Generalization Analysis for CCKD: A Stability and Curvature Perspective

Baoyun Peng
Academy of Military Science, Beijing
pengbaoyun13@alunmi.nudt.edu.cn

Abstract

This document provides a theoretical analysis of the generalization properties of **Correlation Congruence Knowledge Distillation (CCKD)** [1]. Building upon the Neural Tangent Kernel (NTK) framework, we leverage stability analysis of optimization dynamics to derive a novel geometric condition under which CCKD provably induces norm reduction in the associated Reproducing Kernel Hilbert Space (RKHS). This allows us to establish a formal generalization bound demonstrating how CCKD's feature alignment can lead to superior generalization compared to standard Knowledge Distillation (KD) [2], clarifying the interplay between feature matching, loss landscape curvature, and model complexity.

1 Preliminaries and Notation

In this section, we establish the mathematical framework and notation used throughout our analysis.

Definition 1.1 (Setup). *Let $\mathcal{X} \subseteq \mathbb{R}^p$ be the input space and $\mathcal{Y} = \{1, \ldots, K\}$ be the label space. Data is drawn from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We are given a training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$. A neural network is a function $h : \mathcal{X} \to \mathbb{R}^K$. We consider two networks: a pre-trained, fixed teacher network h_T and a student network h_S.*

Definition 1.2 (Neural Tangent Kernel (NTK) and RKHS). *For a student network family parameterized by $\boldsymbol{\Phi}$, the Neural Tangent Kernel (NTK) is $\Theta(\boldsymbol{x}, \boldsymbol{x}') = tr(\nabla_{\boldsymbol{\Phi}} h(\boldsymbol{x})^\top \nabla_{\boldsymbol{\Phi}} h(\boldsymbol{x}'))$. It induces a unique Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\Theta$ of functions, with norm $\|\cdot\|_{\mathcal{H}_\Theta}$ and inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_\Theta}$ [3, 4].*

Definition 1.3 (Feature Gram Matrix). *For a network with a feature extractor $f : \mathcal{X} \to$*

$\mathbb{R}^d$ *(the layer before the final classifier), the feature Gram matrix on the training set S is a matrix $[\boldsymbol{K}_S(h)]_{ij} = \langle f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\rangle \in \mathbb{R}^{n\times n}$. We denote the Gram matrix of the student as $\boldsymbol{K}_S(h_S)$ and the pre-computed Gram matrix of the teacher as $\boldsymbol{K}_T$.*

Definition 1.4 (Learning Objectives and Risk)**.** *Let ℓ_{CE} be the Cross-Entropy loss for ground-truth labels and ℓ_{KL} be the Kullback-Leibler divergence for soft targets from the teacher. Let $\alpha \in [0,1]$ be the distillation trade-off parameter and $\mu > 0$ be the weight decay regularization parameter.*

1. ***Standard Knowledge Distillation (KD)** combines the standard classification loss with the distillation loss from the teacher's logits. The learning objective is to find a student function h_S^{KD} that minimizes the regularized empirical risk $J_{KD}(h)$:*

$$h_S^{KD} := \arg\min_{h\in\mathcal{H}_\Theta} \left\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} [(1-\alpha)\ell_{CE}(h(\boldsymbol{x}_i), y_i) + \alpha\ell_{KL}(h(\boldsymbol{x}_i), h_T(\boldsymbol{x}_i))] +}_{\textit{Empirical Distillation Risk } \hat{R}_n(h)} \underbrace{\frac{\mu}{2}\|h\|^2_{\mathcal{H}_\Theta}}_{\textit{Regularization}} \right\} \tag{1}$$

2. ***Correlation Congruence Knowledge Distillation (CCKD)** extends standard KD by adding an explicit feature alignment term. This term penalizes the Frobenius norm of the difference between the student's and teacher's feature Gram matrices. The objective is to find h_S^{CCKD} by minimizing $J_{CCKD}(h)$:*

$$h_S^{CCKD} := \arg\min_{h\in\mathcal{H}_\Theta} \left\{ \hat{R}_n(h) + \frac{\mu}{2}\|h\|^2_{\mathcal{H}_\Theta} + \underbrace{\frac{\lambda}{n^2}\|\boldsymbol{K}_S(h) - \boldsymbol{K}_T\|_F^2}_{\textit{Feature Alignment Loss } G(h)} \right\} \tag{2}$$

where $\lambda > 0$ is the hyperparameter controlling the strength of the feature alignment.

