we can have many (unlabeled) training samples fed into the knowledge risk, which hence may not satisfy these assumptions. Thus, we analyze knowledge-informed over-parameterized neural networks under relaxed assumptions (Section 4).

Regularization. In the broad context of regularization, (Wei et al., 2019) shows that over-parameterized neural networks with l_2-regularization can achieve a larger margin and thus better generalization, (Blanc et al., 2020) proves that SGD with label noise is equivalent to an implicit regularization term, while (Wei et al., 2020) shows that the drop-out operation for neural networks has both explicit and implicit regularization effects. These regularizers are usually imposed on the network weights, whereas the knowledge-based regularizer in informed machine learning also incorporates inputs and directly regularizes the network output.

3. Informed Neural Network

Notations: We use the expression $[L]$ to denote the set $\{1, 2, \cdots, L\}$ for a positive integer L. Denote the indicator function as $\mathbb{1}(x) = 1$ if $x > 0$, and $\mathbb{1}(x) = 0$ otherwise. $\mathbb{E}$ is the expectation operator and $\mathbb{P}$ is a probability measure. $\mathbb{R}^d$ is d-dimensional real number space. $\mathcal{N}(x, \sigma^2)$ is the Gaussian distribution with mean x and variance σ^2. Denote $|\mathcal{A}|$ as the size of a set $\mathcal{A}$. For a vector x, $\|x\|$ is l_2-norm and $[x]_j$ is the jth entry. For a matrix $\boldsymbol{X}$, $\|\boldsymbol{X}\|_2$ represents the spectral norm, and $\|\boldsymbol{X}\|$ is the Frobenius norm. $\mathcal{B}(x, \tau) = \{y \mid \|x - y\| \leq \tau\}$ is the neighborhood domain.

3.1. Preliminaries of Neural Networks

Consider a supervised learning task to learn a relationship mapping the input $x \in \mathcal{X} \subseteq \mathbb{R}^b$ to its output $y \in \mathcal{Y} \subseteq \mathbb{R}^d$. The pair of input and output (x, y) follows a joint distribution $\mathbb{P}_{XY}$. More concretely, we consider a fully-connected DNN with an input layer, $L \geq 1$ hidden layers, and an output layer. Each hidden layer has m neurons, followed by ReLu activation denoted as $\sigma(\cdot)$. Denote $\boldsymbol{W}_0 \in \mathbb{R}^{b \times m}$ as the weights for the input layer, $\boldsymbol{W}_l \in \mathbb{R}^{m \times m}$ as

the weights for the l-th layer for $l \in [L]$, and $\boldsymbol{V} \in \mathbb{R}^{d\times m}$ as the weights for the output layer. We denote the output of the l-th layer as $h_l = \sigma\left(\boldsymbol{W}_l h_{l-1}\right)$, for $l \in [L]$, where h_0 is the input x. The output of the neural network can be expressed as $h_{\boldsymbol{W}} = \boldsymbol{V} h_L$, where $\boldsymbol{W} = \{\boldsymbol{W}_0, \boldsymbol{W}_1, \cdots, \boldsymbol{W}_L\}$. Thus, the DNN can be expressed as

$$h_{\boldsymbol{W}}(x) = \boldsymbol{V}\sigma\left(\boldsymbol{W}_L\sigma(\boldsymbol{W}_{L-1}\cdots\sigma\left(\boldsymbol{W}_1\sigma(\boldsymbol{W}_0 x)\right))\right). \quad (1)$$

Given a DNN $h_{\boldsymbol{W}}$, the risk for a labeled sample (x, y) is denoted as $r\left(h_{\boldsymbol{W}}(x), y\right)$. The goal of the learning task is to learn a DNN that minimizes the population risk:

$$R(h) = \mathbb{E}\left[r\left(h(x), y\right)\right]. \quad (2)$$

3.2. Integration of Knowledge

We consider a commonly-used informed learning method, i.e., integrating knowledge into the neural network during the training stage (von Rueden et al., 2021). During training, a labeled dataset $S_z = \{(x_1, z_1), \cdots, (x_{n_z}, z_{n_z})\}$ with n_z samples drawn from $\mathbb{P}_{XZ}$ is provided. We assume $x_i, i \in [n]$ are drawn from the distribution $\mathbb{P}_X$, but the training label $z_i \in \mathcal{Y}$ may not be the same as the true label y_i for the input x_i, because the training label may be of low quality (e.g., corrupted, noisy, and/or quantized)(Cannings et al., 2020; Zhou, 2018). Denote $h_{\boldsymbol{W},i} = h_{\boldsymbol{W}}(x_i)$ as the output of the neural network with respect to the input x_i. Based on the labeled dataset, the empirical label-based risk can be written as $\hat{R}_{S_z}(\boldsymbol{W}) = \frac{1}{n_z}\sum_{S_z} r\left(h_{\boldsymbol{W},i}, z_i\right)$.

The domain knowledge includes a knowledge-based model $g(x)$ regarding the input x and a knowledge-based risk function $r_{\mathrm{K}}\left(h_{\boldsymbol{W}}(x), g(x)\right)$ that relates the DNN's output $h_{\boldsymbol{W}}(x)$ to $g(x)$. More concrete examples of risk functions for domain knowledge can be found in Appendix F.

For the ease of analysis, we assume that both the risk function r and the knowledge-based risk function r_{K} are Lipschitz continuous, upper bounded, and strongly convex with respect to the network output, and the eigenvalues of their Hessian matrix regarding the network output lie in $[\rho, 1]$ for $\rho \in (0, 1]$. Note that the incorporated domain knowledge may not necessarily be perfect since it can be obtained based on subjective experiences (e.g., medical prognosis) (Muralidhar et al., 2018; Bica et al., 2020), pre-existing machine learning models (Hinton et al., 2014) or theoretical models which itself can deviate from the real physical world (Institute, 2020).

For training, in addition to the labeled dataset S_z, a dataset S_g with n_g unlabeled samples is generated for knowledge-based supervision. Note that S_g can also include inputs in S_z, and n_g can be sufficiently large since unlabeled samples are typically easier to obtain than labeled ones. The training risk of the informed neural network, which we simply refer to as *informed* risk, is

$$\hat{R}_{\mathrm{I}}(\boldsymbol{W}) = \frac{1-\lambda}{n_z}\sum_{S_z} r(h_{\boldsymbol{W},i}, z_i) + \frac{\lambda}{n_g}\sum_{S_g} r_{\mathrm{K}}(h_{\boldsymbol{W},i}, g_i), \quad (3)$$

where $\lambda \in [0, 1]$ is a hyper-weight, $h_{\boldsymbol{W},i} = h_{\boldsymbol{W}}(x_i)$, and $g_i = g(x_i)$. Note that Eqn. (3) can also be re-written as

$$\hat{R}_{\mathrm{I}}(\boldsymbol{W}) = \sum_{S_z \bigcup S_g}\left[\mu_i r\left(h_{\boldsymbol{W},i}, z_i\right) + \lambda_i r_{\mathrm{K}}\left(h_{\boldsymbol{W},i}, g_i\right)\right] \quad (4)$$

with hyper-parameters chosen as $\mu_i = \frac{1-\lambda}{n_z}\mathbb{1}(x_i \in S_z)$ and $\lambda_i = \frac{\lambda}{n_g}\mathbb{1}(x_i \in S_g)$. Eqn. (4) is used for convergence analysis.

To train the informed DNN, we consider a gradient descent approach in Algorithm 1 shown in Appendix A. This training approach has also been commonly considered in the literature (Allen-Zhu et al., 2019b; Zou & Gu, 2019; Du et al., 2019) for theoretical analysis of standard DNNs without domain knowledge. For the sake of analysis, we also define a hypothesis space $\mathcal{H} = \left\{h_{\boldsymbol{W}} \mid \boldsymbol{W} \in \mathcal{B}\left(\boldsymbol{W}^{(0)}, \tau\right)\right\}$ where $\boldsymbol{W}^{(0)}$ is the initialized weight and τ is the maximum distance between the weights in gradient descent and the initialized weights. We denote $h_l^{(0)}(x), l \in [L]$ as the output of the l-th layer for an input x at initialization.

Remark 1. The considered informed learning is relevant to several other frameworks. For example, it can model weakly-supervised learning (Zhou, 2018; Wang et al., 2020) with a few (possibly imperfectly) labeled samples as well as other weak supervision signals (i.e., knowledge). Besides, by viewing $\{z_i\}$ as hard labels and the knowledge-based model $g(x)$ as soft labels provided by a teacher model, the informed learning captures knowledge distillation (Hinton et al., 2014; Phuong & Lampert, 2019; Rahbar et al., 2020). Thus, our work can complement the existing analysis for the aforementioned learning frameworks from a different and more unified perspective. Additionally, PAC-Bayesian learning optimizes the PAC-Bayesian bound which is a trade-off between the empirical error and a regularization term based on a prior distribution given by knowledge (Guedj, 2019; Amit & Meir, 2018; Germain et al., 2016). But, different from PAC-Bayesian learning which considers random hypothesis, we analyze an over-parameterized neural network with a predetermined architecture.

4. Effects of Domain Knowledge

4.1. Convergence

Since the domain knowledge is integrated into a neural network during training, it is important to analyze the convergence to understand how the label and knowledge supervision jointly determine the network output. While convergence based on gradient descent for over-parameterized

neural networks has been studied extensively (Bahri et al., 2021; Allen-Zhu et al., 2019b; Zou & Gu, 2019; Arora et al., 2019a; Du et al., 2019), the current analysis is *not* suitable to study the convergence of informed over-parameterized neural networks. The reasons are summarized as follows.

• **Inapplicable for multiple supervisions.** Typically, assuming one unique label for each distinct training sample and a large enough network width, the prior studies show that the neural network can fit to the labels, i.e., the network output for each training input converges to the corresponding label (Zhang et al., 2021; Arora et al., 2019a; Zou & Gu, 2019; Oymak & Soltanolkotabi, 2020). But, in our case, one training input can have multiple supervisions from both label and knowledge with possibly different forms of risks. Thus, the network output for an input may not be necessarily determined by a unique label. The convergence of knowledge distillation supervised by both hard and soft labels is studied by (Rahbar et al., 2020), but only the quadratic risk and shallow networks are considered.

• **Strong data separability assumption.** Some prior studies require a lower-bounded distance of any two samples (Allen-Zhu et al., 2019b; Zou & Gu, 2019; Du et al., 2019), but this may not be satisfied for an informed DNN because the input samples for label-based and knowledge-based risks can be very close or even the same. Other studies assume data separability by a neural tangent model (Chen et al., 2021b; Ji & Telgarsky, 2020; Cao & Gu, 2020; Nitanda et al., 2019), but data separability by a neural tangent model is not well defined for training with multiple supervisions in informed DNNs.

To address these challenges, we provide convergence analysis for informed over-parameterized neural networks based on a new data separability assumption of smooth sets. The construction of smooth sets approximates the space $\mathcal{X}$ with discrete pieces, each containing samples that jointly satisfy the smooth properties. The smooth sets are formally defined below, followed by the data separability assumption.

Definition 1 (Smooth sets)**.** Given $\phi > 0$, construct a $\phi-$net (Clarkson, 2006) $\mathcal{X}_\phi = \{x'_k, k \in [N], x'_k \in \mathcal{X}\}$ with $N \sim O(1/\phi^b)$ such that $\forall x'_i, x'_j \in \mathcal{X}_\phi$ and $x'_i \neq x'_j$, $\|x'_i - x'_j\| \geq \phi$ holds, and $\forall x_i \in S_z \bigcup S_g$, there exists at least one $x'_k \in \mathcal{X}_\phi$ satisfying $\|x_i - x'_k\| \leq \phi$. Each input $x'_k \in \mathcal{X}_\phi$, referred to as a representative input, determines a smooth set $\mathcal{C}_{\phi,k} = \{x \in \mathcal{X} \mid \|x - x'_k\| \leq \phi, \|x - x'_j\| \geq \phi/2, \forall j \neq k, x'_k, x'_j \in \mathcal{X}_\phi\}$. The index set of training samples within the kth smooth set is $\mathcal{I}_{\phi,k} = \{i \mid x_i \in S_z \bigcup S_g, x_i \in \mathcal{C}_{\phi,k}\}, k \in [N]$.

Assumption 1 (Data separability by smooth sets)**.** For each smooth set k with representative sample x'_k, there exists a non-empty subset of neuron indices $\mathcal{G}_{k,\alpha} \in [m]$ with size $|\mathcal{G}_{k,\alpha}| = \alpha m, \alpha \in (0, 1]$ such that at initialization, $\forall i \in \mathcal{I}_{\phi,k}$, $\forall j \in \mathcal{G}_{k,\alpha}$, $\mathbb{1}\left(\left[h_L^{(0)}(x_i)\right]_j \geq 0\right) = \mathbb{1}\left(\left[h_L^{(0)}(x'_k)\right]_j \geq 0\right)$, and $\forall j \notin \mathcal{G}_{k,\alpha}$, the pre-activation of the L-th layer $\left|\left[\boldsymbol{W}_L^{(0)} h_{L-1}^{(0)}(x_i)\right]_j\right| \geq \frac{3\sqrt{2\pi}\phi^{b+1}}{16\sqrt{m}}$.

Instead of requiring a lower-bounded distance of any two training samples, the data separability assumption requires that, at initialization, for samples in one smooth set, the outputs of the last hidden layer either have the same signs as those of the representative sample, or their absolute values are larger than a very small threshold. Thus, this data separability assumption is set-wise and addresses the cases where two training inputs are very close or the same, and hence is milder than the one in existing studies (e.g., (Allen-Zhu et al., 2019b)). The parameter α indicates slackness: with larger α, more neurons have the same signs. Actually, data separation by smooth set with $\phi > 0$ in Assumption 1 always exists: when ϕ is small enough such that only one inputs or several same inputs are included in a smooth set, Assumption 1 is satisfied with $\alpha = 1$. Even in this worst case, our assumption is still milder than the data separability assumption considered in (Allen-Zhu et al., 2019b; Zou & Gu, 2019) that excludes the existence of two training samples with the same inputs but different supervisions.

With the data-separability assumption by smooth sets, we are ready to show the labels and knowledge jointly determine the network output for training inputs. We introduce the notation *effective label*, as formally defined below.

Definition 2 (Effective label)**.** For the k-th smooth set, define the effective label as $y_{\text{eff},k} = \arg\min_h \sum_{i \in \mathcal{I}_{\phi,k}} \{\mu_i r(h, z_i) + \lambda_i r_K(h, g_i)\}$ with μ_i, λ_i defined in Eqn. (3) and h in the space of network output, and the effective optimal risk as $r_{\text{eff},k} = \sum_{i \in \mathcal{I}_{\phi,k}} \{\mu_i r(y_{\text{eff},k}, z_i) + \lambda_i r_K(y_{\text{eff},k}, g_i)\}$.

Next, we show the convergence analysis. Note that the proof based on the data separability by smooth sets (Assumption 1) invalidates the proofs in previous studies, and we need new lemmas that lead to novel convergence to effective labels in Definition 2. In particular, in Lemma B.1, to approximate the outputs in the smooth set k by the output of the representative input x'_k, we need to bound the difference of the outputs with respect to x'_k and an input in the smooth set k. Also, based on Assumption 1, we derive in Lemma B.4 the gradient lower bound which relies on the number of smooth sets N instead of the sample size $n_z + n_g$ in the previous analysis. This makes the network width m in our analysis directly rely on the smooth set size ϕ. Moreover, in Lemma B.5, we prove based on the definition of smooth sets that the first-order approximation error of the total informed risk depends on the difference between the risk and effective risk in Definition 2. This is important to prove the convergence to the effective labels. The details of the convergence analysis are deferred to Appendix B.4.

Theorem 4.1. *Assume that the network width satisfies* $m \geq \Omega\left(\phi^{-11b-4}L^{15}d\rho^{-4}\bar{\lambda}^{-4}\alpha^{-4}\log^3(m)\right)$*, and the step size is set as* $\eta = O(\frac{d}{L^2 m})$*. With Assumptions 1 satisfied, for any* $\epsilon > 0$ *and* $\phi \leq \widetilde{O}\left(\epsilon L^{-9/2}\log^{-3}(m)\right)$*, we have with probability at least* $1 - O(\phi)$*, by gradient descent after* $T = O\left(\frac{L^2}{\phi^{1+2b}\rho\bar{\lambda}\alpha}\log(\epsilon^{-1}\log(\phi^{-1}))\right)$ *steps, the informed risk in Eqn. (4) is bounded as:* $\hat{R}_{\mathrm{I}}(\boldsymbol{W}^{(T)}) - \hat{R}_{\mathrm{eff}} \leq O(\epsilon)$*, where* $\hat{R}_{\mathrm{eff}} = \sum_{k=1}^{N} r_{\mathrm{eff},k}$*,* $\bar{\lambda} = \Omega(\min(1-\lambda, \lambda)\mathbb{1}(\lambda \in (0,1)) + \mathbb{1}(\lambda \in \{0,1\}))$*. Also, the DNN outputs satisfy:*

$$\sum_{S_z \bigcup S_g} (\mu_i + \lambda_i)\left\|h_{\boldsymbol{W}^{(T)}}(x_i) - y_{\mathrm{eff},k(x_i)}\right\|^2 \leq O(\epsilon),$$

where $k(x_i)$ *is the index of the smooth set that includes* x_i*,* $\mu_i = \frac{1-\lambda}{n_z}\mathbb{1}(x_i \in S_z)$ *and* $\lambda_i = \frac{\lambda}{n_g}\mathbb{1}(x_i \in S_g)$*.*

Remark 2. The convergence analysis in Theorem 4.1 addresses the limitations mentioned at the beginning of this section. First, instead of fitting a unique label for each input, the informed neural network with multiple supervisions converges to effective labels. Second, the data separability assumption is enough for convergence analysis of informed neural networks. Another observation is that with smaller ϕ and smaller α, Assumption 1 becomes milder, but a larger network width and more training steps are needed to guarantee convergence.

Additionally, different from previous convergence analysis where the width m increases directly with the sample size, the network width m in our analysis depends on the smooth set size ϕ and is non-decreasing with sample size (i.e., m may not always increase with the sample size). To see this, given a construction of smooth sets by size ϕ that meets Assumption 1, if we continue to add (either labeled or knowledge-supervised) training samples that lie in the existing smooth sets and satisfy Assumption 1, the width m remains the same, and smaller ϕ (larger m) is needed to guarantee the convergence only when the added samples violate Assumption 1 under the current ϕ. The large network width needed for analysis is due to the limitation of over-parameterization techniques, while in practice a much smaller network width is enough. Albeit beyond the scope of our study, addressing the gap between theory and practice is clearly important and still active research in the community (Bahri et al., 2021).

Remark 3. We can get more insights about the effects of labels and knowledge from the conclusion that the network outputs converge to the corresponding *effective* labels in Definition 2. On the one hand, if knowledge is applied to the samples within the same smooth sets as labeled samples, knowledge-based supervision and label-based supervision jointly determine the network output together: knowledge serves as *a regularization for labels* in this case. On the other hand, if a smooth set only contains knowledge-supervised samples, the network output is determined solely by knowledge: knowledge *supplements labeled samples* (albeit possibly imperfectly) to provide additional supervision.

4.2. Generalization

We now formally analyze how the domain knowledge affects the generalization performance. From our convergence analysis, there are two different effects of knowledge (Remark 3). We characterize the two effects by formally defining knowledge imperfectness and knowledge-regularized label imperfectness. Before this, we list some notations for further analysis. Given a $\phi-$net $\mathcal{X}_\phi$ (Definition 1), $\mathcal{U}_\phi(S_z) = \{k \in [N] \mid \exists x \in S_z, x \in \mathcal{C}_{\phi,k}\}$ is the index collection of smooth sets that contain at least one labeled sample, and $\mathcal{X}_\phi(S_z) = \bigcup_{k\in\mathcal{U}_\phi}(S_z)\mathcal{C}_{\phi,k}$ is the region covered by the smooth sets in $\mathcal{U}_\phi(S_z)$. $S'_g = S_g \bigcap \mathcal{X}_\phi(S_z)$ is the knowledge supervised dataset with samples share the common smooth sets with labeled samples in S_z while the samples in $S''_g = S_g \setminus S'_g$ lie in smooth sets without labeled samples. Denote $n'_g = |S'_g|$ and $n''_g = |S''_g|$.

Definition 3 (Knowledge imperfectness)**.** Let $h^*_{\mathrm{K}} = \min_h \frac{1}{n''_g}\sum_{S''_g}[r_{\mathrm{K}}(h(x_i), g(x_i))]$ be the optimal hypothesis for the knowledge-based risk on the dataset S''_g. The imperfectness of domain knowledge K applied to the dataset S''_g is defined as $\widehat{Q}_{\mathrm{K},S''_g} = \frac{1}{n''_g}\sum_{x_i\in S''_g} r(h^*_{\mathrm{K}}(x_i), y_i)$ where y_i is the true label of x_i. Correspondingly, let $\bar{h}^*_{\mathrm{K}} = \min_h \mathbb{E}\left[r_{\mathrm{K}}(h(x), g(x))\right]$ be the optimal hypothesis for the expected knowledge-based risk, and the expected imperfectness of domain knowledge K is defined as $Q_{\mathrm{K}} = \mathbb{E}\left[r(\bar{h}^*_{\mathrm{K}}(x), y)\right]$.

The (empirical or expected) knowledge imperfectness is defined as the risk under the hypothesis optimally learned by knowledge-based supervision. Thus, it measures the extent to which the domain knowledge is inconsistent with the true labels, measured in terms of the risk over the hypothesis set $\mathcal{H}$. Besides knowledge-based supervision, the network outputs for some smooth sets that contain both samples for knowledge risks and labeled samples are jointly determined by label-based and knowledge-based supervisions. Thus, we define knowledge-regularized label imperfectness below.

Definition 4 (Knowledge-regularized label imperfectness)**.** Let $h^*_{\mathrm{R},\beta} = \arg\min_h \frac{1-\beta}{n_z}\sum_{S_z} r(h(x_i), z_i) + \frac{\beta}{n'_g}\sum_{S'_g} r_{\mathrm{K}}(h(x_i), g(x_i))$ be the optimal hypothesis for the knowledge-regularized risk and $\beta \in [0,1]$. The knowledge-regularized label imperfectness is $\widehat{Q}_{\mathrm{R},S_z,S'_g}(\beta) = \frac{1}{n_z}\sum_{S_z} r(h^*_{\mathrm{R},\beta}(x_i), y_i)$, where y_i is the true label regarding x_i. Correspondingly, with $\bar{h}^*_{\mathrm{R},\beta} = \arg\min_h \mathbb{E}[\frac{1-\beta}{n_z}\sum_{S_z} r(h(x_i), z_i) + \frac{\beta}{n'_g}\sum_{S'_g} r_{\mathrm{K}}(h(x_i), g(x_i))]$ being the optimal hypothesis for the regularized risk, the expected knowledge regularized

label imperfectness is $Q_{\mathrm{R}}(\beta) = \mathbb{E}\left[r(\bar{h}^*_{\mathrm{R},\beta}(x), y)\right]$.

Like knowledge imperfectness, knowledge-regularized label imperfectness indicates the risk of the hypothesis optimally learned by joint supervision from labels and knowledge. We see that when $\beta = 0$, $\widehat{Q}_{\mathrm{R}}(0)$ (or $Q_{\mathrm{R}}(0)$) is the imperfectness of pure label-based supervision. Thus, the gain due to knowledge is $\Delta\widehat{Q}_{\mathrm{R},\beta} = \widehat{Q}_{\mathrm{R}}(0) - \widehat{Q}_{\mathrm{R}}(\beta)$ (or $\Delta Q_{\mathrm{R},\beta} = Q_{\mathrm{R}}(0) - Q_{\mathrm{R}}(\beta)$ for the expected version). We show in the following theorem how the two types of imperfectness affect the population risk trained on the informed risk in Eqn. (3). The details are deferred to Appendix B.5.1.

Theorem 4.2. *With $\boldsymbol{W}^{(T)}$ trained on Eqn. (3), $\phi \leq \widetilde{O}\left(\epsilon^2 L^{-9/2}\log^{-3}(m)\right), \phi \leq (\sqrt{\epsilon}/n_z)^{1/b}$, and other assumptions the same as Theorem 4.1, with probability at least $1 - O(\phi) - \delta, \delta \in (0,1)$, the population risk satisfies*

$$R\left(h_{\boldsymbol{W}^{(T)}}\right) \leq O(\sqrt{\epsilon}) + (1-\lambda)\widehat{Q}_{\mathrm{R},S_z,S'_g}(\beta_\lambda) + \lambda\widehat{Q}_{\mathrm{K},S''_g} + O\left(\Phi + \sqrt{\log(1/\delta)}\right)\left(\frac{1-\lambda}{\sqrt{n_z}} + \frac{\lambda}{\sqrt{n_g}}\right),$$

where $\beta_\lambda = \frac{\lambda n'_g}{(1-\lambda)n_g + \lambda n'_g}$, $\widehat{Q}_{\mathrm{R},S_z,S'_g}(\beta_\lambda)$ is the knowledge-regularized label imperfectness in Definition 4 and $\widehat{Q}_{\mathrm{K},S''_g}$ is the knowledge imperfectness in Definition 3 applied to S''_g, and $\Phi = O\left(4^L L^{3/2} m^{1/2}\phi^{-b-1/2} d\rho^{-1/2}\bar{\lambda}^{-1/2}\alpha^{-1/2}\right)$.

Remark 4. Theorem 4.2 shows that by training on the informed risk (3), knowledge affects the generation performance in the following two ways.

• **Knowledge for regularization.** When knowledge is applied to sample inputs inside the same smooth sets as labeled samples, it serves as an explicit regularization for label-based supervision, possibly reducing the label imperfectness from $\widehat{Q}_{\mathrm{R},S_z,S'_g}(0)$ to $\widehat{Q}_{\mathrm{R},S_z,S'_g}(\beta_\lambda)$.

• **Knowledge for supplementing labels.** The generalization error is in the order of $O\left(\frac{1-\lambda}{\sqrt{n_z}} + \frac{\lambda}{\sqrt{n_g}}\right)$. When no knowledge is used ($\lambda = 0$), the order is as large as $O\left(\frac{1}{\sqrt{n_z}}\right)$. If knowledge is applied ($\lambda > 0$), then the generalization error decreases with the increasing of knowledge-supervised sample size n_g. Thus, when knowledge is applied to smooth sets without labeled samples, it serves as an (possibly imperfect) supplement for labels, while introducing knowledge imperfectness $\widehat{Q}_{\mathrm{K},S''_g}$.

The hyper-parameter λ can be used to balance the introduced imperfectness and generalization error from label and knowledge supervision. However, by the risk bound, it is hard to use one hyper-parameter λ to control the two effects of knowledge, which will be further discussed in the next section.

5. A Generalized Training Objective

In the informed risk in Eqn. (3), only one hyper-weight λ is present, controlling the two different effects of knowledge (Remark 4). To better reap the benefits of knowledge, we consider a generalized informed risk in Eqn.(5) by introducing another hyper-weight β, which introduces more flexibility to govern the roles of domain knowledge.

$$\hat{R}_{\mathrm{I},G}(\boldsymbol{W}) = \frac{(1-\lambda)(1-\beta)}{n_z}\sum_{S_z} r\left(h_{\boldsymbol{W},i}, z_i\right) + \frac{(1-\lambda)\beta}{n'_g}\sum_{S'_g} r_{\mathrm{K}}(h_{\boldsymbol{W},i}, g_i) + \frac{\lambda}{n''_g}\sum_{S''_g} r_{\mathrm{K}}(h_{\boldsymbol{W},i}, g_i), \tag{5}$$

where $\beta, \lambda \in [0,1], h_{\boldsymbol{W},i} = h_{\boldsymbol{W}}(x_i), g_i = g(x_i)$.

