REVISITING LOCAL NEIGHBORHOOD METHODS IN MACHINE LEARNING

Sarath Shekkizhar, Antonio Ortega

Department of Electrical and Computer Engineering
University of Southern California, Los Angeles
email: shekkizh@usc.edu, aortega@usc.edu

ABSTRACT

Several machine learning methods leverage the idea of locality by using k-nearest neighbor (KNN) techniques to design better pattern recognition models. However, the choice of KNN parameters such as k is often made experimentally, e.g., via cross-validation, leading to local neighborhoods without a clear geometric interpretation. In this paper, we replace KNN with our recently introduced polytope neighborhood scheme - Non Negative Kernel regression (NNK). NNK formulates neighborhood selection as a sparse signal approximation problem and is adaptive to the local distribution of samples in the neighborhood of the data point of interest. We analyze the benefits of local neighborhood construction based on NNK. In particular, we study the generalization properties of local interpolation using NNK and present data dependent bounds in the non asymptotic setting. The applicability of NNK in transductive few shot learning setting and for measuring distance between two datasets is demonstrated. NNK exhibits robust, superior performance in comparison to standard locally weighted neighborhood methods.

Index Terms— Neural networks, polytope interpolation, local methods, generalization, leave one out, k-nearest neighbor.

1. INTRODUCTION

Local neighborhood methods such as k-nearest neighbor (KNN) [1] and Nadarya-Watson estimator (weighted k-nearest neighbor estimator) [2, 3] are some of the most popular non-parametric learning methods with application in density estimation, classification and regression [4]. What is meant by *local* in these methods is based on the choice of an appropriate feature space for data representation and a similarity kernel or distance. Thus, the use of a weighted KNN method requires a careful choice of k and of the weights assigned to the selected neighbors. Theoretical results in [4, 5, 6] suggest that the value of k in the asymptotic regime, where the number of samples (N) goes to infinity, should be such that $k \to \infty$ and $k/N \to 0$. In practice, for N finite, [7, 8] recommend k to be set to a fractional power of the dataset size. This approach can yield poor performance and, as a general rule of thumb, the choice of k is often made using task specific cross validation. Other works such as [9, 10] make use of labels in the training dataset to obtain an adaptive choice of k. However, these methods lack a geometrical interpretation relating the resulting *locality* (e.g., the value of k obtained via cross validation) and the intrinsic dimension of the data samples. Further, these methods do not extend to scenarios where no label information exists. Nevertheless, the simplicity and empirical success of local neighborhood methods such as KNN makes them a popular choice in machine learning. They can even be used for modern deep learning systems, with the goal of achieving regularized classification models [11, 12], as well as semi-supervised [13, 14] and unsupervised [15] learning systems, amongst others.

This paper takes as a starting point our recently proposed local neighborhood construction, Non Negative Kernel regression (NNK) [16], and explores novel applications of local methods. NNK finds a first approximation of the neighborhood using KNN, but instead of using the resulting points directly, it optimizes and reweighs this selection, leading to a sparser and stable set of neighbors having a geometric interpretation. In this work, we leverage the geometrical properties of the NNK solution to theoretically bound its generalization from the Bayes estimator and its leave-one-out estimate. Our analysis makes explicit the relationship between generalization, the smoothness of the functional values (e.g., labels) at nearby points, and the local distribution of data. Experimentally, we evaluate the application of locally weighted NNK estimators in a transductive few shot learning scenario [17] and propose a new distance measure in the space of datasets based on the theoretical properties of NNK[1].

2. PRELIMINARIES AND BACKGROUND

2.1. Notation

Throughout the paper, lowercase (e.g., x and θ), lowercase bold (e.g., $\boldsymbol{x}$ and $\boldsymbol{\theta}$), uppercase (e.g., X and Y), and uppercase bold (e.g., $\boldsymbol{K}$ and $\boldsymbol{\Phi}$) letters denote scalars, vectors, random variables, and matrices, respectively. We use $\mathbb{I} : \{0,1\} \to \{0,1\}$ to indicate the truth value of an expression. $\mathcal{D}_{train} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \dots (\boldsymbol{x}_N, y_N)\}$ is the set of training data obtained from an unknown distribution in $X \times Y$ and $\mathcal{D}^i_{train}$ the set obtained by removing the point $(\boldsymbol{x}_i, y_i)$ from $\mathcal{D}_{train}$. We use μ to represent the marginal distribution of X, supp(μ) its non zero support in X, and $\eta(\boldsymbol{x}) = \mathbb{E}(Y|X = \boldsymbol{x})$ the conditional mean of Y. The risk or generalization error associated with a function $\hat{\eta} : X \to Y$ is represented as $\mathcal{R}_{gen}(\hat{\eta}) = \mathbb{E}[l(\hat{\eta}(\boldsymbol{x}), y)]$, where $l(\hat{\eta}(\boldsymbol{x}), y)$ corresponds to the error between actual and estimated value of y.

2.2. NNK interpolation

In NNK, neighborhood selection is formulated as a signal representation problem, where each data point is to be approximated using a dictionary formed by its neighbors [16]. This problem formulation leads, for each data point, to an adaptive and principled approach to the choice of neighbors and their weights. While KNN is used as

Our work was supported by a grant under DARPA's LwLL program (FA8750-19-2-1005)

[1] A longer version of the work is posted on ArXiv.org with experiments in additional scenarios, such as explainability and model selection in deep learning models, demonstrating robustness and superior performance of NNK over conventional KNN-based approaches [18]

an initialization, NNK performs an optimization akin to orthogonal matching pursuit [19] in kernel space resulting in a *stable* representation with a *geometric* interpretation. The Kernel Ratio Interval (KRI) theorem in [16] states, for a given data point i and similarity kernel $\boldsymbol{K} \in [0, 1]$, a necessary and sufficient condition for *both* j and k to be NNK neighbors of i:

$$\boldsymbol{K}_{j,k} < \frac{\boldsymbol{K}_{i,j}}{\boldsymbol{K}_{i,k}} < \frac{1}{\boldsymbol{K}_{j,k}}. \tag{1}$$

