Generalization Analysis for CCKD: A Stability and Curvature Perspective

Abstract

This document analyzes the generalization of CCKD theoretically. We build on the Neural Tangent Kernel (NTK) framework and use stability analysis of optimization dynamics to derive a new geometric condition. Under this condition, CCKD reduces the norm in the RKHS, which lets us prove a formal generalization bound. This shows how CCKD's feature alignment improves generalization over standard KD [**?**], clarifying the link between feature matching, loss curvature, and model complexity.

1 Preliminaries and Notation

We define key concepts and establish the critical link between feature Gram matrices and NTK, laying the groundwork for the core logical chain.

1.1 Basic Setup

Definition 1 (Input/Label Space and Models)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^p$ (input space) and $\mathcal{Y} = \{1, \ldots, K\}$ (label space). Data is sampled from an unknown distribution $\mathcal{D} \sim \mathcal{X} \times \mathcal{Y}$, with training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$. We study two networks:*

- *A **fixed, fully trained teacher** $h_T : \mathcal{X} \to \mathbb{R}^K$ (with generalization error converged to a stable minimum).*
- *A **trainable student** $h_S : \mathcal{X} \to \mathbb{R}^K$ (to be optimized via KD/CCKD).*

1.2 NTK, RKHS, and Representer Theorem

Definition 2 (Neural Tangent Kernel (NTK) and RKHS)**.** *For a student network parameterized by $\boldsymbol{\Phi}$, the NTK is a kernel function that captures gradient similarity:*

$$\Theta(\boldsymbol{x}, \boldsymbol{x}') = tr\left(\nabla_{\boldsymbol{\Phi}} h(\boldsymbol{x})^\top \nabla_{\boldsymbol{\Phi}} h(\boldsymbol{x}')\right)$$

The NTK matrix on training set S is $\boldsymbol{\Theta}_S = [\Theta(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n}$.

The kernel Θ uniquely defines a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\Theta$ [?, ?]. For any $h \in \mathcal{H}_\Theta$, the RKHS norm is defined as $\|h\|_{\mathcal{H}_\Theta}^2 = \langle h, h\rangle_{\mathcal{H}_\Theta}$. By the Representer Theorem, any minimizer of a regularized empirical risk in $\mathcal{H}_\Theta$ can be expressed as $h(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i \Theta(\boldsymbol{x}_i, \boldsymbol{x})$ with $\|h\|_{\mathcal{H}_\Theta}^2 = \sum_{i,j=1}^n \alpha_i\alpha_j\Theta(\boldsymbol{x}_i, \boldsymbol{x}_j)$, which is a critical metric for complexity via Rademacher theory.

1.3 Feature Gram Matrix and NTK Decomposition

Definition 3 (Feature Gram Matrix and NTK Decomposition)**.** *For networks with a **linear final layer**, any function $h : \mathcal{X} \to \mathbb{R}^K$ can be decomposed as $h(\boldsymbol{x}) = W f(\boldsymbol{x}) + b$, where $f : \mathcal{X} \to \mathbb{R}^d$ is the **pre-final feature extractor** (the output of the network before the final linear layer), $W \in \mathbb{R}^{K\times d}$ is the final layer weight matrix, and $b \in \mathbb{R}^K$ is the bias.*

In the NTK regime with a fixed network architecture, the feature extractor f is implicitly defined by the network's structure and the parameters that realize a given function $h \in \mathcal{H}_{\Theta_S}$. While the mapping $h \mapsto f$ may not be unique in general, we work within a fixed parameterization where the feature Gram matrix $\boldsymbol{K}_S(h) = [f(\boldsymbol{x}_i)^T f(\boldsymbol{x}_j)]_{i,j}$ is a well-defined object associated with each training outcome.

For a function h with associated feature extractor f, the feature Gram matrix on training set S is:

$$\boldsymbol{K}_S(h) = [\langle f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\rangle]_{i,j=1}^n \in \mathbb{R}^{n\times n}$$

We use $\boldsymbol{K}_S(h_S)$ for the student and $\boldsymbol{K}_T = \boldsymbol{K}_S(h_T)$ for the teacher.