In Eqn. (5), the two hyper-parameters λ and β can jointly control the knowledge effects (and the introduced imperfectness) when knowledge is applied. The hyperparameter β is used to controls the knowledge regularization strength. By Remark 4, knowledge-supervised samples in S'_g serve as an explicit regularization for label-based supervision while introducing knowledge-regularized label imperfectness $Q_{\mathrm{R}}(\beta)$. Thus, when β is larger, more effects from S'_g are incorporated and the regularization effect from knowledge is stronger. Also, we use λ to adjust the effect of supplementing labels and the introduction of Q_{K}. By Remark 4, S''_g serves as an supplement for labels while introducing the knowledge imperfectness Q_{K}. Thus, with larger λ, more effects from S''_g are incorporated, which means we incorporate more effects of data supplement from knowledge and also knowledge imperfectness Q_{K} but less effect of knowledge regularization and knowledge-regularized label imperfectness Q_{R}. The benefit of the training objective in Eqn. (5) will be explained formally in Theorem 5.1 and Corollary 5.2.

Compared with the objective in Eqn. (3) with only one hyper-parameter λ, Eqn. (5) introduces another hyper-parameter β to independently adjust the degree of the knowledge regularization, making Eqn. (5) more general and flexible. To train on Eqn. (5), we need to separate dataset for knowledge supervision into two datasets S'_g and S''_g based on whether an input is close to a labeled input and assign different hyper-weights to them. The knowledge-based dataset separation is determined by ϕ in Definition 1. Specifically, when the network width goes to infinity (ϕ goes to zero), S'_g shares the same inputs as S_z, but S'_g and S_z are supervised by knowledge and labels, respectively. We have $S''_g = S_g \setminus S'_g = S_g \setminus S_z$ which supplements the labels as shown in Remark 4. Note that when the knowledge is perfect and knowledge-supervised samples are sufficient, we do not need labeled samples, i.e., $S_z = \emptyset$ and we set $\lambda = 1, \beta = 1$. Then, we have $S''_g = S_g$ and Eqn. (5) becomes a purely knowledge-based risk. When no knowledge

is applied, we set $\lambda = 0, \beta = 0$, and Eqn. (5) becomes a purely lable-based risk. In general cases when labels and knowledge are both used, hyper-parameters λ and β are used to control the effects of knowledge.

5.1. Population Risk

Note that Eqn. (5) can also be written as the form of Eqn. (4) with hyper-parameters chosen as $\mu_i = \frac{(1-\lambda)(1-\beta)}{n_z}\mathbb{1}(x_i \in S_z)$ and $\lambda_i = \frac{(1-\lambda)\beta}{n'_g}\mathbb{1}(x_i \in S'_g) + \frac{\lambda}{n''_g}\mathbb{1}(x_i \in S''_g)$, so Theorem 4.1 for convergence still holds. Next, we bound the population risk based on the generalized informed risk. The details are given in Appendix B.5.2.

Theorem 5.1. *Assume that $\boldsymbol{W}^{(T)}$ trained on Eqn. (5) and other assumptions are the same with those of Theorem 4.1, setting $\phi : \phi \leq \widetilde{O}\left(\epsilon^2 L^{-9/2}\log^{-3}(m)\right)$ and $\phi \leq (\sqrt{\epsilon}/n_z)^{1/b}$, with probability at least $1-O(\phi)-\delta, \delta \in (0,1)$, the population risk satisfies*

$$R(h_{\boldsymbol{W}^{(T)}}) \leq O(\sqrt{\epsilon}) + (1-\lambda)\widehat{Q}_{\mathrm{R},S_z,S'_g}(\beta) + \lambda\widehat{Q}_{\mathrm{K},S''_g} + O\left(\Phi + \sqrt{\log(1/\delta)}\right)\left(\frac{1-\lambda}{\sqrt{n_z}} + \frac{\lambda}{\sqrt{n''_g}}\right),$$

where β and λ are trade-off hyper-parameters in Eqn. (5)

Additionally, to obtain more insights for sampling complexity, we further bound the population risk in terms of expected imperfectness, at the expense of some tightness. The proof details are deferred to Appendix B.5.3.

Corollary 5.2. *With the same assumptions as in Theorem 5.1, with probability at least $1-O(\phi)-\delta, \delta \in (0,1)$, the population risk satisfies*

$$R(h_{\boldsymbol{W}^{(T)}}) \leq O(\sqrt{\epsilon}) + (1-\lambda)Q_{\mathrm{R}}(\beta) + \lambda Q_{\mathrm{K}} + O\left(\Phi + \log^{1/4}(1/\delta)\right)\sqrt{\frac{1-\lambda}{\sqrt{n_z}} + \frac{\lambda}{\sqrt{n''_g}}},$$

where $Q_{\mathrm{R}}(\beta)$ is the expected knowledge-regularized label imperfectness in Definition 4, Q_{K} is the expected knowledge imperfectness in Definition 3.

Remark 5. Theorem 5.1 and Corollary 5.2 show that by training on the generalized informed risk in Eqn. (5), label and knowledge supervision jointly affect the population risk while introducing a combination of knowledge-regularized label imperfectness $Q_{\mathrm{R}}(\beta)$ and knowledge imperfectness Q_{K}. The effect of knowledge regularization is controlled by β and the trade-off between the two imperfectness terms and the trade-off between the two generalization errors $\frac{1-\lambda}{\sqrt{n_z}}$ and $\frac{\lambda}{\sqrt{n''_g}}$ are both controlled by λ. Thus, this gives us more flexibility to adjust how much domain knowledge is incorporated when it plays different roles in informed learning as discussed in Remark 4. Also, as shown by the population risk bounds, we can tune the two hyper-parameters separately — we can first tune β to minimize $Q_{\mathrm{R}}(\beta)$, and then tune λ to balance $Q_{\mathrm{R}}(\beta)$ and Q_{K}, and also balance the generalization errors due to sizes of datasets.

5.2. Sampling Complexity

We discuss the choices of hyper-parameters β and λ in different cases to guarantee a small population risk, and give the sampling complexity in each case, whose details are deferred to Appendix B.5.4.

Corollary 5.3 (Sampling Complexity)**.** *With the same set of assumptions as in Corollary 5.2 and setting $\beta^* = \arg\min_{\beta\in[0,1]} Q_{\mathrm{R}}(\beta)$, with probability at least $1-O(\phi)-\delta, \delta \in (0,1)$, to guarantee a population risk no larger than $\sqrt{\epsilon}$, we have the following cases:*

(a) If $Q_{\mathrm{K}} \leq \sqrt{\epsilon}$, set $\lambda = 1$, the sampling complexity for labels is $n_z = 0$ and the sampling complexity for knowledge-supervision is $n_g \sim O(1/(\epsilon^2-\epsilon^3))$.

(b) If $Q_{\mathrm{K}} > \sqrt{\epsilon}$ and $\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{R}}(\beta^*)} \geq 1$, set $\lambda = \frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}}$, the sampling complexity for labels is $n_z \sim O\left(\left(1/\epsilon - 1/\left(\sqrt{\epsilon}Q_{\mathrm{K}}\right)\right)^2\right)$ and the sampling complexity for knowledge-supervision is $n_g \sim O(1/\left((\epsilon-\epsilon^2)Q_{\mathrm{K}}^2\right))$.

(c) If $\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{R}}(\beta^*)} < 1$, a population risk as low as $\sqrt{\epsilon}$ cannot be achieved no matter what λ is and how many samples are used.

Remark 6. In practice, unlabeled samples are typically cheaper to obtain than *labeled* samples. If $Q_{\mathrm{K}} \leq \sqrt{\epsilon}$, the domain knowledge is good enough for supervision, and thus we can perform purely knowledge-based training without any labeled samples and guarantee a population risk no larger than $\sqrt{\epsilon}$ with $n''_g \sim O(1/\epsilon^2)$, and hence $n_g \sim O(1/(\epsilon^2-\epsilon^3))$. When the knowledge imperfectness $Q_{\mathrm{K}} > \sqrt{\epsilon}$, we discuss the following two cases. First, if $\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{R}}(\beta^*)} \geq 1$, we can choose λ from $\left[1 - \frac{\sqrt{\epsilon}}{Q_{\mathrm{R}}(\beta^*)}, \frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}}\right]$ to control the risk from knowledge and label imperfectness as low as $\sqrt{\epsilon}$. We thus choose the largest $\lambda = \frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}}$ to reduce the label sampling complexity. In this case, knowledge is not good enough, but label imperfectness is not too large. Thus, we can guarantee a population risk no larger than $\sqrt{\epsilon}$ with labeled samples $n_z \sim O\left(\left(1/\epsilon - 1/\left(\sqrt{\epsilon}Q_{\mathrm{K}}\right)\right)^2\right)$ and knowledge supervised samples $n_g \sim O(1/\left((\epsilon-\epsilon^2)Q_{\mathrm{K}}^2\right))$. Finally, if $\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{R},\beta^*}} < 1$, we cannot guarantee a population risk less than $\sqrt{\epsilon}$ no matter what λ is and how many samples are used since the neither knowledge nor labels are of high enough quality.

In summary, the extreme cases are: Case (a) where the knowledge supervision alone is nearly perfect, and Case (c) where the knowledge and labels are both of low quality. Usually, we are in Case (b) where knowledge is imperfect but labels (after knowledge regularization) are good enough. In contrast, DNNs without using domain knowledge requires the label imperfectness $Q_{\mathrm{R},0}$ not to exceed $\sqrt{\epsilon}$; otherwise, the population risk cannot be guaranteed to be no greater than $\sqrt{\epsilon}$. The informed DNNs relaxes this requirement by requiring $\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{R}}(\beta^*)} \geq 1$. In addition, the incorporation of domain knowledge reduces the labeled sampling complexity from $n_z \sim O(\frac{1}{\epsilon^2})$ in the traditional no-knowledge setting to $n_z \sim O\left(\left(1/\epsilon - 1/\left(\sqrt{\epsilon}Q_{\mathrm{K}}\right)\right)^2\right)$. In other words, the incorporation of knowledge is equivalent to $O(\frac{2}{\epsilon^{3/2}Q_{\mathrm{K}}} - \frac{1}{\epsilon Q_{\mathrm{K}}^2})$ labeled samples, establishing a quantitative comparison between knowledge supervision and labeled samples.

6. Further Discussions

Summary of analysis. The convergence analysis in Theorem 4.1 introduces the concept of smooth sets and explains how the neural network output behaves by training on an informed risk. The generalization analysis in Theorem 4.2 explicitly shows the two different effects the domain knowledge has on the population risk (i.e., regularizing labels and supplementing labels). Based on this observation, we propose a generalized informed risk in Eqn. 5 to get more flexibility to control the two effects of knowledge, which is validated by Theorem 5.1 and its Corollary 5.2. Finally, the sampling complexity in Corollary 5.3 shows the effects of joint knowledge and label supervision in a quantitative way.

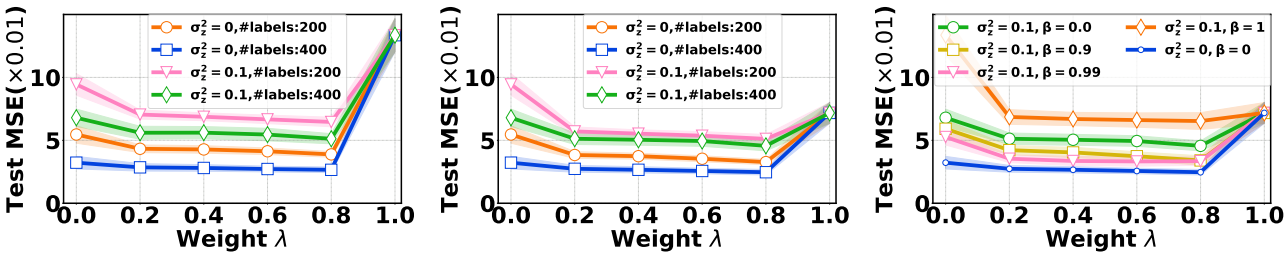
Understanding knowledge distillation from the perspective of informed learning. Knowledge distillation is extremely useful in practice (e.g., for model compression (Hinton et al., 2014)). Here, we show how our analysis complement the existing understanding of knowledge distillation (Hinton et al., 2014; Phuong & Lampert, 2019; Rahbar et al., 2020; Dao et al., 2021; Ji & Zhu, 2020) from the perspective of hard label and teacher's knowledge imperfectness. In our formulation, hard labels are $\{z_i\}$ in the labeled dataset, whose imperfectness (non-softness) is measured by $Q_{\mathrm{R}}(0)$. In Theorems 4.2, 5.1, and Corollary 5.2, by viewing the teacher model $g(x)$ as domain knowledge, we show the teacher benefits the student training by providing a regularization gain $\Delta Q_{\mathrm{R},\beta}$, and reducing the sampling complexity of hard labels by Corollary 5.3. The knowledge-regularized label imperfectness $Q_{\mathrm{R},\beta}$ can be less than pure lable imperfectness $Q_{\mathrm{R}}(0)$ because the soft label can smooth the network output within each smooth set. But, given the teacher (knowledge) imperfectness Q_{K}, there exists a trade-off between hard label and teacher supervision.

Importantly, our results are in line with the observations and also complement the analysis in (Ji & Zhu, 2020). Specifically, (Ji & Zhu, 2020) uses NTK to show that the soft labels provided by a teacher model (knowledge) are easier to learn than hard labels while hard labels can correct imperfect teachers pointwise, exhibiting a trade-off between hard labels and the imperfect teacher. We define the hard label and teacher (knowledge) imperfectness, and show that for a neural network with finite width, hard labels and teacher's knowledge compensate for each other within each smooth set. In consistency with our results, (Rahbar et al., 2020) based on NTK also presents a trade-off between labels and the imperfect teacher. The teacher model imperfectness is also observed by (Dao et al., 2021) which measures the teacher imperfectness by the squared norm of the difference of the soft label and the true Bayesian class probability. Note, however, that our analysis *cannot* adequately explain the benefit of knowledge distillation for the perspective of feature learning due to the inherent limitations of over-parameterization techniques, which are further discussed in (Allen-Zhu & Li, 2020).

7. Numerical Results

7.1. Problem Setup

We consider an informed DNN with domain knowledge in the form of constraints to learn a Bohachevsky function. The learning task is to learn a relationship $y(x)$. The learner is provided with a dataset with labeled samples $S_z = \{(x_i, z_i), i \in [n_z]\}$, having possibly noisy labels $z_i = y(x_i) + n_i, n_i \sim \mathcal{N}(0, \sigma_z^2)$, and an unlabeled dataset $S_g = \{(x_i), i \in [n_g]\}$. Additionally, the learner is informed with the constraint knowledge, which includes an upper bound $g_{\mathrm{ub}}(x)$ and an lower bound $g_{\mathrm{lb}}(x)$ on the true label corresponding to input x, i.e. $g_{\mathrm{lb}}(x) \leq y(x) \leq g_{\mathrm{ub}}(x)$. A neural network $h_{\boldsymbol{W}}(x)$ is used for learning and the metric of interest is the mean square error (MSE) of the network output $h_{\boldsymbol{W}}(x)$ with respect to the true label $y(x)$ on a test dataset S_t, which is expressed as $\hat{R}_{S_t}(h_{\boldsymbol{W}}) = \frac{1}{2|S_t|}\sum_{(x_i,y_i)\in S_t}\left(h_{\boldsymbol{W}}(x_i) - y_i\right)^2$. Assume that the relationship to be learned is governed by a multi-dimensional Bohachevsky function $y(x) = x\boldsymbol{A}\boldsymbol{A}^\top x^\top - c\cos\left(a^\top x\right) + c$, where $\boldsymbol{A}$ is a $b \times b$ matrix, a is a b-dimensional vector and c is a constant. The constraint knowledge includes an upper bound model $g_{\mathrm{ub}}(x) = x\boldsymbol{A}\boldsymbol{A}^\top x^\top + ub$ with $ub \geq 2c$, and an lower bound model $g_{\mathrm{lb}}(x) = x\boldsymbol{A}\boldsymbol{A}^\top x^\top + lb$. with $lb \leq 0$. While it is not strongly convex and hence deviates from the assumptions in our theoretical analysis, we use ReLU as the knowledge-based risk function, i.e., the knowledge-based risk is written as $r_{\mathrm{K}}(h_{\boldsymbol{W}}(x)) = \mathrm{relu}\left(h_{\boldsymbol{W}}(x) - g_{\mathrm{ub}}(x)\right) + \mathrm{relu}\left(g_{\mathrm{lb}}(x) - h_{\boldsymbol{W}}(x)\right)$. If $ub - lb$ is larger, the uncertainty of the label given the knowledge is larger — the knowledge imperfectness is higher. We choose (lb, ub) as $(0, 0.6)$ and



(a) (b) (c)

Figure 1. Test MSE under different hyper-parameters. $\sigma_z^2 = 0$ means using perfect labels; $\sigma_z^2 = 0.1$ means using imperfect labels with noise variance 0.1; knowledge imperfectness is determined by ub and lb in the problem setting. (a) Training on the standard informed objective Eqn. (3) using knowledge with high imperfectness; (b) Training on the standard informed objective Eqn. (3) using knowledge with low imperfectness; (c) Training on the generalized informed objective Eqn. (5) using knowledge with low imperfectness and 400 labels.

$(0, 0.8)$ respectively to show the performances under low and high knowledge imperfectness. More details of the setup are in Appendix G.1.

7.2. Results

The curves of test MSE with different knowledge and label settings are shown in Fig. 1. In all the three figures, test MSE in $\lambda = 0$ approximately measures knowledge-regularized label imperfectness in Definition. 4, while test MSE in $\lambda = 1$ approximately measures knowledge imperfectness in Definition 3. We first use the training objective Eqn.(3) in Fig. 1(a) and Fig. 1(b) to show the effect of adjusting λ, which controls the knowledge effects (see Remark 4). From both Fig. 1(a) and Fig. 1(b), we see that the test MSE is smaller when there are more labeled samples and when label noise variance is lower. Importantly, domain knowledge helps reduce the MSE compared with pure label-supervised learning, especially for the cases with fewer labels and high label noise variance. Also, by comparing Fig. 1(a) and Fig. 1(b), we can find that the test MSE is lower when the knowledge imperfectness is lower. Additionally, Fig. 1(c) gives the test MSEs training on the generalized objective (5) under different β when the labeled dataset size is 400, showing that the test risk can be reduced by adjusting β which controls the knowledge regularization effect (see Remark 5). We can find that by properly adjusting β, the test MSEs under label noise are very close to that without label noise (the blue line). When $\beta = 1$, the test MSE is the highest since no labeled data is used to provide supervision.

More results, including another application of learning to manage wireless spectrum, are available in Appendix G.2.

8. Conclusion

In this paper, we consider an informed DNN with domain knowledge integrated with its training risk function. We quantitatively demonstrate that domain knowledge can improve the generalization performance and reduce the sampling complexity, while also impacting the point to which the network output converges. Our analysis also reveals that knowledge affects the generalization performance in two ways: regularizing the label supervision, and supplementing the labeled samples. Finally, we discuss how an informed DNN relates to other learning frameworks.

Acknowledgment

This work was supported in part by the U.S. NSF under CNS-1910208.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *NeurIPS*, 2019a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019b.

Amit, R. and Meir, R. Meta-learning by adjusting priors based on extended pac-bayes theory. In *ICML*, pp. 205–214, 2018.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, 2019a.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. *NeurIPS*, 2019b.

Bahri, Y., Gu, Q., Karbasi, A., and Sedghi, H. Over-parameterization: Pitfalls and opportunities. In *ICML Workshop*, 2021. URL

https://icml.cc/Conferences/2021/ScheduleMultitrack?event=8357.

Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., et al. Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Technical report, USDOE Office of Science (SC), Washington, DC (United States), 2019.

Bamler, R., Salehi, F., and Mandt, S. Augmenting and tuning knowledge graph embeddings. In *UAI*, pp. 508–518. PMLR, 2020.

Bardenet, R., Brendel, M., Kégl, B., and Sebag, M. Collaborative hyperparameter tuning. In *ICML*, 2013.

Bartlett, P., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *NeurIPS*, 2017.

Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J., et al. Interaction networks for learning about objects, relations and physics. In *NeurIPS*, pp. 4502–4510, 2016.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Beck, C., Weinan, E., and Jentzen, A. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *Journal of Nonlinear Science*, 29(4):1563–1619, 2019.

Benaim, S. and Wolf, L. One-shot unsupervised cross domain translation. In *NeurIPS*, pp. 2104–2114, 2018.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *ICLR*, 2020. URL https://openreview.net/forum?id=BJg866NFvB.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pp. 483–513. PMLR, 2020.

Borghesi, A., Baldo, F., and Milano, M. Improving deep learning models via constraint-based domain knowledge: a brief survey. *arXiv preprint arXiv:2005.10691*, 2020.

Cannings, T. I., Fan, Y., and Samworth, R. J. Classification with imperfect training labels. *Biometrika*, 107(2):311–330, 2020.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *NeurIPS*, 2019.

Cao, Y. and Gu, Q. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *AAAI*, 2020.

Chen, Y., Gao, R., Liu, F., and Zhao, D. Modulenet: Knowledge-inherited neural architecture search. *IEEE Transactions on Cybernetics*, 2021a.

Chen, Z., Cao, Y., Zou, D., and Gu, Q. How much over-parameterization is sufficient to learn deep relu networks? *ICLR*, 2021b.

Chiang, M., Hande, P., and Lan, T. *Power control in wireless cellular networks*. Now Publishers Inc, 2008.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *ICCV*, pp. 4794–4802, 2019.

Clarkson, K. L. Building triangulations using ε-nets. In *STOC*, 2006.

Cui, W., Shen, K., and Yu, W. Spatial deep learning for wireless scheduling. *IEEE Journal on Selected Areas in Communications*, 37(6):1248–1261, 2019.

Dao, T., Kamath, G. M., Syrgkanis, V., and Mackey, L. Knowledge distillation as semiparametric inference. *ICLR*, 2021.

Deng, C., Ji, X., Rainey, C., Zhang, J., and Lu, W. Integrating machine learning with human knowledge. *iScience*, 23(11):101656, 2020.

Diligenti, M., Gori, M., and Sacca, C. Semantic-based regularization for learning and inference. *Artificial Intelligence*, 244:143–165, 2017a.

Diligenti, M., Roychowdhury, S., and Gori, M. Integrating prior knowledge into deep learning. In *ICMLA*, pp. 920–923, 2017b.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *ICML*, pp. 1675–1685, 2019.

Fang, Y., Kuan, K., Lin, J., Tan, C., and Chandrasekhar, V. Object detection meets knowledge graphs.(2017). In *IJCAI*, 2017.

Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., and Anandkumar, A. Born again neural networks. In *ICML*, pp. 1607–1616, 2018.

Gao, H., Shou, Z., Zareian, A., Zhang, H., and Chang, S.-F. Low-shot learning via covariance-preserving adversarial augmentation networks. In *NeurIPS*, pp. 975–985, 2018a.

Gao, T., Liu, H., Liu, J., Rajan, H., and Gao, H. A global convergence theory for deep relu implicit networks via over-parameterization. In *ICML*, 2021.

Gao, Y., Xu, H., Lin, J., Yu, F., Levine, S., and Darrell, T. Reinforcement learning from imperfect demonstrations. *ICLR Workshop*, 2018b.

Garcia, V. and Bruna, J. Few-shot learning with graph neural networks. *ICLR*, 2018.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. Pac-bayesian theory meets bayesian inference. *NeurIPS*, 2016.

Goldsmith, A. *Wireless Communications*. Cambridge University Press, 2005.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Gore, A. D. and Karandikar, A. Link scheduling algorithms for wireless mesh networks. *IEEE Communications Surveys & Tutorials*, 13(2):258–273, 2010.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Guedj, B. A primer on pac-bayesian learning. *Proceedings of the 2nd congress of the Société Mathématique de France*, pp. 391–414, 2019.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Dulac-Arnold, G., et al. Deep q-learning from demonstrations. *AAAI*, 2018.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *Neurips Deep Learning Workshop*, 2014.

Hong, M. and Luo, Z.-Q. Signal processing and optimal resource allocation for the interference channel. In *Academic Press Library in Signal Processing*, volume 2, pp. 409–469. 2014.

Humbird, K. D., Peterson, J. L., and McClarren, R. G. Deep neural network initialization with decision trees. *IEEE transactions on neural networks and learning systems*, 30 (5):1286–1295, 2018.

Husken, M. and Goerick, C. Fast learning for problem classes using knowledge based network initialization. In *IEEE IJCNN*, 2000.

Institute, T. A. T. Physics-informed machine learning, 2020. https://www.turing.ac.uk/research/theory-and-method-challenge-fortnights/physics-informed-machine-learning.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *NeurIPS*, 2018.

Ji, G. and Zhu, Z. Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. *NeurIPS*, 2020.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *ICLR*, 2020.

Karpatne, A., Watkins, W., Read, J., and Kumar, V. Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*, 2017.

Khanduri, P., Yang, H., Hong, M., Liu, J., Wai, H. T., and Liu, S. Decentralized learning for overparameterized problems: A multi-agent kernel approximation approach. In *International Conference on Learning Representations*, 2021.

Khoo, Y., Lu, J., and Ying, L. Solving parametric pde problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

Klautau, A., Batista, P., González-Prelcic, N., Wang, Y., and Heath, R. W. 5g mimo data for machine learning: Application to beam-selection using deep learning. In *ITA*, pp. 1–9, 2018.

Kurata, G., Xiang, B., and Zhou, B. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In *NAACL: Human Language Technologies*, pp. 521–526, 2016.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *NeurIPS*, 2019.

Liang, F., Shen, C., Yu, W., and Wu, F. Towards optimal power control via ensembling deep neural networks. *IEEE Transactions on Communications*, 68(3): 1760–1776, 2019.

Lu, L., Meng, X., Mao, Z., and Karniadakis, G. E. Deepxde: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, 2021.

Maher, M. and Sakr, S. Smartml: A meta learning-based framework for automated selection and hyperparameter tuning for machine learning algorithms. In *EDBT*, 2019.

Marino, K., Salakhutdinov, R., and Gupta, A. The more you know: Using knowledge graphs for image classification. *CVPR*, 2016.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. Information bottleneck learning using privileged information for visual recognition. In *CVPR*, June 2016.

Muralidhar, N., Islam, M. R., Marwah, M., Karpatne, A., and Ramakrishnan, N. Incorporating prior domain knowledge into deep neural networks. In *IEEE Big Data*, pp. 36–45, 2018.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The role of over-parametrization in generalization of neural networks. In *ICLR*, 2018.

Nitanda, A., Chinot, G., and Suzuki, T. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Oymak, S. and Soltanolkotabi, M. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Pfrommer, J., Zimmerling, C., Liu, J., Kärger, L., Henning, F., and Beyerer, J. Optimisation of manufacturing process parameters using deep neural networks as surrogate models. *Procedia CiRP*, 72:426–431, 2018.

Phuong, M. and Lampert, C. Towards understanding knowledge distillation. In *ICML*, pp. 5142–5151, 2019.

Rahbar, A., Panahi, A., Bhattacharyya, C., Dubhashi, D., and Chehreghani, M. H. On the unreasonable effectiveness of knowledge distillation: Analysis in the kernel regime. *arXiv preprint arXiv:2003.13438*, 2020.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Ramsey, C. L. and Grefenstette, J. J. Case-based initialization of genetic algorithms. In *ICGA*, pp. 84–91, 1993.

Robinson, J., Jegelka, S., and Sra, S. Strength from weakness: Fast learning using weak supervision. In *ICML*, pp. 8127–8136, 2020.

Sanayei, S. and Nosratinia, A. Antenna selection in mimo systems. *IEEE Communications magazine*, 42(10):68–73, 2004.

Sharmanska, V., Quadrianto, N., and Lampert, C. H. Learning to rank using privileged information. In *ICCV*, 2013.

Shin, Y., Darbon, J., and Karniadakis, G. E. On the convergence and generalization of physics informed neural networks. *arXiv preprint arXiv:2004.01806*, 2020.

Silvestri, M., Lombardi, M., and Milano, M. Injecting domain knowledge in neural networks: a controlled experiment on a constrained problem. pp. 266–282, 2021.

Song, C., Ramezani-Kebrya, A., Pethick, T., Eftekhari, A., and Cevher, V. Subquadratic overparameterization for shallow neural networks. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *NeurIPS*, 2021.