Geometrically, KRI reduces to a series of hyper plane conditions, one per NNK neighbor, which applied inductively lead to a convex polytope around each data point $\boldsymbol{x}$, denoted $\text{NNK}_{poly}(\boldsymbol{x})$. Our proposed unbiased NNK interpolation at $\boldsymbol{x}$ is defined as

$$\hat{\eta}(\boldsymbol{x}) = \sum_{i \in \text{NNK}_{poly}(\boldsymbol{x})} \frac{\theta_i \; y_i}{\sum_{j \in \text{NNK}_{poly}(\boldsymbol{x})} \theta_j} \tag{2}$$

where $\text{NNK}_{poly}(\boldsymbol{x})$ is the convex polytope formed by $\hat{k}$ neighbors identified by NNK and $\boldsymbol{\theta}$ denotes a $\hat{k}$ length vector of non zero values obtained from the solution to the data approximation objective, i.e.,

$$\begin{aligned} \boldsymbol{\theta}^* &= \min_{\boldsymbol{\theta}_S \geq 0} ||\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\Phi}_S \boldsymbol{\theta}_S||^2 \\ &= \min_{\boldsymbol{\theta}_S \geq 0} \boldsymbol{K}_{*,*} - 2\boldsymbol{\theta}_S^\top \boldsymbol{K}_{S,*} + \boldsymbol{\theta}_S^\top \boldsymbol{K}_{S,S} \boldsymbol{\theta}_S \end{aligned} \tag{3}$$

where $\boldsymbol{\Phi}_S = [\boldsymbol{\phi}(\boldsymbol{x}_1) \dots \boldsymbol{\phi}(\boldsymbol{x}_k)]$ corresponds to the kernel space representation of the k nearest neighbors of $\boldsymbol{x}$. $\boldsymbol{K}_{S,*}$ corresponds to the kernel evaluated between the neighbors (set S) and $\boldsymbol{x}$.

3. THEORETICAL ANALYSIS OF LOCAL INTERPOLATION WITH NNK

3.1. A general bound on NNK classifier

In this section, we study the generalization risk associated with the NNK estimator of equation (2) under a general assumption of smoothness. Our analysis follows a similar setup and proof style as the simplicial interpolation analysis in [20], but adapted to NNK interpolation. Note that simplicial interpolation [20] is impractical for high dimensional data, a typical setting in modern machine learning, while a simpler method such as KNN does not have the geometric properties required for our analysis. Further, a simplicial interpolation, even when feasible, leads to an arbitrary choice of the containing simplex when data lies on one of the simplicial faces. This situation becomes increasingly common in high dimensions, worsening interpolation complexity. By relaxing the simplex constraint of [20] to convex polytope structures, such as those obtained using NNK, we obtain robust interpolation estimates that are dependent on the intrinsic dimension of the space around each training data.

In summary, NNK combines some of the best features of existing methods, providing a theoretical interpretation and performance guarantees as the simplicial interpolation [20], while being practical and realizable with a complexity comparable to that of KNN-based schemes. We first study NNK in a regression setting and then extend the results for classification[2]. We assume each y_i is corrupted by independent noise and hence can deviate from the Bayes optimal estimate $\eta(\boldsymbol{x}_i)$. Note that the result does not make specific assumptions about y and holds for any signal (class label, cluster or set membership) associated with each data point $\boldsymbol{x}$.

[2]All proofs related to theoretical statements in this section are included in the supplementary material

In a **regression** setting, the generalization error of function $\hat{\eta}$ is given by the mean squared error, i.e., $\mathcal{R}_{gen}(\hat{\eta}) = \mathbb{E}[(\hat{\eta}(\boldsymbol{x}) - y)^2]$. Statistically, the Bayes estimate corresponding to the conditional mean $\eta(\boldsymbol{x})$ is the optimal predictor and bounds other estimators as $\mathbb{E}[\mathcal{R}(\hat{\eta}, \boldsymbol{x}) - \mathcal{R}(\eta, \boldsymbol{x})] \leq \mathbb{E}[(\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x}))^2]$. In Theorem 1, we present a data dependent bound on the excess risk of NNK as compared to the Bayes estimator, in a non-asymptotic setting.

Theorem 1. *For a conditional distribution $\hat{\eta}(\boldsymbol{x})$ obtained using unbiased NNK interpolation given training data $D_{train} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \dots (\boldsymbol{x}_N, y_N)\}$ in $\mathbb{R}^d \times [0, 1]$, the excess mean square risk is given by*

$$\begin{aligned} \mathbb{E}[(\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x}))^2 | D_{train}] &\leq \mathbb{E}[\mu(\mathbb{R}^d \backslash \mathcal{C})] + A^2 \mathbb{E}[\delta^{2\alpha}] \\ &+ \frac{2A'}{\mathbb{E}[\hat{k}] + 1} \mathbb{E}[\delta^{\alpha'}] + \frac{2}{\mathbb{E}[\hat{k}] + 1} \mathbb{E}[(Y - \eta(\boldsymbol{x}))^2] \end{aligned} \tag{4}$$

under the following assumptions

1. *μ is the marginal distribution of $X \in \mathbb{R}^d$ and $\mathcal{C} = Hull(\boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2) \dots \boldsymbol{\phi}(\boldsymbol{x}_N))$ is the convex hull of the training data in transformed kernel space.*
2. *The conditional distribution η is Holder (A, α) smooth in kernel space.*
3. *The conditional variance $var(Y|X = \boldsymbol{x})$ satisfies (A', α') smoothness condition.*
4. *The maximum diameter of the polytope formed with NNK neighbors for any data in $\mathcal{C}$ is represented as $\delta = \max_{\boldsymbol{x} \in \mathcal{C}} diam(NNK_{poly}(\boldsymbol{x}))$, where $NNK_{poly}(\boldsymbol{x})$ denotes the convex polytope around $\boldsymbol{x}$ formed by $\hat{k}$ neighbors of NNK.*

Remark 1. The first term in the bound corresponds to extrapolation, where the test data falls outside the interpolation area (i.e., outside of the convex hull of points, $\mathcal{C}$) while the last term corresponds to label noise. The remaining terms capture the dependence of the interpolation on the size of each polytope defined for test data, and the smoothness of the y_i's over this region (i.e., within each polytope). Note that a smaller δ, arising when test samples are covered by smaller polytopes, leads to a risk closer to optimal. This is important because NNK leads to a polytope having smallest diameter or volume among all polytopes obtained with exactly $\hat{k}$ points chosen from the k nearest neighbors (as guaranteed by the conditions of (1)). From the theorem, this corresponds to a better risk bound. The bound associated with simplicial interpolation is a special case, where each simplex enclosing the data point is a fixed size polytope containing $d + 1$ vertices. Thus, in our approach the number of points (neighbors) forming the polytope is variable (dependent on local data topology), while in the simplicial case it is fixed and depends on the dimension of the space. Though the latter bound seems better (excess risk is inversely related to $\hat{k}$), the diameter of a simplex increases with d making the excess risk possibly sub optimal compared to NNK.

