COMS 6998: Theoretical Foundations of LLMs Spring 2025

Progressive distillation learns leaky k-juntas

Final project

Shiv Kampani

1 Introduction

Knowledge distillation is the idea of using a (larger) teacher model to train a small student model at a particular task. At a high level, the logits of the teacher model provide the student with more information about the problem than hard labels. In a fascinating recent paper titled "Progressive distillation induces an implicit curriculum," Panigrahi et al. proved that progressive knowledge distillation (using multiple teacher model checkpoints) is more sample efficient than "one-shot distillation" for the sparse-parity problem. Sparse parity is an instance of a class of problems called k-juntas (I first heard about learning juntas from Rocco Servedio and his famous paper on the same topic). In this short brief, I prove a similar result as Panigrahi et al. but for a sub-class of k-juntas that I call (degree-1) leaky k-juntas. A "useful" example from this class is sparse majority.

2 Preliminaries

2.1 Analysis of boolean functions

Before proceeding to the proof of the central theorem of the project, we provide an overview of Fourier analysis of boolean functions of the form $f : \{\pm 1\}^n \to \{\pm 1\}$. The space of all of these functions is called the boolean hypercube or the Hamming cube. Before restricting the range of the functions to $\{\pm 1\}$, consider the vector space formed by the functions $f : \{\pm 1\}^n \to \mathbb{R}$.

Fact 1. Every function $f : \{\pm 1\}^n \to \mathbb{R}$ can be expressed as a linear sum of **monomials** χ_S:

$$\forall S \subseteq [n], \quad \chi_S(x) = \prod_{i \in S} x_i$$

We use this fact to define the Fourier spectrum of the f. By convention, $\chi_\varnothing(x) = 1$. We will also refer to $\chi_S(x)$ as the **character** for set S.

Definition 1. (Fourier expansion) The **Fourier expansion** of $f : \{\pm 1\}^n \to \mathbb{R}$ is:

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \prod_{i \in S} x_i = \sum_{S \subseteq [n]} \widehat{f}(S) \chi_S(x)$$

We call $\widehat{f}(S)$ the **Fourier coefficient** of f on set S.

We define an inner product on the vector space of f.

Definition 2 (Inner product)**.** $\forall f, g : \{\pm 1\}^n \to \mathbb{R}$, define $\langle f, g \rangle = \mathbb{E}_{x \sim \mathcal{U}}[f(x) \cdot g(x)]$. Loosely speaking, the inner product can be thought of as the *correlation* between f, g in expectation over uniformly random bit-strings.

Fact 2. The set of characters forms an **orthonormal basis** for the space of $f : \{\pm 1\}^n \to \mathbb{R}$.

As a corollary, we can obtain the Fourier coefficient: $\widehat{f}(S) = \langle f, \chi_S \rangle$. The idea behind this is that the character for S is orthogonal to all other characters for sets $T \neq S$.

Theorem (Plancherel)**.** $\forall f, g : \{\pm 1\}^n \to \mathbb{R}$ we have:

$$\langle f, g \rangle = \sum_{S \subseteq [n]} \widehat{f}(S) \cdot \widehat{g}(S)$$

Theorem (Parseval)**.** When $f = g$, we have a special case of Plancherel:

$$\langle f, f \rangle = ||f||_2^2 = \sum_{S \subseteq [n]} \widehat{f}(S)^2$$

$||f||_2^2$ is also known as $L_2(f)$ or the L-2 norm of f. We may also define the L-1 norm $L_1(f) = \sum_{S \subseteq [n]} |\widehat{f}(S)|$. We will also refer to $\widehat{f}(S)^2$ as the **Fourier weight** of f on S.
Now, let's fix the range of the functions by focusing on $f : \{\pm 1\}^n \to \{\pm 1\}$. An interesting fact emerges.

Fact 3. $\forall f : \{\pm 1\}^n \to \{\pm 1\}$, $L_2(f) = 1$.

This means that we can think of the **Fourier weights** of f as a probability distribution over sets S. High Fourier weight means that the character χ_S is highly *correlated* to $f(x)$.

We will focus on k-junta type problems, of which sparse parity is an instance.

Definition 3 (k-junta)**.** A k-junta problem is defined on n bits, but the result, $f(x)$ depends only on a subset $S \subseteq [n]$, where $|S| = k$. By definition, (n, k)-sparse parity is a k-junta problem.