*For networks with a **linear final layer**, the NTK decomposes into two additive components:*

$$\Theta(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{K}_{feat}(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{K}_{deep}(\boldsymbol{x}, \boldsymbol{x}')$$

where:

- $\boldsymbol{K}_{feat}(\boldsymbol{x}, \boldsymbol{x}') = \langle f(\boldsymbol{x}), f(\boldsymbol{x}')\rangle$ *(feature component, linked to the Gram matrix),*
- $\boldsymbol{K}_{deep}(\boldsymbol{x}, \boldsymbol{x}')$ *(deep component, from gradients of non-final layers, e.g., convolutions).*

The function $G(h) = \frac{\lambda}{n^2}\|\boldsymbol{K}_S(h) - \boldsymbol{K}_T\|_F^2$ is well-defined on $\mathcal{H}_{\Theta_S}$ under our framework.

*For a function $h \in \mathcal{H}_{\Theta_S}$, we define its **effective kernel matrix** on training set S as:*

$$\boldsymbol{K}_S^{eff}(h) = [\langle h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)\rangle]_{i,j=1}^n \in \mathbb{R}^{n\times n}$$

where $\langle h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)\rangle = \sum_{k=1}^K h_k(\boldsymbol{x}_i)h_k(\boldsymbol{x}_j)$ is the inner product of output vectors. This matrix captures the correlation structure of function outputs, directly reflecting the model's behavioral characteristics.

Remark 1 (Connection to NTK)**.** *In the NTK regime, by the Representer Theorem, any vector-valued function $h \in \mathcal{H}_{\Theta_S}$ ($h : \mathcal{X} \to \mathbb{R}^K$) can be expressed as:*

$$h(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \Theta_S(\boldsymbol{x}_i, \boldsymbol{x})$$

where $\boldsymbol{\alpha}_i \in \mathbb{R}^K$ are coefficient vectors determined by the training objective.

The effective kernel matrix can then be written in terms of the NTK matrix $\boldsymbol{\Theta}_S$ and the coefficients:

$$\boldsymbol{K}_S^{\mathit{eff}}(h) = \boldsymbol{\Theta}_S^T G \boldsymbol{\Theta}_S$$

where $G = AA^T \in \mathbb{R}^{n \times n}$ is the Gram matrix of the coefficient vectors, with $A = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]^T \in \mathbb{R}^{n \times K}$.

This decomposition is crucial: while $\boldsymbol{\Theta}_S$ is fixed in the NTK regime, $\boldsymbol{K}_S^{\mathit{eff}}(h)$ varies through the coefficient Gram matrix G, which depends on the training method. This allows us to rigorously discuss "effective kernel alignment" without contradicting the fixed nature of the NTK.

1.4 Learning Objectives

Definition 4 (KD/CCKD Objectives)**.** *Let ℓ_{CE} (cross-entropy), ℓ_{KL} (KL divergence), $\alpha \in [0, 1]$ (distillation weight), $\mu > 0$ (weight decay), and $\lambda > 0$ (feature alignment strength).*

1. ***Standard KD****: Minimizes a mix of classification and distillation loss:*

$$\begin{aligned} J_{KD}(h) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} [(1-\alpha)\ell_{CE}(h(\boldsymbol{x}_i), y_i)}_{\hat{R}_n(h)} \\ + \alpha \ell_{KL}(h(\boldsymbol{x}_i), h_T(\boldsymbol{x}_i))] \\ + \frac{\mu}{2} \|h\|_{\mathcal{H}_\Theta}^2 \end{aligned} \tag{1}$$

The optimal KD student is $h_S^{KD} = \arg\min_{h \in \mathcal{H}_\Theta} J_{KD}(h)$.

2. ***CCKD****: Adds a feature alignment term to KD's objective:*

$$J_{CCKD}(h) = J_{KD}(h) + \underbrace{\frac{\lambda}{n^2} \|\boldsymbol{K}_S(h) - \boldsymbol{K}_T\|_F^2}_{G(h)}$$

The optimal CCKD student is $h_S^{CCKD} = \arg\min_{h \in \mathcal{H}_\Theta} J_{CCKD}(h)$.

2 Assumptions

We group our assumptions by their role: establishing the foundational analysis framework, postulating the core mechanisms that drive our results, and ensuring the technical regularity required for the proofs.