Corollary 1.1. *Based on an additional assumption that supp(μ) belongs to a convex and bounded region of $\mathbb{R}^d$, the excess mean square risk converges asymptotically as*

$$\limsup_{N \to \infty} \mathbb{E}[(\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x}))^2] \leq \mathbb{E}[(Y - \eta(\boldsymbol{x}))^2] \tag{5}$$

Remark 2. The asymptotic risk of NNK interpolation method in the regression setting is bounded, similar to the 1-nearest neighbor method, by twice the Bayes risk. The rate of convergence of the proposed method depends on the kernel function: how *close* two data

points need to be for them to be indistinguishable depends on the parameters chosen for the similarity kernel.

Now, we turn our attention to a **binary classification** setting, where the domain of Y is reduced to $\{0, 1\}$. The risk associated to a classifier $\hat{f}$ is defined as $\mathcal{R}_{gen}(\hat{f}) = \mathbb{E}[P(\hat{f}(\boldsymbol{x}) \neq y)]$. Similar to regression, this risk can be associated with that of the Bayes optimal classifier $f^* = \mathbb{I}(P(Y = 1|X = \boldsymbol{x}) > 0.5)$ as $\mathbb{E}[\mathcal{R}(\hat{\eta}, \boldsymbol{x}) - \mathcal{R}(f^*(\boldsymbol{x}), \boldsymbol{x})] \leq \mathbb{E}[P(\hat{f}(\boldsymbol{x}) \neq f^*(\boldsymbol{x}))]$.

Corollary 1.2 presents a bound on the excess risk associated with the plug-in NNK classifier $\hat{f}(\boldsymbol{x}) = \mathbb{I}(\hat{\eta}(\boldsymbol{x}) > 0.5)$ for $\mathcal{D}_{train} = \{(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_N, y_N)\}$ in $\mathbb{R}^d \times \{0, 1\}$ using Corollary 1.1 and the relationship between classification risk and regression risk [4].

Corollary 1.2. *A plug-in NNK classifier under the assumptions of Corollary 1.1 has excess classifier risk bounded as*

$$\limsup_{N \to \infty} \mathbb{E}[\mathcal{R}(\hat{f}(\boldsymbol{x})) - \mathcal{R}(f(\boldsymbol{x}))] \leq 2\sqrt{\mathbb{E}[(Y - \eta(\boldsymbol{x})^2]} \quad (6)$$

Remark 3. The classification bound presented here makes no assumptions on the margin associated to the classification boundary and is thus only a weak bound. The bound can be improved exponentially as in [20] with stronger assumptions such as h-hard margin boundary condition [21].

3.2. Leave one out stability

The leave one out (LOO) procedure (also known as deleted estimate or U-method) is an important statistical measure with a long history in machine learning [22]. Unlike empirical error, it is *almost unbiased* [23] and is often used for model (hyperparameter) selection. The LOO error associated with NNK interpolation is given by

$$\mathcal{R}_{loo}(\hat{\eta}|\mathcal{D}_{train}) = \frac{1}{N} \sum_{i=1}^{N} l(\hat{\eta}(\boldsymbol{x}_i)|\mathcal{D}^i_{train}, y_i) \quad (7)$$

where the NNK interpolation estimator in the summation for $\boldsymbol{x}_i$ is based on all training points except $\boldsymbol{x}_i$.

Theoretical results by Rogers, Devroye and Wagner [24, 25] about generalization of k-nearest neighbor methods using LOO performance are relevant to our proposed NNK algorithm. The number of neighbors $\hat{k}$ around each data point $\boldsymbol{x}$ in our method is dependent on the local distribution of data and replaces the fixed k from their results by an expected value, $\mathbb{E}[\hat{k}]$.

Theorem 2. *The leave one out performance of the NNK interpolation classifier given γ, the maximum number of distinct points that can have the same nearest neighbor, is bounded as*

$$P(|\mathcal{R}_{loo}(\hat{\eta}|\mathcal{D}_{train}) - \mathcal{R}_{gen}(\hat{\eta})| > \epsilon) \leq 2e^{-N\epsilon^2/18} + 6e^{-N\epsilon^3/\left(108\ \mathbb{E}[\hat{k}](2+\gamma)\right)}$$

Remark 4. The value of γ is common to both KNN and NNK settings and is dependent on the dimension of the space where the data is embedded. The exact evaluation of γ is difficult in practice but bounds do exist for this measure in the sphere covering literature [26, 27]. The theorem allows us to relate the LOO risk of a model to the generalization error. Unlike the bound based on hyperparameter k in KNN methods, the bound for NNK is adaptive to the training data, capturing the distribution characteristics of the dataset.

4. NEIGHBORHOOD METHODS IN MACHINE LEARNING

We briefly group nearest neighbor methods in machine learning into three categories based on the type of data (**L**abeled, **UnL**abeled) at the decision point $\boldsymbol{x}$ and associated neighbors forming the polytope $\text{NNK}_{poly}(\boldsymbol{x})$ in Table 1. In this work, we center our experiments

Data − Neighbor	Applications	Reference
L − L	Generalization and robustness analysis, Curriculum learning	Section 3, [18]
UL − L	Semi Supervised Learning, Transductive learning, Explainable predictions	Section 4.1
UL − UL	Clustering, Two sample statistic, Distance between datasets	Section 4.2

Table 1: Overview of local neighborhood methods based on the availability of labels (**L**abeled, **UnL**abeled) at data point and corresponding neighbors with few applications in each category. The last column in Table links to relevant materials in our work corresponding to the setting in each group.

around the range normalized cosine kernel defined as,

$$\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2}\left(1 + \frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{\|\boldsymbol{x}_i\|\ \|\boldsymbol{x}_j\|}\right) \quad (8)$$

though our theoretical statements and claims make no assumption on the type of kernel, other than it be positive with range $[0, 1]$. In particular, we transform the input data using the non linear mapping $\boldsymbol{h_w}$ corresponding to deep neural networks (DNN) parameterized by $\boldsymbol{w}$ to modify our kernel definition as

$$\mathcal{K}_{DNN}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2}\left(1 + \frac{\langle \boldsymbol{h_w}(\boldsymbol{x}_i), \boldsymbol{h_w}(\boldsymbol{x}_j) \rangle}{\|\boldsymbol{h_w}(\boldsymbol{x}_i)\|\ \|\boldsymbol{h_w}(\boldsymbol{x}_j)\|}\right) \quad (9)$$

The theoretical analysis in Section 3 and the empirical study for model selection using LOO in our longer version [18] demonstrate the advantages of NNK in the labeled data setting (L − L). In this paper, we focus on experimental results demonstrating advantages of NNK over KNN for the (UL − L) and (UL − UL) cases[3]. Through these experiments, our aim is to show that simple local methods can achieve good results, even though additional training or parameter turning may be needed to be competitive with state of the art.