Sun, H., Chen, X., Shi, Q., Hong, M., Fu, X., and Sidiropoulos, N. D. Learning to optimize: Training deep neural networks for interference management. *IEEE Transactions on Signal Processing*, 66(20):5438–5453, 2018.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *CVPR*, pp. 1199–1208, 2018.

Towell, G. G. and Shavlik, J. W. Knowledge-based artificial neural networks. *Artificial intelligence*, 70(1-2):119–165, 1994.

Van Rijn, J. N. and Hutter, F. Hyperparameter importance across datasets. In *ACM SIGKDD*, pp. 2367–2376, 2018.

Vapnik, V. and Vashist, A. A new learning paradigm: Learning using privileged information. *Neural networks*, 22 (5-6):544–557, 2009.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., Pfrommer, J., Pick, A., et al. Informed machine learning-a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *NeurIPS*, pp. 9712–9724, 2019.

Wei, C., Kakade, S., and Ma, T. The implicit and explicit regularization effects of dropout. In *ICML*, 2020.

Willard, J., Jia, X., Xu, S., Steinbach, M. S., and Kumar, V. Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, 2020.

Xu, J., Zhang, Z., Friedman, T., Liang, Y., and Broeck, G. A semantic loss function for deep learning with symbolic knowledge. In *ICML*, pp. 5502–5511, 2018.

Yang, G. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

Yin, Y., Le Guen, V., Dona, J., de Bézenac, E., Ayed, I., Thome, N., and Gallinari, P. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, 2021.

Zappone, A., Di Renzo, M., and Debbah, M. Wireless networks design in the era of deep learning: Model-based, ai-based, or both? *IEEE Transactions on Communications*, 67(10):7331–7376, 2019. doi: 10.1109/TCOMM.2019.2924010.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, Y., Tang, H., and Jia, K. Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data. In *ECCV*, pp. 233–248, 2018.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

Zou, D. and Gu, Q. An improved analysis of training over-parameterized deep neural networks. In *NeurIPS*, 2019.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

Appendix

A. Training Algorithm

To train the knowledge-informed DNN, we consider a gradient descent approach in Algorithm 1. This training approach has also been commonly considered in the literature (Allen-Zhu et al., 2019b; Zou & Gu, 2019; Du et al., 2019) for theoretical analysis of standard DNNs without domain knowledge.

Algorithm 1 Informed Neural Network Training by Gradient Descent

Initialization: Initialize each entry of weights $\boldsymbol{W}_0^{(0)}$, $\boldsymbol{W}_l^{(0)}, l \in [L]$ independently by $\mathcal{N}\left(0, \frac{2}{m}\right)$ and each entry of $\boldsymbol{V}^{(0)}$ independently by $\mathcal{N}\left(0, \frac{1}{d}\right)$.
for $t = 0, \cdots, T-1$ **do**
  Update the weights as $\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} + \eta \bigtriangledown_{\boldsymbol{W}} \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)$.
end for
Output: $\boldsymbol{W}^{(T)}$.

B. Notations, Key Lemmas and Proofs of Main Results in Section 4 and Section 5

B.1. Further Notations

Before the proofs, we list some additional notations as below. Denote $n' = |S_z| + |S_g| = n_z + n_g$. We assign the samples in the dataset S_z with indices from 1 to n_z and the samples in the dataset S_g with indices from $n_z + 1$ to n'. The informed risk of an informed DNN in Eqns. (3),(5) can be re-written as

$$\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right) = \sum_{i=1}^{n'} \left[\mu_i r\left(h_{\boldsymbol{W}}\left(x_i\right), z_i\right) + \lambda_i r_{\mathrm{K}}\left(h_{\boldsymbol{W}}\left(x_i\right), g(x_i)\right)\right], \tag{6}$$

where $\sum_{i=1}^{n'}\left(\mu_i + \lambda_i\right) = 1$. Thus, in Eqn. (3), we have $\mu_i = \frac{1-\lambda}{n_z}\mathbb{1}(x_i \in S_z)$ and $\lambda_i = \frac{\lambda}{n_g}\mathbb{1}(x_i \in S_g)$; in Eqn. (5), we have $\mu_i = \frac{(1-\lambda)(1-\beta)}{n_z}\mathbb{1}(x_i \in S_z)$ and $\lambda_i = \frac{(1-\lambda)\beta}{n_g'}\mathbb{1}(x_i \in S_g') + \frac{\lambda}{n_g''}\mathbb{1}(x_i \in S_g'')$. We prove convergence for the above three risks.

For any input $x_i, i \in [n']$, we denote the DNN output with respect to weight $\boldsymbol{W}$ as $h_{\boldsymbol{W},i} = h_{\boldsymbol{W}}(x_i)$. To express the output of the ReLu activation of the l-th layer for an input sample x_i, for $l \in [L]$ and $i \in [n]$, we denote a diagonal matrix $\boldsymbol{D}_{l,i}$ with its j-th (for $j \in [m]$) diagonal entry as $\mathbb{1}\left(\left[\boldsymbol{W}_l h_{l-1}\right]_j \geq 0\right)$. Thus, given the input x_i, the DNN outpoput can be expressed as

$$h_{\boldsymbol{W},i} = \boldsymbol{V}\boldsymbol{D}_{L,i}\boldsymbol{W}_L\boldsymbol{D}_{L-2,i}\cdots\boldsymbol{D}_{0,i}\boldsymbol{W}_0 x_i. \tag{7}$$

Also, we denote the informed risk for hypothesis $h \in \mathcal{H}$ and input x_i as

$$r_{\mathrm{I},i} = \mu_i r\left(h(x_i), z_i\right) + \lambda_i r_{\mathrm{K}}\left(h(x_i), g(x_i)\right) \tag{8}$$

The gradient of informed risk with respect to the hypothesis output is

$$u_i(h(x_i)) = \bigtriangledown_h r_{\mathrm{I},i}(h(x_i)) = \mu_i \bigtriangledown_h r\left(h(x_i), z_i\right) + \lambda_i \bigtriangledown_h r_{\mathrm{K}}\left(h(x_i), g(x_i)\right). \tag{9}$$

After constructing the smooth sets, denote for the kth smooth set, the sum of indices as $M_k = \sum_{\mathcal{I}_{\phi,k}}\left(\mu_i + \lambda_i\right)$. Denote the sum risk of the kth smooth set for hypothesis $h \in \mathcal{H}$ as

$$\bar{r}_{\mathrm{I},k}(h(x_i)) = \sum_{i \in \mathcal{I}_{\phi,k}} r_{\mathrm{I},i}(h(x_i)). \tag{10}$$

Thus, the effective label given in Definition 2 is written as $y_{\mathrm{eff},k} = \arg\min_h \bar{r}_{\mathrm{I},k}(h)$ with h in the space of network output, and the optimal effective risk is written as $r_{\mathrm{eff},k} = \bar{r}_{\mathrm{I},k}(y_{\mathrm{eff},k})$.

We then give some key technical lemmas which are the foundations for our further analysis. The proofs for these lemmas are shown in Appendix C.

B.2. Forward Perturbation Regarding Inputs

The forward perturbation for weights in the weight update range is proved in (Allen-Zhu et al., 2019b), However, to characterize the smooth sets, it is important to prove forward perturbation for inputs in a smooth set, which is given as follows.

Lemma B.1. *For any $i \in \mathcal{I}_{\phi,k}, k \in [N]$, let $h_{l,k} = h_l(x'_k)$, $h_{l,i} = h_l(x_i)$, and $f_{l,k} = \boldsymbol{W}_l h_{l-1}(x'_k)$, $f_{l,i} = \boldsymbol{W}_l h_{l-1}(x_i)$ and denote $\boldsymbol{D}'_{l,i,k} \in \mathbb{R}^{m\times m}$ as the diagonal matrix with $[\boldsymbol{D}'_{l,i,k}]_{j,j} = \mathbb{1}([f^{(0)}_{l,i}]_j \geq 0) - \mathbb{1}([f^{(0)}_{l,k}]_j \geq 0)$. Assuming $\phi \leq O(L^{-9/2}\log^{-3}(m)\log^{-3/4}(1/\phi))$, we have with probability at least $1-\phi$ over the randomness of $\boldsymbol{W}^{(0)}$,*
(a) At initialization, $\|\boldsymbol{D}'_{l,i,k}\|_0 \leq O(m\phi^{2/3}L\log^{1/2}(1/\phi))$
(b) For $\boldsymbol{W} \in \mathcal{B}(\boldsymbol{W}^{(0)}, \tau)$ with $\tau \leq O(\phi^{3/2})$ we have $\|h_{l,i} - h_{l,k}\| \leq O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)})$ and $\|f_{l,i} - f_{l,k}\| \leq O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)})$.

The proof of Lemma B.1 is given in Section C.1. The forward perturbation regarding inputs indicates the smoothness property of neural networks with respect to inputs. For compactness, we absorb the logarithmically increasing terms into $\widetilde{O}$ and denote $\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)) = O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)})$, $\widetilde{O}(L^{-9/2}\log^{-3}(m)) = O(L^{-9/2}\log^{-3}(m)\log^{-3/4}(1/\phi))$ in the following analysis.

B.3. Properties of Strong Convexity

Since our analysis is based on strongly convex risk functions, we give some key properties of strongly convex functions.

Lemma B.2 (Properties of Strong Convexity)**.** *If a strongly convex function $r(h)$ has a minimum value of $r(h^*) = r_{\min}$ and the eigenvalues of its Hessian matrix lie in $[\rho, 1]$, then we have $\|\bigtriangledown r(h)\|^2 \leq 2\left(r\left(h\right) - r_{\min}\right)$, $\|\bigtriangledown r(h)\|^2 \geq 2\rho\left(r\left(h\right) - r_{\min}\right)$ and $\|h^* - h\| \leq \frac{2}{\rho}\left\|\bigtriangledown r\left(h\right)\right\|$.*

Lemma B.3. *If the risk functions r and r_{K} are strongly convex with their eigenvalues of Hessian matrices in $[\rho, 1]$, then we have for hypothesis $h \in \mathcal{H}$, if $\|h(x_i) - h(x'_k)\| \leq \widetilde{O}(L^{5/2}\phi\log^{1/2}(m))$ for $i \in \mathcal{I}_{\phi,k}, k \in [N]$, the sum risk gradient for a smooth set with $\phi \leq \widetilde{O}(L^{-9/2}\log^{-3}(m))$ with respect to h satisfies,*

$$\begin{aligned}
&\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i(h(x_i))\|^2 \geq 2M_k\rho\left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k}\right) - M_k^2\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\\
&\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i(h(x_i))\|^2 \leq 2M_k\left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k}\right) + M_k^2\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\\
&\sum_{i\in\mathcal{I}_{\phi,k}}\left(\mu_i+\lambda_i\right)\left\|h\left(x_i\right) - y_{\mathrm{eff},k}\right\|^2 \leq \frac{1}{\rho^2}O\left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k} + M_k\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right),
\end{aligned}$$

where u_i is defined in Eqn. (9) and $M_k = \sum_{\mathcal{I}_{\phi,k}}\left(\mu_i+\lambda_i\right)$.

Lemma B.2 and B.3 are proved in Section C.2.

B.4. Proof of Theorem 4.1.

In this section, we prove the convergence for informed risks in Eqn. (3), Eqn. (5). First, the gradient lower bound, semi-smoothness of the risk function, and initialized risk bound are proved.

Lemma B.4 (Gradient Lower Bound)**.** *. For any $\boldsymbol{W}$: $\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\| \leq \tau$ $\tau = O(N^{-9/2}\phi^{3/2}\rho^{3/2}\bar{\lambda}^{3/2}\alpha^{3/2}L^{-15/2}\log^{-3/2}(m))$ and $\phi \leq \widetilde{O}(L^{-9/2}\log^{-3}(m))$, with Assumption 1 satisfied, we have with probability at least $1 - O(\phi)$ over the randomness of $\boldsymbol{W}^{(0)}$, the gradient of label-based data risk satisfies*

$$\left\|\bigtriangledown_{\boldsymbol{W}}\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right)\right\|_F^2 \geq \Omega\left(\frac{\alpha m\phi\rho\bar{\lambda}}{dN^2}\right)\left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right) - \hat{R}_{\mathrm{eff}} - \widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right).$$

where $\hat{R}_{\mathrm{eff}} = \sum_{k=1}^N r_{\mathrm{eff},k}$, and $\bar{\lambda}$ is a parameter with lower bound $\Omega(\min(1-\lambda,\lambda)\mathbb{1}(\lambda\in(0,1)) + \mathbb{1}(\lambda\in\{0,1\}))$.

The proof of Lemma B.4 can be found in Section C.3.

Lemma B.5. *For any* $\boldsymbol{W}$, $\boldsymbol{W}' \in \mathcal{B}\left(\boldsymbol{W}^{(0)}, \tau\right)$, $\tau \in \left[\Omega(d^{3/2}m^{-3/2}L^{-5/2}\log^{-3/2}(m)), O(L^{-9/2}[\log^{-3}(m)])\right]$ *and* $\phi \leq \widetilde{O}(L^{-9/2}\log^{-3}(m))$, *with probability at least* $1 - O(\phi)$ *over the randomness of* $\boldsymbol{W}^{(0)}$, *we have*

$$\begin{aligned}\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}'\right) \leq & \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right) + \left\langle \bigtriangledown_{\boldsymbol{W}} \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right), \boldsymbol{W}' - \boldsymbol{W} \right\rangle + O(L^2 m/d)\left\|\widehat{\boldsymbol{W}}\right\|^2 \\ & + \left(\sqrt{\left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right) - \hat{R}_{\mathrm{eff}} - \widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)}\right) O\left(N^{1/2}\tau^{1/3}L^{5/2}\sqrt{m\log(m)}d^{-1/2}\right)\left\|\widehat{\boldsymbol{W}}\right\|\end{aligned}$$

The proof of Lemma B.5 can be found in Section C.4.

Lemma B.6. *If* $m \geq \Omega\left(L\log(NL\phi^{-1})\right)$ *and* $\phi \leq \widetilde{O}(L^{-9/2}\log^{-3}(m))$, *with probability at least* $1 - O(\phi)$ *over the randomness of* $\boldsymbol{W}^{(0)}$, *at initialization, we have for any* $x_i, i \in [n']$,

$$\begin{aligned}\left\|h_{\boldsymbol{W}^{(0)},i}\right\| &\leq O\left(\log^{1/2}(1/\phi)\right), \\ \text{and } \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(0)}\right) - \hat{R}_{\mathrm{eff}} &\leq O\left(\log^{1/2}(1/\phi)\right).\end{aligned}$$

The proof of Lemma B.6 can be found in Section C.5.

Proof of Theorem 4.1.

Proof. **Convergence of the informed risk.** We first assume $\tau = \frac{\Gamma}{\sqrt{m}}$ with $\Gamma = Nd^{1/2}\phi^{-1/2}\rho^{-1/2}\bar{\lambda}^{-1/2}\alpha^{-1/2}$. Hence, with the choice of m, we have $\tau = O(N^{-9/2}\phi^{3/2}\rho^{3/2}\bar{\lambda}^{3/2}\alpha^{3/2}L^{-15/2}\log^{-3/2}(m))$. We get the recursion inequality based on gradient descent. By the weight update rule of gradient descent, we have $\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)} = -\eta \bigtriangledown \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t-1)}\right)$. Let $\Psi = \widetilde{O}(L^{5/2}\phi\log^{1/2}(m))$. By Lemma B.5, we have

$$\begin{aligned}& \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t+1)}\right) - \hat{R}_{\mathrm{eff}} - \Psi \\ \leq & \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right) - \hat{R}_{\mathrm{eff}} - \Psi - \left(\eta - O(\eta^2 L^2 m/d)\right)\left\|\bigtriangledown \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\right\|^2 \\ & + \eta\sqrt{2N\left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right) - \hat{R}_{\mathrm{eff}} - \Psi\right)} O\left(\tau^{1/3}L^{5/2}\sqrt{m\log(m)}d^{-1/2}\right)\left\|\bigtriangledown \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\right\| \\ \leq & \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right) - \hat{R}_{\mathrm{eff}} - \Psi - \Omega\left(\eta\right)\left\|\bigtriangledown \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\right\|^2 \\ & + \eta\Omega\left(N^{3/2}\tau^{1/3}L^{5/2}\log^{1/2}(m)\phi^{-1/2}\rho^{-1/2}\bar{\lambda}^{-1/2}\alpha^{-1/2}\right)\left\|\bigtriangledown \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\right\|^2 \\ \leq & \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right) - \hat{R}_{\mathrm{eff}} - \Psi - \Omega\left(\eta\right)\left\|\bigtriangledown \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\right\|^2,\end{aligned} \tag{11}$$

where the second inequality holds by the choice of $\eta = O(\frac{d}{L^2 m})$ such that $O(\eta L^2 m/d) = O(1)$ and the gradient lower bound in Lemma B.4, and the last inequality holds by the choice of $m \geq \Omega\left(N^{11}L^{15}d\phi^{-4}\rho^{-4}\bar{\lambda}^{-4}\alpha^{-4}\log^3(m)\right)$ such that $\Omega\left(N^{3/2}\tau^{1/3}L^{5/2}\log^{1/2}(m)\phi^{-1/2}\rho^{-1/2}\bar{\lambda}^{-1/2}\alpha^{-1/2}\right) \leq O(1)$.

Further, by Lemma B.4, we have

$$\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t+1)}\right) - \hat{R}_{\mathrm{eff}} - \Psi \leq \left(1 - \Omega\left(\frac{\eta\alpha m\phi\rho\bar{\lambda}}{dN^2}\right)\right)\left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right) - \hat{R}_{\mathrm{eff}} - \Psi\right).$$

Based on the iteration of the recursion inequality, with probability at least $1 - O(\phi)$, we have

$$\begin{aligned}& \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right) - \hat{R}_{\mathrm{eff}} - \Psi \\ \leq & \left(1 - \Omega\left(\frac{\eta\alpha m\phi\rho\bar{\lambda}}{dN^2}\right)\right)^t \left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(0)}\right) - \hat{R}_{\mathrm{eff}} - \Psi\right) \\ \leq & \left(1 - \Omega\left(\frac{\eta\alpha m\phi\rho\bar{\lambda}}{dN^2}\right)\right)^t O(\log^{1/2}(1/\phi)),\end{aligned}$$

where the last inequality comes from Lemma B.6. Then, by taking logarithm, we get

$$\ln\left(\left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)-\hat{R}_{\mathrm{eff}}-\Psi\right)\right)\leq t\ln\left(1-\Omega\left(\frac{\eta\alpha m\phi\rho\bar{\lambda}}{dN^2}\right)\right)+\frac{1}{2}\ln\log(1/\phi)$$
$$\leq -t\Omega\left(\frac{\eta\alpha m\phi\rho\bar{\lambda}}{dN^2}\right)+\frac{1}{2}\ln\log(1/\phi).$$

Since $\eta=O(\frac{d}{L^2m})$, after $T=O\left(\frac{L^2N^2}{\phi\rho\bar{\lambda}\alpha}\ln(\epsilon^{-1}\log(\phi^{-1}))\right)$ iterations, for any $\epsilon>0$, we have

$$\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(T)}\right)-\hat{R}_{\mathrm{eff}}\leq O(L^{5/2}\phi\log^{1/2}(1/\phi)\log^{1/2}m)+\epsilon.$$

By setting ϕ as $\phi\log^{1/2}(1/\phi))\leq\epsilon L^{-5/2}\log^{-1/2}m$ (which satisfies the assumption of ϕ in Theorem 4.1), we can bound $\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)-\hat{R}_{\mathrm{eff}}$ by a small positive quantity ϵ.

Verify the weight update range. Now, we verify that the assumption $\left\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\right\|\leq\frac{\Gamma}{\sqrt{m}}$ holds. Denote $\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)=\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)-\hat{R}_{\mathrm{eff}}-\Psi$. By Eqn. (11), we have

$$\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t+1)}\right)-\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\leq-\Omega\left(\eta\right)\left\|\triangledown\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\right\|^2.$$

Then, we have

$$\sqrt{\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t+1)}\right)}-\sqrt{\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)}=\frac{\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t+1)}\right)-\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)}{\sqrt{\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t+1)}\right)}+\sqrt{\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)}}$$
$$\leq\frac{-\Omega\left(\eta\right)\left\|\triangledown\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\right\|^2}{2\sqrt{\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)}}\leq-O\left(\frac{m^{1/2}\phi^{1/2}\rho^{1/2}\bar{\lambda}^{1/2}\alpha^{1/2}}{d^{1/2}N}\right)\left\|\eta\triangledown\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(t)}\right)\right\|$$

where the last inequality follows from Lemma B.4.

By the triangle inequality, for any $t\in[T]$, we have

$$\begin{aligned}\left\|\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(0)}\right\|\leq&\sum_{s=0}^{t}\left\|\eta\triangledown\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(s)}\right)\right\|\\ \leq&O\left(\frac{d^{1/2}N}{m^{1/2}\phi^{1/2}\rho^{1/2}\bar{\lambda}^{1/2}\alpha^{1/2}}\right)\sqrt{\bar{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(0)}\right)}\leq O\left(\frac{\Gamma}{\sqrt{m}}\right),\end{aligned}\tag{12}$$

where $\Gamma=Nd^{1/2}\phi^{-1/2}\rho^{-1/2}\bar{\lambda}^{-1/2}\alpha^{-1/2}$. Hence, with the choice of m, we have $\tau=O(N^{-9/2}\phi^{3/2}\rho^{3/2}\bar{\lambda}^{3/2}\alpha^{3/2}L^{-15/2}\log^{-3/2}(m))$.

Convergence of network output. By Lemma B.3,we have

$$\sum_{k=1}^{N}\sum_{i\in\mathcal{I}_{\phi,k}}\left(\mu_i+\lambda_i\right)\left\|h_{\boldsymbol{W}^{(T)}}\left(x_i\right)-y_{\mathrm{eff},k}\right\|^2\leq\frac{1}{\rho^2}O\left(\left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(T)}\right)-\hat{R}_{\mathrm{eff}}\right)+\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)\leq O(\epsilon).$$

Denoting $k(x_i)$ as the index of the cell containing x_i and rearranging the above summation, we have

$$\sum_{x_i\in S_z}\mu_i\left\|h_{\boldsymbol{W}^{(T)}}\left(x_i\right)-y_{\mathrm{eff},k(x_i)}\right\|^2+\sum_{x_j\in S_g}\lambda_j\left\|h_{\boldsymbol{W}^{(T)}}\left(x_j\right)-y_{\mathrm{eff},k(x_j)}\right\|^2\leq O(\epsilon).$$

□

B.5. Proof of Generalization

In this section, we prove the generalization bound based on Rademacher complexity. We first present the bound of Rademacher complexity for neural networks.

Lemma B.7. *[Theorem 3.3 in (Bartlett et al., 2017),Lemma A.3 in (Chen et al., 2021b)] If risk functions are 1-Lipschitz continuous, with probability at least* $1 - L\exp(-\Omega(m))$*, the Rademacher complexity* $\mathfrak{R}_S(\mathcal{F})$ *for the risk set* $\mathcal{F} = \left\{r\left(h_{\boldsymbol{W}}\left(x\right), y\right) : (x, y) \in \mathcal{X} \times \mathcal{Y}, \left\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\right\| \leq \tau\right\}$, $\tau = \frac{\Gamma}{\sqrt{m}}$ *with* $\Gamma = Nd^{1/2}\phi^{-1/2}\rho^{-1/2}\bar{\lambda}^{-1/2}\alpha^{-1/2}$ *given a dataset* S *of* n *samples is bounded as*

$$\mathfrak{R}_S(\mathcal{F}) \leq \Phi/\sqrt{n}, \tag{13}$$

where $\Phi = O\left(4^L L^{3/2} m^{1/2}\phi^{-b-1/2} d\rho^{-1/2}\bar{\lambda}^{-1/2}\alpha^{-1/2}\right)$.

Then, we need to bound the error between effective labels in Definition 2 and the output of optimal hypothesis in Definitions 3 and 4.

Lemma B.8. *Consistent with Definition 2, assume that for any smooth set* $k \in \mathcal{U}_\phi(S_z)$ *(containing at least one labeled sample),* $y_{\text{eff},k}$ *equivalently minimizes* $\sum_{i\in I_{\phi,k}} \frac{1-\beta}{n_z}\mathbb{1}(x_i \in S_z) r\left(h, z_i\right) + \frac{\beta}{n'_g}\mathbb{1}(x_i \in S'_g) r_{\mathrm{K}}\left(h, g_i\right)$*, and for any smooth set* $k \in [N] \setminus \mathcal{U}_\phi(S_z)$ *(not containing labeled sample),* $y_{\text{eff},k}$ *equivalently minimizes* $\sum_{i\in I_{\phi,k}} \frac{1}{n''_g} r_{\mathrm{K}}\left(h, g_i\right)$.
(a) *Letting* h^*_{K} *and* $h^*_{\mathrm{R},\beta}$ *be the optimal hypothesis for empirical risks in Definitions 3 and 4, respectively, we have with probability at least* $1 - O(\phi)$ *over the randomness of* $\boldsymbol{W}^{(0)}$*,*

$$\frac{1}{n''_g}\sum_{S''_g}\left\|h^*_{\mathrm{K},i} - y_{\text{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right),$$

and

$$\frac{1-\beta}{n_z}\sum_{S_z}\left\|h^*_{\mathrm{R},\beta,i} - y_{\text{eff},k(x_i)}\right\|^2 + \frac{\beta}{n'_g}\sum_{S'_g}\left\|h^*_{\mathrm{R},\beta,i} - y_{\text{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right),$$

where $\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) = O(L^{5/4}\phi^{1/2}\log^{1/4}(1/\phi)\log^{1/4} m)$.
(b) *Letting* $\bar{h}^*_{\mathrm{K}}$ *and* $\bar{h}^*_{\mathrm{R},\beta}$ *be the optimal hypothesis for the expected risks in Definitions 3 and 4, respectively, we have with probability at least* $1 - O(\phi) - \delta$*,*

$$\frac{1}{n''_g}\sum_{S''_g}\left\|\bar{h}^*_{\mathrm{K},i} - y_{\text{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + O(\sqrt{\frac{\log(1/\delta)}{n''_g}}),$$

and

$$\frac{1-\beta}{n_z}\sum_{S_z}\left\|\bar{h}^*_{\mathrm{R},\beta,i} - y_{\text{eff},k(x_i)}\right\|^2 + \frac{\beta}{n'_g}\sum_{S'_g}\left\|\bar{h}^*_{\mathrm{R},\beta,i} - y_{\text{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + O(\sqrt{\frac{\log(1/\delta)}{n_z}}),$$

where $\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) = O(L^{5/4}\phi^{1/2}\log^{1/4}(1/\phi)\log^{1/4} m)$.

Proof of Lemma B.8 is given in Section C.6.

B.5.1. Proof of Theorem 4.2

Proof. By generalization bound with Rademacher complexity and Lemma B.7, the population risk is bounded with probability at least $1 - \delta, \delta \in (0, 1)$ as

$$\begin{aligned}
&R(\boldsymbol{W}^{(T)}) \\
=&(1-\lambda)R(\boldsymbol{W}^{(T)}) + \lambda R(\boldsymbol{W}^{(T)}) \\
\leq&\frac{1-\lambda}{n_z}\sum_{S_z} r(h_{\boldsymbol{W}^{(T)},i}, y_i) + \frac{\lambda}{n_g}\sum_{S_g} r(h_{\boldsymbol{W}^{(T)},i}, y_i) + O\left(\Phi + \sqrt{\log(1/\delta)}\right)\left((1-\lambda)\sqrt{\frac{1}{n_z}} + \lambda\sqrt{\frac{1}{n_g}}\right)
\end{aligned} \tag{14}$$

For the empirical risk, we have

$$\begin{aligned}
&\frac{1-\lambda}{n_z}\sum_{S_z} r(h_{\boldsymbol{W}^{(T)},i}, y_i) + \frac{\lambda}{n_g}\sum_{S_g} r(h_{\boldsymbol{W}^{(T)},i}, y_i) \\
\leq &\frac{1-\lambda}{n_z}\sum_{S_z}\left(r(y_{\text{eff},k(x_i)}, y_i) + \left\|h_{\boldsymbol{W}^{(T)},i} - y_{\text{eff},k(x_i)}\right\|\right) + \frac{\lambda}{n_g}\sum_{S_g}\left[r(y_{\text{eff},k(x_i)}, y_i) + \left\|h_{\boldsymbol{W}^{(T)},i} - y_{\text{eff},k(x_i)}\right\|\right] \\
\leq &\sqrt{\epsilon} + \frac{1-\lambda}{n_z}\sum_{S_z} r(y_{\text{eff},k(x_i)}, y_i) + \frac{\lambda}{n_g}\sum_{S_g} r(y_{\text{eff},k(x_i)}, y_i)
\end{aligned} \tag{15}$$

where the first inequality holds because of the 1-Lipschitz of risk functions such that $r(h_{\boldsymbol{W}^{(t)},i}, y_i) - r(y_{\text{eff},k(x_i)}, y_i) \leq \left\|h_{\boldsymbol{W}^{(t)},i} - y_{\text{eff},k(x_i)}\right\|$, and the second inequality follows from the convergence of network output in Theorem 4.1.