2.1 Foundational Framework

Assumption 1 (NTK Regime and Teacher Quality)**.** *We operate within the standard Neural Tangent Kernel (NTK) framework.*

1. ***NTK Regime****: The student's training stays in the NTK regime, where its kernel Θ_S is fixed and its function h_S evolves within the corresponding RKHS, $\mathcal{H}_{\Theta_S}$.*
2. ***Teacher Boundedness****: The teacher h_T is a well-behaved function with a bounded RKHS norm, $\|h_T\|_{\mathcal{H}_{\Theta_T}} \leq B_T$.*

2.2 Core Mechanism Assumptions

These two assumptions form the conceptual core of our argument, each powering one of the main logical paths.

Assumption 2 (Feature-to-Behavior Proxy (Bi-Lipschitz))**.** *The alignment of feature Gram matrices serves as a valid proxy for the alignment of the model's output behavior. Formally, there exist constants $L, c > 0$ such that for any two student functions h_1, h_2 within a bounded subset of the RKHS:*

$$\begin{aligned} c\|\boldsymbol{K}_S(h_1) - \boldsymbol{K}_S(h_2)\|_F &\leq \|\boldsymbol{K}_S^{\mathit{eff}}(h_1) - \boldsymbol{K}_S^{\mathit{eff}}(h_2)\|_F \\ &\leq L\|\boldsymbol{K}_S(h_1) - \boldsymbol{K}_S(h_2)\|_F \end{aligned} \tag{2}$$

This bi-Lipschitz condition ensures that feature alignment and behavioral alignment are equivalent up to constants. The lower bound is crucial for proving that a strict improvement in feature alignment implies a strict improvement in behavioral alignment. This assumption is justified for networks with stable linear final layers (where the weight matrix W has bounded condition number).

Assumption 3 (Geometric Condition for Norm Reduction)**.** *At the KD optimum h_S^{KD}, the introduction of the feature-matching gradient term $G(h)$ pushes the solution "towards the origin" in the RKHS. Formally:*

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

As explained by the optimization intuition, this geometric alignment ensures that minimizing the CCKD objective also reduces model complexity (Lemma ??).

Remark 2 (Geometric Intuition of Assumption ??)**.** *The condition in Assumption ?? states that:*

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

*This has a clear geometric interpretation. The term $(H_{J_{KD}})^{-1}\nabla_h G$ represents the **natural gradient** direction for reducing the feature alignment loss G, which accounts for the curvature of the main KD loss landscape. The assumption posits that this direction is positively correlated with the position vector of the KD solution h_S^{KD} itself.*

In the proof of Lemma ??, we show that the derivative of the squared norm is $\phi'(0) = -2\langle h_S^{KD}, (H_{J_{KD}})^{-1}\nabla_h G\rangle$. Therefore, this assumption is equivalent to stating that adding the feature alignment term causes the solution to initially move towards the origin in the RKHS, thus reducing its norm.

***When does this hold?** This condition is expected to hold when the teacher's feature structure represents a "simpler" solution (in the sense of having a smaller RKHS norm) than what standard KD produces. Intuitively, if the teacher has learned a more parsimonious or regularized representation, then aligning the student's features to match the teacher's should guide the student towards a simpler solution as well. This is consistent with the view of knowledge distillation as a form of regularization, where the teacher provides a better-regularized target.*

2.3 Technical Regularity Conditions

These are standard assumptions required to ensure the loss landscapes are well-behaved and that our mathematical tools are applicable.

Assumption 4 (Convexity, and Lipschitz Continuity)**.** *1. **Differentiability and Convexity**: The loss components $\hat{R}_n(h)$ and $G(h)$ are convex and twice Fréchet differentiable in a neighborhood of h_S^{KD}. This ensures the existence of unique minimizers and allows for perturbation analysis via the implicit function theorem.*

*2. **Lipschitz Continuity**: The empirical risk $\hat{R}_n(h)$ is L_R-Lipschitz with respect to the model's outputs. This is a standard requirement for applying Rademacher complexity bounds.*

2.4 Lemma 1: CCKD Ensures Feature Alignment

Lemma 1 (Feature Alignment Guarantee)**.** *For any $\lambda > 0$:*

1. $G(h_S^{CCKD}) \leq G(h_S^{KD})$ (non-strict reduction of feature alignment loss).