4.1. A simple few shot framework (UL − L)

In few shot learning (FSL), one is given a set of *base* data $\mathcal{D}_{base} = \{(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_N, y_N)\}$ where $y_i \in \{1, 2, \ldots C_{base}\}$ and a *support* data $\mathcal{D}_{sup}$ for C_{novel} classes with a small m number (e.g., a common setting in FSL uses $m = 1/5$) of training examples for each class. The aim of few shot learning is to construct a model that can perform well on the C_{novel} classes. This setup is referred to as m-shot C_{novel}- way learning system. Due to the limited availability of data in the novel classes, a FSL model needs to exploit the base dataset for training so that it can be successful in transferring to the novel class with good classification performance. One approach

[3] Source code for all experiments is available at github.com/STAC-USC

adapted by several FSL systems [28, 29, 30] is to train a model with $\mathcal{D}_{base}$ for feature extraction followed by a 1-nearest neighbor classification, or other simple classifiers, on features extracted on $\mathcal{D}_{sup}$.

In this work, we study a simple few shot learner based on local NNK interpolation, where each unlabeled data point is classified using labeled data neighbors obtained using a deep feature extractor trained on the base dataset. We focus on the transductive FSL setting where unlabeled test data is available during model construction. We iteratively refine the predictions on the unlabeled test data by selecting for each point a pseudo label, i.e., the label for which the prediction has most confidence, and using these pseudo labels as additional support data. Note that this process does not involve expensive training of the neural network model or fine tuning of additional parameters to improve performance. Algorithm 1 describes the proposed method for transductive classification of test data queries $\boldsymbol{X}_Q$ using NNK. For comparison, we also evaluate the proposed algorithm by replacing the NNK interpolation classifier in the framework with a locally weighted KNN classifier. The proposed FSL framework can be adapted to a semi-supervised inductive classifier setting by pseudo labeling the available unlabeled data (instead of the test data as done for augmentation in the transductive case) followed by classification of queries using this augmented support dataset.

Algorithm 1: NNK Transductive Few Shot Learning

Input : Neural Network $\boldsymbol{h_w}$, Datasets $\mathcal{D}_{base}, \mathcal{D}_{sup}$, Test queries $\boldsymbol{X}_Q$, No. of Neighbors k

Train $\boldsymbol{h_w}$ using $\mathcal{D}_{base}$
while $\boldsymbol{X}_Q$ *not empty* **do**
  for $\boldsymbol{x}_i \in \boldsymbol{X}_Q$ **do**
    /* $\mathcal{N}_{\boldsymbol{x}_i}$: $\boldsymbol{h_w}(\boldsymbol{x}_i)$ neighbors in $\mathcal{D}_{sup}$ */
    $\boldsymbol{y}_i$ = Label $\boldsymbol{x}_i$ using $\mathcal{N}_{\boldsymbol{x}_i}$ in NNK interpolator (2)
  end
  /* Pseudo label confident predictions in each class - $(\boldsymbol{X}_Q^*, \boldsymbol{Y}_Q^*)$ */
  $\mathcal{D}_{sup} = \mathcal{D}_{sup} \cup (\boldsymbol{X}_Q^*, \boldsymbol{Y}_Q^*)$
  $\boldsymbol{X}_Q = \boldsymbol{X}_Q - \boldsymbol{X}_Q^*$
end

Output: Class predictions $\boldsymbol{Y}_Q$

Experiment Setup: We apply our proposed FSL framework on two standard benchmark datasets *mini*-Imagenet[28] and *tiered*-Imagenet[31] which are subsets of ImageNet[32] dataset with 100 and 608 classes respectively. All images are resized to 84×84 via rescaling, cropping. For few shot evaluation, we follow a setting similar to [13, 14, 29, 30, 33] where we draw random samples of $1/5$-shot 5- way tasks: each task has 5 novel classes with $1/5$ labeled (support) data and is tested on 15 queries per class.

We use wide residual network architecture [34] as our model backbone with 28 convolutional layers and a widening factor of 10. We do not perform any hyperparameter search and restrict ourselves to the settings from [29] for training the model using $\mathcal{D}_{base}$. The network is trained in batches of 256 for 90 epochs with data augmentation from [35] and initial learning rate 0.1 which is reduced by a factor of 10 at fixed schedules. We perform early stopping using 1-nearest neighbor classification on randomly sampled set of 5-validation classes. For both NNK and weighted KNN classifier, we set a max k value and resort to using the entire support set when the number of examples in $\mathcal{D}_{sup}$ is smaller than k.

Method	*mini*-ImageNet		*tiered*-ImageNet	
	1-shot	5-shot	1-shot	5-shot
SimpleShot [29]	63.50	80.33	69.75	85.31
KNN ($k = 5$)	74.73	81.29	76.39	84.32
NNK ($k = 5$)	73.25	80.88	79.86	86.42
KNN ($k = 20$)	66.67	76.83	70.19	79.21
NNK ($k = 20$)	74.44	85.09	80.64	88.41
KNN ($k = 50$)	51.59	63.43	55.36	65.92
NNK ($k = 50$)	**74.99**	**85.05**	**80.73**	**88.61**
Requiring extra training / hyperparameter tuning				
Fine-tuning [30]	65.73	78.40	73.34	85.50
EPNet [14]	70.74	84.34	78.50	88.36
LaplacianShot [13]	74.86	84.13	80.18	87.56
PT+MAP [33]	**82.92**	**88.82**	**85.41**	**90.44**

Table 2: 1-shot and 5-shot accuracy (in %, higher is better) for 5-way classification on *mini*-ImageNet and *tiered*-ImageNet averaged over 600 runs. Results from KNN, NNK transductive classification are compared to an inductive method SimpleShot (CL2N)[29] and listed performances from recently studied transductive methods such as [30, 14, 13, 33]. We see that NNK outperforms KNN as k increases, while achieving robust performance. Further, our simple framework is comparable to, and often better than, recent and more complex transductive FSL algorithms that require additional training, fine-tuning of hyperparameters or preprocessing as in [33].

Table 2 presents our results using local neighborhood based FSL in comparison with recent FSL learners in literature. We do not compare to approaches that are semi supervised (extra unlabeled data per class in $\mathcal{D}_{sup}$) or perform data augmentation, as such approaches make use of additional data statistics or induce specific bias through augmentation. Further, we note that prior methods report results with various network architectures; to eliminate the effect of network backbone in FSL models, we compare our framework only with models having Wide-ResNet-28-10 backbone.