For simplicity in our subsequent analysis, we will use the compact notation:

- $J_{KD}(h) = \hat{R}_n(h) + \frac{\mu}{2}\|h\|^2$
- $J_{CCKD}(h) = J_{KD}(h) + \lambda G(h)$

2 Assumptions

Assumption 1 (Kernel Regime)**.** *Training occurs in the NTK regime where the kernel Θ is fixed. Function h_S evolves within $\mathcal{H}_{\Theta_S}$ [5].*

Assumption 2 (Smoothness and Local Convexity)**.** *Under Assumption 1, the loss functional with respect to the network's output is convex. We assume the standard empirical risk $\hat{R}_n(h)$ is convex and twice continuously differentiable. We further assume that the* *feature alignment loss $G(h)$ is twice continuously differentiable and, crucially, is **convex in a neighborhood of the KD solution** h_S^{KD}.*

Remark 2.1 (On Local vs. Global Minimizers)**.** *The assumption of local convexity for $G(h)$ is a standard technique when analyzing complex, non-convex objectives. This assumption implies that the full CCKD objective, $J_{CCKD}(h)$, is strongly convex in a neighborhood of h_S^{KD}. The existence and smoothness of a unique local minimizer $h(\lambda)$ that emerges from perturbing the KD solution $h(0) = h_S^{KD}$ is then guaranteed by the implicit function theorem, a cornerstone of sensitivity analysis in optimization [6]. We denote this local minimizer by h_S^{CCKD}.*

Our approach of analyzing the solution path $h(\lambda)$ as the hyperparameter λ varies from zero is analogous to the well-established "regularization path" analysis in classical machine learning, famously used for LASSO [7]. Our analysis, therefore, pertains to the specific local minimizer found by tracking this path, which is often the solution obtained in practice when fine-tuning with a small regularization parameter λ.

Furthermore, analyzing the local loss landscape and its curvature around a specific solution is a common theme in modern deep learning theory, particularly within the NTK framework [8]. Because both h_S^{CCKD} (for small λ) and h_S^{KD} reside within this same region of local strong convexity, the standard optimality inequality $J_{CCKD}(h_S^{CCKD}) \leq J_{CCKD}(h_S^{KD})$ holds, preserving the validity of our subsequent proofs.

Assumption 3 (Hessian-Transformed Geometric Alignment)**.** *Let $H_{J_{KD}}(h) = H_{\hat{R}_n}(h) + \mu I$ be the Hessian of the KD objective. At the KD solution h_S^{KD}, we assume:*

$$\langle h_S^{KD}, (H_{J_{KD}}(h_S^{KD}))^{-1} \nabla_h G(h_S^{KD}) \rangle_{\mathcal{H}_{\Theta_S}} > 0 \tag{3}$$

Remark 2.2 (Interpretation of Assumption 3)**.** *This assumption posits that CCKD's norm-reducing force acts effectively along directions where the original KD problem is "soft" or where its loss landscape is relatively flat.*

Assumption 4 (Lipschitz Loss)**.** *$\hat{R}_n(h)$ is L_R-Lipschitz w.r.t a function's outputs on the training set [9].*

3 Main Results with Detailed Proofs

Lemma 3.1 (CCKD-Induced Norm Reduction)**.** *Under Assumptions 1-3, there exists a $\lambda_0 > 0$ such that for all $0 < \lambda < \lambda_0$, the CCKD solution satisfies:*

$$\|h_S^{CCKD}\|_{\mathcal{H}_{\Theta_S}} < \|h_S^{KD}\|_{\mathcal{H}_{\Theta_S}} \tag{4}$$

Detailed Proof. **Step 1: Setup for Perturbation Analysis.** Let $h(\lambda) = h_S^{\text{CCKD}}$ be the minimizer of $J_{\text{CCKD}}(h)$ for a given $\lambda \geq 0$. Note that $h(0) = h_S^{\text{KD}}$. The function $h(\lambda)$ is implicitly defined by the first-order optimality condition $\nabla_h J_{\text{CCKD}}(h(\lambda)) = 0$. That is:

$$\nabla_h \hat{R}_n(h(\lambda)) + \mu h(\lambda) + \lambda \nabla_h G(h(\lambda)) = 0 \tag{5}$$

Our goal is to understand how $\|h(\lambda)\|^2$ behaves as λ increases from 0. We will analyze the sign of its derivative at $\lambda = 0$.