*2. If $\nabla_h G(h_S^{KD}) \neq 0$, the inequality is **strict** for small $\lambda > 0$ (CCKD strictly improves feature alignment).*

Proof. **Part 1: Non-Strict Inequality**. By definition of minimizers: $J_{\mathrm{KD}}(h_S^{\mathrm{KD}}) \leq J_{\mathrm{KD}}(h_S^{\mathrm{CCKD}})$ and $J_{\mathrm{CCKD}}(h_S^{\mathrm{CCKD}}) \leq J_{\mathrm{CCKD}}(h_S^{\mathrm{KD}})$.

Expanding:

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

Subtracting the second from the first:

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

which yields $G(h_S^{\mathrm{CCKD}}) \leq G(h_S^{\mathrm{KD}})$ for all $\lambda > 0$.

Part 2: Strict Inequality for Small λ. Let $h(\lambda)$ be the minimizer of J_{CCKD} as a function of λ, with $h(0) = h_S^{\mathrm{KD}}$. Define $\psi(\lambda) = G(h(\lambda))$. We will show that if $\nabla_h G(h_S^{\mathrm{KD}}) \neq 0$, then $\psi'(0) < 0$, implying strict reduction for small $\lambda > 0$.

The optimality condition for $h(\lambda)$ is:

$$\nabla_h J_{\mathrm{KD}}(h(\lambda)) + \lambda \nabla_h G(h(\lambda)) = 0$$

At $\lambda = 0$, this gives $\nabla_h J_{\mathrm{KD}}(h_S^{\mathrm{KD}}) = 0$ (KD optimality).

The Hessian operator $H_{J_{\mathrm{KD}}}(h_S^{\mathrm{KD}})$ is positive definite by Assumption **??** and invertible by Assumption **??**. Thus, we can solve for the initial direction $h'(0)$:

$$h'(0) = -\left(H_{J_{\mathrm{KD}}}(h_S^{\mathrm{KD}})\right)^{-1} \nabla_h G(h_S^{\mathrm{KD}})$$

The derivative of ψ at $\lambda = 0$ is:

$$\begin{aligned} \psi'(0) &= \left\langle \nabla_h G(h_S^{\mathrm{KD}}), h'(0) \right\rangle_{\mathcal{H}_{\Theta_S}} \\ &= -\left\langle \nabla_h G(h_S^{\mathrm{KD}}), \left(H_{J_{\mathrm{KD}}}(h_S^{\mathrm{KD}})\right)^{-1} \nabla_h G(h_S^{\mathrm{KD}}) \right\rangle_{\mathcal{H}_{\Theta_S}} \end{aligned} \tag{3}$$

Since $H_{J_{\mathrm{KD}}}$'s inverse is positive definite and $\nabla_h G(h_S^{\mathrm{KD}}) \neq 0$, the inner product is strictly positive, implying $\psi'(0) < 0$. By the continuity of $\psi(\lambda)$ (from Assumption **??**), this negative derivative ensures that $\psi(\lambda) < \psi(0)$ for small $\lambda > 0$, proving the strict inequality.

□

2.5 Lemma 2: CCKD Enables Effective Kernel Alignment

Lemma 2 (Effective Kernel Alignment via Feature Control)**.** *Under Assumptions* **??** *and* **??***, CCKD achieves strictly better alignment of the student's effective kernel matrix to the teacher's, compared to KD:*

$$\|\boldsymbol{K}_S^{\mathit{eff}}(h_S^{\mathit{CCKD}}) - \boldsymbol{K}_T^{\mathit{eff}}(h_T)\|_F < \|\boldsymbol{K}_S^{\mathit{eff}}(h_S^{\mathit{KD}}) - \boldsymbol{K}_T^{\mathit{eff}}(h_T)\|_F$$

Proof. The proof relies on the bi-Lipschitz relationship between feature Gram matrices and effective kernel matrices established in Assumption **??**, combined with the strict improvement in feature alignment from Lemma **??**.

From Lemma **??**, for small $\lambda > 0$:

$$\|\boldsymbol{K}_S(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_T\|_F < \|\boldsymbol{K}_S(h_S^{\mathrm{KD}}) - \boldsymbol{K}_T\|_F$$

Let $\Delta_{\mathrm{feat}} = \|\boldsymbol{K}_S(h_S^{\mathrm{KD}}) - \boldsymbol{K}_T\|_F - \|\boldsymbol{K}_S(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_T\|_F > 0$ denote the strict improvement in feature alignment.