Since with $\beta_\lambda = \frac{\lambda n_g'}{(1-\lambda)n_g + \lambda n_g'}$, the training objective in Eqn.(3) can also be written as $\hat{R}_{\text{I}}(\boldsymbol{W}) = \left(1-\lambda+\frac{\lambda n_g'}{n_g}\right)\left(\frac{1-\beta_\lambda}{n_z}\sum_{S_z} r(h_{\boldsymbol{W},i}, z_i) + \frac{\beta_\lambda}{n_g'}\sum_{S_g'} r_{\text{K}}(h_{\boldsymbol{W},i}, g_i)\right) + \frac{\lambda n_g''}{n_g}\sum_{S_g''} r_{\text{K}}(h_{\boldsymbol{W},i}, g_i)$, for any smooth set $k \in \mathcal{U}_\phi(S_z)$ (containing at least one labeled sample), $y_{\text{eff},k}$ equivalently minimizes $\sum_{i\in I_{\phi,k}} \frac{1-\beta_\lambda}{n_z}\mathbb{1}(x_i \in S_z) r(h, z_i) + \frac{\beta_\lambda}{n_g'}\mathbb{1}(x_i \in S_g') r_{\text{K}}(h, g_i)$ by Definition 2. Thus, the bounds of the differences between optimal hypothesis and effective labels in Lemma B.8 hold for $\beta = \beta_\lambda$ and $h^*_{\text{R},\beta_\lambda}$. Next, we can bound the total effective risk in terms of label and knowledge imperfectness, with probability at least $1 - O(\phi)$,

$$\begin{aligned}
&\frac{1-\lambda}{n_z}\sum_{S_z} r(y_{\text{eff},k(x_i)}, y_i) + \frac{\lambda}{n_g}\sum_{S_g} r(y_{\text{eff},k(x_i)}, y_i) \\
= &\frac{1-\lambda}{n_z}\sum_{S_z} r(y_{\text{eff},k(x_i)}, y_i) + \frac{\lambda}{n_g}\sum_{S_g'} r(y_{\text{eff},k(x_i)}, y_i) + \frac{\lambda}{n_g}\sum_{S_g''} r(y_{\text{eff},k(x_i)}, y_i) \\
\leq &\frac{1-\lambda}{n_z}\sum_{S_z} r(h^*_{\text{R},\beta_\lambda,i}, y_i) + \frac{\lambda}{n_g}\sum_{S_g'} r(h^*_{\text{R},\beta_\lambda,i}, y_i) + \frac{\lambda}{n_g}\sum_{S_g''} r(h_{\text{K},i}, y_i) \\
&+ \frac{1-\lambda}{n_z}\sum_{S_z}\left\|h^*_{\text{R},\beta_\lambda,i} - y_{\text{eff},k(x_i)}\right\| + \frac{\lambda}{n_g}\sum_{S_g'}\left\|h^*_{\text{R},\beta_\lambda,i} - y_{\text{eff},k(x_i)}\right\| + \frac{\lambda}{n_g}\sum_{S_g''}\left\|h^*_{\text{K},i} - y_{\text{eff},k(x_i)}\right\| \\
\leq &(1-\lambda)\widehat{Q}_{\text{R},S_z,S_g'}(\beta_\lambda) + \lambda\widehat{Q}_{\text{K},S_g''} + O(\sqrt{\epsilon})
\end{aligned} \tag{16}$$

where $\beta_\lambda = \frac{\lambda n_g'}{(1-\lambda)n_g + \lambda n_g'}$, the first inequality comes from the Lipschitz continuity of risk functions, and the last inequality holds by Lemma B.8 and the assumption $\phi \leq \widetilde{O}\left(\epsilon^2 L^{-9/2}\log^{-3}(m)\right)$ such that $\frac{1-\lambda}{n_z}\sum_{S_z}\left\|h^*_{\text{R},\beta_\lambda,i} - y_{\text{eff},k(x_i)}\right\| + \frac{\lambda}{n_g}\sum_{S_g'}\left\|h^*_{\text{R},\beta_\lambda,i} - y_{\text{eff},k(x_i)}\right\| \leq \left(1-\lambda+\frac{\lambda n_g'}{n_g}\right)\sqrt{\frac{1-\beta_\lambda}{n_z}\sum_{S_z}\|h^*_{\text{R},\beta_\lambda,i} - y_{\text{eff},k(x_i)}\|^2 + \frac{\beta_\lambda}{n_g'}\sum_{S_g'}\|h^*_{\text{R},\beta_\lambda,i} - y_{\text{eff},k(x_i)}\|^2} \leq O(\sqrt{\epsilon})$ and $\frac{\lambda}{n_g}\sum_{S_g'}\left\|h^*_{\text{R},\beta_\lambda,i} - y_{\text{eff},k(x_i)}\right\| \leq \frac{\lambda n_g''}{n_g}\sqrt{\frac{1}{n_g''}\sum_{S_g''}\|h^*_{\text{K},i} - y_{\text{eff},k(x_i)}\|^2} \leq O(\sqrt{\epsilon})$. In the last inequality of (16), we absorb $\frac{\lambda}{n_g}\sum_{S_g'} r(h^*_{\text{R},\beta_\lambda,i}, y_i)$ into $O(\sqrt{\epsilon})$ because the risk functions are upper bounded and S_g' is the set of samples sharing the same smooth sets with S_z, and so $\frac{\lambda}{n_g}\sum_{S_g'} r(h^*_{\text{R},\beta_\lambda,i}, y_i) \leq O(\frac{n_g'}{n_g}) \leq O(n_z\phi^b) \leq O(\sqrt{\epsilon})$.

Substituting Eqn.(16) and (15) into Eqn. (14), the population risk is bounded with probability at least $1 - O(\phi) - \delta, \delta \in (0,1)$ as

$$\begin{aligned}
&R(\boldsymbol{W}^{(T)}) = (1-\lambda)R(\boldsymbol{W}^{(T)}) + \lambda R(\boldsymbol{W}^{(T)}) \\
\leq &\sqrt{\epsilon} + (1-\lambda)\widehat{Q}_{\text{R},S_z,S_g'}(\beta_\lambda) + \lambda\widehat{Q}_{\text{K},S_g''} + O\left(\Phi + \sqrt{\log(1/\delta)}\right)\left((1-\lambda)\sqrt{\frac{1}{n_z}} + \lambda\sqrt{\frac{1}{n_g}}\right).
\end{aligned}$$

□

B.5.2. PROOF OF THEOREM 5.1

Proof. Based on the construction of smooth sets in Definition 1, denote $\mathcal{X}' = \mathcal{X}_\phi(S_z) = \bigcup_{k\in\mathcal{U}_\phi}(S_z)\mathcal{C}_{\phi,k}$ as the region covered by the smooth sets containing at least one sample in S_z, and let $\mathcal{X}'' = \mathcal{X}/\mathcal{X}'$. Let $\mathbb{P}_{\mathcal{X}'} = \int_{\mathcal{X}',\mathcal{Y}} p(y \mid x)\mathrm{d}y p(x)\mathrm{d}x$ and $\mathbb{P}_{\mathcal{X}''} = \int_{\mathcal{X}'',\mathcal{Y}} p(y \mid x)\mathrm{d}y p(x)\mathrm{d}x$ where $p(x)$ and $p(y \mid x)$ are probability densities. Then we have $P_{\mathcal{X}'} \leq O(n_z/N) = O(n_z\phi^b) = O(\sqrt{\epsilon})$ by the assumption that $\phi \leq (\sqrt{\epsilon}/n_z)^{1/b}$, and $\mathbb{E}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right] = \mathbb{E}_{\mathbb{P}_{\mathcal{X}'}}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right]\mathbb{P}_{\mathcal{X}'} + \mathbb{E}_{\mathbb{P}_{\mathcal{X}''}}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right]\mathbb{P}_{\mathcal{X}''}$. By generalization bound with Rademacher complexity and Lemma B.7, the population risk is bounded with probability at least $1-\delta, \delta \in (0,1)$ as

$$
\begin{aligned}
&R(\boldsymbol{W}^{(T)}) = (1-\lambda)\mathbb{E}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right] + \lambda\mathbb{E}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right] \\
=&(1-\lambda)\mathbb{E}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right] + \lambda\mathbb{E}_{\mathbb{P}_{\mathcal{X}''}}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right] + \lambda\left(\mathbb{E}_{\mathbb{P}_{\mathcal{X}'}}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right] - \mathbb{E}_{\mathbb{P}_{\mathcal{X}''}}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right]\right)\mathbb{P}_{\mathcal{X}'} \\
\leq&\frac{(1-\lambda)}{n_z}\sum_{S_z} r\left(h_{\boldsymbol{W}^{(T)},i}, y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(h_{\boldsymbol{W}^{(T)},i}, y_i\right) + \lambda O\left(\sqrt{\epsilon}\right) + O\left(\Phi + \sqrt{\log(1/\delta)}\right)\left((1-\lambda)\sqrt{\frac{1}{n_z}} + \lambda\sqrt{\frac{1}{n_g''}}\right).
\end{aligned} \tag{17}
$$

Then for the empirical risk, we have

$$
\begin{aligned}
&\frac{1-\lambda}{n_z}\sum_{S_z} r\left(h_{\boldsymbol{W}^{(T)},i}, y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(h_{\boldsymbol{W}^{(T)},i}, y_i\right) \\
\leq&\frac{1-\lambda}{n_z}\sum_{S_z} r\left(y_{\mathrm{eff},k(x_i)} y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(y_{\mathrm{eff},k(x_i)}, y_i\right) + \frac{1-\lambda}{n_z}\sum_{S_z}\left\|h_{\boldsymbol{W}^{(T)},i} - y_{\mathrm{eff},i}\right\| + \frac{\lambda}{n_g''}\sum_{S_g''}\left\|h_{\boldsymbol{W}^{(T)},i} - y_{\mathrm{eff},i}\right\| \\
\leq&O(\sqrt{\epsilon}) + \frac{1-\lambda}{n_z}\sum_{S_z} r\left(y_{\mathrm{eff},k(x_i)} y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(y_{\mathrm{eff},k(x_i)}, y_i\right),
\end{aligned} \tag{18}
$$

where the first inequality holds because of the Lipschitz continuity of risk functions such that $r(h_{\boldsymbol{W}^{(T)},i}, y_i) - r(y_{\mathrm{eff},k(x_i)}, y_i) \leq \left\|h_{\boldsymbol{W}^{(T)},i} - y_{\mathrm{eff},k(x_i)}\right\|$, and the second inequality follows from the convergence of network output in Theorem 4.1 (By Theorem 4.1, we have $\frac{(1-\lambda)(1-\beta)}{n_z}\sum_{S_z}\left\|h_{W^{(T)},i} - y_{\mathrm{eff},i}\right\|^2 \leq O(\epsilon)$ and $\frac{\lambda}{n_g''}\sum_{S_g''}\left\|h_{W^{(T)},i} - y_{\mathrm{eff},i}\right\|^2 \leq O(\epsilon)$, and so $\frac{1-\lambda}{n_z}\sum_{S_z}\left\|h_{W^{(T)},i} - y_{\mathrm{eff},i}\right\| \leq \sqrt{\frac{1-\lambda}{1-\beta}}O(\sqrt{\epsilon}) = O(\sqrt{\epsilon})$ and $\frac{\lambda}{n_g''}\sum_{S_g''}\left\|h_{W^{(T)},i} - y_{\mathrm{eff},i}\right\| \leq \sqrt{\lambda}O(\sqrt{\epsilon}) = O(\sqrt{\epsilon})$).

Next, we bound the empirical risk in terms of label and knowledge imperfectness as follows:

$$
\begin{aligned}
&\frac{1-\lambda}{n_z}\sum_{S_z} r\left(y_{\mathrm{eff},k(x_i)} y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(y_{\mathrm{eff},k(x_i)}, y_i\right) \\
\leq&\frac{1-\lambda}{n_z}\sum_{S_z} r(h^*_{\mathrm{R},\beta,i}, y_i) + \frac{\lambda}{n_g''}\sum_{S_g''} r(h_{\mathrm{K},i}, y_i) + \frac{1-\lambda}{n_z}\sum_{S_z}\left\|h^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\| + \frac{\lambda}{n_g''}\sum_{S_g''}\left\|h^*_{\mathrm{K},i} - y_{\mathrm{eff},k(x_i)}\right\| \\
\leq&(1-\lambda)\widehat{Q}_{\mathrm{R},S_z,S_g'}(\beta) + \lambda\widehat{Q}_{\mathrm{K},S_g''} + O(\sqrt{\epsilon})
\end{aligned} \tag{19}
$$

where the first inequality comes from the Lipschitz continuity of risk functions, and the concavity of squared root, the last inequality holds by Definitions 3 and 4, and Lemma B.8 and the assumption of ϕ such that $\phi\log^{1/2}(1/\phi) \leq O\left(\epsilon^2 L^{-5/2}\log^{-1/2}(m)\right)$. Concretely, by Lemma B.8 (a), we have $\frac{1-\beta}{n_z}\sum_{S_z}\left\|h^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right)$ and $\frac{1}{n_g''}\sum_{S_g''}\left\|h^*_{\mathrm{K},i} - y_{\mathrm{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right)$, and so it holds that $\frac{1-\lambda}{n_z}\sum_{S_z}\left\|h^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\| \leq \frac{1-\lambda}{\sqrt{1-\beta}}\widetilde{O}\left(L^{5/8}\phi^{1/4}\log^{1/8}(m)\right)$ and $\frac{\lambda}{n_g''}\sum_{S_g''}\left\|h^*_{\mathrm{K},i} - y_{\mathrm{eff},k(x_i)}\right\| \leq \lambda\widetilde{O}\left(L^{5/8}\phi^{1/4}\log^{1/8}(m)\right)$. Thus we obtain the last inequality of (19) by the assumption $\phi\log^{1/2}(1/\phi) \leq O\left(\epsilon^2 L^{-5/2}\log^{-1/2}(m)\right)$.

Substituting Eqns. (19) and (18) into Eqn. (17), we have

$$\begin{aligned}&R(\boldsymbol{W}^{(T)}) = (1-\lambda)\mathbb{E}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right] + \lambda\mathbb{E}\left[r\left(h_{\boldsymbol{W}^{(T)}}(x), y\right)\right]\\ \leq& O(\sqrt{\epsilon}) + (1-\lambda)\widehat{Q}_{\mathrm{R},S_z,S_g'}(\beta) + \lambda\widehat{Q}_{\mathrm{K},S_g''} + O\left(\Phi + \sqrt{\log(1/\delta)}\right)\left((1-\lambda)\sqrt{\frac{1}{n_z}} + \lambda\sqrt{\frac{1}{n_g''}}\right).\end{aligned}$$

□

B.5.3. Proof of Corollary 5.2

Proof. First, following (17), with probability at least $1 - O(\phi) - \delta, \delta \in (0,1)$, it holds that

$$\begin{aligned}R(\boldsymbol{W}^{(T)}) \leq& \frac{(1-\lambda)}{n_z}\sum_{S_z} r\left(h_{\boldsymbol{W}^{(T)},i}, y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(h_{\boldsymbol{W}^{(T)},i}, y_i\right) + \lambda O\left(\sqrt{\epsilon}\right)\\ &+ O\left(\Phi + \sqrt{\log(1/\delta)}\right)\left((1-\lambda)\sqrt{\frac{1}{n_z}} + \lambda\sqrt{\frac{1}{n_g''}}\right)\end{aligned} \tag{20}$$

With the same reason as in Eqn. (18), we have

$$\begin{aligned}&\frac{1-\lambda}{n_z}\sum_{S_z} r\left(h_{\boldsymbol{W}^{(T)},i}, y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(h_{\boldsymbol{W}^{(T)},i}, y_i\right)\\ \leq& O(\sqrt{\epsilon}) + \frac{1-\lambda}{n_z}\sum_{S_z} r\left(y_{\mathrm{eff},k(x_i)}, y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(y_{\mathrm{eff},k(x_i)}, y_i\right).\end{aligned} \tag{21}$$

Then, unlike in the proof of Theorem 5.1, we need to bound the risk in (21) in terms of expected label and knowledge imperfectness. Thus, replacing $h^*_{\mathrm{R},\beta}$ and h^*_{K} in Eqn. (19) with $\bar{h}^*_{\mathrm{R},\beta}$ and $\bar{h}^*_{\mathrm{K}}$, we have

$$\begin{aligned}&\frac{1-\lambda}{n_z}\sum_{S_z} r\left(y_{\mathrm{eff},k(x_i)} y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(y_{\mathrm{eff},k(x_i)}, y_i\right)\\ \leq& (1-\lambda)\widehat{Q}_{\mathrm{R},S_z,S_g'}(\beta) + \lambda\widehat{Q}_{\mathrm{K},S_g''} + \frac{1-\lambda}{n_z}\sum_{S_z}\left\|\bar{h}^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\| + \frac{\lambda}{n_g''}\sum_{S_g''}\left\|\bar{h}^*_{\mathrm{K},i} - y_{\mathrm{eff},k(x_i)}\right\|\end{aligned} \tag{22}$$

By Lemma B.8 (b), it holds that $\frac{1}{n_g''}\sum_{S_g''}\left\|\bar{h}^*_{\mathrm{K},i} - y_{\mathrm{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + O(\sqrt{\frac{\log(1/\delta)}{n_g''}})$ and $\frac{1-\beta}{n_z}\sum_{S_z}\left\|\bar{h}^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + O(\sqrt{\frac{\log(1/\delta)}{n_z}})$. Thus we have $\frac{\lambda}{n_g''}\sum_{S_g''}\left\|\bar{h}^*_{\mathrm{K},i} - y_{\mathrm{eff},k(x_i)}\right\| \leq \lambda\left(\widetilde{O}\left(L^{5/8}\phi^{1/4}\log^{1/8}(m)\right) + O\left((\frac{\log(1/\delta)}{n_g''})^{\frac{1}{4}}\right)\right) \leq \lambda\left(O(\epsilon) + O\left((\frac{\log(1/\delta)}{n_g''})^{\frac{1}{4}}\right)\right)$ and $\frac{1-\lambda}{n_z}\sum_{S_z}\left\|\bar{h}^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\| \leq \frac{1-\lambda}{\sqrt{1-\beta}}\left(\widetilde{O}\left(L^{5/8}\phi^{1/4}\log^{1/8}(m)\right) + O\left((\frac{\log(1/\delta)}{n_z})^{\frac{1}{4}}\right)\right) \leq (1-\lambda)\left(O(\epsilon) + O\left((\frac{\log(1/\delta)}{n_z})^{\frac{1}{4}}\right)\right)$. Therefore, continuing with (22), it holds that

$$\begin{aligned}&\frac{1-\lambda}{n_z}\sum_{S_z} r\left(y_{\mathrm{eff},k(x_i)} y_i\right) + \frac{\lambda}{n_g''}\sum_{S_g''} r\left(y_{\mathrm{eff},k(x_i)}, y_i\right)\\ \leq& (1-\lambda)\widehat{Q}_{\mathrm{R},S_z,S_g'}(\beta) + \lambda\widehat{Q}_{\mathrm{K},S_g''} + O(\sqrt{\epsilon}) + O\left((1-\lambda)\left(\frac{\log(1/\delta)}{n_z}\right)^{\frac{1}{4}} + \lambda\left(\frac{\log(1/\delta)}{n_g''}\right)^{\frac{1}{4}}\right)\\ \leq& O(\sqrt{\epsilon}) + (1-\lambda)Q_{\mathrm{R}}(\beta) + \lambda Q_{\mathrm{K}} + O\left((1-\lambda)\left(\frac{\log(1/\delta)}{n_z}\right)^{\frac{1}{4}} + \lambda\left(\frac{\log(1/\delta)}{n_g''}\right)^{\frac{1}{4}}\right),\end{aligned} \tag{23}$$

where the second inequality holds by Lemma B.8 and the last inequality holds by McDiarmid's inequality. Finally, substituting Eqns. (23) and (21) into Eqn. (20), with probability at least $1-O(\phi)-\delta, \delta\in(0,1)$, it holds that

$$R(\boldsymbol{W}^{(T)}) \leq O(\sqrt{\epsilon}) + (1-\lambda)Q_{\mathrm{R}}(\beta) + \lambda Q_{\mathrm{K}} + O\left(\Phi + \log^{1/4}(1/\delta)\right)\sqrt{\frac{1-\lambda}{\sqrt{n_z}} + \frac{\lambda}{\sqrt{n_g''}}}.$$

This completes the proof. □

B.5.4. Proof of Corollary 5.3

Proof. **Proof of (a).** If $Q_{\mathrm{K}} \leq \sqrt{\epsilon}$ and λ is set as 1, it holds by Corollary 5.2 that

$$\begin{aligned} R(\boldsymbol{W}^{(T)}) &\leq O(\sqrt{\epsilon}) + Q_{\mathrm{K}} + O\left(\Phi + \log^{1/4}(1/\delta)\right)\left(\frac{1}{n_g''}\right)^{1/4} \\ &\leq O(\sqrt{\epsilon}) + O\left(\frac{1}{n_g''}\right)^{1/4}, \end{aligned}$$

where in the last inequality we absorb the scales of the last term by O notation. Thus, $n_g'' \leq O(\frac{1}{\epsilon^2})$ guarantees that $R(\boldsymbol{W}^{(T)}) \leq \sqrt{\epsilon}$. In the proof of Theorem 5.1, we prove that the probability that a sample belongs to the region covered by the smooth sets containing at least one labeled sample is $P_{\mathcal{X}'} = O(n_z/N) = O(n_z\phi^b) = O(\sqrt{\epsilon})$. Thus we have $P_{\mathcal{X}''} = 1 - P_{\mathcal{X}'} = 1 - O(\sqrt{\epsilon})$, and so $n_g = \frac{n_g''}{P_{\mathcal{X}''}} = n_g''/(1-O(\epsilon)) \sim O(1/(\epsilon^2-\epsilon^3))$.

Proof of (b). If $Q_{\mathrm{K}} > \sqrt{\epsilon}$ and $\lambda = \frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}}$, then by Corollary 5.2, we have

$$\begin{aligned} R(\boldsymbol{W}^{(T)}) &\leq O(\sqrt{\epsilon}) + Q_{\mathrm{R}}(\beta^*) - \frac{Q_{\mathrm{R}}(\beta^*)}{Q_{\mathrm{K}}}\sqrt{\epsilon} + \sqrt{\epsilon} + O\left(\sqrt{(1-\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}})\frac{1}{\sqrt{n_z}} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}}\frac{1}{\sqrt{n_g''}}}\right) \\ &\leq O(\sqrt{\epsilon}) + O\left(\sqrt{(1-\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}})\frac{1}{\sqrt{n_z}} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}}\frac{1}{\sqrt{n_g''}}}\right), \end{aligned}$$

where the second inequality holds because $\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{R}}(\beta^*)} \geq 1$ such that $Q_{\mathrm{R}}(\beta^*) - \frac{Q_{\mathrm{R}}(\beta^*)}{Q_{\mathrm{K}}}\sqrt{\epsilon} \leq \sqrt{\epsilon}$. Then to guarantee $R(\boldsymbol{W}^{(T)}) \leq \sqrt{\epsilon}$, we require that $(1-\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}})\frac{1}{\sqrt{n_z}} \leq \epsilon$ and $\frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}}\frac{1}{\sqrt{n_g''}} \leq \epsilon$. Thus, we have $n_z \sim O\left((1/\epsilon - 1/(\sqrt{\epsilon}Q_{\mathrm{K}}))^2\right)$, $n_g'' \sim O(\frac{1}{\epsilon Q_{\mathrm{K}}^2})$ and $n_g = n_g''/(1-O(\epsilon)) \sim O(1/((\epsilon-\epsilon^2)Q_{\mathrm{K}}^2))$.

Proof of (c). We prove (c) by contradiction. If $R(\boldsymbol{W}^{(T)}) \leq \sqrt{\epsilon}$, we have $(1-\lambda)Q_{\mathrm{R}}(\beta^*) \leq \sqrt{\epsilon}$ and $\lambda Q_{\mathrm{K}} \leq \sqrt{\epsilon}$. Then $\frac{\sqrt{\epsilon}}{Q_{\mathrm{R}}(\beta^*)} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}} \geq 1-\lambda+\lambda = 1$. This is contradictory to the condition $\frac{\sqrt{\epsilon}}{Q_{\mathrm{R}}(\beta^*)} + \frac{\sqrt{\epsilon}}{Q_{\mathrm{K}}} \leq 1$. Thus completes the proof.

□

C. Proofs of Lemmas in Appendix B

We now show the proofs of lemmas in Appendix B, while the proofs of lemmas newly introduced in this section are deferred to Appendix D.

C.1. Proof of Lemma B.1

In this section, we prove the forward perturbation with respect to inputs. We first recall some important notations. For the smooth set $k \in [N]$, layer $l \in [L]$, let $h_{l,k} = h_l(x_k')$, $h_{l,i} = h_l(x_i)$ be the activated output of lth layer, and $f_{l,k} = \boldsymbol{W}_l h_{l-1}(x_k')$, $f_{l,i} = \boldsymbol{W}_l h_{l-1}(x_i)$ be the pre-activated output of lth layer for some weight $\boldsymbol{W} \in \mathcal{B}\left(\boldsymbol{W}^{(0)}, \tau\right)$. At initialization, denote $h_{l,k}^{(0)} = h_l^{(0)}(x_k')$, $h_{l,i}^{(0)} = h_l^{(0)}(x_i)$, $f_{l,k}^{(0)} = \boldsymbol{W}_l^{(0)} h_{l-1}^{(0)}(x_k')$, $f_{l,i}^{(0)} = \boldsymbol{W}_l^{(0)} h_{l-1}^{(0)}(x_i)$, the diagonal matrices $\boldsymbol{D}_{l,k}^{(0)} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{D}_{l,i}^{(0)} \in \mathbb{R}^{m\times m}$ with $\left[\boldsymbol{D}_{l,k}^{(0)}\right]_{j,j} = \mathbb{1}([f_{l,k}^{(0)}]_j \geq 0)$ and $\left[\boldsymbol{D}_{l,i}^{(0)}\right]_{j,j} = \mathbb{1}([f_{l,i}^{(0)}]_j \geq 0)$ for $i \in \mathcal{I}_{\phi,k}, j \in [m]$. Then we denote for initialization $f_{l,i}' = f_{l,i}^{(0)} - f_{l,k}^{(0)}$ and the diagonal matrix $\boldsymbol{D}_l' \in \mathbb{R}^{m\times m}$ with $[\boldsymbol{D}_l']_{j,j} = \left[\boldsymbol{D}_{l,k}^{(0)}\right]_{j,j} - \left[\boldsymbol{D}_{l,i}^{(0)}\right]_{j,j}$, omitting the notation (0) and i,k.