Step 2: Differentiating the Squared Norm. Let $\phi(\lambda) = \|h(\lambda)\|^2 = \langle h(\lambda), h(\lambda) \rangle$. Differentiating with respect to λ using the product rule for inner products gives:

$$\phi'(\lambda) = \frac{d}{d\lambda} \langle h(\lambda), h(\lambda) \rangle = 2 \left\langle h(\lambda), \frac{dh(\lambda)}{d\lambda} \right\rangle = 2 \langle h(\lambda), h'(\lambda) \rangle \tag{6}$$

Step 3: Finding $h'(\lambda)$ via Implicit Differentiation. To find $h'(\lambda)$, we apply the implicit function theorem [6] by differentiating the entire optimality condition (Eq. 5) with respect to λ. Using the chain rule for functional derivatives and the product rule:

$$\frac{d}{d\lambda} \left[\nabla_h \hat{R}_n(h(\lambda)) \right] + \frac{d}{d\lambda} [\mu h(\lambda)] + \frac{d}{d\lambda} [\lambda \nabla_h G(h(\lambda))] = 0 \tag{7}$$

The terms expand as follows:

- $H_{\hat{R}_n}(h(\lambda)) h'(\lambda)$
- $\mu h'(\lambda)$
- $1 \cdot \nabla_h G(h(\lambda)) + \lambda H_G(h(\lambda)) h'(\lambda)$

Combining these, we get:

$$(H_{\hat{R}_n}(h(\lambda)) + \mu I + \lambda H_G(h(\lambda))) h'(\lambda) = -\nabla_h G(h(\lambda)) \tag{8}$$

Step 4: Evaluating at $\lambda = 0$. We are interested in the behavior at the starting point $\lambda = 0$, where $h(0) = h_S^{\text{KD}}$. Setting $\lambda = 0$ simplifies the above equation to:

$$(H_{\hat{R}_n}(h_S^{\text{KD}}) + \mu I) h'(0) = -\nabla_h G(h_S^{\text{KD}}) \tag{9}$$

By definition, $H_{J_{\text{KD}}}(h_S^{\text{KD}}) = H_{\hat{R}_n}(h_S^{\text{KD}}) + \mu I$. Since J_{KD} is μ-strongly convex, its Hessian is invertible [10]. We can solve for $h'(0)$:

$$h'(0) = -(H_{J_{\text{KD}}}(h_S^{\text{KD}}))^{-1} \nabla_h G(h_S^{\text{KD}}) \tag{10}$$

Step 5: Determining the Sign of the Norm's Derivative. Now substitute $h'(0)$

from Eq. 10 back into the derivative of the squared norm (Eq. 6) at $\lambda = 0$:

$$\phi'(0) = 2\langle h(0), h'(0)\rangle = -2\left\langle h_S^{\mathrm{KD}}, (H_{J_{\mathrm{KD}}}(h_S^{\mathrm{KD}}))^{-1}\nabla_h G(h_S^{\mathrm{KD}})\right\rangle_{\mathcal{H}_{\Theta_S}} \tag{11}$$

Step 6: Concluding with Assumption 3. By Assumption 3, the inner product term is strictly positive. Therefore, $\phi'(0) < 0$. Under Assumption 2, all objectives are smooth, implying that $h(\lambda)$ is a continuously differentiable function of λ. Consequently, $\phi(\lambda) = \|h(\lambda)\|^2$ is also continuously differentiable. Since $\phi'(0)$ is strictly negative, by the definition of a derivative, there must exist a neighborhood around 0, say $(-\lambda_0, \lambda_0)$, such that for any λ in that neighborhood, the sign of $\frac{\phi(\lambda)-\phi(0)}{\lambda}$ is the same as the sign of $\phi'(0)$. For $\lambda \in (0, \lambda_0)$, this means $\phi(\lambda) - \phi(0) < 0$, which implies $\left\|h_S^{\mathrm{CCKD}}\right\|^2 < \left\|h_S^{\mathrm{KD}}\right\|^2$. □