4.2. Distance between datasets (UL − UL)

The empirical performance of deep neural networks in transfer learning and domain adaptation setting has generated renewed interest in the field [36, 37]. In a simple scenario, a model is trained on a dataset (possibly labeled) and then applied to unseen data or fine-tuned to the new data. In this context it would be useful to develop a practical tool to identify in advance when and if a model will transfer well to a particular dataset. In this section, we introduce an asymmetric metric to characterize distance between datasets as a first step towards capturing the likelihood of success in model transfer. The distance measure is label independent and can be obtained for any two datasets (different modalities and domains) provided a kernel can be defined to quantify similarity of samples across datasets. The asymmetric nature of our distance is justified by the fact that transfer from a simple to difficult data is more difficult than the other way.

Definition 1. *Given dataset samples* $\mathcal{D}_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2 \dots \boldsymbol{x}_M\}$ *and* $\mathcal{D}_2 = \{\boldsymbol{y}_1, \boldsymbol{y}_2 \dots \boldsymbol{y}_N\}$*, the NNK distance between the datasets for a given kernel* $\boldsymbol{K} \in [0, 1]$ *is defined as*

$$NNK(\mathcal{D}_1|\mathcal{D}_2) = \frac{1}{M} \sum_{\boldsymbol{x}_i \in \mathcal{D}_1} \min_{\boldsymbol{\theta}_S \geq 0} \boldsymbol{K}_{i,i} - \boldsymbol{\theta}_S^\top \boldsymbol{K}_{S,i} + \boldsymbol{\theta}_S^\top \boldsymbol{K}_{S,S} \boldsymbol{\theta}_S \tag{10}$$

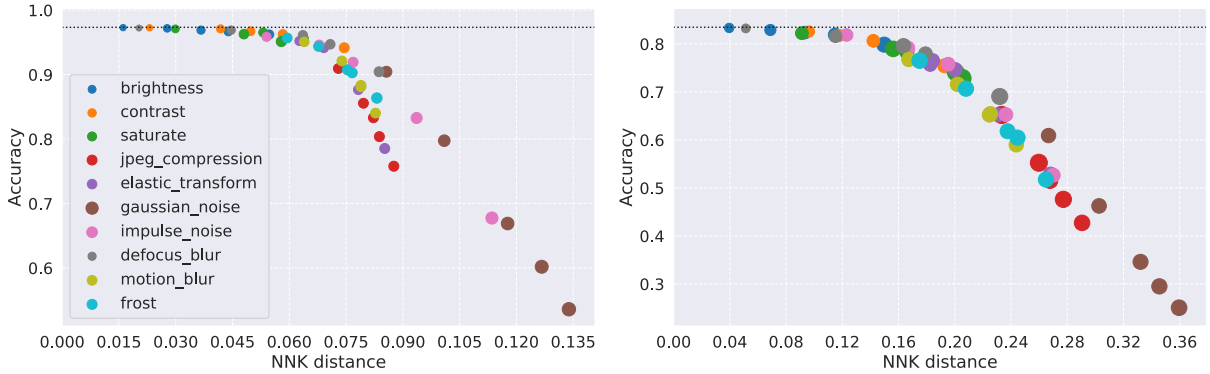


Fig. 1: (Left) CIFAR-10 and (Right) CIFAR-100. Wide-ResNet-28-10 model accuracy vs NNK distance (10) between clean dataset and 5 different noise levels of various corrupted datasets. The accuracy on the clean dataset is denoted by dashed line and the size of the scatter point corresponds to the standard deviation of the terms within the summation in each distance. We see that the proposed distance $\text{NNK}(\mathcal{D}_{\text{clean}}|\mathcal{D}_{\text{corrupt}})$ is indicative of the model's generalization with performance decreasing with increasing distance.

where set $S = \{\boldsymbol{y}_{s_1}, \dots \boldsymbol{y}_{s_k}\}$ corresponds to the set of k nearest neighbors of $\boldsymbol{x}_i$ from $\mathcal{D}_2$. $\boldsymbol{K}_{S,i}$ corresponds to the kernel similarity evaluated between the neighbors and $\boldsymbol{x}_i$.

Let $\phi_a(\boldsymbol{x}_i)$ denote the approximation obtained with the data interpolation in NNK equation (3). Then, the minimization objective $\mathcal{J}(\boldsymbol{x}_i)$ associated with a data $\boldsymbol{x}_i$ in equation (10) can be rewritten as

$$\begin{aligned}\mathcal{J}(\boldsymbol{x}_i) &= \boldsymbol{K}_{i,i} - \boldsymbol{\theta}_S^\top \boldsymbol{K}_{S,i} + \boldsymbol{\theta}_S^\top \boldsymbol{K}_{S,S}\boldsymbol{\theta}_S \\ &= ||\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\theta}_S^\top \boldsymbol{\Phi}_S||^2 = ||\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\phi}_a(\boldsymbol{x}_i)||^2 \end{aligned} \tag{11}$$

where set S corresponds to set of neighbors from $\mathcal{D}_2$ that is used to estimate $\boldsymbol{x}_i$ in $\mathcal{D}_1$. Thus, NNK distance is the difference between the actual observation and the approximated value of the observation based on another set of samples. Intuitively, the average value of $\mathcal{J}(\boldsymbol{x}_i)$ captures the extent to which dataset $\mathcal{D}_2$ fits the dataset $\mathcal{D}_1$.

Proposition 1. *The NNK distance from Definition 1 is asymmetric, non-parametric and satisfies*

1. ***Positivity:*** $NNK(\mathcal{D}_1|\mathcal{D}_2) \geq 0$
2. ***Identity:*** $NNK(\mathcal{D}_1|\mathcal{D}_2) = 0 \iff \mathcal{D}_1 \subseteq \mathcal{D}_2$
3. ***Triangle Inequality:*** $NNK(\mathcal{D}_1|\mathcal{D}_2) \leq NNK(\mathcal{D}_1|\mathcal{D}_3) + NNK(\mathcal{D}_3|\mathcal{D}_2)$

Note that a similar definition of distance using k-nearest neighbors is not straightforward. This is in part due to the fact that k and the weights in KNN are not explicitly chosen to minimize an approximation objective. For example, consider a KNN distance definition where we use the interpolation of equation (11), by replacing the weights obtained from NNK optimization with the relative similarity between $\boldsymbol{x}_i$ and its KNN neighbors i.e., $\boldsymbol{\theta}_{\text{KNN}} = \boldsymbol{K}_{S,i}/(\mathbf{1}^\top \boldsymbol{K}_{S,i})$:

$$KNN(\mathcal{D}_1|\mathcal{D}_2) = \frac{1}{M}\sum_{\boldsymbol{x}_i \in \mathcal{D}_1} \boldsymbol{K}_{i,i} - 2\boldsymbol{\theta}_{\text{KNN}}^\top \boldsymbol{K}_{S,i} + \boldsymbol{\theta}_{\text{KNN}}^\top \boldsymbol{K}_{S,S}\boldsymbol{\theta}_{\text{KNN}} \tag{12}$$

This KNN distance lacks basic properties of distance. For example, it is easy to see that the distance grows as k increases (since we add additional terms corresponding to points that are farther away). Also, consider the case where both the datasets are the same, ideally we would want distance between them to be zero but the defined KNN distance is nonzero for all values of $k > 1$. This hints at the fact that it may not be possible to design a suitable distance without optimizing the set of neighboring points and weights (as in NNK).