Lemma C.1. *If $f'_{l,i}$ can be written as $f'_{l,i} = f'_{l,i,1} + f'_{l,i,2}$ with $\|f'_{l,i,1}\| \leq O(L^{3/2}\phi\log^{1/2}(1/\phi))$ and $\|f'_{l,i,2}\|_\infty \leq O(L\phi^{2/3}\log^{1/2}(1/\phi)m^{-1/2})$, then with probability at least $1-\exp(-\Omega(m\phi^{2/3}L))$ over the randomness of $\boldsymbol{W}^{(0)}$, we have*

$$\|\boldsymbol{D}'_l f^{(0)}_{l,i}\|_0 \leq \|\boldsymbol{D}'_l\|_0 \leq O(m\phi^{2/3}L\log^{1/2}(1/\phi)),$$

$$\|\boldsymbol{D}'_l f^{(0)}_{l,i}\| \leq O(\phi L^{3/2}\log^{1/2}(1/\phi)).$$

Proof Lemma C.1 is given in Section D.1.

Proof of Lemma B.1

Proof. We first prove the following three conclusions by induction under the assumptions in Lemma B.1: for $i \in \mathcal{I}_{\phi,k}, k \in [N]$, with probability at least $1-O(\phi)$,
(a)$f'_{l,i}$ at initialization can be written as $f'_{l,i,1} + f'_{l,i,2}$ with $\|f'_{l,i,1}\| \leq O(L^{3/2}\phi\log^{1/2}(1/\phi))$ and $\|f'_{l,i,2}\|_\infty \leq O(L\phi^{2/3}m^{-1/2}\log^{1/2}(1/\phi))$.
(b) At initialization, $\|\boldsymbol{D}'_l f^{(0)}_{l,i}\|_0 \leq \|\boldsymbol{D}'_l\|_0 \leq O(m\phi^{2/3}L\log^{1/2}(1/\phi))$, $\|\boldsymbol{D}'_l f^{(0)}_{l,i}\| \leq O(\phi L^{3/2}\log^{1/2}(1/\phi))$.
(c)$\|h^{(0)}_{l,i} - h^{(0)}_{l,k}\| \leq O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)})$ and $\|f^{(0)}_{l,i} - f^{(0)}_{l,k}\| \leq O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)})$.

When $l=0$, we have $\|h^{(0)}_{0,i} - h^{(0)}_{0,k}\| = \|x_i - x_k\| \leq O(\phi)$. Since $\left[\boldsymbol{W}^{(0)}_1\right]_{j,j} \sim \mathcal{N}\left(0, \frac{2}{m}\right), j \in [m]$, we have $\|f^{(0)}_{1,i} - f^{(0)}_{1,k}\| \leq O(\phi\log^{1/2}(1/\phi))$ with probability at least $1-O(\phi)$ over the randomness of $\boldsymbol{W}^{(0)}_1$. By Lemma C.1, the above three conclusions hold. Then we assume the conclusions (a) holds for layer $a, a \leq l-1$ and prove (a)(b)(c) hold for l.

First, we re-write $f'_{l,i}$ as

$$\begin{aligned}
&f'_{l,i} = f^{(0)}_{l,i} - f^{(0)}_{l,k} = \boldsymbol{W}^{(0)}_l\left(\boldsymbol{D}^{(0)}_{l-1,k} + \boldsymbol{D}'_{l-1}\right)\left(f^{(0)}_{l-1,k} + f'_{l-1,i}\right) - \boldsymbol{W}^{(0)}_l\boldsymbol{D}^{(0)}_{l-1,k}f^{(0)}_{l-1,k} \\
=&\boldsymbol{W}^{(0)}_l\boldsymbol{D}'_{l-1}\left(f^{(0)}_{l-1,k} + f'_{l-1,i}\right) + \boldsymbol{W}^{(0)}_l\boldsymbol{D}^{(0)}_{l-1,k}f'_{l-1,i} \\
=&\cdots \\
=&\sum_{a=2}^{l}\left(\prod_{b=a+1}^{l}\boldsymbol{W}^{(0)}_b\boldsymbol{D}^{(0)}_{b-1,k}\right)\boldsymbol{W}^{(0)}_a\boldsymbol{D}'_{a-1}\left(f^{(0)}_{a-1,k} + f'_{a-1,i}\right) + \boldsymbol{W}^{(0)}_1(x_i - x'_k)
\end{aligned}$$

By Lemma C.1, and the inductive assumption (a) for layer $a, a \leq l-1$, we have with probability at least $1-\exp(-\Omega\left(m\phi^{2/3}L\right))$,

$$\|\boldsymbol{D}'_a\left(f^{(0)}_{a,k} + f'_{a,i}\right)\|_0 \leq O(m\phi^{2/3}L\log^{1/2}(1/\phi)), \tag{24}$$

$$\|\boldsymbol{D}'_a\left(f^{(0)}_{a,k} + f'_{a,i}\right)\| \leq O(\phi L^{3/2}\log^{1/2}(1/\phi)), \tag{25}$$

so (b) holds for layer l. Then let $q_a = \left(\prod_{b=a+1}^{l}\boldsymbol{W}^{(0)}_b\boldsymbol{D}^{(0)}_{b-1,k}\right)\boldsymbol{W}^{(0)}_a\boldsymbol{D}'_{a-1}\left(f^{(0)}_{a-1,k} + f'_{a-1,i}\right)$. By Eqn.(24), (25), and Claim 8.5 ($s = O(m\phi^{2/3}L)$) in (Allen-Zhu et al., 2019b), with probability at least $1-\exp(-\Omega\left(m\phi^{2/3}L\log(m)\right))$, we can write $q_a = q_{a,1} + q_{a,2}$ with

$$\|q_{a,1}\| \leq O(\phi^{4/3}L^2\log(m)\log^{3/4}(1/\phi)) \quad \text{and} \quad \|q_{a,2}\|_\infty \leq O(\phi L^{3/2}\sqrt{\log(m)/m}\log^{1/2}(1/\phi)). \tag{26}$$

Let $f'_{l,i,1} = \sum_{a=2}^{l} q_{a,1} + \boldsymbol{W}^{(0)}_1(x_i - x'_k)$ and $f'_{l,i,2} = \sum_{a=2}^{l} q_{a,2}$. Then we have $f'_{l,i} = f'_{l,i,1} + f'_{l,i,2}$. Since $\|\boldsymbol{W}^{(0)}_1(x_i - x'_k)\| \leq O(\phi\sqrt{\log(1/\phi)})$ with probability at least $1-\phi, \phi \in (0,1)$, by triangle inequality, we can write

$$\begin{aligned}
&\|f^{(0)}_{l,i} - f^{(0)}_{l,k}\| = \|f'_{l,i}\| = \|f'_{l,i,1} + f'_{l,i,2}\| \\
\leq&O(\phi^{4/3}L^3\log(m)\log^{3/4}(1/\phi) + \phi\sqrt{\log(1/\phi)}) + O(\phi L^{5/2}\sqrt{\log(m)}\log^{1/2}(1/\phi)) \\
\leq&O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)}),
\end{aligned}$$

where the first inequality comes from inequalities 26, and the last inequality holds by the assumption $\phi \leq O(L^{-9/2}\log^{-3}(m)\log^{-3/4}(1/\phi))$. Also, by the requirement of ϕ, we have with $\|f'_{l,i,1}\| \leq O(\phi^{4/3}L^3\log(m)\log^{3/4}(1/\phi) + \phi\sqrt{\log(1/\phi)}) \leq O(L^{3/2}\phi\log^{1/2}(1/\phi))$ and $\|f'_{l,i,2}\|_\infty \leq O(\phi L^{5/2}\sqrt{\log(m)/m}\log^{1/2}(1/\phi)) \leq O(L\phi^{2/3}m^{-1/2}\log^{1/2}(1/\phi))$. Thus (a) holds for layer l. And by Lemma C.1, we have with probability at least $1-\phi$,

$$\begin{aligned}\|h_{l,i}^{(0)} - h_{l,k}^{(0)}\| &= \|(\boldsymbol{D}_{l,k}^{(0)} + \boldsymbol{D}_l')\left(f_{l,k}^{(0)} + f_{l,i}'\right) - \boldsymbol{D}_{l,k}^{(0)} f_{l,k}^{(0)}\| \\ &\leq \|\boldsymbol{D}_l' f_{l,k}^{(0)}\| + \|(\boldsymbol{D}_{l,k}^{(0)} + \boldsymbol{D}_l')f_{l,i}'\| \leq O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)}),\end{aligned}$$

where the second inequality comes from Lemma C.1. Thus, conclusion (c) holds for layer l.

Finally, by Lemma 8.2 in (Allen-Zhu et al., 2019b) which gives forward perturbation regarding weights, we have with probability at least $1 - O(\phi)$,

$$\begin{aligned}\|f_{l,i} - f_{l,k}\| &\leq \|f_{l,i}^{(0)} - f_{l,k}^{(0)}\| + \|f_{l,i} - f_{l,i}^{(0)}\| + \|f_{l,k} - f_{l,k}^{(0)}\| \\ &\leq O(L^{3/2}\phi\log^{1/2}(1/\phi)) + O(\tau L^{5/2}\sqrt{\log(m)}) \leq O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)}),\end{aligned}$$

where the last probability holds by the assumption $\tau \leq O(\phi^{3/2})$. Similarly, we have with probability at least $1 - O(\phi)$,

$$\begin{aligned}\|h_{l,i} - h_{l,k}\| &\leq \|h_{l,i}^{(0)} - h_{l,k}^{(0)}\| + \|h_{l,i} - h_{l,i}^{(0)}\| + \|h_{l,k} - h_{l,k}^{(0)}\| \\ &\leq O(L^{3/2}\phi\log^{1/2}(1/\phi)) + O(\tau L^{5/2}\sqrt{\log(m)}) \leq O(L^{5/2}\phi\sqrt{\log(m)\log(1/\phi)}).\end{aligned}$$

□

C.2. Proofs of Lemma B.2 and Lemma B.3

Proof of Lemma B.2

Proof. By the mean value theorem, r can be represented as

$$r(h') = r(h) + \bigtriangledown r(h)^\top (h'-h) + \frac{1}{2}(h'-h)^\top \bigtriangledown^2 r(z)(h'-h), \tag{27}$$

where z lies in the line segment between h' and h.

Since the maximum eigenvalue of the Hessian matrix of r is bounded by 1, for any output of the neural network h and h' , we have

$$r(h') \leq r(h) + \bigtriangledown r(h)^\top (h'-h) + \frac{1}{2}\|h'-h\|_2^2 \tag{28}$$

Let $h' = h - \bigtriangledown r(h)$. We have

$$r_{\min} \leq r(h') \leq r(h) - \frac{1}{2}\| \bigtriangledown r(h)\|_2^2. \tag{29}$$

Thus, we get the first inequality of the lemma $\|\bigtriangledown r(h)\|^2 \leq 2(r(h) - r_{\min})$.

By strong convexity, for any h and h' in the domain of risk function r, we have

$$\begin{aligned}r(h') &\geq r(h) + \bigtriangledown r(h)^\top (h'-h) + \frac{\rho}{2}\|h'-h\|^2 \\ &\geq r(h) - \frac{1}{2\rho}\| \bigtriangledown r(h)\|^2,\end{aligned} \tag{30}$$

where the first inequality comes from strong convexity and the second inequality holds by choosing $h' = -\frac{\bigtriangledown r(h)}{\rho}$ that minimizes the right hand side. Then letting h' in the left hand side equals to h^* such that $r(h^*) = r_{\min}$, we get the second inequality of the lemma $\|\bigtriangledown r(h)\|^2 \geq 2\rho(r(h) - r_{\min})$.

Also, letting h' in Eqn. (30) be h^*, we have

$$r_{\min} = r\left(h^*\right) \geq r\left(h\right) - \left\|\bigtriangledown r\left(h\right)\right\| \left\|\left(h^* - h\right)\right\| + \frac{\rho}{2}\|h^* - h\|^2 \tag{31}$$

By the fact that $r_{\min} \leq r\left(h\right)$, we have

$$\|h^* - h\| \leq \frac{2}{\rho}\left\|\bigtriangledown r\left(h\right)\right\|. \tag{32}$$

We thus get the third inequality. □

Proof of Lemma B.3

Proof. Denote the risk of the kth cell $k \in [N]$ with respect to the input $x'_k \in \mathcal{X}_\phi$ for hypothesis $h \in \mathcal{H}$ as

$$\bar{r}^{\circ}_{\mathrm{I},k} = \sum_{i \in \mathcal{I}_{\phi,k}} r_{\mathrm{I},i}(h(x'_k)).$$

Recall that $M_k = \sum_{i\in\mathcal{I}_{\phi,k}}(\mu_i+\lambda_i)$. By 1-Lipschitz continuity of risk functions and their gradients, we have with probability at least $1 - O(\phi)$,

$$\left|\bar{r}_{\mathrm{I},k} - \bar{r}^{\circ}_{\mathrm{I},k}\right| \leq \sum_{i\in\mathcal{I}_{\phi,k}} \left(\mu_i + \lambda_i\right)\|h(x_i) - h(x'_k)\| \leq \widetilde{O}(M_k L^{5/2}\phi\log^{1/2}(m)),$$

$$\text{and } \left\|\bigtriangledown_h \bar{r}_{\mathrm{I},k} - \bigtriangledown_h \bar{r}^{\circ}_{\mathrm{I},k}\right\| \leq \sum_{i\in\mathcal{I}_{\phi,k}} \left(\mu_i + \lambda_i\right)\|h(x_i) - h(x'_k)\| \leq \widetilde{O}(M_k L^{5/2}\phi\log^{1/2}(m)).$$

Since the eigenvalues of $\bigtriangledown_h^2 \bar{r}^{\circ}_{\mathrm{I},k}$ is no less than $M_k\rho$ and $r^*_{\mathrm{eff},k}$ is the minimum value of $\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{I},i}(h)$, by Lemma B.2, we have

$$\|\bigtriangledown_h \bar{r}^{\circ}_{\mathrm{I},k}\|^2 \geq 2M_k\rho\left(\bar{r}^{\circ}_{\mathrm{I},k} - r^*_{\mathrm{eff},k}\right) \geq 2M_k\rho\left(\bar{r}_{\mathrm{I},k} - r^*_{\mathrm{eff},k}\right) - \widetilde{O}(M_k^2\rho L^{5/2}\phi\log^{1/2}(m)).$$

Therefore, we have

$$\begin{aligned}\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i(h(x_i))\|^2 = \|\bigtriangledown_h \bar{r}_{\mathrm{I},k}\|^2 &\geq \|\bigtriangledown_h \bar{r}^{\circ}_{\mathrm{I},k}\|^2 - \widetilde{O}(M_k^2\, L^{5/2}\phi\log^{1/2}(m)) \\ &\geq 2M_k\rho\left(\bar{r}_{\mathrm{I},k} - r^*_{\mathrm{eff},k}\right) - \widetilde{O}(M_k^2 L^{5/2}\phi\log^{1/2}(m))\end{aligned}$$

Also, since the eigenvalues of $\bigtriangledown_h^2 \bar{r}^{\circ}_{\mathrm{I},k}$ is no larger than M_k, we have

$$\|\bigtriangledown_h \bar{r}^{\circ}_{\mathrm{I},k}\|^2 \leq 2M_k\left(\bar{r}^{\circ}_{\mathrm{I},k} - r_{\mathrm{eff},k}\right) \leq 2M_k\left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k}\right) + \widetilde{O}(M_k^2 L^{5/2}\phi\log^{1/2}(m)).$$

Therefore, it holds that

$$\begin{aligned}\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i(h(x_i))\|^2 = \|\bigtriangledown_h \bar{r}_{\mathrm{I},k}\|^2 &\leq \|\bigtriangledown_h \bar{r}^{\circ}_{\mathrm{I},k}\|^2 + \widetilde{O}(M_k^2\rho L^{5/2}\phi\log^{1/2}(m)) \\ &\leq 2M_k\left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k}\right) + \widetilde{O}(M_k^2 L^{5/2}\phi\log^{1/2}(m)).\end{aligned}$$

Applying Lemma B.2 for $\bar{r}^{\circ}_{\mathrm{I},k}$, we have

$$\left\|h\left(x'_k\right) - y_{\mathrm{eff},k}\right\|^2 \leq \frac{4}{M_k^2\rho^2}\left\|\bigtriangledown_h \bar{r}^{\circ}_{\mathrm{I},k}\right\|^2 \leq \frac{1}{M_k\rho^2} O\left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k} + \widetilde{O}(M_k L^{5/2}\phi\log^{1/2}(m))\right).$$

By applying Lemma B.1 to $h(x_i)$, we have

$$\begin{aligned}&\left\|h\left(x_i\right) - y_{\mathrm{eff},k}\right\|^2 \leq 2\left\|h\left(x_i\right) - h\left(x'_k\right)\right\|^2 + 2\left\|h\left(x'_k\right) - y_{\mathrm{eff},k}\right\|^2 \\ \leq &\frac{1}{M_k\rho^2} O\left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k} + \widetilde{O}(M_k L^{5/2}\phi\log^{1/2}(m))\right) + \widetilde{O}(L^5\phi^2\log(m)).\end{aligned}$$

Taking weighted summation in the cell $\mathcal{I}_{\phi,k}$, since $\sum_{i\in\mathcal{I}_{\phi,k}}(\mu_i + \lambda_i) = M_k$, we have

$$\sum_{i\in\mathcal{I}_{\phi,k}} \left(\mu_i + \lambda_i\right)\left\|h\left(x_i\right) - y_{\mathrm{eff},k}\right\|^2 \leq \frac{1}{\rho^2} O\left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k} + \widetilde{O}(M_k L^{5/2}\phi\log^{1/2}(m))\right).$$

□

C.3. Proof of Lemma B.4

Lemma C.2. *Suppose that* $m \geq \Omega(N^2 d^2 \phi^{-1})$. *For any* $u_i : \|u_i\| \leq \mu_i + \lambda_i, i \in [n']$ *and* $v_j \sim \mathcal{N}(0, (1/d)\mathbf{I})$, $w_j \sim \mathcal{N}(0, (2/m)\mathbf{I})$, *with Assumption 1 satisfied, with probability at least* $1 - O(\phi)$, *we have*

$$\sum_{j=1}^{m} \| \sum_{k=1}^{N} \sum_{i \in \mathcal{I}_{\phi,k}} \langle u_i, v_j \rangle \sigma' \left(\left\langle w_j, h_{L-1}^{(0)}(x_i) \right\rangle \right) h_{L-1}^{(0)}(x_i) \|^2 \geq \Omega\left(\frac{\alpha m \phi}{Nd}\right) \left(\sum_{k=1}^{N} \| \sum_{i \in \mathcal{I}_{\phi,k}} u_i \|^2 - \widetilde{O}(L^{5/2} \phi \log^{1/2}(m)) \right).$$

Lemma C.2 is proved in Section D.2

Lemma C.3 (Lemma 8.7, Lemma 8.2c in (Allen-Zhu et al., 2019b))**.** *For any* $\boldsymbol{W} \in \mathcal{B}\left(\boldsymbol{W}^{(0)}, \tau\right)$, *with probability at least* $1 - \exp\left(-O\left(m\tau^{2/3} L\right)\right)$,

$$\left\| \boldsymbol{V} \boldsymbol{D}_{i,L} - \boldsymbol{V} \boldsymbol{D}_{i,L}^{(0)} \right\|_2 \leq O\left(\tau^{1/3} L^2 \sqrt{m \log(m)/d}\right),$$

and $\forall l \in [L]$,

$$\left\| h_{i,l} - h_{i,l}^{(0)} \right\| \leq O\left(\tau L^{5/2} \sqrt{\log(m)}\right).$$

Proof of Lemma B.4

Proof. Denote $u_i = u_i(h_{\boldsymbol{W}}(x_i))$. The gradient of the empirical informed risk can be expressed as

$$\bigtriangledown_{W_l} \hat{R}_{\mathrm{I}}(\boldsymbol{W}) = \sum_{i=1}^{n'} \left(u_i \boldsymbol{V} \boldsymbol{D}_{L,i} \boldsymbol{W}_{L,i} \cdots \boldsymbol{W}_{l+1,i} \boldsymbol{D}_l\right)^\top h_{l-1,i}. \tag{33}$$

Let $\boldsymbol{G} = \bigtriangledown_{W_L} \hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(0)}\right) = \sum_{k=1}^{N} \sum_{i \in \mathcal{I}_{\phi,k}} \left(u_i \boldsymbol{V} \boldsymbol{D}_{L,i}^{(0)}\right)^\top h_{L-1,i}^{(0),\top}$. By Lemma C.2, with probability at least $1 - O(\phi)$, we have

$$\begin{aligned}
\|\boldsymbol{G}\|_F^2 \geq & \Omega\left(\frac{\alpha m \phi}{Nd}\right) \left(\sum_{k=1}^{N} \| \sum_{i \in \mathcal{I}_{\phi,k}} u_i \|^2 - \widetilde{O}(L^{5/2} \phi \log^{1/2}(m)) \right) \\
\geq & \Omega\left(\frac{\alpha m \phi \rho}{dN}\right) \left(\sum_{k=1}^{N} M_k \left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k}\right) - \widetilde{O}(L^{5/2} \phi \log^{1/2}(m)) \right) \\
\geq & \Omega\left(\frac{\alpha m \phi \rho \bar{\lambda}}{dN^2}\right) \left(\hat{R}_{\mathrm{I}} - \hat{R}_{\mathrm{eff}} - \widetilde{O}(L^{5/2} \phi \log^{1/2}(m)) \right),
\end{aligned} \tag{34}$$

where $\hat{R}_{\mathrm{eff}} = \sum_{k=1}^{N} r_{\mathrm{eff},k}$ and the second inequality comes from Lemma B.3 and the last inequality holds because $\sum_{k=1}^{N} M_k \left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k}\right) \geq \bar{M} \sum_{k=1}^{N} \left(\bar{r}_{\mathrm{I},k} - r_{\mathrm{eff},k}\right)$ with $\bar{M} = \min_k M_k$, $M_k = \sum_{\mathcal{I}_{\phi,k}} (\mu_i + \lambda_i)$, and $N\bar{M} = \bar{\lambda}$.

Here, we need to discuss more about $\bar{\lambda}$ which is different for different objectives. Denote $\bar{p}_z = \min_k |S_z \cap S_{\mathcal{I}_{\phi,k}}|$, $\bar{p}_g = \min_k |S_g \cap S_{\mathcal{I}_{\phi,k}}|$ and $\bar{p}_{g/z} = \min_k |(S_g \setminus S_z) \cap S_{\mathcal{I}_{\phi,k}}|$. When $\lambda \neq 1$ or $\lambda \neq 0$, $\bar{\lambda} = N \min\left\{ \frac{(1-\lambda)\bar{p}_z}{n_z} + \frac{\lambda \bar{p}_g}{n_g}, \frac{\lambda \bar{p}_g}{n_g} \right\} \geq \Omega(\lambda)$ for objective (3),$\bar{\lambda} = N \min\left\{ \frac{(1-\lambda)(1-\beta)\bar{p}_z}{n_z} + \frac{(1-\lambda)\beta \bar{p}_g}{n_g}, \frac{\lambda \bar{p}_g}{n'_g} \right\} \geq \Omega(\min(1-\lambda, \lambda))$ for objective (5)[1]. Beside, the cases when $\lambda = 0$ or $\lambda = 1$ mean the corresponding datasets are empty (e.g. when $\lambda = 1$ in (5), $S_z = \emptyset$ and $S'_g = \emptyset$), so we have $\bar{\lambda} = 1$. In conclusion, we have $\bar{\lambda} = \Omega(\min(1-\lambda, \lambda)\mathbb{1}(\lambda \in (0,1)) + \mathbb{1}(\lambda \in \{0,1\}))$ for two objectives.

[1]Here, $S_{\mathcal{I}_{\phi,k}}$ is the set of samples with their indices in $\mathcal{I}_{\phi,k}$. Thus, there exists a constant C such that $n_z \leq CN\bar{p}_z$, $n_g \leq CN\bar{p}_g$, $(n_g - n_z) \leq CN\bar{p}_{g/z}$ where C relies on the input distribution.

Next we bound the difference of $\|\boldsymbol{G}\|$ and $\| \bigtriangledown_{W_L} \hat{R}_{\mathrm{I}}(\boldsymbol{W})\|$ with $\boldsymbol{W} \in \mathcal{B}(\boldsymbol{W}^{(0)}, \tau)$. By definition, we have

$$\begin{aligned}
&\left\|\boldsymbol{G} - \bigtriangledown_{W_L} \hat{R}_{\mathrm{I}}(\boldsymbol{W})\right\|_F = \left\|\sum_{k=1}^{N}\sum_{i\in\mathcal{I}_{\phi,k}} \left(u_i \boldsymbol{V}\boldsymbol{D}_{L,i}^{(0)}\right)^\top h_{L-1,i}^{(0),\top} - \sum_{k=1}^{N}\sum_{i\in\mathcal{I}_{\phi,k}} \left(u_i \boldsymbol{V}\boldsymbol{D}_{L,i}\right)^\top h_{L-1,i}^\top\right\|_F \\
\leq &\underbrace{\left\|\sum_{k=1}^{N}\sum_{i\in\mathcal{I}_{\phi,k}} \left(u_i \boldsymbol{V}\boldsymbol{D}_{L,i}^{(0)} - u_i \boldsymbol{V}\boldsymbol{D}_{L,i}\right)^\top h_{L-1,i}^{(0),\top}\right\|_F}_{(a)} + \underbrace{\left\|\sum_{k=1}^{N}\sum_{i\in\mathcal{I}_{\phi,k}} \left(u_i \boldsymbol{V}\boldsymbol{D}_{L,i}\right)^\top \left(h_{L-1,i}^{(0)} - h_{L-1,i}\right)^\top\right\|_F}_{(b)}
\end{aligned} \tag{35}$$

For the term (a) in the above inequality, denoting $h_{L-1,k} = h_{L-1}(x_k')$ and letting $(a)_k$ be the kth item in the summation, we have

$$\begin{aligned}
(a)_k \leq &\left\|\sum_{i\in\mathcal{I}_{\phi,k}} \left(\boldsymbol{D}_{L,k}^{(0)} - \boldsymbol{D}_{L,k}\right) \boldsymbol{V}^\top u_i^\top h_{L-1,k}^{(0),\top}\right\|_F \\
&+ \left\|\sum_{i\in\mathcal{I}_{\phi,k}} \left(\boldsymbol{D}_{L,i}^{(0)} - \boldsymbol{D}_{L,i}\right) \boldsymbol{V}^\top u_i^\top h_{L-1,i}^{(0),\top} - \sum_{i\in\mathcal{I}_{\phi,k}} \left(\boldsymbol{D}_{L,k}^{(0)} - \boldsymbol{D}_{L,k}\right) \boldsymbol{V}^\top u_i^\top h_{L-1,k}^{(0),\top}\right\|_F \\
\leq &O\left(\tau^{1/3} L^2 \sqrt{m\log(m)/d}\right) \left\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\right\| + \sum_{i\in\mathcal{I}_{\phi,k}} \left\|\boldsymbol{V}\left(\boldsymbol{D}_{L,i}^{(0)} - \boldsymbol{D}_{L,i}\right)\right\|_2 \|u_i\| \left\|h_{L-1,i}^{(0)}\right\| \\
&+ \sum_{i\in\mathcal{I}_{\phi,k}} \left\|\boldsymbol{V}\left(\boldsymbol{D}_{L,k}^{(0)} - \boldsymbol{D}_{L,k}\right)\right\|_2 \|u_i\| \left\|h_{L-1,k}^{(0)}\right\| \\
\leq &O\left(\tau^{1/3} L^2 \sqrt{m\log(m)/d}\right) \left(\left\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\right\| + M_k\right),
\end{aligned}$$

where the second inequality comes from Lemma C.3 and Cauchy-Schwartz inequality, and the last inequality comes from Lemma C.3 and $\sum_{i\in\mathcal{I}_{\phi,k}} \|u_i\| \leq M_k$ and Lemma B.6 such that $\left\|h_{L-1,i}^{(0)}\right\| \leq \tilde{O}(1)$ with probability at least $1 - O(\phi)$.