Lemma 3.2 (Feature Alignment Guarantee). *For any $\lambda > 0$, $G(h_S^{CCKD}) \leq G(h_S^{KD})$. If $\nabla_h G(h_S^{KD}) \neq 0$, the inequality is strict for small enough $\lambda > 0$.*

Detailed Proof. **Part 1: Non-strict Inequality.** By definition, h_S^{KD} minimizes $J_{\mathrm{KD}}(h)$ and h_S^{CCKD} minimizes $J_{\mathrm{CCKD}}(h)$. This gives two inequalities:

$$J_{\mathrm{KD}}(h_S^{\mathrm{KD}}) \leq J_{\mathrm{KD}}(h_S^{\mathrm{CCKD}}) \tag{12}$$

$$J_{\mathrm{CCKD}}(h_S^{\mathrm{CCKD}}) \leq J_{\mathrm{CCKD}}(h_S^{\mathrm{KD}}) \tag{13}$$

Expanding Eq. 13 using the definition $J_{\mathrm{CCKD}} = J_{\mathrm{KD}} + \lambda G$:

$$J_{\mathrm{KD}}(h_S^{\mathrm{CCKD}}) + \lambda G(h_S^{\mathrm{CCKD}}) \leq J_{\mathrm{KD}}(h_S^{\mathrm{KD}}) + \lambda G(h_S^{\mathrm{KD}})$$

Substituting Eq. 12 into this inequality:

$$J_{\mathrm{KD}}(h_S^{\mathrm{KD}}) + \lambda G(h_S^{\mathrm{CCKD}}) \leq J_{\mathrm{KD}}(h_S^{\mathrm{CCKD}}) + \lambda G(h_S^{\mathrm{CCKD}}) \leq J_{\mathrm{KD}}(h_S^{\mathrm{KD}}) + \lambda G(h_S^{\mathrm{KD}})$$

This simplifies to $\lambda G(h_S^{\mathrm{CCKD}}) \leq \lambda G(h_S^{\mathrm{KD}})$. Since $\lambda > 0$, we have $G(h_S^{\mathrm{CCKD}}) \leq G(h_S^{\mathrm{KD}})$.

Part 2: Strict Inequality. Let $\psi(\lambda) = G(h(\lambda))$. We analyze its derivative at $\lambda = 0$:

$$\psi'(\lambda) = \langle \nabla_h G(h(\lambda)), h'(\lambda)\rangle \tag{14}$$

At $\lambda = 0$, using $h'(0)$ from Eq. 10:

$$\psi'(0) = \left\langle \nabla_h G(h_S^{\mathrm{KD}}), -(H_{J_{\mathrm{KD}}}(h_S^{\mathrm{KD}}))^{-1}\nabla_h G(h_S^{\mathrm{KD}})\right\rangle \tag{15}$$

Let $v = \nabla_h G(h_S^{\mathrm{KD}})$. Then $\psi'(0) = -\langle v, (H_{J_{\mathrm{KD}}})^{-1}v\rangle$. Since J_{KD} is μ-strongly convex, its Hessian $H_{J_{\mathrm{KD}}}$ is positive definite with eigenvalues at least μ. Its inverse $(H_{J_{\mathrm{KD}}})^{-1}$ is also positive definite. Therefore, for any non-zero vector v, the inner product $\langle v, (H_{J_{\mathrm{KD}}})^{-1}v\rangle$

is strictly positive. If $\nabla_h G(h_S^{\mathrm{KD}}) \neq 0$, then $\psi'(0) < 0$. By the same continuity and Mean Value Theorem argument used in Lemma 3.1, this implies $G(h(\lambda)) < G(h(0))$ for $\lambda \in (0, \lambda_0)$. □