By the triangle inequality:

$$\begin{aligned} &\|\boldsymbol{K}_S(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_S(h_S^{\mathrm{KD}})\|_F \\ &\leq \|\boldsymbol{K}_S(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_T\|_F + \|\boldsymbol{K}_T - \boldsymbol{K}_S(h_S^{\mathrm{KD}})\|_F \\ &= 2\|\boldsymbol{K}_S(h_S^{\mathrm{KD}}) - \boldsymbol{K}_T\|_F - \Delta_{\mathrm{feat}} \end{aligned} \tag{4}$$

Now we apply the bi-Lipschitz condition from Assumption **??**. By the lower bound:

$$\|\boldsymbol{K}_S^{\mathrm{eff}}(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_S^{\mathrm{eff}}(h_S^{\mathrm{KD}})\|_F \geq c\|\boldsymbol{K}_S(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_S(h_S^{\mathrm{KD}})\|_F$$

By the reverse triangle inequality applied to effective kernel matrices:

$$\begin{aligned} &\left|\|\boldsymbol{K}_S^{\mathrm{eff}}(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_T^{\mathrm{eff}}\|_F - \|\boldsymbol{K}_S^{\mathrm{eff}}(h_S^{\mathrm{KD}}) - \boldsymbol{K}_T^{\mathrm{eff}}\|_F\right| \\ &\qquad \leq \|\boldsymbol{K}_S^{\mathrm{eff}}(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_S^{\mathrm{eff}}(h_S^{\mathrm{KD}})\|_F \end{aligned}$$

The key insight is that the bi-Lipschitz condition establishes a two-way relationship: improvements in feature space translate to improvements in effective kernel space, and vice versa. Since we have a strict improvement $\Delta_{\mathrm{feat}} > 0$ in feature alignment, and the bi-Lipschitz constants c, L are finite and positive, the mapping preserves the strict inequality.

More formally, suppose for contradiction that $\|\boldsymbol{K}_S^{\mathrm{eff}}(h_S^{\mathrm{CCKD}}) - \boldsymbol{K}_T^{\mathrm{eff}}\|_F \geq \|\boldsymbol{K}_S^{\mathrm{eff}}(h_S^{\mathrm{KD}}) -$

$\boldsymbol{K}_T^{\text{eff}}\|_F$. Then by the lower bound of the bi-Lipschitz condition applied in reverse (using the fact that the effective kernel distance cannot improve without the feature distance improving), we would have $\|\boldsymbol{K}_S(h_S^{\text{CCKD}}) - \boldsymbol{K}_T\|_F \geq \|\boldsymbol{K}_S(h_S^{\text{KD}}) - \boldsymbol{K}_T\|_F$, contradicting Lemma **??**.

Therefore, the strict inequality holds:

$$\|\boldsymbol{K}_S^{\text{eff}}(h_S^{\text{CCKD}}) - \boldsymbol{K}_T^{\text{eff}}(h_T)\|_F < \|\boldsymbol{K}_S^{\text{eff}}(h_S^{\text{KD}}) - \boldsymbol{K}_T^{\text{eff}}(h_T)\|_F$$

□

2.6 Lemma 3: CCKD Reduces RKHS Norm

Lemma 3 (RKHS Norm Reduction)**.** *Under all Assumptions (**??**–**??**), there exists $\lambda_0 > 0$ such that for all $0 < \lambda < \lambda_0$:*

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

Proof. Let $h(\lambda)$ be the minimizer of J_{CCKD}. The Fréchet differentiability and convexity from Assumption **??** together with the invertibility from Assumption **??** allow us to apply the implicit function theorem [**?**]. This guarantees that $h(\lambda)$ is a unique, continuously differentiable function of λ near $\lambda = 0$.

We analyze the squared norm $\phi(\lambda) = \|h(\lambda)\|^2_{\mathcal{H}_{\Theta_S}}$. Its derivative is $\phi'(\lambda) = 2\langle h(\lambda), h'(\lambda)\rangle_{\mathcal{H}_{\Theta_S}}$.