For the term (b), it holds that

$$\begin{aligned}
(b)_k \leq &\left\|\sum_{i\in\mathcal{I}_{\phi,k}} \left(\boldsymbol{V}\boldsymbol{D}_{L,k}\right)^\top u_i^\top \left(h_{L-1,k}^{(0)} - h_{L-1,k}\right)^\top\right\|_F \\
&+ \left\|\sum_{i\in\mathcal{I}_{\phi,k}} \left(\boldsymbol{V}\boldsymbol{D}_{L,i}\right)^\top u_i^\top \left(h_{L-1,i}^{(0)} - h_{L-1,i}\right)^\top - \sum_{i\in\mathcal{I}_{\phi,k}} \left(\boldsymbol{V}\boldsymbol{D}_{L,k}\right)^\top u_i^\top \left(h_{L-1,k}^{(0)} - h_{L-1,k}\right)^\top\right\|_F \\
\leq &O\left(\tau L^{5/2} \sqrt{m\log(m)/d}\right) \left\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\right\| + \sum_{i\in\mathcal{I}_{\phi,k}} \|\boldsymbol{V}\boldsymbol{D}_{L,i}\|_2 \|u_i\| \left\|h_{L-1,i}^{(0)} - h_{L-1,i}\right\| \\
&+ \sum_{i\in\mathcal{I}_{\phi,k}} \|\boldsymbol{V}\boldsymbol{D}_{L,k}\|_2 \|u_i\| \left\|h_{L-1,k}^{(0)} - h_{L-1,k}\right\| \\
\leq &O\left(\tau L^{5/2} \sqrt{m\log(m)/d}\right) \left(\left\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\right\| + M_k\right),
\end{aligned}$$

where the second inequality comes from Lemma C.3 and Cauchy-Schwartz inequality, and the last inequality comes from Lemma C.3 and $\sum_{i\in\mathcal{I}_{\phi,k}} \|u_i\| \leq M_k$ and Lemma B.6 such that $\left\|h_{L-1,i}^{(0)}\right\| \leq \tilde{O}(1)$ with probability at least $1 - O(\phi)$.

Therefore, we can bound Eqn. (35) as

$$\begin{aligned}
&\left\|\boldsymbol{G}-\bigtriangledown_{W_L}\hat{R}_{\mathrm{I}}(\boldsymbol{W})\right\|_F \leq O\left(\tau^{1/3}L^{5/2}\sqrt{m\log(m)/d}\right)\left(\sum_{k=1}^{N}\left\|\sum_{i\in\mathcal{I}_{\phi,k}}u_i\right\|+1\right)\\
\leq& O\left(N^{1/2}\tau^{1/3}L^{5/2}\sqrt{m\log(m)/d}\right)\left(\sqrt{\left(\hat{R}_{\mathrm{I}}-\hat{R}_{\text{eff}}+\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)}+1/\sqrt{N}\right)\\
\leq& O\left(N^{1/2}\tau^{1/3}L^{5/2}\sqrt{m\log(m)/d}\right)\left(\sqrt{\left(\hat{R}_{\mathrm{I}}-\hat{R}_{\text{eff}}-\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)}\right),
\end{aligned}\tag{36}$$

where the second inequality holds by Lemma B.3 and the last inequality holds because ϕ is small enough such that $\hat{R}_{\mathrm{I}}(\boldsymbol{W})-\hat{R}_{\text{eff}}+\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\leq 2(\hat{R}_{\mathrm{I}}(\boldsymbol{W})-\hat{R}_{\text{eff}}-\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)))$.

Combining Eqn. (36) with Eqn. (34), we have

$$\begin{aligned}
&\left\|\bigtriangledown_{W_L}\hat{R}_{\mathrm{I}}(\boldsymbol{W})\right\|_F \geq \|\boldsymbol{G}\|_F-\left\|\boldsymbol{G}-\bigtriangledown_{W_L}\hat{R}_{\mathrm{I}}(\boldsymbol{W})\right\|_F\\
\geq&\Omega\left(\sqrt{\frac{\alpha m\phi\rho\bar{\lambda}}{dN^2}}-O\left(\tau^{1/3}N^{1/2}L^{5/2}\sqrt{m\log(m)/d}\right)\right)\left(\hat{R}_{\mathrm{I}}-\hat{R}_{\text{eff}}-\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)\\
\geq&\Omega\left(\sqrt{\frac{\alpha m\phi\rho\bar{\lambda}}{dN^2}}\left(\hat{R}_{\mathrm{I}}-\hat{R}_{\text{eff}}-\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)\right),
\end{aligned}$$

where the last inequality holds by the choice of $m\geq\Omega\left(N^{11}L^{15}d\phi^{-4}\rho^{-4}\bar{\lambda}^{-4}\alpha^{-4}\log^3(m)\right)$ and the weight update range in the proof of Theorem 4.1 such that $\tau^{1/3}=O(N^{-3/2}\phi^{1/2}\rho^{1/2}\bar{\lambda}^{1/2}\alpha^{1/2}L^{-5/2}\log^{-1/2}(m))$.

$$\left\|\bigtriangledown_{\boldsymbol{W}}\hat{R}_{\mathrm{I}}(\boldsymbol{W})\right\|_F^2\geq\left\|\bigtriangledown_{W_L}\hat{R}_{\mathrm{I}}(\boldsymbol{W})\right\|_F^2\geq\Omega\left(\frac{\alpha m\phi\rho\bar{\lambda}}{dN^2}\right)\left(\hat{R}_{\mathrm{I}}-\hat{R}_{\text{eff}}-\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right).$$

□

C.4. Proof of Lemma B.5

Proof of Lemma B.5

Proof. Since the maximum eigenvalue of the second order derivation of the informed risk function $r_{\mathrm{I},i}(h_{\boldsymbol{W},i})$ with respect to h is less than $\mu_i+\lambda_i$, we have

$$r_{\mathrm{I},i}(h_{\boldsymbol{W}',i})-r_{\mathrm{I},i}(h_{\boldsymbol{W},i})\leq u_i(h_{\boldsymbol{W},i})^\top(h_{\boldsymbol{W}',i}-h_{\boldsymbol{W},i})+O\left(\frac{\mu_i+\lambda_i}{2}\|h_{\boldsymbol{W}',i}-h_{\boldsymbol{W},i}\|^2\right).\tag{37}$$

Then denote $\widehat{\boldsymbol{W}}=\boldsymbol{W}'-\boldsymbol{W}$. We have

$$\begin{aligned}
&\sum_{i\in\mathcal{I}_{\phi,k}}r_{\mathrm{I},i}(h_{\boldsymbol{W}',i})-r_{\mathrm{I},i}(h_{\boldsymbol{W},i})-\left\langle\bigtriangledown_{\boldsymbol{W}}r_{\mathrm{I},i}(h_{\boldsymbol{W},i}),\widehat{\boldsymbol{W}}\right\rangle\\
\leq&\sum_{i\in\mathcal{I}_{\phi,k}}u_i(h_{\boldsymbol{W},i})^\top\left(h_{\boldsymbol{W}',i}-h_{\boldsymbol{W},i}-\left\langle\bigtriangledown_{\boldsymbol{W}}h_{\boldsymbol{W},i},\widehat{\boldsymbol{W}}\right\rangle\right)+O\left(\sum_{i\in\mathcal{I}_{\phi,k}}\frac{\mu_i+\lambda_i}{2}\|h_{\boldsymbol{W}',i}-h_{\boldsymbol{W},i}\|^2\right)\\
\leq&\left(\sum_{i\in\mathcal{I}_{\phi,k}}u_i(h_{\boldsymbol{W},i})\right)^\top\left(h_{\boldsymbol{W}',k}-h_{\boldsymbol{W},k}-\left\langle\bigtriangledown_{\boldsymbol{W}}h_{\boldsymbol{W},k},\widehat{\boldsymbol{W}}\right\rangle\right)+O\left(\sum_{i\in\mathcal{I}_{\phi,k}}\frac{\mu_i+\lambda_i}{2}\|h_{\boldsymbol{W}',i}-h_{\boldsymbol{W},i}\|^2\right)\\
&+\sum_{i\in\mathcal{I}_{\phi,k}}\|u_i(h_{\boldsymbol{W},i})\|\left\|\left[\left(h_{\boldsymbol{W}',i}-h_{\boldsymbol{W},i}-\left\langle\bigtriangledown_{\boldsymbol{W}}h_{\boldsymbol{W},i},\widehat{\boldsymbol{W}}\right\rangle\right)-\left(h_{\boldsymbol{W}',k}-h_{\boldsymbol{W},k}-\left\langle\bigtriangledown_{\boldsymbol{W}}h_{\boldsymbol{W},k},\widehat{\boldsymbol{W}}\right\rangle\right)\right]\right\|,
\end{aligned}\tag{38}$$

where Cauchy-Schwartz inequality is used in the last inequality. By Theorem 4 in (Allen-Zhu et al., 2019b), we have with probability at least $1-\exp(-\Omega(m\tau^{2/3}L))$,

$$\begin{aligned}&\left\|h_{\boldsymbol{W}',i}-h_{\boldsymbol{W},i}-\left\langle\bigtriangledown_{\boldsymbol{W}}h_{\boldsymbol{W},i},\widehat{\boldsymbol{W}}\right\rangle\right\|\\ \leq&O\left(\tau^{1/3}L^{5/2}\sqrt{m\log(m)}d^{-1/2}\right)\left\|\widehat{\boldsymbol{W}}\right\|+O\left(L^2\sqrt{m/d}\left\|\widehat{\boldsymbol{W}}\right\|^2\right)\end{aligned}$$

By Claim 11.2 in (Allen-Zhu et al., 2019b), we have

$$\left\|h_{\boldsymbol{W}',i}-h_{\boldsymbol{W},i}\right\|\leq O(L\sqrt{m/d})\left\|\widehat{\boldsymbol{W}}\right\|.$$

Thus since $\|u_i(h_{\boldsymbol{W},i})\|\leq O(\mu_i+\lambda_i)$, we have

$$\begin{aligned}&\sum_{i\in\mathcal{I}_{\phi,k}}r_{\mathrm{I},i}(h_{\boldsymbol{W}',i})-r_{\mathrm{I},i}(h_{\boldsymbol{W},i})-\left\langle\bigtriangledown_{\boldsymbol{W}}r_{\mathrm{I},i}(h_{\boldsymbol{W},i}),\widehat{\boldsymbol{W}}\right\rangle\\ \leq&O\left(\sqrt{M_k\left(\bar{r}_{\mathrm{I},k}-r_{\mathrm{eff},k}+\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)}+M_k\right)O\left(\tau^{1/3}L^{5/2}\sqrt{m\log(m)}d^{-1/2}\right)\left\|\widehat{\boldsymbol{W}}\right\|\\ &+O(M_kL^2m/d)\left\|\widehat{\boldsymbol{W}}\right\|^2.\end{aligned}\tag{39}$$

where the inequality comes from Lemma B.3. Taking summation over $i\in[N]$, we have

$$\begin{aligned}&\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}'\right)-\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right)\leq\left\langle\bigtriangledown_{\boldsymbol{W}}\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right),\boldsymbol{W}'-\boldsymbol{W}\right\rangle+O(L^2m/d)\left\|\widehat{\boldsymbol{W}}\right\|^2\\ &+\left(\sqrt{\left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right)-\hat{R}_{\mathrm{eff}}+\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)}+1/\sqrt{N}\right)O\left(N^{1/2}\tau^{1/3}L^{5/2}\sqrt{m\log(m)}d^{-1/2}\right)\left\|\widehat{\boldsymbol{W}}\right\|\\ \leq&\left\langle\bigtriangledown_{\boldsymbol{W}}\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right),\boldsymbol{W}'-\boldsymbol{W}\right\rangle+O(L^2m/d)\left\|\widehat{\boldsymbol{W}}\right\|^2\\ &+\left(\sqrt{\left(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right)-\hat{R}_{\mathrm{eff}}-\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)}\right)O\left(N^{1/2}\tau^{1/3}L^{5/2}\sqrt{m\log(m)}d^{-1/2}\right)\left\|\widehat{\boldsymbol{W}}\right\|,\end{aligned}$$

where the second inequality comes from the choice of ϕ such that $\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right)-\hat{R}_{\mathrm{eff}}+\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\leq 2(\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}\right)-\hat{R}_{\mathrm{eff}}-\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)))$ and $1/\sqrt{N}\leq\sqrt{\phi}\leq\sqrt{\left(\hat{R}_{\mathrm{I}}-\hat{R}_{\mathrm{eff}}+\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)}$.

□

C.5. Proof of Lemma B.6

Proof. By Lemma 7.1 in (Allen-Zhu et al., 2019b), with probability at least $1-O(NL)\exp\left(-\Omega\left(m/L\right)\right)$, we have $\forall k\in[N],\left\|h_{k,L}^{(0)}\right\|\leq 2$.. Thus by Lemma B.1, we have with probability at least $1-O(\phi)$,

$$\forall k\in[N],\forall i\in\mathcal{I}_{\phi,k},\left\|h_{i,L}^{(0)}\right\|\leq 2+\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)).$$

Then since each entry of $\boldsymbol{V}$ satisfies $\mathcal{N}\left(0,\frac{1}{d}\right)$ and $O(NL)\exp\left(-\Omega\left(m/L\right)\right)\leq O(\phi)$, we have with probability at least $1-O(\phi)$,

$$\left\|h_{\boldsymbol{W}^{(0)},i}\right\|=\left\|\boldsymbol{V}h_{i,L}^{(0)}\right\|\leq 2\left\|h_{i,L}^{(0)}\right\|\sqrt{\log(1/\phi)}=O\left(\sqrt{\log(1/\phi)}\right).$$

Let $r_{\mathrm{I},i}(y_{\mathrm{eff},k(x_i)})=\mu_i r(y_{\mathrm{eff},k(x_i)},y_i)+\lambda_i r_K(y_{\mathrm{eff},k(x_i)},g(x_i))$. Thus with probability at least $1-O(\phi)$, by 1-Lipschitz continuity of risk functions, we have

$$\begin{aligned}&r_{\mathrm{I},i}^{(0)}-r_{\mathrm{I},i}(y_{\mathrm{eff},k(x_i)})\leq(\mu_i+\lambda_i)\left\|h_{\boldsymbol{W}^{(0)},i}-y_{\mathrm{eff},k(x_i)}\right\|\\ \leq&(\mu_i+\lambda_i)\left(\left\|h_{\boldsymbol{W}^{(0)},i}\right\|+\left\|y_{\mathrm{eff},k(x_i)}\right\|\right)\\ \leq&O\left((\mu_i+\lambda_i)\log^{1/2}(1/\phi)\right).\end{aligned}$$

Taking summation for $i \in [n']$, we have

$$\hat{R}_{\mathrm{I}}\left(\boldsymbol{W}^{(0)}\right) - \hat{R}_{\mathrm{eff}} \leq O\left(\log^{1/2}(1/\phi)\right).$$

□

C.6. Proof of Lemma B.8

Proof. **Proof of (a)**: Denote $\mathcal{U}_g' = \mathcal{U}_\phi(S_z) = \{k \in [N] \mid \exists x \in S_z, x \in \mathcal{C}_{\phi,k}\}$ as the index collection of smooth sets that contain at least one labeled sample, and $\mathcal{U}_g'' = [N] \setminus \mathcal{U}_g'$ as the index collection of smooth sets that only contain knowledge-supervised samples. Denote $h^*_{\mathrm{K},i} = h^*_{\mathrm{K}}(x_i)$ for notation simplicity, and recall that $x_k' \in \mathcal{X}_\phi$ is the representative input of the smooth set k, so we have

$$\begin{aligned}
&\frac{1}{n_g''}\sum_{S_g''} r_{\mathrm{K}}\left(h^*_{\mathrm{K},i}, g_i\right) = \frac{1}{n_g''}\sum_{k\in\mathcal{U}_g''}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(h^*_{\mathrm{K},i}, g_i\right)\\
\geq&\frac{1}{n_g''}\sum_{k\in\mathcal{U}_g''}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(h^*_{\mathrm{K}}(x_k'), g_i\right) - \widetilde{O}\left(L^{5/2}\phi\log^{1/2}(m)\right)\\
\geq&\frac{1}{n_g''}\sum_{k\in\mathcal{U}_g''}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right) + \frac{1}{n_g''}\sum_{k\in\mathcal{U}_g''}\left\langle \bigtriangledown_h \sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right), h^*_{\mathrm{K}}(x_k') - y_{\mathrm{eff},k}\right\rangle\\
&+\frac{\rho}{2n_g''}\sum_{k\in\mathcal{U}_g''}\left|\mathcal{I}_{\phi,k}\right|\left\|h^*_{\mathrm{K}}(x_k') - y_{\mathrm{eff},k}\right\|^2 - \widetilde{O}\left(L^{5/2}\phi\log^{1/2}(m)\right),
\end{aligned} \tag{40}$$

where the first inequality holds by Lemma B.1 and Lipschitz continuity of the risk function, and the second inequality holds by the strongly convexity of $\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(h, g_i\right)$ with respect to h. By subtracting $\frac{1}{n_g''}\sum_{k\in\mathcal{U}_g''}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right)$ from both sides of (40), we have

$$\begin{aligned}
&\frac{1}{n_g''}\sum_{k\in\mathcal{U}_g''}\left|\mathcal{I}_{\phi,k}\right|\left\|h^*_{\mathrm{K}}(x_k') - y_{\mathrm{eff},k}\right\|^2\\
\leq&\frac{2}{\rho n_g''}\sum_{k\in\mathcal{U}_g''}\left\|\bigtriangledown_h \sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right)\right\|\left\|h^*_{\mathrm{K}}(x_k') - y_{\mathrm{eff},k}\right\| + \widetilde{O}\left(L^{5/2}\phi\log^{1/2}(m)\right)\\
\leq&O\left(\frac{1}{n_g''}\sum_{k\in\mathcal{U}_g''}\left\|\bigtriangledown_h \sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right)\right\|\right) + \widetilde{O}\left(L^{5/2}\phi\log^{1/2}(m)\right)\\
\leq&\sum_{k\in\mathcal{U}_g''}\left[\sqrt{\frac{2|\mathcal{I}_{\phi,k}|}{n_g''}\left(r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right) - r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right)\right)} + \frac{|\mathcal{I}_{\phi,k}|}{n_g''}\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right)\right] + \widetilde{O}\left(L^{5/2}\phi\log^{1/2}(m)\right)\\
=&\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + \widetilde{O}(L^{5/2}\phi\log^{1/2}(m)) \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right),
\end{aligned} \tag{41}$$

where the first inequality holds since $\left\{h^*_{\mathrm{K},i}, i \in S_g''\right\}$ minimizes $\frac{1}{n_g''}\sum_{S_g''} r_{\mathrm{K}}\left(h(x_i), g_i\right)$, the second inequality holds since $\|h^*_{\mathrm{K}}(x_k') - y_{\mathrm{eff},k}\| \leq \|h^*_{\mathrm{K}}(x_k')\| + \|y_{\mathrm{eff},k}\| \leq \widetilde{O}(1)$ by Lemma B.6 and Lemma 8.2(c) in (Allen-Zhu et al., 2019b), and the third inequality holds by applying Lemma B.3 for $\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right)$ with $r_{\mathrm{eff},k} = \sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right)$. Therefore, by Lemma B.1, we have

$$\frac{1}{n_g''}\sum_{S_g''}\left\|h^*_{\mathrm{K},i} - y_{\mathrm{eff},k(x_i)}\right\|^2 \leq \frac{1}{n_g''}\sum_{k\in\mathcal{U}_g''}\left|\mathcal{I}_{\phi,k}\right|\left\|h^*_{\mathrm{K}}(x_k') - y_{\mathrm{eff},k}\right\|^2 + \widetilde{O}\left(L^{5/2}\phi\log^{1/2}(m)\right) \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right). \tag{42}$$

Similarly, denote $r_{\mathrm{R},\beta}(h(x_i)) = \frac{1-\beta}{n_z} r(h(x_i), y_i)\mathbb{1}\left(x_i \in S_z\right) + \frac{\beta}{n'_g} r_K(h(x_i), g(x_i))\mathbb{1}\left(x_i \in S'_g\right)$. We have

$$\begin{aligned}
&\sum_{S_z \bigcup S'_g} r_{\mathrm{R},\beta}\left(h^*_{\mathrm{R},\beta}(x_i)\right) \geq \sum_{k\in\mathcal{U}'_g}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{R},\beta}\left(h^*_{\mathrm{R},\beta}(x'_k)\right) - \widetilde{O}\left(L^{5/2}\phi\log^{1/2}(m)\right)\\
\geq &\sum_{k\in\mathcal{U}'_g}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{R},\beta}\left(y_{\mathrm{eff},k}\right) - \sum_{k\in\mathcal{U}'_g}\left\|\bigtriangledown h \sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{R},\beta}(y_{\mathrm{eff},k})\right\| \left\|h^*_{\mathrm{R},\beta}(x'_k) - y_{\mathrm{eff},k}\right\| \\
&+ \frac{\rho}{2}\sum_{k\in\mathcal{U}'_g} M_k \left\|h^*_{\mathrm{R},\beta}(x'_k) - y_{\mathrm{eff},k}\right\|^2 - \widetilde{O}\left(L^{5/2}\phi\log^{1/2}(m)\right),
\end{aligned} \tag{43}$$

where $M_k = \sum_{i\in\mathcal{I}_{\phi,k}}\left[\frac{1-\beta}{|S_z|}\mathbb{1}\left(x_i \in S_z\right) + \frac{\beta}{|S'_g|}\mathbb{1}\left(x_i \in S'_g\right)\right]$, the first inequality holds by Lemma B.1 and Lipschitz continuity of the risk function, and the second inequality holds by the strongly convexity of $r_{\mathrm{R},\beta}\left(h^*_{\mathrm{R},\beta}\right)$ with respect to $h^*_{\mathrm{R},\beta}$. Then, subtracting $\sum_{k\in\mathcal{U}'_g}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{R},\beta}\left(y_{\mathrm{eff},k}\right)$ from both sides of (43), similarly as Eqn. (41),it holds that

$$\begin{aligned}
&\sum_{k\in\mathcal{U}'_g} M_k \left\|h^*_{\mathrm{R},\beta}(x'_k) - y_{\mathrm{eff},k}\right\|^2\\
\leq &\frac{2}{\rho}\sum_{k\in\mathcal{U}'_g}\left\|\bigtriangledown h \sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{R},\beta}\left(y_{\mathrm{eff},k}\right)\right\| \left\|h^*_{\mathrm{R},\beta}(x'_k) - y_{\mathrm{eff},k}\right\| + \widetilde{O}\left(L^{5/2}\phi\log^{1/4}(m)\right)\\
\leq &\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right),
\end{aligned} \tag{44}$$

where the first inequality holds because $\sum_{S_z\bigcup S'_g} r_{\mathrm{R},\beta}\left(h^*_{\mathrm{R},\beta}(x_i)\right) - \sum_{k\in\mathcal{U}'_g}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{R},\beta}\left(y_{\mathrm{eff},k}\right) \leq 0$, and the second inequality holds since $\|h^*_{\mathrm{K}}(x'_k) - y_{\mathrm{eff},k}\| \leq \widetilde{O}(1)$ by Lemma B.6 and Lemma 8.2(c) in (Allen-Zhu et al., 2019b) and then applying Lemma B.3. Therefore, by Lemma B.1, we have

$$\begin{aligned}
\frac{1-\beta}{n_z}\sum_{S_z}\left\|h^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\|^2 + \frac{\beta}{n'_g}\sum_{S'_g}\left\|h^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\|^2 &\leq O\left(\sum_{k\in\mathcal{U}'_g} M_k\left\|h^*_{\mathrm{R},\beta}(x'_k) - y_{\mathrm{eff},k}\right\|^2\right)\\
&\leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right).
\end{aligned}$$

Proof of (b): Replacing h^*_{K} in Eqn. (40) with $\bar{h}^*_{\mathrm{K}}$ and applying the second and third inequality in Eqn. (41), we have

$$\begin{aligned}
&\frac{1}{n''_g}\sum_{k\in\mathcal{U}''_g}\left|\mathcal{I}_{\phi,k}\right|\left\|\bar{h}^*_{\mathrm{K}}(x'_k) - y_{\mathrm{eff},k}\right\|^2\\
\leq &\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + \frac{1}{n''_g}\sum_{S''_g} r_{\mathrm{K}}\left(\bar{h}^*_{\mathrm{K}}(x_i), g_i\right) - \frac{1}{n''_g}\sum_{k\in\mathcal{U}''_g}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{K}}\left(y_{\mathrm{eff},k}, g_i\right)\\
\leq &\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + \mathbb{E}\left[r_{\mathrm{K}}\left(\bar{h}^*_{\mathrm{K}}(x), g(x)\right)\right] - \mathbb{E}\left[r_{\mathrm{K}}\left(y_{\mathrm{eff},k(x)}, g(x)\right)\right] + O(\sqrt{\frac{\log(1/\delta)}{n''_g}})\\
\leq &\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + O(\sqrt{\frac{\log(1/\delta)}{n''_g}}),
\end{aligned}$$

where the second inequality follows from McDiarmid's inequality and the last inequality is because $\bar{h}^*_{\mathrm{K}}(x)$ minimizes $\mathbb{E}\left[r_{\mathrm{K}}\left(h, , g(x)\right)\right]$. Therefore, by Lemma B.1 and with the same reason as (42), we have

$$\frac{1}{n''_g}\sum_{S''_g}\left\|\bar{h}^*_{\mathrm{K},i} - y_{\mathrm{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + O(\sqrt{\frac{\log(1/\delta)}{n''_g}}).$$

Similarly, replacing $h_{\mathrm{R},\beta}$ in Eqns. (43) with $h^*_{\mathrm{R},\beta}$ and with the same reason as the second inequality in (44), we have

$$\begin{aligned}&\sum_{k\in\mathcal{U}'_g} M_k \left\|\bar{h}^*_{\mathrm{R},\beta}(x'_k) - y_{\mathrm{eff},k}\right\|^2\\ \leq&\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + \sum_{S_z \bigcup S'_g} r_{\mathrm{R},\beta}\left(\bar{h}^*_{\mathrm{R},\beta}(x_i)\right) - \sum_{k\in\mathcal{U}'_g}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{R},\beta}\left(y_{\mathrm{eff},k}\right).\end{aligned} \tag{45}$$

Continuing with (45) and by McDiarmid's inequality, we have

$$\begin{aligned}&\sum_{k\in\mathcal{U}'_g} M_k \left\|\bar{h}^*_{\mathrm{R},\beta}(x'_k) - y_{\mathrm{eff},k}\right\|^2\\ \leq&\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + \mathbb{E}\left[\sum_{S_z \bigcup S'_g} r_{\mathrm{R},\beta}\left(\bar{h}^*_{\mathrm{R},\beta}(x_i)\right)\right] - \mathbb{E}\left[\sum_{k\in\mathcal{U}'_g}\sum_{\mathcal{I}_{\phi,k}} r_{\mathrm{R},\beta}\left(y_{\mathrm{eff},k}\right)\right] + O(\sqrt{\frac{\log(1/\delta)}{n_z}})\\ \leq&\widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + O(\sqrt{\frac{\log(1/\delta)}{n_z}}),\end{aligned}$$

where the second inequality is because $\bar{h}^*_{\mathrm{R},\beta}(x)$ minimizes $\mathbb{E}\left[\sum_{S_z \bigcup S'_g} r_{\mathrm{R},\beta}\left(\bar{h}^*_{\mathrm{R},\beta}(x_i)\right)\right]$. And thus by Lemma B.1, we have

$$\frac{1-\beta}{n_z}\sum_{S_z}\left\|\bar{h}^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\|^2 + \frac{\beta}{n'_g}\sum_{S'_g}\left\|\bar{h}^*_{\mathrm{R},\beta,i} - y_{\mathrm{eff},k(x_i)}\right\|^2 \leq \widetilde{O}\left(L^{5/4}\phi^{1/2}\log^{1/4}(m)\right) + O(\sqrt{\frac{\log(1/\delta)}{n_z}}).$$

□

D. Proof of Lemmas in Appendix C

D.1. Proof of Lemma C.1

Proof. We simply use $\boldsymbol{D}'$ to denote $\boldsymbol{D}'_l$. If for some $j \in [m]$, $[\boldsymbol{D}']_{j,j} \neq 0$, then it holds that

$$|[f'_{l,i}]_j| = |[f'_{l,i,1}]_j + [f'_{l,i,2}]_j| > |[f^{(0)}_{l,k}]_j|. \tag{46}$$

Let $\xi : \xi \leq \frac{1}{2\sqrt{m}}$ and $\xi \leq 2\|f'_{l,i,2}\|_\infty$ be a parameter to be chosen later. We then discuss the zero norm of $\boldsymbol{D}'$ in the following two cases.

