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


**Jianyi Yang and Shaolei Ren**

**University of California, Riverside**



**ICML 2022**


ICML
International Conference
On Machine Learning

UC
RIVERSIDE

# Ways to Integrate Knowledge in Machine Learning

- **Augment training data**
  - Generate data by demonstration
  - Image pre-processing (cropping, flipping, scaling )
- **Determine the learning architecture ( hypothesis set )**
  - Attention mechanism in transformer
  - Neural architecture based on knowledge graph
- **Calibrate the machine learning output**
  - Refine the predicted results by consistency check
- ***Supervise model training (considered)***
  - Pseudo label Generation
  - Knowledge distillation
  - Learning to solve PDEs
  - Weakly-supervised learning
  - Learning using Privileged Knowledge
  - PAC-Bayesian Learning

Knowledge-supervised risk

$$\hat{R}_{\mathrm{I}}(h) = \frac{1-\lambda}{n_z}\sum_{S_z} r(h(x_i), z_i) + \frac{\lambda}{n_g}\sum_{S_g} r_{\mathrm{K}}(h(x_j), g_j)$$

Label-supervised risk

$h$: learning model

$S_z$: labeled dataset

$S_g$: knowledge-supervised dataset

$\lambda$: tradeoff weight

# Overview

- A novel proof of convergence jointly determined by labels and knowledge.
- A metric to measure the imperfectness of knowledge and labels.
- Generalization bound related to the imperfectness.
- Effects and benefits of integrating knowledge.
- A design of a generalized informed training objective.
- Sampling complexity of informed learning under different settings.

# Convergence of Informed Learning by Wide Neural Networks

- **Inapplicability of current convergence analysis of wide neural networks**
  - multiple supervisions.
  - Strong data separability assumptions.
- **The concept of smooth sets and a new data-separability assumption**
  - **Data separability assumption (Informal):** At *initialization*, for each smooth set $k$, and each sample $i$ in this smooth set, neurons at last hidden layer satisfy

$$\mathrm{sign}\left([h_L^{(0)}(x_i)]_j\right) = \mathrm{sign}\left([h_L^{(0)}(x_k')]_j\right) \quad or \quad \left|\left[W_L^{(0)} h_{L-1}^{(0)}(x_i)\right]_j\right| \geq \frac{3\sqrt{2\pi}\phi^{b+1}}{16\sqrt{m}}$$

- **Convergence Theorem**

**Theorem 1 (Informal):** When the network width $m$ is large enough and the size of smooth sets $\phi$ is small enough, if the **data separability assumption** is satisfied, after enough training rounds, the informed training risk $\hat{\boldsymbol{R}}_{\mathbf{I}}(\boldsymbol{W})$ converges to the ***effective risk*** $\hat{\boldsymbol{R}}_{\mathbf{eff}}(\boldsymbol{W})$ plus a small error $\epsilon$, and the network output $\boldsymbol{h}_{\boldsymbol{W}}(\boldsymbol{x}_i)$ converges to ***effective labels*** $\boldsymbol{y}_{\mathbf{eff},\boldsymbol{k}(\boldsymbol{x}_i)}$

- **Effective label:** for the $k$th smooth set, effective label is defined to minimize the total risk in the smooth set:

$$y_{\mathrm{eff},k} = \arg\min_h \sum_{i\in\mathcal{I}_{\phi,k}} \{\mu_i r(h, z_i) + \lambda_i r_K(h, g_i)\}$$

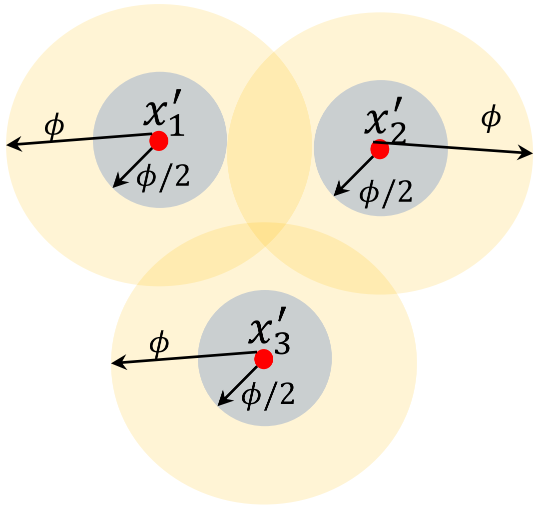


Figure 1. Illustration: Smooth sets discretize the input space.
If an input $x$ belongs to a ***smooth set*** $k$, then

$$||x - x_k'|| \leq \phi, ||x - x_j'|| \geq \frac{\phi}{2}, \forall j \neq k.$$

# Generalization and Effects of Knowledge

## ➤ Definition of imperfectness (Informal)

- Knowledge imperfectness

  $Q_{\mathrm{K}} = R(h_K^*)$ $h_K^*$: optimal neural network purely supervised by knowledge

- knowledge regularized label imperfectness.