Experiment Setup: We evaluate the proposed distance metric with CIFAR-10 and CIFAR-100 datasets [38] and their corrupted variants from [39]. As in FSL experiment, we use a Wide-ResNet-28-10 architecture for constructing a deep learning model for each CIFAR dataset. The network is trained using only training dataset with data augmentation in batches of 128 for 200 epochs with learning rate 0.1 and weight decay $5e^{-4}$ as used in [40]. We use features extracted at the penultimate layer of the trained network to measure NNK distance between the clean test dataset and various corrupted versions of the test dataset. We observe a power law relationship between the proposed distance metric and the model performance in both datasets with accuracy decreasing as NNK distance increases.

Figure 1 shows the relationship between the defined NNK distance and the performance of the model on the datasets. The distance allows us to understand the predictive performance of a pre-trained model on a new dataset (corrupted test dataset $\mathcal{D}_{\text{corrupt}}$), given its performance on past dataset ($\mathcal{D}_{\text{clean}}$). We believe such a measure could improve the process of transfer learning, where one can choose a particular model from a pool of pre-trained models that is more adaptive to the new dataset for fine tuning.

5. CONCLUSION

In this work, we presented a theoretical study of the Non Negative Kernel regression (NNK) framework and its performance in deep learning settings. In particular, we studied the performance of neighborhood methods in a transductive few shot learning scenario where we show that NNK approach offers a competitive and robust baseline in comparison to other methods, without requiring additional training and parameter tuning. Further, we introduced a notion of distance between datasets and observed its relationship to model performance as the distance increases. Our ultimate goal is for the NNK framework to encourage the study and use of interpretable and robust neighborhood methods in machine learning. Prototypes, subsampling, and approximation schemes are key to scaling neighborhood methods to large datasets ($N > 10^6$). We leave as future work, the design and study of computationally scalable NNK techniques leveraging the geometrical aspects of the framework.

6. REFERENCES

[1] Thomas Cover and Peter Hart, "Nearest neighbor pattern classification," *IEEE Trans. on Inf. Theory*, 1967.

[2] Elizbar A Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, 1964.

[3] Geoffrey S Watson, "Smooth regression analysis," *Sankhyā: The Indian J. of Statistics, Series A*, 1964.

[4] Gérard Biau and Luc Devroye, *Lectures on the Nearest Neighbor method*, Springer, 2015.

[5] Charles J Stone, "Consistent nonparametric regression," *The Annals of Statistics*, 1977.

[6] Richard J Samworth et al., "Optimal weighted nearest neighbour classifiers," *The Annals of Statistics*, 2012.

[7] Luc Devroye, László Györfi, and Gábor Lugosi, *A probabilistic theory of pattern recognition*, Springer, 2013.

[8] Samory Kpotufe, "k-nn regression adapts to local intrinsic dimension," in *Advances in Neural Inf. Process. Syst.*, 2011.

[9] Oren Anava and Kfir Levy, "k^*-nearest neighbors: From global to local," in *Advances in Neural Inf. Process. Syst.*, 2016.

[10] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das, "Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance," *IEEE Trans. on Neural Networks and Learning Syst.*, 2018.

[11] Nicolas Papernot and Patrick McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," *arXiv:1803.04765*, 2018.

[12] Myriam Bontonou, Carlos Lassance, Ghouthi Boukli Hacene, Vincent Gripon, Jian Tang, and Antonio Ortega, "Introducing graph smoothness loss for training deep learning architectures," in *IEEE Data Science Workshop*, 2019.

[13] Imtiaz Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed, "Laplacian regularized few-shot learning," in *Int. Conf. on Machine Learning*. PMLR, 2020.

[14] Pau Rodríguez, Issam Laradji, Alexandre Drouin, and Alexandre Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *Proc. of the European Conf. on Computer Vision*, 2020.

[15] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool, "SCAN: Learning to classify images without labels," in *Proc. of the European Conf. on Computer Vision*, 2020.

[16] Sarath Shekkizhar and Antonio Ortega, "Graph construction from data by Non-Negative Kernel Regression," in *Int. Conf. on Acoustics, Speech and Signal Process.* IEEE, 2020.

[17] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, 2020.

[18] Sarath Shekkizhar and Antonio Ortega, "DeepNNK: Explaining deep models and their generalization using polytope interpolation," *arXiv:2007.10505*, 2020.

[19] J.A Tropp and A.C Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. on info. theory*, 2007.

[20] Mikhail Belkin, Daniel J Hsu, and Partha Mitra, "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate," in *Advances in Neural Inf. Process. Syst.*, 2018.

[21] Pascal Massart, Élodie Nédélec, et al., "Risk bounds for statistical learning," *The Annals of Statistics*, 2006.

[22] André Elisseeff and Massimiliano Pontil, "Leave-one-out error and stability of learning algorithms with applications," *NATO Science Series III: Computer and Systems Sciences*, 2003.

[23] Aleksandr Luntz, "On estimation of characters obtained in statistical procedure of recognition," *Technicheskaya Kibernetica*, 1969.

[24] William H Rogers and Terry J Wagner, "A finite sample distribution-free performance bound for local discrimination rules," *The Annals of Statistics*, 1978.

[25] Luc Devroye and Terry Wagner, "Distribution-free inequalities for the deleted and holdout error estimates," *IEEE Trans. on Inf. Theory*, 1979.

[26] C.A. Rogers, "Covering a sphere with spheres," *Mathematika*, 1963.

[27] Long Chen, "New analysis of the sphere covering problems and optimal polytope approximation of convex bodies," *J. of Approximation Theory*, 2005.

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., "Matching networks for one shot learning," in *Advances in Neural Inf. Process. Syst.*, 2016.