First, we consider the case that $|[f^{(0)}_{l,k}]_j| \leq \xi$. In this case, (46) is easy to be satisfied. Denote $S_1 = \left\{j \in [m] \mid |[f^{(0)}_{l,k}]_j| \leq \xi\right\}$. Since $[f^{(0)}_{l,k}]_j \sim \mathcal{N}(0, \frac{2}{m})$, we have $\mathbb{P}\{[f^{(0)}_{l,k}]_j \leq \xi\} \leq O(\xi\sqrt{m})$. Since $|S_1| = \sum_{j=1}^m \mathbb{1}([f^{(0)}_{l,k}]_j \leq \xi)$, we have $\mathbb{E}[\exp(|S_1|)] \leq \exp(\xi m^{3/2}(e-1))$. Thus, by Chernoff bound, $\mathbb{P}(|S_1| \geq 2\xi m^{3/2}) \leq \frac{\mathbb{E}[\exp(|S_1|)]}{\exp(\xi m^{3/2})} \leq \exp(\xi m^{3/2}(e-3))$. Hence, with probability at least $1 - \exp(-\Omega(m^{3/2}\xi))$, we have

$$|S_1| \leq O(\xi m^{3/2}).$$

Then, for $j \in S_1$ such that $[\boldsymbol{D}']_{j,j} \neq 0$, we have $|[\boldsymbol{D}' f^{(0)}_{l,i}]_j| \leq |[f^{(0)}_{l,k}]_j| + |[f'_{l,i,1}]_j| + |[f'_{l,i,2}]_j| \leq |[f'_{l,i,1}]_j| + 3\xi/2$. Further, we have

$$\sum_{j\in S_1}[\boldsymbol{D}' f^{(0)}_{l,i}]_j^2 \leq O(\|f'_{l,i,1}\|^2 + \xi^2|S_1|) \leq O(\|f'_{l,i,1}\|^2 + \xi^2|S_1|) \leq O(\|f'_{l,i,1}\|^2 + \xi^3 m^{3/2}).$$

Second, we consider the case that $|[f^{(0)}_{l,k}]_j| > \xi$. Denote $S_2 = \left\{j \in [m] \mid |[f^{(0)}_{l,k}]_j| > \xi, [\boldsymbol{D}']_{j,j} \neq 0\right\}$. Then, (46) requires that

$$|[f'_{l,i,1}]_j| = |[f'_{l,i}]_j - |[f'_{l,i,2}]_j|| \geq |[f'_{l,i}]_j| - |[f'_{l,i,2}]_j| \geq |[f^{(0)}_{l,k}]_j| - |[f'_{l,i,2}]_j| \geq \xi - \|f'_{l,i,2}\|_\infty \geq \xi/2.$$

Thus we have

$$|S_2| \leq \frac{4\|f'_{l,i,1}\|^2}{\xi^2}.$$

Then since for $j \in S_2$ such that $[\boldsymbol{D}']_{j,j} \neq 0$, the signs of $[f^{(0)}_{l,k}]_j + [f'_{l,i,1}]_j + [f'_{l,i,2}]_j$ and $[f^{(0)}_{l,k}]_j$ are opposite, we have

$$|[\boldsymbol{D}' f^{(0)}_{l,i}]_j| = |[f^{(0)}_{l,k}]_j + [f'_{l,i,1}]_j + [f'_{l,i,2}]_j| \leq |[f'_{l,i,1}]_j + [f'_{l,i,2}]_j| \leq |[f'_{l,i,1}]_j| + \xi/2 \leq 2|[f'_{l,i,1}]_j|.$$

Therefore, it holds that

$$\sum_{j\in S_s} [\boldsymbol{D}' f^{(0)}_{l,i}]_j^2 \leq 4 \sum_{j\in S_2} |[f'_{l,i,1}]_j|^2 \leq 4\|f'_{l,i,1}\|^2.$$

Combining the two cases, we have

$$\|\boldsymbol{D}\|_0 \leq |S_1| + |S_2| \leq O\left(\xi m^{3/2} + \frac{4\|f'_{l,i,1}\|^2}{\xi^2}\right),$$

$$\|\boldsymbol{D}' f^{(0)}_{l,i}\|^2 \leq O(\|f'_{l,i,1}\|^2 + \xi^3 m^{3/2}).$$

Choosing $\xi = \max\left\{2\|f'_{l,i,2}\|_\infty, \Theta(\frac{\|f'_{l,i,1}\|^{2/3}}{m^{1/2}})\right\}$, and recalling $\|f'_{l,i,1}\| \leq O(L^{3/2}\phi\log^{1/2}(1/\phi))$ and $\|f'_{l,i,2}\|_\infty \leq O(L\phi^{2/3}/m^{1/2}\log^{1/2}(1/\phi))$, we get $\|\boldsymbol{D}\|_0 \leq O(mL\phi^{2/3}\log^{1/2}(1/\phi))$. Choosing $\xi = 2\|f'_{l,i,2}\|_\infty$, we get $\|\boldsymbol{D}' f^{(0)}_{l,i}\| \leq O(\phi L^{3/2}\log^{1/2}(1/\phi))$. □

D.2. Proof of Lemma C.2

Lemma D.1 (Lemma B.1 in (Zou et al., 2020)). *Assume $m > \Omega(L\log(NL))$. For any $x'_i, x'_j \in \mathcal{X}_\phi$, $i, j \in [N], l \in [L]$, with probability at least $1 - \exp(-O(m/L))$ over the randomness of $\boldsymbol{W}^{(0)}$, it holds that $1/2 \leq \|h_l(x'_i)\| \leq 2$ and $\left\|h_l(x'_i)/\|h_l(x'_i)\| - h_l(x'_j)/\left\|h_l(x'_j)\right\|\right\| \geq \phi/2$, where $h_l(x'_i)$ is the output of the $l-$th layer at initialization.*

Denote $b_i = h_{\boldsymbol{W}^{(0)},L-1}(x_i)$ and $\bar{b}_i = b_i/\|b_i\|$ for $x_i \in \mathcal{X}$, and $b'_i = h_{\boldsymbol{W}^{(0)},L-1}(x'_i)$ and $\bar{b}'_i = b'_i/\|b'_i\|$ for $x'_i \in \mathcal{X}_\phi$. By Lemma D.1, we have $\forall i \notin \mathcal{I}_{\phi,k}, \left\|\bar{b}_i - \bar{b}'_k\right\| \geq \phi/4$. Moreover, by Lemma B.1, we have $\forall i \in \mathcal{I}_{\phi,k}, \left\|b_i - b'_j\right\| \leq \widetilde{O}(L^{5/2}\phi\log^{1/2}(m))$.

Then we construct several sets for the vector $w \in \mathcal{R}^m$ subject to $\mathcal{N}(0,(2/m)\mathbf{I})$. Given $\bar{b}'_k$, we construct an orthogonal matrix $Q_k = [\bar{b}'_k, Q'_k] \in \mathcal{R}^{m\times m}$ and let $q_k = Q_k^\top w \sim \mathcal{N}(0,(2/m)\mathbf{I})$. In this way, the vector w is decomposed as two orthogonal vector: $w = Q_k q_k = q_k^{(1)}\bar{b}'_k + Q'_k q'_k$ where $q_k^{(1)}$ is the first element of q_k. Letting $\gamma = \sqrt{2\pi}\phi/(32N\sqrt{m})$, we construct the set

$$\mathcal{W}_k = \left\{w \in \mathcal{R}^d \mid |q_k^{(1)}| \leq \gamma, |\left\langle Q'_k q'_k, \bar{b}'_j\right\rangle| \geq 2\gamma, \forall j \neq k\right\}, \tag{47}$$

where $[q_k^{(1)}, q'_k] = q_k$.

Lemma D.2 (Lemma C.1 in (Zou & Gu, 2019)). *For any $\mathcal{W}_j$ and $\mathcal{W}_k$, $j \neq k$, we have $\mathcal{W}_j \bigcap \mathcal{W}_k = \emptyset$ and $\mathbb{P}(w \in \mathcal{W}_k) \geq \frac{\phi}{N32\sqrt{2e}}$.*

Lemma D.3. *Let $f(w_j) = \sum_{k=1}^N \sum_{i\in\mathcal{I}_{\phi,k}} a_i\sigma'(\langle w_j, b_i\rangle)b_i$ where $w_j, j \in [m]$ is drawn from $\mathcal{N}(0,(2/m)\mathbf{I})$, $|a_i| \leq O\left((\mu_i+\lambda_i)/\sqrt{d}\right)$. If for each smooth set k, there exists a subset $\mathcal{G}_{k,\alpha} \in [m]$ with size $\alpha m, \alpha \in (0,1)$ such that $\forall i \in \mathcal{I}_{\phi,k}$, $\forall j \in \mathcal{G}_{k,\alpha}$, $\sigma'(\langle w_j, b_i\rangle) = \sigma'(\langle w_j, b'_k\rangle)$ and $\forall j \notin \mathcal{G}_{k,\alpha}$, $|\langle w_j, b_i\rangle| \geq \frac{3\sqrt{2\pi}\phi}{16N\sqrt{m}}$, we have for any $j \in \mathcal{G}_{k,\alpha}$, $\mathbb{P}\left(\|f(w_j)\| \geq |A_k|/4 - M_k/\sqrt{d}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)) \mid w_j \in \mathcal{W}_k\right) \geq 1/2$ where $A_k = \sum_{i\in\mathcal{I}_{\phi,k}} a_i, k \in [N]$.*

Proof. For $j \in \mathcal{G}_{k,\alpha}$, let $q_k = Q_k^\top w_j \sim \mathcal{N}(0,(2/m)\mathbf{I})$. Then we have $w_j = Q_k q_k = q_k^{(1)}\bar{b}'_k + Q'_k q'_k$. We decompose

$f(w_j)$ as

$$\begin{aligned} f(w_j) &= \sum_{i\in\mathcal{I}_{\phi,k}} a_i\sigma'(\langle w, b_i\rangle)b_i + \sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle w, b_i\rangle)b_i \\ &= \sum_{i\in\mathcal{I}_{\phi,k}} a_i\sigma'(\langle w, b'_k\rangle)b_i + \sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle w, b_i\rangle)b_i \\ &= \sum_{i\in\mathcal{I}_{\phi,k}} a_i\sigma'(q_k^{(1)})b_i + \sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle w, b_i\rangle)b_i, \end{aligned} \tag{48}$$

where the second equality holds by the assumption $\forall j \in \mathcal{G}_{k,\alpha}$, $\sigma'(\langle w_j, b_i\rangle) = \sigma'(\langle w_j, b'_k\rangle)$.

Then for the second term of (48), if $j \in \mathcal{G}_{k',\alpha}$, we have for $i \in \mathcal{I}_{\phi,k'}$, $\sigma'(\langle w_j, b_i\rangle) = \sigma'(\langle w_j, b'_{k'}\rangle)$ and thus

$$\begin{aligned} &\sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle w, b_i\rangle)b_i = \sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle w, b'_{k'}\rangle)b_i \\ =&\sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(q_k^{(1)}\left\langle \bar{b}'_k, b'_{k'}\right\rangle + \langle Q'_k q'_k, b'_{k'}\rangle)b_i \\ =&\sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle Q'_k q'_k, b'_{k'}\rangle)b_i \end{aligned}$$

where the last equality holds by the condition $w_j \in \mathcal{W}_k$ such that for $k' \neq k$, $|\langle Q'_k q'_k, b'_{k'}\rangle| \geq 2\gamma\|b'_{k'}\| \geq |q_k^{(1)}|\|b'_{k'}\| \geq |q_k^{(1)} < \bar{b}'_k, b'_{k'} > |$ and thus the sign is determined by $\langle Q'_k q'_k, b'_{k'}\rangle$. Therefore, if $j \in \mathcal{G}_{k',\alpha}$, we can write (48) as

$$f(w_j) = \sum_{i\in\mathcal{I}_{\phi,k}} a_i\sigma'(q_k^{(1)})b_i + \sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle Q'_k q'_k, b'_{k'}\rangle)b_i. \tag{49}$$

In the other case with $j \notin \mathcal{G}_{k',\alpha}$, by assumption $\forall i \in \mathcal{I}_{\phi,k'}, |\langle w_j, b_i\rangle| \geq \frac{3\sqrt{2\pi}\phi}{16N\sqrt{m}} = 6\gamma$, we have with probability at least $1 - \exp(-O(m/L))$, $\left|\left\langle w_j, \bar{b}_i\right\rangle\right| = |\langle w_j, b_i\rangle| \frac{1}{\|b_i\|} \geq 3\gamma$ by Lemma D.1. Then $\forall i \in \mathcal{I}_{\phi,k'}$, we have

$$\begin{aligned} &|\left\langle Q'_k q'_k, \bar{b}_i\right\rangle| = |\left\langle w_j, \bar{b}_i\right\rangle - \left\langle q_k^{(1)}\bar{b}'_k, \bar{b}_i\right\rangle| \\ \geq&|\left\langle w_j, \bar{b}_i\right\rangle| - |\left\langle q_k^{(1)}\bar{b}'_k, \bar{b}_i\right\rangle| \geq |\left\langle w_j, \bar{b}_i\right\rangle| - |q_k^{(1)}| \\ \geq&|\left\langle w_j, \bar{b}_i\right\rangle| - \gamma \geq 2\gamma \\ \geq&2\gamma\|b'_{k'}\| \geq |q_k^{(1)}|\|b'_{k'}\| \geq |q_k^{(1)} < \bar{b}'_k, b'_{k'} > |, \end{aligned}$$

where the first inequality comes from triangle inequality, the second inequality holds by $|\left\langle \bar{b}'_k, \bar{b}_i\right\rangle| \leq 1$, and the last inequality holds by the condition $w_j \in \mathcal{W}_k$. Therefore if $j \notin \mathcal{G}_{k',\alpha}$, we can write the second term in (48) as

$$\begin{aligned} &\sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle w, b_i\rangle)b_i \\ =&\sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(q_k^{(1)}\left\langle \bar{b}'_k, b_i\right\rangle + \langle Q'_k q'_k, b_i\rangle)b_i \\ =&\sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle Q'_k q'_k, b_i\rangle)b_i \end{aligned}$$

Therefore, if $j \notin \mathcal{G}_{k',\alpha}$, we can write (48) as

$$f(w_j) = \sum_{i\in\mathcal{I}_{\phi,k}} a_i\sigma'(q_k^{(1)})b_i + \sum_{k'\neq k}\sum_{i\in\mathcal{I}_{\phi,k'}} a_i\sigma'(\langle Q'_k q'_k, b_i\rangle)b_i. \tag{50}$$

Note that (49) and (50) are different only in terms of whether $b'_{k'}$ or $b_i, i \in \mathcal{I}_{\phi,k'}$ determines the second term, but for both of them, the second term does not rely on $q_k^{(1)}$. We thus proceed as follows.

Since $q_k^{(1)} > 0$ and $q_k^{(1)} < 0$ occurs with equal probability conditioned on the event $w \in \mathcal{W}_k$, we have

$$\mathbb{P}\left[\|f(w_j)\|_2 \geq \inf_{q_1>0,q_2<0} \max\left\{\|f\left(q_1\bar{b}'_k + Q'_k q'_k\right)\|, \|f\left(q_2\bar{b}'_k + Q'_k q'_k\right)\|\right\} \mid w \in \mathcal{W}_k\right] \geq 1/2.$$

Thus, with probability at least $1/2$ conditioned on the event $w \in \mathcal{W}_k$, we have

$$\begin{aligned}
\|f(w_j)\| &\geq \inf_{q_1>0,q_2<0} \max\left\{\|f\left(q_1\bar{b}'_k + Q'_k q'_k\right)\|, \|f\left(q_2\bar{b}'_k + Q'_k q'_k\right)\|\right\} \\
&\geq \inf_{q_1>0,q_2<0} \left\|f\left(q_1\bar{b}'_k + Q'_k q'_k\right) - f\left(q_2\bar{b}'_k + Q'_k q'_k\right)\right\|/2 \\
&= \|\sum_{i\in\mathcal{I}_{\phi,k}} a_i b_i\|
\end{aligned}$$

Since $|a_i| \leq O\left((\mu_i+\lambda_i)/\sqrt{d}\right)$ and $\|b_i - b'_k\| \leq \widetilde{O}(L^{5/2}\phi\log^{1/2}(m))$ for $i \in \mathcal{I}_{\phi,k}$, we have, $\left|\|\sum_{i\in\mathcal{I}_{\phi,k}} a_i b_i\| - \|\sum_{i\in\mathcal{I}_{\phi,k}} a_i b'_k\|\right| \leq \|\sum_{i\in\mathcal{I}_{\phi,k}} a_i(b_i - b'_k)\| \leq \sum_{i\in\mathcal{I}_{\phi,k}} |a_i|\|b_i - b'_k\| \leq M_k\sqrt{d}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))$. Thus,

$$\begin{aligned}
\|\sum_{i\in\mathcal{I}_{\phi,k}} a_i b_i\| &\geq \|\sum_{i\in\mathcal{I}_{\phi,k}} a_i b'_k\| - \left|\|\sum_{i\in\mathcal{I}_{\phi,k}} a_i b_i\| - \|\sum_{i\in\mathcal{I}_{\phi,k}} a_i b'_k\|\right| \\
&\geq \|\sum_{i\in\mathcal{I}_{\phi,k}} a_i b'_k\| - M_k\sqrt{d}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)) \\
&\geq |A_k|/4 - M_k/\sqrt{d}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)),
\end{aligned}$$

where the last inequality follows from Lemma D.1. The proof is completed. □

Lemma D.4 (Bernstein inequality)**.** *Let $X_1, \cdots, X_n$ be independent zero-mean random variables. If $|X_i| \leq 1$ almost surely for all i, then $\forall t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{1}{2}t^2 / \left(\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{1}{3}t\right)\right).$$

Proof of Lemma C.2

Proof. Denote $b_i = h_{L-1}^{(0)}(x_i)$, so $[h_L^{(0)}(x_i)]_j = \langle w_j, b_i\rangle$. For any fixed $[u_1, \cdots, u_{n'}]$, denote $a_i(v_j) = \langle u_i, v_j\rangle, i \in [n'], j \in [m]$ and $A_k(v_j) = \sum_{i\in\mathcal{I}_{\phi,k}} a_i(v_j)$. Let $f(v_j, w_j) = \sum_{k=1}^N \sum_{i\in\mathcal{I}_{\phi,k}} a_i(v_j)\sigma'(\langle w_j, b_i\rangle)b_i$. Define the event for $k \in [N]$

$$\mathcal{E}_k = \left\{j \in \mathcal{G}_{k,\alpha} : w_j \in \mathcal{W}_k, \|f(v_j, w_j)\| \geq \|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/(4\sqrt{d}) - M_k d^{-1/2}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right\}.$$

Since $v_j \sim \mathcal{N}(0, (1/d)\mathbf{I})$, we have $A_k(v_j) = \left\langle\sum_{i\in\mathcal{I}_{\phi,k}} u_i, v_j\right\rangle \sim \mathcal{N}(0, \|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|^2/d)$. Thus, we have

$$\mathbb{P}\left(\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/\sqrt{d} \leq |A_k(v_j)| \leq 2\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/\sqrt{d}\right) \geq 1/4.$$

Note that when $|A_k(v_j)| \leq 2\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/\sqrt{d}$, we have $\forall i \in \mathcal{I}_{\phi,k}$, $|a_i(v_j)| \leq |A_k(v_j)|/|\mathcal{I}_{\phi,k}| - \widetilde{O}(L^{5/2}\phi\log^{1/2}(m)) \leq 2\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/\sqrt{d}/|\mathcal{I}_{\phi,k}| - \widetilde{O}(L^{5/2}\phi\log^{1/2}(m)) \leq 2(\mu_i+\lambda_i)/\sqrt{d} - 3\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)) \leq 3(\mu_i+\lambda_i)\sqrt{d}$ when ϕ is small enough, so the condition about $|a_i(v_j)|$ in Lemma D.3 is met.

Since Assumption 1 is satisfied, we have for each smooth set k, there exists a subset $\mathcal{G}_{k,\alpha} \in [m]$ with size $\alpha m, \alpha \in (0,1)$ such that $\forall i \in \mathcal{I}_{\phi,k}$, $\forall j \in \mathcal{G}_{k,\alpha}$, $\sigma'(\langle w_j, b_i\rangle) = \sigma'(\langle w_j, b'_k\rangle)$ and $\forall j \notin \mathcal{G}_{k,\alpha}$, $|\langle w_j, b_i\rangle| \geq \frac{3\sqrt{2\pi}\phi}{16N\sqrt{m}}$, so the assumption in Lemma D.3 is satisfied. Then by Lemma D.2, Lemma D.3 and the fact that w_j and v_j are independent, we have for $j \in \mathcal{G}_{k,\alpha}$

$$\begin{aligned}\mathbb{P}(j \in \mathcal{E}_k) = &\mathbb{P}\left\{\|f(v_j, w_j)\| \geq |A_k(v_j)|/4 - M_k d^{-1/2}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m)) \mid w_j \in \mathcal{W}_k\right\} \\ &\cdot \mathbb{P}\left\{w_j \in \mathcal{W}_k\right\}\mathbb{P}\left(\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/\sqrt{d} \leq |A_k(v_j)| \leq 2\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/\sqrt{d}\right) \geq \frac{\phi}{N256\sqrt{2e}} = p_\phi.\end{aligned}$$

and $\mathcal{E}_{k_1}\bigcap\mathcal{E}_{k_2} = \emptyset$ for any $k_1 \neq k_2$.

For smooth set k, denote Bernoulli random variables $\mathbb{1}(j \notin \mathcal{E}_k)$ for $j \in \mathcal{G}_{k,\alpha}$. Then we have $\mathbb{E}\left[\mathbb{1}(j \notin \mathcal{E}_k)\right] = 1 - p_\phi$ and $\mathrm{var}\left[\mathbb{1}(j \notin \mathcal{E}_k)\right] = p_\phi(1-p_\phi)$. By Bernstein inequality in Lemma D.4 for random variables $\mathbb{1}(j \notin \mathcal{E}_k) - (1-p_\phi), j \in [\alpha m]$, it holds that

$$\mathbb{P}\left(\sum_{j\in\mathcal{G}_{k,\alpha}} \mathbb{1}(j \notin \mathcal{E}_k) - \alpha m(1-p_\phi) \geq \frac{\alpha m p_\phi}{2}\right) \leq \exp\left(-\frac{\frac{1}{4}\alpha m p_\phi}{1 - p_\phi + \frac{1}{6}}\right) \leq \exp\left(-\frac{3}{14}\alpha m p_\phi\right).$$

Thus, by union bounds, with probability at least $1 - O(N)\exp\left(-O\left(\alpha m\phi/N\right)\right)$, we have for any $k \in [N]$, $\sum_{j\in\mathcal{G}_{k,\alpha}} \mathbb{1}(j \notin \mathcal{E}_k) \leq \alpha m - \alpha m p_\phi/2$ and

$$\left|\mathcal{G}_{k,\alpha}\bigcap\mathcal{E}_k\right| = \sum_{j\in\mathcal{G}_{k,\alpha}} \mathbb{1}(j \in \mathcal{E}_k) = \alpha m - \sum_{j\in\mathcal{G}_{k,\alpha}} \mathbb{1}(j \notin \mathcal{E}_k) \geq \alpha m p_\phi/2. \tag{51}$$

Therefore, with probability at least $1 - O(\phi)$, it holds that

$$\begin{aligned}&\sum_{j=1}^m \|f(v_j, w_j)\|^2 \geq \sum_{j=1}^m \|f(v_j, w_j)\|^2 \sum_{k=1}^N \mathbb{1}(j \in \mathcal{E}_k) = \sum_{k=1}^N\sum_{j=1}^m \|f(v_j, w_j)\|^2 \mathbb{1}(j \in \mathcal{E}_k) \\ \geq &\sum_{k=1}^N \sum_{j\in\mathcal{G}_{k,\alpha}} \|f(v_j, w_j)\|^2 \mathbb{1}(j \in \mathcal{E}_k) = \sum_{k=1}^N \sum_{j\in\mathcal{G}_{k,\alpha}, j\in\mathcal{E}_k} \|f(v_j, w_j)\|^2 \\ \geq &\sum_{k=1}^N \sum_{j\in\mathcal{G}_{k,\alpha}, j\in\mathcal{E}_k} \left(\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/(4\sqrt{d}) - M_k d^{-1/2}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)^2 \\ \geq &\sum_{k=1}^N \sum_{j\in\mathcal{G}_{k,\alpha}, j\in\mathcal{E}_k} \left(\frac{1}{16d}\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|^2 - M_k^2 d^{-1}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right) \\ = &\sum_{k=1}^N \left(\frac{1}{16d}\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|^2 - M_k^2 d^{-1}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right)\left|\mathcal{G}_{k,\alpha}\bigcap\mathcal{E}_k\right| \\ \geq &\frac{\alpha m p_\phi}{2}\sum_{k=1}^N \left(\frac{1}{16d}\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|^2 - M_k^2 d^{-1}\widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right) \\ \geq &\Omega\left(\frac{\alpha m\phi}{Nd}\right)\left(\sum_{k=1}^N \|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|^2 - \widetilde{O}(L^{5/2}\phi\log^{1/2}(m))\right),\end{aligned} \tag{52}$$

where the first inequality comes from the fact that $\mathcal{E}_{k_1}\bigcap\mathcal{E}_{k_2} = \emptyset$ such that $\sum_{k=1}^N \mathbb{1}(j \in \mathcal{E}_k) \leq 1$ and the second inequality comes from the fact that $\mathcal{G}_{k,\alpha} \in [m]$, and the third inequality holds by the definition of event $\mathcal{E}_k$, and the forth inequality comes from the fact that $(a-b)^2 \geq a^2 - 2ab$ and $\|\sum_{i\in\mathcal{I}_{\phi,k}} u_i\|/(4\sqrt{d}) \leq M_k d^{-1/2}/4$, and the fifth inequality comes from (51) and the last inequality holds by the fact that $\sum_{k=1}^N M_k^2 \leq (\sum_{k=1}^N M_k)^2 = 1$. □

E. Preliminaries on Informed Machine Learning

Informed machine learning is rapidly emerging as a broad paradigm that incorporates domain knowledge, either directly or indirectly, to augment the purely data-driven approach and better accomplish a machine learning task. We provide a summary of how domain knowledge is integrated with machine learning (von Rueden et al., 2021).

- *Training Dataset*. A straightforward approach to utilizing domain knowledge is to generate (sometimes synthetic) data and enlarge the otherwise limited training dataset. For example, based on the simple knowledge of image invariance, cropping(Gao et al., 2018a), scaling(Zhang et al., 2018), flipping(Benaim & Wolf, 2018) and many other image pre-processing methods have been used to augment the training data for image classification tasks. As another example, in reinforcement learning (e.g., robot control and autonomous driving) where initial pre-training is crucial to avoid arbitrarily bad decisions in the real world, simulated environments can be built based on domain knowledge, providing simulations or demonstrations to generate training data (Gao et al., 2018b; Hester et al., 2018). Additionally, generative models constructed based on specific knowledge have been shown useful for increasing training data to improve model performance and robustness (Gao et al., 2018a; Goodfellow et al., 2016).

- *Hypothesis Set.* The goal of a machine learning task is to search for an optimal hypothesis that correctly expresses the relationships between input and output. To reduce the training complexity, the target hypothesis set (decided by, e.g., different neural architectures) should contain the optimal hypothesis and preferably be small enough. Thus, domain knowledge can be employed for hypothesis set selection. For example, (Chen et al., 2021a) makes use of the prior knowledge from the existing neural architectures to design new architectures (and hence, new hypothesis sets) for DNNs. As implicit domain knowledge, long short-term memory recurrent neural networks are commonly used for time series prediction (Goodfellow et al., 2016). Also, the structure of a knowledge graph helps to determine the hypothesis set of graph learning (Marino et al., 2016; Battaglia et al., 2018), while (Towell & Shavlik, 1994) maps the domain knowledge represented in propositional logic into neural networks.