  $Q_{\mathrm{R}} = R(h_{R,\beta}^*)$ $h_{R,\beta}^*$: optimal neural network supervised by labels and knowledge regularization with regularization strength $\beta$

## ➤ Generalization Bound (Informal)

With convergence assumptions satisfied and small enough $\phi$, the population risk is bounded with probability at least $1 - O(\phi) - \delta$ as

Knowledge imperfectness

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

Training loss

Knowledge-regularized label imperfectness

Error from data finiteness

## ➤ Effects of Knowledge

- An explicit regularization for label-based supervision.
- A (possibly imperfect) supplement for labels.

# A Generalized Training Objective

## ➤ Generalized Informed Training Objective

$$\hat{R}_{\mathrm{I,G}} = \frac{(1-\lambda)(1-\beta)}{n_z}\sum_{S_z} r(h(x_i), z_i) + \frac{(1-\lambda)\beta}{n'_g}\sum_{S'_g} r_{\mathrm{K}}(h(x_i), g_i) + \frac{\lambda}{n''_g}\sum_{S''_g} r_{\mathrm{K}}(h(x_i), g_i)$$

Label-based supervision

Knowledge supervision for regularization

Knowledge supervision for data supplementing

■ Introduce another hyper-parameter $\beta$ to control the knowledge regularization strength.

## ➤ Generalization Bound (Informal)

With convergence assumptions satisfied and small enough $\phi$, the population risk is bounded with probability at least $1 - O(\phi) - \delta$ as

Knowledge imperfectness

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

Training loss

Knowledge-regularized label imperfectness

Error from data finiteness

■ Take-aways: generalized informed risk is more flexible to adjust how much knowledge is incorporated when it plays different effects ($\beta$ to adjust knowledge regularization, $\lambda$ to balance the two effects ).

# Sampling Complexity

**Theorem of sampling complexity (Informal):**

To achieve a population risk less than $\sqrt{\epsilon}$.

- **(a) Knowledge is nearly perfect. ($\boldsymbol{Q_K \leq \sqrt{\epsilon}}$)**

  Sampling complexity for labeled data: $n_z$=0;
  Sampling complexity for Knowledge-supervision: $n_g \sim O(1/(\epsilon^2 - \epsilon^3))$

Labels are usually more expensive. Only use knowledge supervision when knowledge is very good.

---

- **(b) Knowledge is imperfect ($\boldsymbol{Q_K > \sqrt{\epsilon}}$), but labels are good enough ($\boldsymbol{\frac{\sqrt{\epsilon}}{Q_K} + \frac{\sqrt{\epsilon}}{Q_R} \geq 1}$).**

  Sampling complexity for labeled data: $n_z \sim \boldsymbol{O}\left(\left(\frac{1}{\epsilon} - \frac{1}{\sqrt{\epsilon} Q_K}\right)^2\right)$;
  Sampling complexity for Knowledge-supervision: $n_g \sim \boldsymbol{O}\left(\frac{1}{(\epsilon - \epsilon^2) Q_K^2}\right)$

The incorporation of knowledge is equivalent to $\boldsymbol{O(2/(\epsilon^{1.5} Q_K) - 1/(\epsilon Q_K^2))}$ labeled samples

---

- **(c) Knowledge and labels are both of low quality ($\boldsymbol{\frac{\sqrt{\epsilon}}{Q_K} + \frac{\sqrt{\epsilon}}{Q_R} < 1}$).**

  A generalization error less than $\sqrt{\epsilon}$ cannot be achieved.

A requirement for knowledge and label imperfectness to achieve low risk.

# Thank you!

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

**Jianyi Yang and Shaolei Ren**

jyang239@ucr.edu sren@ece.ucr.edu

**Poster： Hall E #300**