Theorem 3.3 (Comparative Generalization Bound)**.** *Under the stated assumptions, with probability at least $1-\delta$:*

$$R(h_S^{CCKD}) \leq R(h_S^{KD}) - \lambda\Delta_G - \frac{\mu}{2}\Delta_{\|\cdot\|^2} + \frac{C}{\sqrt{n}}(\|h_S^{CCKD}\| + \|h_S^{KD}\|) + 4L_R\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{16}$$

where $\Delta_G = G(h_S^{KD}) - G(h_S^{CCKD})$, $\Delta_{\|\cdot\|^2} = \|h_S^{KD}\|^2 - \|h_S^{CCKD}\|^2$, and $C = 2L_R\sqrt{K \cdot \max_i \Theta_S(\boldsymbol{x}_i, \boldsymbol{x}_i)}$ is a constant independent of n.

Detailed Proof. **Step 1: Rademacher Complexity Bounds.** A standard result from generalization theory states that for a function class $\mathcal{F}_B = \{h \in \mathcal{H}_\Theta : \|h\|_{\mathcal{H}_\Theta} \leq B\}$, with probability at least $1-\delta/2$, for any $h \in \mathcal{F}_B$:

$$R(h) \leq \hat{R}_n(h) + 2L_R\hat{\mathfrak{R}}_n(\mathcal{F}_B) + \text{const} \cdot \sqrt{\frac{\ln(2/\delta)}{n}}$$

For a ball of radius B in an RKHS, the empirical Rademacher complexity is bounded as $\hat{\mathfrak{R}}_n(\mathcal{F}_B) \leq \frac{B\sqrt{\mathrm{tr}(\boldsymbol{\Theta}_S)}}{n}$ [9]. Crucially, the trace of the kernel matrix typically scales with the number of samples, i.e., $\mathrm{tr}(\boldsymbol{\Theta}_S) \leq n \cdot \max_i \Theta_S(\boldsymbol{x}_i, \boldsymbol{x}_i)$. This leads to a complexity term of:

$$2L_R\frac{B\sqrt{n \cdot \max_i \Theta_S(\boldsymbol{x}_i, \boldsymbol{x}_i)}}{n} = \frac{2L_R B\sqrt{\max_i \Theta_S(\boldsymbol{x}_i, \boldsymbol{x}_i)}}{\sqrt{n}} \tag{17}$$

Letting $C = 2L_R\sqrt{K \cdot \max_i \Theta_S(\boldsymbol{x}_i, \boldsymbol{x}_i)}$ (where the $\sqrt{K}$ factor comes from the multi-class extension), and replacing the fixed radius B with the specific norms of our solutions, we can establish two bounds that hold simultaneously with probability at least $1-\delta$ by a union bound:

$$R(h_S^{\mathrm{CCKD}}) \leq \hat{R}_n(h_S^{\mathrm{CCKD}}) + \frac{C}{\sqrt{n}}\|h_S^{\mathrm{CCKD}}\| + 2L_R\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{18}$$

$$\hat{R}_n(h_S^{\mathrm{KD}}) \leq R(h_S^{\mathrm{KD}}) + \frac{C}{\sqrt{n}}\|h_S^{\mathrm{KD}}\| + 2L_R\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{19}$$

Step 2: CCKD Optimality Condition. This step remains unchanged. From the optimality of h_S^{CCKD}:

$$\hat{R}_n(h_S^{\mathrm{CCKD}}) \leq \hat{R}_n(h_S^{\mathrm{KD}}) - \lambda\Delta_G - \frac{\mu}{2}\Delta_{\|\cdot\|^2} \tag{20}$$

Step 3: Combining the Inequalities. We chain the inequalities. Start with Eq. 18, substitute Eq. 20, and then substitute Eq. 19:

$$R(h_S^{\text{CCKD}}) \leq \left(\hat{R}_n(h_S^{\text{KD}}) - \lambda\Delta_G - \frac{\mu}{2}\Delta_{\|\cdot\|^2}\right) + \frac{C}{\sqrt{n}}\|h_S^{\text{CCKD}}\| + 2L_R\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{21}$$

$$\leq \left(R(h_S^{\text{KD}}) + \frac{C}{\sqrt{n}}\|h_S^{\text{KD}}\| + 2L_R\sqrt{\frac{\ln(2/\delta)}{2n}}\right) \tag{22}$$

$$- \lambda\Delta_G - \frac{\mu}{2}\Delta_{\|\cdot\|^2} + \frac{C}{\sqrt{n}}\|h_S^{\text{CCKD}}\| + 2L_R\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{23}$$