From the proof of Lemma **??**, we have the expression for the initial direction:

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

Evaluating the derivative of the norm at $\lambda = 0$:

$$\begin{aligned} \phi'(0) &= 2\langle h(0), h'(0)\rangle_{\mathcal{H}_{\Theta_S}} \\ &= -2\left\langle h_S^{\text{KD}}, \left(H_{J_{\text{KD}}}(h_S^{\text{KD}})\right)^{-1} \nabla_h G(h_S^{\text{KD}})\right\rangle_{\mathcal{H}_{\Theta_S}} \end{aligned} \tag{5}$$

By Assumption **??**, the inner product term is strictly positive. Therefore, $\phi'(0) < 0$. Since $\phi(\lambda)$ is continuous (by the differentiability of $h(\lambda)$), this negative derivative implies that for a small $\lambda_0 > 0$, we have $\phi(\lambda) < \phi(0)$ for all $0 < \lambda < \lambda_0$. This proves that the RKHS norm is reduced. □

2.7 Theorem 1: CCKD Reduces Rademacher Complexity

Theorem 3 (Rademacher Complexity Bound for CCKD)**.** *Under all Assumptions (**??**–**??**), the Rademacher complexity of the function class containing the CCKD solution is smaller*

than that for the KD solution:

$$\hat{\mathfrak{R}}_n(\mathcal{F}_{\|h_S^{CCKD}\|}) < \hat{\mathfrak{R}}_n(\mathcal{F}_{\|h_S^{KD}\|})$$

where $\mathcal{F}_B = \{h \in \mathcal{H}_{\Theta_S} : \|h\|_{\mathcal{H}_{\Theta_S}} \leq B\}$ *is the RKHS ball of radius* B*.*

Proof. The empirical Rademacher complexity of an RKHS ball $\mathcal{F}_B$ is bounded by $\hat{\mathfrak{R}}_n(\mathcal{F}_B) \leq \frac{CB}{\sqrt{n}}$, where C is a constant depending on the Lipschitz constant L_R (from Assumption **??**) and properties of the kernel Θ_S [**?**]. Since complexity is monotonic with the radius B, and Lemma **??** proves $\|h_S^{\mathrm{CCKD}}\|_{\mathcal{H}_{\Theta_S}} < \|h_S^{\mathrm{KD}}\|_{\mathcal{H}_{\Theta_S}}$, the result follows directly. □

2.8 Theorem 2: CCKD Has a Tighter Generalization Bound

Theorem 4 (Comparative Generalization Bound)**.** *Under all Assumptions (**??**–**??**), with probability at least* $1 - \delta$*:*

$$R(h_S^{CCKD}) \leq \hat{R}_n(h_S^{CCKD}) + \frac{2C\|h_S^{CCKD}\|_{\mathcal{H}_{\Theta_S}}}{\sqrt{n}} + 2L_R\sqrt{\frac{\ln(2/\delta)}{2n}}$$

This upper bound on the true risk $R(h_S^{CCKD})$ *is tighter than the corresponding bound for KD. The improvement comes from two sources:*

1. *A smaller empirical risk term:* $\hat{R}_n(h_S^{CCKD}) \leq \hat{R}_n(h_S^{KD}) - \lambda\Delta_G - \frac{\mu}{2}\Delta_{\|\cdot\|^2}$*, where* Δ_G *and* $\Delta_{\|\cdot\|^2}$ *are the non-negative gains from feature alignment and norm reduction.*
2. *A smaller complexity term: The term* $\frac{2C\|h\|_{\mathcal{H}_{\Theta_S}}}{\sqrt{n}}$ *is smaller for CCKD due to the norm reduction shown in Lemma* **??***.*

Proof. The standard Rademacher complexity bound states that for any h, with probability at least $1 - \delta$:

$$R(h) \leq \hat{R}_n(h) + 2\hat{\mathfrak{R}}_n(\mathcal{F}_{\|h\|}) + 2L_R\sqrt{\frac{\ln(2/\delta)}{2n}}$$

The Lipschitz constant L_R is guaranteed by Assumption **??**.

First, we establish the empirical risk relationship. By the optimality of h_S^{CCKD} for J_{CCKD}:

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

Rearranging gives the bound on the empirical risk:

$$\begin{aligned}\hat{R}_n(h_S^{\mathrm{CCKD}}) \leq \hat{R}_n(h_S^{\mathrm{KD}}) &- \lambda(G(h_S^{\mathrm{KD}}) - G(h_S^{\mathrm{CCKD}})) \\ &- \frac{\mu}{2}(\|h_S^{\mathrm{KD}}\|^2 - \|h_S^{\mathrm{CCKD}}\|^2)\end{aligned} \tag{6}$$

Since both gain terms are non-negative (by Lemmas 1 and 3), we have $\hat{R}_n(h_S^{\mathrm{CCKD}}) \leq \hat{R}_n(h_S^{\mathrm{KD}})$.