[29] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten, "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," *arXiv:1911.04623*, 2019.

[30] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto, "A baseline for few-shot image classification," in *Int. Conf. on Learning Representations*, 2020.

[31] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua Tenenbaum, Hugo Larochelle, and R. Zemel, "Meta-learning for semi-supervised few-shot classification," in *Int. Conf. on Learning Representations*, 2018.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "ImageNet: Large Scale Visual Recognition Challenge," *Int. J. of Computer Vision*, 2015.

[33] Yuqing Hu, Vincent Gripon, and Stéphane Pateux, "Leveraging the feature distribution in transfer-based few-shot learning," *arXiv:2006.03806*, 2020.

[34] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," in *British Mach. Vision Conf.*, 2016.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*, 2016.

[36] Gabriela Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv:1702.05374*, 2017.

[37] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, 2020.

[38] Alex Krizhevsky et al., "Learning multiple layers of features from tiny images," 2009.

[39] Dan Hendrycks and Thomas Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *Int. Conf. on Learning Representations*, 2019.

[40] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le, "Autoaugment: Learning augmentation strategies from data," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*, 2019.

Supplementary material

6.1. Proof of Theorem 1

Proof. The proof follows a similar argument as in the simplicial interpolation bound in [20]. The expected excess mean squared risk can be partitioned based on disjoint sets as[4]

$$\begin{aligned}\mathbb{E}[(\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}))^2] &= \mathbb{E}[(\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}))^2|X\notin\mathcal{C}]P(X\notin\mathcal{C}) + \mathbb{E}[(\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}))^2|X\in\mathcal{C}]P(X\in\mathcal{C}) \\ &\leq \mathbb{E}[(\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}))^2|X\notin\mathcal{C}]P(X\notin\mathcal{C}) + \mathbb{E}[(\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}))^2|X\in\mathcal{C}] \end{aligned} \tag{13}$$

For points outside the convex hull, NNK extrapolates and no guarantees can be made on the regression without further assumptions. Thus, $(\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}))^2 \leq 1$ which reduces the first term on the left of equation (13) to that of theorem as $P(X\notin\mathcal{C}) = \mathbb{E}[\mu(\mathbb{R}^d\backslash\mathcal{C})]$.

Let $\boldsymbol{\theta}$ denote a $\hat{k}$ length vector of non zero values obtained from the solution to NNK interpolation objective (3). We represent the normalized weights used to obtain the NNK estimate (2) as $w_i = \frac{\theta_i}{\sum_{i=1}^{\hat{k}}\theta_i}$. The normalized weights $\boldsymbol{w}$ follow a Dirichlet$(1,1\ldots1)$ distribution with $\hat{k}$ concentration parameters.

$$\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}) = \sum_{i=1}^{\hat{k}} w_i(y_i-\eta(\boldsymbol{x})) = \sum_{i=1}^{\hat{k}} w_i(y_i-\eta(\boldsymbol{x}_i)+\eta(\boldsymbol{x}_i)-\eta(\boldsymbol{x})) = \sum_{i=1}^{\hat{k}} w_i\epsilon_i + \sum_{i=1}^{\hat{k}} w_i b_i \tag{14}$$

where $\epsilon_i = y_i-\eta(\boldsymbol{x}_i)$ corresponds to Bayesian estimator errors (noise) in the training data and $b_i = \eta(\boldsymbol{x}_i)-\eta(\boldsymbol{x})$ is related to bias associated to the NNK estimator. By smoothness assumption on η we have

$$|b_i| = |\eta(\boldsymbol{x}_i)-\eta(\boldsymbol{x})| \leq A||\boldsymbol{\phi}(\boldsymbol{x}_i)-\boldsymbol{\phi}(\boldsymbol{x})||^{\alpha} \leq A\delta^{\alpha} \tag{15}$$

The inequality with δ is a direct consequence of its definition, i.e., the maximum distance between the any two vertices forming the NNK$_{poly}(\boldsymbol{x})$. Since b_i and ϵ_i are independent, the second term on the right of equation (13) can be further partitioned as

$$\mathbb{E}[(\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}))^2|X\in\mathcal{C}] = \mathbb{E}[\left(\sum_{i=1}^{\hat{k}} w_i\epsilon_i\right)^2|X\in\mathcal{C}] + \mathbb{E}[\left(\sum_{i=1}^{\hat{k}} w_i b_i\right)^2|X\in\mathcal{C}] \tag{16}$$

Using Jensen's inequality $\left(\sum_{i=1}^{\hat{k}} w_i b_i\right)^2 \leq \sum_{i=1}^{\hat{k}} w_i b_i^2$ and the bound from equation (15), we have

$$\mathbb{E}[\left(\sum_{i=1}^{\hat{k}} w_i b_i\right)^2|X\in\mathcal{C}] \leq \mathbb{E}[\sum_{i=1}^{\hat{k}} w_i b_i^2|X\in\mathcal{C}] \leq \mathbb{E}[\sum_{i=1}^{\hat{k}} w_i A^2\delta^{2\alpha}|X\in\mathcal{C}] = A^2\delta^{2\alpha} \tag{17}$$

Let $\nu(\boldsymbol{x}) = var(Y|X=\boldsymbol{x})$. Under independence assumption on noise, the ϵ terms in equation (16) can be rewritten as

$$\begin{aligned}\mathbb{E}[\left(\sum_{i=1}^{\hat{k}} w_i\epsilon_i\right)^2|X\in\mathcal{C}] &= \mathbb{E}[\sum_{i=1}^{\hat{k}} w_i^2\epsilon_i^2|X\in\mathcal{C}] = \mathbb{E}_K\left[\sum_{i=1}^{\hat{k}}\mathbb{E}_{X|K}[w_i^2|X\in\mathcal{C}]\mathbb{E}_{X|K}[\epsilon_i^2|X\in\mathcal{C}]\right] \\ &\leq \mathbb{E}_K\left[\frac{2}{(\hat{k}+1)(\hat{k})}\sum_{i=1}^{\hat{k}}\nu(\boldsymbol{x}_i)\right] \leq \mathbb{E}_K\left[\frac{2}{(\hat{k}+1)(\hat{k})}\sum_{i=1}^{\hat{k}}\nu(\boldsymbol{x}) + |\nu(\boldsymbol{x}_i)-\nu(\boldsymbol{x})|\right]\end{aligned}$$

where we use the fact that w_i follows Dirichlet distribution. Now, the smoothness assumption on $var(Y|X)$ gives us

$$|\nu(\boldsymbol{x}_i)-\nu(\boldsymbol{x})| \leq A'||\boldsymbol{\phi}(\boldsymbol{x}_i)-\boldsymbol{\phi}(\boldsymbol{x})||^{\alpha'} \leq A'\delta^{\alpha'} \tag{18}$$

$$\implies \mathbb{E}[\left(\sum_{i=1}^{\hat{k}} w_i\epsilon_i\right)^2|X\in\mathcal{C}] \leq \frac{2}{(\mathbb{E}_K[\hat{k}]+1)}\left(\nu(\boldsymbol{x})+A'\delta^{\alpha'}\right) \tag{19}$$