Collecting terms yields the desired result:

$$R(h_S^{\text{CCKD}}) \leq R(h_S^{\text{KD}}) - \lambda\Delta_G - \frac{\mu}{2}\Delta_{\|\cdot\|^2} + \frac{C}{\sqrt{n}}(\|h_S^{\text{KD}}\| + \|h_S^{\text{CCKD}}\|) + 4L_R\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{24}$$

□

Theorem 3.4 (Sufficient Condition for Superior Generalization Bound)**.** *For a sufficiently large sample size n, there exists a non-empty range for the hyperparameter λ, such that for any λ chosen from this range, the generalization bound for h_S^{CCKD} is strictly tighter than that for h_S^{KD}.*

Detailed Proof. From the bound in Theorem 3.3, the generalization bound for h_S^{CCKD} is strictly tighter if the sum of the non-$R(h_S^{\text{KD}})$ terms is negative. This requires the positive improvement terms to outweigh the complexity and statistical error terms:

$$\lambda\Delta_G + \frac{\mu}{2}\Delta_{\|\cdot\|^2} > \frac{C}{\sqrt{n}}(\|h_S^{\text{CCKD}}\| + \|h_S^{\text{KD}}\|) + 4L_R\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{25}$$

We analyze the order of magnitude of each side with respect to λ and n.

Left-Hand Side (LHS) Analysis: We analyze the behavior of the LHS, $\lambda\Delta_G + \frac{\mu}{2}\Delta_{\|\cdot\|^2}$, for small $\lambda > 0$, based on the perturbation analysis in the preceding lemmas.

- **Norm Reduction Term:** The proof of Lemma 3.1 is based on the first-order Taylor expansion of $\phi(\lambda) = \|h(\lambda)\|^2$. This shows that $\Delta_{\|\cdot\|^2} = \|h_S^{\text{KD}}\|^2 - \|h_S^{\text{CCKD}}\|^2 \approx -\lambda\phi'(0)$. Since $\phi'(0)$ is a strictly negative constant (by Assumption 3), $\Delta_{\|\cdot\|^2}$ is of order $O(\lambda)$. Consequently, the term $\frac{\mu}{2}\Delta_{\|\cdot\|^2}$ is of order $O(\lambda)$.
- **Feature Alignment Term:** Similarly, the proof of Lemma 3.2 shows that $\Delta_G = G(h_S^{\text{KD}}) - G(h_S^{\text{CCKD}})$ is of order $O(\lambda)$. Therefore, the term $\lambda\Delta_G$ is $\lambda \cdot O(\lambda) = O(\lambda^2)$.

The entire LHS is the sum $O(\lambda) + O(\lambda^2)$. For sufficiently small λ, the linear term dominates, and thus the LHS is of order $O(\lambda)$.

Right-Hand Side (RHS) Analysis: The norms $\|h_S^{\text{CCKD}}\|$ and $\|h_S^{\text{KD}}\|$ are bounded, and the constants C and L_R are independent of n and λ. The entire RHS is therefore the sum of a complexity term and a statistical term, both of which scale with the sample size as $O(n^{-1/2})$.

Condition for Improvement: For the inequality to hold, the LHS must be larger than the RHS. This translates to the condition:

$$O(\lambda) > O(n^{-1/2})$$

This implies that for a guaranteed improvement, λ must be chosen such that $\lambda > c_1/\sqrt{n}$ for some problem-dependent constant $c_1 > 0$. At the same time, our perturbation analysis is only valid for λ within a neighborhood of zero defined by some constant $\lambda_0 > 0$.