- *Model Training.* Domain knowledge can be integrated, either implicitly or explicitly, with the model training procedure in various ways. First, domain knowledge can assist with the initialization of training. For example, (Ramsey & Grefenstette, 1993) provides a case-based method to initialize genetic algorithms (i.e., generating the initial population based on different cases), while (Husken & Goerick, 2000; Kurata et al., 2016; Humbird et al., 2018) initialize neural network training with various domain knowledge such as label co-occurrence and decision trees. Second, domain knowledge can be used to better tune the hyper-parameters (Bardenet et al., 2013; Van Rijn & Hutter, 2018; Maher & Sakr, 2019; Bamler et al., 2020). In (Bardenet et al., 2013), implicit knowledge from previous training is incorporated to improve hyper-parameter tuning, and (Van Rijn & Hutter, 2018) extracts knowledge from multiple datasets to determine the most important hyper-parameters. In addition, a more explicit way to integrate domain knowledge is to directly modify the training objective function (i.e., risk function) based on rigorous characterization of the model output (von Rueden et al., 2021). For example, in (Muralidhar et al., 2018), the knowledge of constraints is incorporated into neural networks expressing the knowledge based loss by the ReLu function. For another example, when learning to optimally schedule transmissions for rate maximization in multi-user wireless networks, the communication channel capacity can be added as domain knowledge to the standard label-based loss to guide scheduling decisions; in physics, the analytical expression of a partial differential equation can be utilized as domain knowledge on top of labeled data to better learn the solution to the equation given different inputs; more examples are shown in Section F. Such integration of explicit and rigorous domain knowledge can significantly benefit machine learning tasks (e.g., fewer labels needed than otherwise). Thus, it is crucial and being actively studied in informed machine learning (von Rueden et al., 2021; Willard et al., 2020), which is also the focus of our work. Note that using domain knowlege to generate pseudo labeled data to augment the training dataset is a special case of integrating domain knowledge into the training risk function (i.e., the knowledge-based risk is the same as the data-based risk, except that its labels are generated based on domain knowledge).

- *Final Hypothesis.* Domain knowledge can also be used for consistency check on the final learnt hypothesis or model (von Rueden et al., 2021). For example, (Karpatne et al., 2017) employs physics domain knowledge to construct the final model, (Pfrommer et al., 2018) builds simulators to validate results of learned model, and (Fang et al., 2017) leverages semantic consistency is used to refine the predicted probabilities.

F. Application Examples

We now present a few application examples to explain domain knowledge-informed DNNs.

F.1. Learning for resource management in communications networks

Optimizing resource management is crucial to improve the system performance in communications networks (Chiang et al., 2008; Goldsmith, 2005; Zappone et al., 2019). Well-known examples include power allocation (Hong & Luo, 2014; Chiang et al., 2008; Liang et al., 2019), link scheduling (Gore & Karandikar, 2010; Cui et al., 2019), antenna or beam selection (Sanayei & Nosratinia, 2004; Klautau et al., 2018), among others. While many of the problems were studied using theoretical model-based approaches in the past, machine learning has been increasingly employed, in view of the rapidly growing complexity of communications technologies that theoretical models are often incapable of capturing accurately (Zappone et al., 2019). Let us take power allocation in multi-user wireless interference networks as an example. The recent work (Sun et al., 2018) uses a pure data-driven approach for power allocation to maximize the sum rate: a labeled dataset containing channel state information (CSI) and the corresponding power allocation decisions is collected in advance, and a neural network is trained to learn the optimal power allocation. On the other hand, Shannon-based transmission rate has been extensively as an analytical objective function to optimize power allocation, and (Liang et al., 2019) exploits this domain knowledge to train an ensemble of neural networks that directly learn the optimal power allocation for Shannon rate maximization.

The data-driven approach (Sun et al., 2018) can maximize the practically achievable rate (if labels are collected from real systems), but is significantly constrained by the limited amount of training samples. Meanwhile, the knowledge-based approach (Liang et al., 2019) can utilize a large number of input samples (at the expense of higher training complexity), but the resulting power allocation decisions may not maximize the sum rate in real systems. The reason is that the Shannon formula for interference channels, albeit commonly used for analysis, only represents an approximation of the achievable rate which is subject to finite channel code lengths and modulation schemes (Goldsmith, 2005). In other words, even an oracle DNN that minimizes this knowledge-based loss may not maximize the achievable rate in practice.

To reap the benefit of both labeled data and domain knowledge, informed machine learning can be adopted, resulting in a new informed loss as follows:

$$\min_{h\in\mathcal{H}}(1-\gamma)\left\{\frac{1}{n}\sum_{(x,y)\in S}\left[h(x)-y\right]^2\right\}+\gamma\cdot\left\{-\frac{1}{\tilde{n}}\sum_{x\in\tilde{S}_X}\text{Shannon\_rate}[h(x)]+\text{constant}\right\}, \tag{53}$$

where x is the input (e.g., channel state information), $h(x)$ is the learned power allocation given x, the two loss terms represent label-based loss and knowledge-based loss, and n and $\tilde{n}$ are the numbers of labeled data samples and (possibly unlabled) knowledge samples, respectively. The detailed Shannon formula for wireless networks can be found in (Liang et al., 2019; Goldsmith, 2005).

F.2. Image classification based on semantic knowledge

Typical image classifiers rely on labeled training data, but labels can be difficult and expensive to collect in practice (Goodfellow et al., 2016). As a result, few-shot learning (Wang et al., 2020; Sung et al., 2018; Garcia & Bruna, 2018) that only needs a small number of labeled samples has been proposed. Informed machine learning under our consideration can be viewed as few-shot learning. Concretely, semantic knowledge formulated as the first-order logic clauses/sentences (Xu et al., 2018; Diligenti et al., 2017b) can be incorporated to improve learning performance given limited labeled samples. An example logic clause is "if it is an animal and has wings, then it is a bird". By a logic clause K, a knowledge-based loss can be defined as $F_K(h(x), g(x))$ for an (possibly unlabled) input image x and a certain logic clause $g(x)$ that the output class $h(x)$ needs to satisfy. Then, combining the standard label-based loss with knowledge-based loss, the model performance can be improved by minimizing the informed loss Eqn. (3) given limited labeled samples.

F.3. Learning to solve PDEs in scientific and engineering fields

Partial differential equations (PDEs) are classic problems in many scientific and engineering fields, such as physics and mechanical engineering, but are notoriously difficult to solve in most practical settings (Institute, 2020; Baker et al., 2019). In recent years, physics knowledge-informed machine learning has been suggested as a promising approach to augment or even

replace classic PDE solution approaches (Deng et al., 2020; Willard et al., 2020; Khoo et al., 2021; Beck et al., 2019; Raissi et al., 2019; Lu et al., 2021). For example, (Raissi et al., 2019; Lu et al., 2021) proposes a physics-informed neural network (PINN) to solve PDEs by minimizing the PDE residual and penalties of boundary/initial conditions, which correspond to the knowledge-based loss $F_K(h, g(h))$ in our framework. Additionally, we can combine the knowledge-based loss with labeled-based loss, achieving faster convergence and better performances in practice (especially when the PDE-based knowledge does not perfectly represent the real physical world). Take magnetic field strength estimation for magnetic materials as an example. If a few measured magnetic field strengths are provided as labels combined with the knowledge of Maxwell equations, the model trained by minimizing the informed loss can perform better in the real world. The measured labels can partly correct the imperfectness of physics knowledge, while the knowledge can improve the generalization in the presence of limited labels.

F.4. Knowledge distillation and transfer

Knowledge distillation (Hinton et al., 2014; Furlanello et al., 2018; Phuong & Lampert, 2019; Allen-Zhu & Li, 2020) is an important technique to transfer prior knowledge from a pre-trained neural network (a.k.a. teacher network) to another network (a.k.a. student network), with the same or different architectures. Typically, given an (possibly unlabled) input, knowledge distillation is performed by matching the output of the student network with the output of the teacher network. In addition, labeled samples can also be included to introduce a label-based loss. Thus, by formulating $g(X)$ as the output of the teacher network, knowledge distillation can be viewed as a particular instance of informed machine learning, where the knowledge comes from a teacher network and is usually assumed to be perfect.

G. Numerical Results

We consider two specific applications — learning a multi-dimensional Bohachevsky function and learning to manage wireless spectrum.

G.1. Settings of Learning with Constraint Knowledge in Section 7

We consider an informed DNN with domain knowledge in the form of constraints to learn a Bohachevsky function. The learning task is to learn a relationship $y(x)$. The learner is provided with a dataset with labeled samples $S_z = \{(x_i, z_i), i \in [n_z]\}$, having possibly noisy labels

$$z_i = y(x_i) + n_i, n_i \sim \mathcal{N}(0, \sigma_z^2),$$

and an unlabeled dataset $S_g = \{(x_i), i \in [n_g]\}$. Additionally, the learner is informed with the constraint knowledge, which includes an upper bound $g_{\rm ub}(x)$ and an lower bound $g_{\rm lb}(x)$ on the true label corresponding to input x, i.e. $g_{\rm lb}(x) \leq y(x) \leq g_{\rm ub}(x)$. A neural network $h_{\boldsymbol{W}}(x)$ is used to learn the relationship $y(x)$, and the metric of interest is the mean square error (MSE) of the network output $h_{\boldsymbol{W}}(x)$ with respect to the true label $y(x)$ on a test dataset S_t, which is expressed as

$$\hat{R}_{S_t}(h_{\boldsymbol{W}}) = \frac{1}{2|S_t|} \sum_{(x_i, y_i) \in S_t} \text{mse}\left(h_{\boldsymbol{W}}(x_i), y_i\right),$$

where $\text{mse}\left(h_{\boldsymbol{W}}(x_i), y_i\right) = \left(h_{\boldsymbol{W}}(x_i) - y_i\right)^2$ with y_i as the true test label with respect to x_i. Assume that the relationship to be learned is governed by a multi-dimensional Bohachevsky function

$$y(x) = x\boldsymbol{A}\boldsymbol{A}^\top x^\top - c\cos\left(a^\top x\right) + c,$$

where $\boldsymbol{A}$ is a $b \times b$ matrix, a is a b-dimensional vector and c is a constant. The learner has no access to the values of these parameters or the exact form of the relationship, but is empowered with the constraint knowledge in the form of an upper bound model

$$g_{\rm ub}(x) = x\boldsymbol{A}\boldsymbol{A}^\top x^\top + ub$$

with $ub \geq 2c$, and an lower bound model

$$g_{\rm lb}(x) = x\boldsymbol{A}\boldsymbol{A}^\top x^\top + lb.$$

with $lb \leq 0$. While it is not strongly convex and hence deviates from the assumptions in our theoretical analysis, we use ReLU as the knowledge-based risk function, i.e., the knowledge-based risk is written as

$$r_{\rm K}(h_{\boldsymbol{W}}(x)) = \text{relu}\left(h_{\boldsymbol{W}}(x) - g_{\rm ub}(x)\right) + \text{relu}\left(g_{\rm lb}(x) - h_{\boldsymbol{W}}(x)\right).$$

And the label supervised risk given a sample pair (x, z) is $r(h_{\boldsymbol{W}}(x)) = \mathrm{mse}\left(h_{\boldsymbol{W}}(x), z\right)$.

To show the performance under different levels of imperfectness, we consider labels with different noise variances and different knowledge-informed constraints. For training, the labeled dataset S_z contains $n_z \in \{200, 400\}$ labeled samples with label noise variance $\sigma_z^2 \in \{0, 0.1\}$, and the unlabeled dataset S_g for the knowledge risk contains $n_g = 1000$ input samples. The parameters for knowledge-informed constraint models include $lb = 0$ and $ub \in \{0.6, 0.8\}$. Naturally, the higher variance σ_z^2, the worse label quality; and the greater ub, the worse knowledge quality. The test dataset S_t contains 1000 samples with labels calculated as $y_i = y(x_i), x_i \in S_t$.

For training, we use a neural network with two hidden layers, each having 2048 neurons and ReLU activations. Note that for the large network width needed for analysis to gain insights is not necessary in practice. The network is initialized based on Algorithm A. The training procedure is performed by Adam optimizer for 3000 steps with batch size 100. The learning rate is set as 10^{-6} for the first 2000 steps, 5×10^{-5} for the following 500 steps, and 10^{-5} for the remaining 500 steps. We run the network training with 10 random seeds. We run the simulations on a HPC cluster with GPUs of type P100.

G.2. Learning for Resource Management in Wireless Networks

We apply an informed DNN to the problem of learning for resource management in wireless networks — wireless link scheduling in interference channels. We first describe problem setup, then present our method by informed DNN, and finally show the experiment results.

G.2.1. Problem Setup

Link scheduling is a classic and important problem in wireless interference channels, with the objective of maximizing the sum throughput of wireless links. Consider a time-slotted wireless network consisting of a transmitter-receiver set $\mathcal{U} = \{1, 2, \cdots, N\}$ with N links (i.e., transmitter-receiver pairs) subject to cross-link interference. At the beginning of each time slot, the scheduler needs to decide a subset of links $\mathcal{U}_{\mathrm{S}} \subseteq \mathcal{U}$ to transmit depending on the channel state information (CSI).

We assume Rayleigh fading channels with interference across different links. If a link $u \in \mathcal{U}$ is scheduled, the channel gain is $g_{u,u}$ subject to Rayleigh fading. For notational convenience, we omit the time slot index. Multiple links can be scheduled at the same time slot, creating interference to each other. For example, if link u and link v are scheduled simultaneously, the interference channel gain from the transmitter u to receiver v is $g_{u,v}$, and the interference channel gain from the transmitter v to receiver u is $g_{v,u}$. Thus, the received signal at receiver u can be expressed as $g_{u,u}s_u + \sum_{v \in \mathcal{U}_{\mathrm{S}}/u} g_{v,u}s_v + \mathrm{noise}_u$, where $\mathrm{noise}_u \sim \mathcal{N}\left(0, \sigma_{\mathrm{n}}^2\right)$ is an additive white Gaussian noise and the transmit signals s_u and s_v are normalized with unit power. Considering a centralized setting as in (Liang et al., 2019), the scheduler has access to the direct transmit channel gains as well as interference channel gains at the beginning of each time slot, which are contained in a $N \times N$ dimensional CSI vector $x = [g_{1,1}, \cdots, g_{1,N}, g_{2,1} \cdots, g_{N-1,N}, g_{N,1}, \cdots, g_{N,N}]$.

The scheduling decision can be represented by a N dimensional scheduling vector y. Specifically, if the link u is scheduled, then the u-th entry of y is one, and zero otherwise. By the Shannon rate formula in the communications theory (Goldsmith, 2005), the achievable rate for link u can be expressed as

$$C_{\mathrm{Shannon}}^{u}(x, y, \mu) = \log\left(1 + \frac{\mu y(u)\|g_{u,u}\|^2}{\sigma_{\mathrm{n}}^2 + \sum_{v \in \mathcal{U}/u} y(v)\|g_{v,u}\|^2}\right), \tag{54}$$

where $\mu(0, 1]$ is a parameter subject to real communication systems, with $C_{\mathrm{Shannon}}^{u}(x, y, 1)$ representing the standard Shannon rate (i.e., when $\mu = 1$). The sum rate is $C_{\mathrm{Shannon}}(x, y) = \sum_{u \in \mathcal{U}} C_{\mathrm{Shannon}}^{u}(x, y)$.

In practice, given the CSI vector x and the corresponding decision vector y, the real sum rate is denoted as $C_{\mathrm{real}}(x, y) = \sum_{u \in \mathcal{U}} C_{\mathrm{real}}^{u}(x, y)$. The real rate is difficult to express analytically in view of the complex factors in real environments including various schemes of modulation, finite channel coding and quality of service (QoS) guarantee In fact, except for a few special cases, the *exact* channel capacity for general interference channels (even for two links) is still an open problem. Thus, while the Shannon rate is useful and has been utilized to design various systems, it only represents an approximation of the practically achievable rate

Next, we formulate the link scheduling problem as

$$\max_{y} \sum_{u \in \mathcal{U}} C^{u}_{\text{real}}(x, y), \quad \text{s.t.} \quad y(u) \in \{0, 1\}, u \in \mathcal{U}. \tag{55}$$

The scheduling objective is the real sum rate in a practical environment. The challenge of this problem is that the real rate in terms of the CSI x and scheduling decision y is too complex to express precisely, let alone the longstanding challenges of deriving the exact interference channel capacity (Goldsmith, 2005).

G.2.2. Informed DNN for Wireless Link Scheduling

DNNs have strong representation power to learn the optimal scheduling decisions given CSI input (Sun et al., 2018), but they typically require a large number of labeled samples (x, y) for training. On the other hand, domain knowledge (i.e., Shannon rate formula) is also useful, but it may not capture the real achievable rate in practice (Liang et al., 2019). Thus, informed DNN, which exploits domain knowledge to complement labeled samples, has the potential to reap the benefits of both approaches.

Concretely, we use a DNN to represent the relationship between the scheduling decision y and CSI x. Given N links, the input dimension is $N \times N$, which is the dimension of vectorized CSI x and the scheduling decision y is a $N-$dimensional binary vector. The training is based on a labeled dataset $S_y = \{(x_i, y_i), i = 1, 2, \cdots, n_y\}$ collected from real systems or field studies, where y_i is the true label (i.e., optimal scheduling decision) given x_i, along with the domain knowledge of Shannon rate. Also, we use $Y_{\text{comb}} \in \{0, 1\}^{I_{\max} \times N}, I_{\max} = 2^N - 1$ to represent all the possible decision combinations. Denote $I(y)$ as the index of a scheduling decision y in Y_{comb}, i.e. $y = [Y_{\text{comb}}]_{I(y)}$. The output dimension of the DNN is $I_{\max} = 2^N - 1$ with each entry representing an index for a scheduling decision.

The label-based risk is the cross-entropy loss between the output of the DNN and one-hot encoding labels, which is expressed as

$$\hat{R}_{S_y}(\boldsymbol{W}) = \frac{1}{n_y} \sum_{i=1}^{n_y} \text{cross\_entropy}\left(\text{softmax}(h_{\boldsymbol{W}}(x_i)), \text{one\_hot}(I(y_i))\right), \tag{56}$$

where $\text{one\_hot}(I(y_i))$ is the one-hot encoding of the index of y_i. Given an CSI input x and setting $\mu = \mu_{\text{K}}$ based on domain experience, we can compute the sum rate of all possible scheduling decisions by the Shannon equation in Eqn. (54) as $C_{\text{Shannon}}(x, [Y_{\text{comb}}]_j, \mu_{\text{K}}), j \in [I_{\max}]$ and get the vector of sum rate as $\boldsymbol{c}(x) = \left[C_{\text{Shannon}}\left(x, [Y_{\text{comb}}]_1, \mu_{\text{K}}\right), \cdots, C_{\text{Shannon}}\left(x, [Y_{\text{comb}}]_{I_{\max}}, \mu_{\text{K}}\right)\right]$. Taking the softmax operation on $T\boldsymbol{c}(x)$ with T as a scaling hyper-parameter, we get $\text{softmax}\left(T\boldsymbol{c}(x)\right)$, which is essentially soft encoding of scheduling decisions based on the Shannon rate knowledge. Therefore, given an input dataset $S_g = \{x_i, i = 1, 2, \cdots, n_g\}$, the knowledge-based risk is designed as

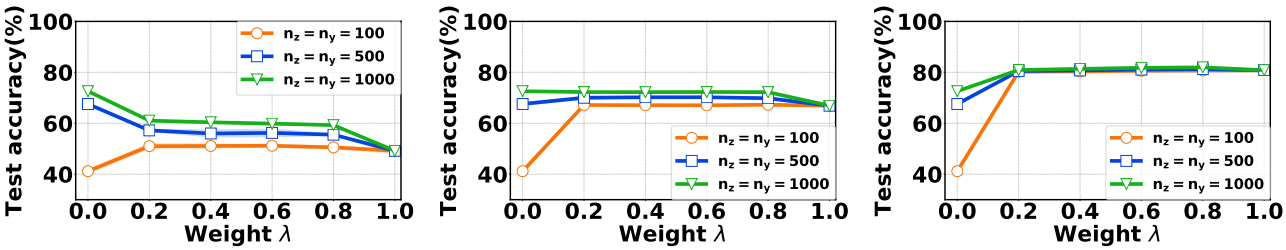
$$\hat{R}_{\text{K}}(\boldsymbol{W}) = -\frac{1}{n_g} \sum_{i=1}^{n_g} \text{cross\_entropy}\left(\text{softmax}(h_{\boldsymbol{W}}(x_i)), \text{softmax}\left(T\boldsymbol{c}(x)\right)\right). \tag{57}$$

Thus, the DNN can be trained to minimize the informed risk combining both label-based and knowledge-based risks: $\hat{R}_{\text{I}}(\boldsymbol{W}) = (1-\lambda)\hat{R}_{S_y}(\boldsymbol{W}) + \lambda \hat{R}_{\text{K}}(\boldsymbol{W})$. That is, the informed DNN uses hard labels for direct supervision, while exploiting domain knowledge in the form of soft labels for indirect supervision on unlabeled inputs. After training the network, the scheduling decision for CSI x is calculated as $y_{\boldsymbol{W}}(x) = [Y_{\text{comb}}]_{I_{\boldsymbol{W}}(x)}$ with $I_{\boldsymbol{W}}(x) = \arg\max_{j \in [I_{\max}]} \left[h_{\boldsymbol{W}(x)}\right]_j$

G.2.3. Results

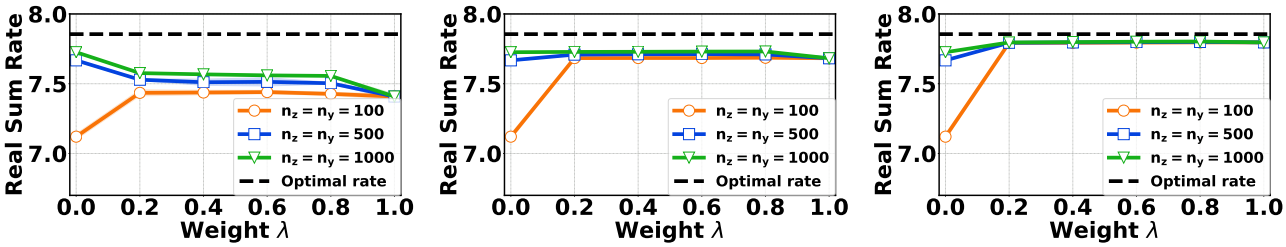
Now, we show the simulation results for the wireless link scheduling problem based on our informed DNN. We first give the simulation settings and then show the results of classification accuracy as well as the sum rate.

Simulation Settings. For illustration, we consider a simulation scenario with $N = 4$ wireless links for scheduling, which is a reasonable setting for many practical ad hoc networks (Goldsmith, 2005). Given the CSI, the scheduler needs to choose one out of 15 scheduling combinations. To evaluate the performance of our informed DNN when the domain knowledge of Shannon rate is not perfect, we construct a synthetic dataset as the ground truth. The direct link channel gain of a wireless link $g_{u,u}$ is subject to Rayleigh distribution, with an expected power gain of 100 dB. The cross-link interference channel gain is also subject to Rayleigh distribution with an expected power gain of 10 dB. The labels in the labeled training dataset



(a) $\mu_{\mathrm{K}} = 0.1$ (b) $\mu_{\mathrm{K}} = 1.0$ (c) $\mu_{\mathrm{K}} = 0.4$

Figure 2. Test accuracy under different knowledge qualities and numbers of labels.



(a) $\mu_{\mathrm{K}} = 0.1$ (b) $\mu_{\mathrm{K}} = 1.0$ (c) $\mu_{\mathrm{K}} = 0.4$

Figure 3. Sum rate under different knowledge qualities and numbers of labels.

and test dataset are generated by a pseudo-real rate expression to reflect some practical constraints:

$$C_{\mathrm{pseudo-real}} = C^{u}_{\mathrm{Shannon}}(x, y, \mu_{\mathrm{R}}), \tag{58}$$

which differs from the standard Shannon formula by using a factor $\mu_{\mathrm{R}} \in (0, 1)$ to account for achievable rate degradation. Note that the pseudo-real rate is only defined to generate synthetic real rate different from the standard Shannon rate for evaluation purposes. In practice, the achievable rate is even more complex. In the simulations, we set $\mu_{\mathrm{R}} = 0.5$ to generate the training and testing labels as ground truth, while the value of $\mu_{\mathrm{R}} = 0.5$ is not available to the learner.

Based on the pseudo-real rate expression, we find the optimal labels (i.e., optimal scheduling decision y) via exhaustive search, while labels are actually be collected by field measurement in a practical environment. We have $n_g = 2000$ unlabeled CSI input samples in the training dataset S_g for knowledge-based supervision, and $n_t = 10000$ samples in the test dataset. The test accuracy is defined as the percentage of DNN outputs that are identical to the optimal scheduling decision label, i.e. for samples in the test dataset, $\mathrm{acc} = \sum_{i=1}^{n_t} \mathbb{1}\left(I_{\boldsymbol{W}}(x_i) = I(y_i)\right)/n_t$. We compare the results when the labeled training dataset has 100, 500 and 1000 samples, respectively. Also, we compare the results obtained by setting different parameters $\mu_{\mathrm{K}} \in \{1.0, 0.4, 0.1\}$ in the knowledge-based Shannon rate in Eqn. (54). The parameter $\mu_{\mathrm{K}} \in \{1.0, 0.4, 0.1\}$ results in a test accuracy of $\{71.4\%, 91.2\%, 52.8\%\}$, which is the maximum test accuracy obtained by directly solving the scheduling problem based on Eqn. (54) and can be used to informally indicate the knowledge quality. Thus, $\mu_{\mathrm{K}} = 0.4$ represents the best knowledge quality, whereas $\mu_{\mathrm{K}} = 0.1$ is the worst.

Now we list the settings for training. The neural network has three hidden layers with 512, 1024 and 512 neurons, respectively, followed by ReLu activations. The network is initialized based on Algorithm 1. The training is performed by the Adam optimizer with learning rate 10^{-5} for 2000 steps on a HPC cluster with GPU type P100. We use 5 random seeds for each setting to evaluate the performance error.

Results. The results, including the test accuracy and the test sum rate under different knowledge quality, numbers of labels and weights λ, are shown in Fig. 2 and Fig. 3. The test sum rate is the (pseudo) real sum rate defined in Eqn. (58) with $\mu_{\mathrm{R}} = 0.5$. We can find that the sum rate expectedly increases if the test accuracy increases. From Fig. 2(a) and Fig. 3(a), we see that if the domain knowledge quality is only 52.8% (i.e., $\mu_{\mathrm{K}} = 0.1$), it has bad effects on learning when labels are enough. Nevertheless, it still benefits the performance when there are only 100 labels and, if we place a less weight on the knowledge-based risk, the accuracy and sum rate is higher.

If the knowledge quality is 71.4% (i.e., $\mu_{\mathrm{K}} = 1.0$), as shown in Fig. 2(b) and Fig. 3(b), the domain knowledge has significant

benefits when there are only 100 labeled samples. When there are 500 labeled samples, the domain knowledge and labels complement each other and get a better performance than pure label-based and knowledge-based learning. When the number of labeled samples is even higher and reaches 1000, the integration of domain knowledge cannot benefit the learning further. In Fig. 2(c) and Fig. 3(c), when the domain knowledge quality further improves, we can see that the domain knowledge can still bring benefits even in the presence of 1000 labeled samples.

From these results, we see that labels and domain knowledge can complement each other. The domain knowledge plays an important role when labels are relatively scarce, while labels, even only a few, help improve the learning performance when domain knowledge has a low quality. Additionally, it is important to achieve a balance between label-based supervision and knowledge-based supervision. In general, we place more weight on the knowledge-based risk if knowledge quality is good enough and the number of labels is small, and vice versa.