Combining equations (17) and (19), we obtain the risk bound for NNK interpolation for points within the convex hull of training data $\mathcal{C}$, i.e.,

$$\mathbb{E}[(\hat{\eta}(\boldsymbol{x})-\eta(\boldsymbol{x}))^2|X\in\mathcal{C}] \leq A^2\delta^{2\alpha} + \frac{2}{(\mathbb{E}_K[\hat{k}]+1)}\left(\nu(\boldsymbol{x})+A'\delta^{\alpha'}\right) \tag{20}$$

Equation (20) along with the extrapolation bound for points outside the convex hull $\mathcal{C}$ obtained earlier gives the excess risk bound for NNK estimator and concludes the proof. □

[4] All expectation in this proof are condition on D_{train}. For the sake of brevity, we do not make this conditioning explicit in our statements.

6.2. Proof of Corollary 1.1

Proof. The nearest neighbor convergence lemma of [1] states that for an i.i.d sequence of random variables $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2 \dots \boldsymbol{x}_N\}$ in $\mathbb{R}^d$, the nearest neighbor of $\boldsymbol{x}$ from the set $\mathcal{D}$ converges in probability, $NN(\boldsymbol{x}) \rightarrow_p \boldsymbol{x}$. Equivalently, this would correspond to convergence in kernel representation of the data points. Thus, the solution to NNK data interpolation objective is reduced to 1-nearest neighbor interpolation with $\mathbb{E}_K[\hat{k}] = 1$ and $\limsup_{N\to\infty} \delta = 0$. Now, under the assumption that the $supp(\mu)$ belongs to a bounded and convex region in $\mathbb{R}^d$, the first term on the right of equation (4) corresponding to NNK extrapolation vanishes i.e., $\limsup_{N\to\infty} E_X[\mu(\mathbb{R}^d \backslash \mathcal{C})] = 0$.

Applying the asymptotic vanishing property of δ and $E_X[\mu(\mathbb{R}^d \backslash \mathcal{C})]$ in Theorem 1 gives us the result of Corollary 1.1. □

6.3. Proof of Corollary 1.2

Proof. The excess classification risk associated with a plug-in NNK classifier is related the regression risk (see Theorem 17.1 in [4]) as

$$\mathbb{E}[\mathcal{R}(\hat{f}(\boldsymbol{x})) - \mathcal{R}(f(\boldsymbol{x}))] \leq \mathbb{E}[\mathbb{I}(\hat{f}(\boldsymbol{x}) \neq f(\boldsymbol{x}))] \leq 2\mathbb{E}[|\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x})|] \tag{21}$$

From Corollary 1.1, we have

$$\limsup_{N\to\infty} \mathbb{E}[(\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x}))^2] \leq \mathbb{E}[(Y - \eta(\boldsymbol{x})^2]$$

By Jensen's inequality

$$\limsup_{N\to\infty} (\mathbb{E}[|\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x})|])^2 \leq \limsup_{N\to\infty} \mathbb{E}[(\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x}))^2] \tag{22}$$

Combining with equation (21) gives the required risk bound. □

6.4. Proof of Theorem 2

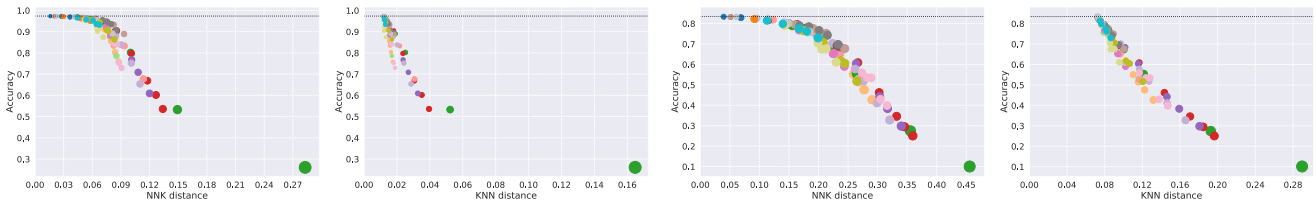
Proof. The proof is based on the k-nearest neighbor result from Theorem 1 in [25] which states that

$$P(|\mathcal{R}_{loo}(\hat{\eta}|\mathcal{D}_{train}) - \mathcal{R}_{gen}(\hat{\eta})| > \epsilon) \leq 2e^{-N\epsilon^2/18} + 6e^{-N\epsilon^3/(108k(2+\gamma))} \tag{23}$$

As in [25], where the result is extended based on the 1-nearest neighbor, here it suffices to replace k by $\mathbb{E}_K[\hat{k}]$ since each data point on average cannot be NNK neighbors to more than $\mathbb{E}_K[\hat{k}]\gamma + 2 \leq \mathbb{E}_K[\hat{k}](\gamma + 2)$ data points, where $gamma$ corresponds to maximum number of data points that can share the same *nearest* neighbor. □

6.5. Additional Experiments

Figure 2 shows the relationship between the defined NNK, KNN distance and the performance of the model on all corruptions used in [39]. We see a similar trend in distance as before with the accuracy of the model decreasing smoothly with increasing NNK distance. As noted in section 4.2, though the performance of the model decreases with proposed KNN distance as well, it is unclear if this distance captures the right measure of complexity between the datasets and remains unclear in terms of interpretability.



(a) CIFAR-10 (b) CIFAR-100

Fig. 2: Wide-ResNet-28-10 model accuracy vs NNK distance (10) , KNN distance (12) between clean dataset and 5 different noise levels of all corruptions from [39]. The accuracy on the clean dataset is denoted by dashed line and the size of the scatter point corresponds to the standard deviation of the minimization objection in equation (10). Apart from the robustness to the choice of k in comparison to KNN distance, we conjecture that NNK distance $\text{NNK}(\mathcal{D}_{\text{clean}}|\mathcal{D}_{\text{corrupt}})$ captures subtle difference in datasets, as can be seen distinctly in the small distance−high accuracy region of the CIFAR-10 plots, and is more principled and interpretable.