For any sufficiently large sample size n, it is possible to make the lower bound $c_1/\sqrt{n}$ smaller than the fixed upper bound λ_0. This guarantees the existence of a non-empty interval $(\frac{c_1}{\sqrt{n}}, \lambda_0)$ for λ. Choosing λ from this interval ensures that the improvement from CCKD's regularization is large enough to overcome the statistical error terms, resulting in a strictly tighter generalization bound. □

4 Discussion and Interpretation

Our analysis provides a formal theoretical justification for the generalization benefits of CCKD within the NTK framework. However, the scope of these results is defined by its underlying assumptions and the nature of the bounds. We discuss the key interpretations and limitations below.

1. **On the Nature of Assumption 3**: Our central theoretical result, the CCKD-induced norm reduction (Lemma 3.1), hinges on this geometric alignment assumption. While we have replaced a direct assumption about norm reduction with this more foundational condition, it is non-trivial. It posits a positive correlation between the solution vector h_S^{KD} and a 'softened' gradient direction of the feature alignment loss, $\nabla_h G$. The softening operator, $(H_{J_{\text{KD}}})^{-1}$, amplifies directions where the original KD loss landscape is flat (i.e., has small eigenvalues). The assumption therefore suggests that CCKD's regularization is most effective when the direction of desired feature alignment coincides with directions of low curvature in the KD loss landscape. Verifying this condition for complex, high-dimensional models remains an open and challenging question.

2. **The Trade-off between λ and n**: Theorem 3.4 and its proof reveal a crucial trade-off. The CCKD hyperparameter λ cannot be chosen in isolation. Our analysis

imposes two simultaneous constraints on λ for a tighter generalization bound to be guaranteed. First, the perturbation analysis in Lemmas 3.1-3.2, which underpins the entire argument for norm reduction, is valid only for λ within a small neighborhood of zero, i.e., $\lambda \in (0, \lambda_0)$, where λ_0 is a constant determined by the local geometry of the loss landscape. Second, to realize a tangible improvement in the generalization bound, the benefit from CCKD (which is of order $O(\lambda)$) must overcome the statistical and complexity error terms (of order $O(n^{-1/2})$). This requires $\lambda > c_1/\sqrt{n}$ for some problem-dependent constant $c_1 > 0$.

For any sufficiently large sample size n, we can make the lower bound $c_1/\sqrt{n}$ smaller than the fixed upper bound λ_0. Thus, a non-empty interval $(\frac{c_1}{\sqrt{n}}, \lambda_0)$ for λ exists. Choosing any λ from this 'sweet spot' guarantees a strictly tighter generalization bound. This formalizes the empirical observation that λ is a sensitive hyperparameter: too small, and the feature-matching effect is lost in statistical noise; too large, and it may drastically alter the solution, potentially moving away from the favorable local landscape around the KD solution where our analysis holds.

3. **Limitations of the NTK Regime**: Our entire analysis is predicated on Assumption 1, which places us in the infinite-width or NTK regime where the network behaves like a linear model in function space with a fixed kernel. While this provides a powerful avenue for theoretical analysis, modern neural networks often operate in a "rich" or "feature learning" regime where the kernel itself evolves during training. Extending this analysis beyond the fixed-kernel setting is a significant challenge. Our results are therefore most directly applicable to overparameterized networks trained with small learning rates, aligning with the conditions of the NTK theory.

4. **Looseness of Generalization Bounds**: The Rademacher complexity-based bounds used in our proof are known to be quantitatively loose in practice, often failing to explain the actual generalization performance of deep neural networks. Therefore, the primary value of Theorem 3.3 and 3.4 is not in predicting the precise generalization error, but rather in providing a rigorous, qualitative comparison. They demonstrate a mechanism through which CCKD can improve upon KD by imposing a form of regularization (RKHS norm reduction) that is provably beneficial from a statistical learning theory perspective.

References

[1] Baoyun Peng, Xiao Jin, Dongsheng Li, Shunfeng Zhou, Yichao Wu, Jiaheng Liu, Zhaoning Zhang, and Yu Liu. Correlation congruence for knowledge distillation.

In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5006–5015. IEEE, 2019.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[3] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[4] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[5] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[6] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.

[7] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[8] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[9] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

[10] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.