Second, we analyze the complexity term. Using the bound from Theorem **??**, we have:

$$2\hat{\mathfrak{R}}_n(\mathcal{F}_{\|h_S^{\mathrm{CCKD}}\|}) \leq \frac{2C\|h_S^{\mathrm{CCKD}}\|_{\mathcal{H}_{\Theta_S}}}{\sqrt{n}} < \frac{2C\|h_S^{\mathrm{KD}}\|_{\mathcal{H}_{\Theta_S}}}{\sqrt{n}}$$

When plugging these two improvements into the general bound for $R(h_S^{\mathrm{CCKD}})$, it becomes evident that the resulting bound is tighter than the one for $R(h_S^{\mathrm{KD}})$. □

2.9 Theorem 3: Existence of a Valid λ Range

Theorem 5 (Valid λ Range)**.** *For a sufficiently large sample size n, there exists a non-empty range of λ values for which the generalization advantage of CCKD is non-trivial.*

Proof. (Proof sketch) The total gain in the generalization bound is roughly $\Delta_{\mathrm{Gain}} \approx \lambda\Delta_G + \frac{\mu}{2}\Delta_{\|\cdot\|^2} + \frac{2C}{\sqrt{n}}(\|h_S^{\mathrm{KD}}\| - \|h_S^{\mathrm{CCKD}}\|)$. From the proofs of Lemma 1 and 3, all gain terms (Δ_G, $\Delta_{\|\cdot\|^2}$, and the norm difference) are of order $O(\lambda)$. Thus, $\Delta_{\mathrm{Gain}} = O(\lambda)$. For this gain to be meaningful, it must be larger than the statistical noise term, which is of order $O(n^{-1/2})$. This requires $\lambda \gg O(n^{-1/2})$. Simultaneously, our perturbation analysis in Lemma **??** requires λ to be smaller than some constant λ_0. Therefore, for large enough n, a non-empty range $\lambda \in (c_1 n^{-1/2}, \lambda_0)$ exists where CCKD provides a guaranteed theoretical improvement. The existence of λ_0 is guaranteed by the implicit function theorem under Assumptions **??** and **??**. □

3 Limitations

The analysis is constrained by its assumptions, which define its practical applicability:

1. **NTK Regime**: Results apply only to networks operating in the NTK regime, where the kernel remains approximately fixed during training. This typically requires wide networks (width $\gg$ depth). In practice, smaller student networks used in knowledge distillation may not fully satisfy this condition, though the analysis may still provide qualitative insights.
2. **Linear Final Layer**: The feature-output relationship (Assumption **??**) and effective kernel alignment (Lemma 2) require a linear final layer. Many modern architectures satisfy this, but the assumption may be restrictive for networks with non-linear output layers.
3. **Bi-Lipschitz Feature-Output Relationship**: Assumption **??** assumes a bi-Lipschitz

relationship between feature Gram matrices and effective kernel matrices, quantified by constants $c, L > 0$. This assumption connects feature-level alignment to output-level behavioral alignment. While theoretically justified for networks with stable linear final layers (bounded weight matrix condition number), its practical validity depends on final layer stability. The constants c, L may vary across different network architectures and training conditions, requiring empirical validation in practice.

4. **Small λ Range**: Valid λ is restricted to $\lambda \in (\frac{c_1}{\sqrt{n}}, \lambda_0)$ (Theorem 3). For large n, the lower bound becomes very small, while the upper bound λ_0 depends on network architecture and training dynamics. This may limit hyperparameter tuning flexibility in practice, though empirical validation can guide λ selection.
5. **Assumption Dependencies**: The results rely on Assumptions **??** and **??**, whose practical validity depends on network architecture and training conditions. Further empirical or theoretical validation of these assumptions would strengthen the analysis.
6. **Loose Bounds**: Rademacher complexity bounds provide qualitative insights about generalization but may not quantitatively predict real-world generalization error due to potentially large constants and asymptotic nature.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[2] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[3] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer Science & Business Media, 2009.

[4] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[5] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.