For (n, k)-sparse parity problems, it is easy to see that $f(x) = \chi_S$, where $S, |S| = k$ is the sparse set of indices over which parity is taken. Thus, all of the Fourier weight of the function is fully concentrated at χ_S. For other functions, non-zero Fourier weight is associated with degree-1 monomials x_i for $i \in S$ (the k-junta set). We call these functions **degree-1 leaky** or simply, leaky. It is easy to see that a sparse majority function is leaky. With high probability, random k-juntas are also degree-1 leaky (but we leave this as a conjecture). Note that parities are not degree-1 leaky.

Definition 4 (Degree-1 leak)**.** $\text{Leak}_i(f) = |\mathbb{E}_{x \sim \mathcal{U}}[f(x)x_i]| = |\widehat{f}(\{i\})|$

Definition 5. $\gamma = \min_i \text{Leak}_i(f) = \min_i |\widehat{f}(\{i\})|$. This is our parameter of interest for the degree-1 leaky functions. We define **degree-1 leaky** functions as those functions f which have $\gamma > 0$.

Fact 4. For sparse majority, $\gamma = \Theta(1/\sqrt{k})$.

Panigrahi et al. showed that progressive distillation is helpful for sparse-parity problems. We were wondering what other kinds of k-junta problems progressive distillation is (provably) helpful for (with regard to sample complexity). Our key result is that progressive distillation is **more sample-efficient** that one-shot distillation for the class of **degree-1 leaky** functions.

2.2 Model specification

We will consider ReLU networks with 1 hidden-layer of size (width) m. Our models are defined identically to Panigrahi et al. We also use the hinge-loss function and the two-stage training. The only difference is the initialization scheme; we initialize parameters as random Gaussians with normalized (w.r.t m, n) variance and no bias. Both teacher ($\mathcal{T}$) and student ($\mathcal{S}$) models look like (step t, σ is ReLU):

$$f_{\mathcal{M}}^{(t)}(x) = \sum_{i=1}^{m} a_i^{(t)} \sigma(\langle w_i^{(t)}, x \rangle + b_i^{(t)})$$

3 Key result

We show that progressive distillation leads to a sample complexity improvement when learning **degree-1 leaky** k-juntas.

Theorem. Consider $f : \{\pm 1\}^n \to \{\pm 1\}$, where f is a degree-1 leaky k-junta. With probability $1-\delta$, the sample complexity for attaining ε loss is $\tilde{O}(c^{-2}\gamma^{-3}\log(mn/\delta) + 2^k n^2 \gamma^{-2}\varepsilon^{-2})$, where c is a small constant. For one-shot distillation, Panigrahi et al.'s lower bound of $\Omega(\varepsilon^{-2})$ carries over.

Proof sketch. The key idea is that we are able to identify the set S from degree-1 leakage through the first checkpoint (Phase 1). Then, it is possible to learn the function from the second checkpoint (Phase 2). We provide an overview of the technical details.

Lemma 1. Define $G_{ij} = \mathbb{E}_{x\sim\mathcal{U}}[\nabla_{w_{ij}} \ell(x, f(x); f_{\mathcal{T}}^{(0)}(x))]$ as the population gradient of the hinge-loss at the first iteration. We show that, for degree-1 leaky f, the gradients reveal important information about S. Formally, we are able to show that:

$$\begin{aligned} (j \in S) \quad & G_{ij} \geq \Theta(\widehat{f}(\{i\})) = \Theta(\gamma) \\ (j \notin S) \quad & G_{ij} = 0 \end{aligned}$$

Lemma 2. Next, we show that the batch gradients $\tilde{G}_{ij}$ are close to population gradients for large enough batch size $B \sim O(\gamma^{-2} \log(mn/\delta))$. The proof involves applying a Hoeffding inequality to bound the absolute difference between batch and population gradients. For batch size mentioned above, we obtain (for tiny constant c):

$$\begin{aligned} (j \in S) \quad & \tilde{G}_{ij} \geq \Theta((1-c)\gamma) \\ (j \notin S) \quad & \tilde{G}_{ij} \leq c\gamma \end{aligned}$$

Through the lemmas above, we are able to show that the batch gradient at initialization reveals information about S.

Lemma 3. We show that the first step of SGD significantly increments weights w_{ij} for $j \in S$, but not when $j \in S$. Specifically, with $\eta \sim \Theta(\gamma)$:

$$\begin{aligned} (j \in S) \quad & |\eta_1 \tilde{G}_{ij}| \geq \Theta((1-c)\gamma^2) \\ (j \notin S) \quad & |\eta_1 \tilde{G}_{ij}| \leq \Theta(c\gamma^2) \end{aligned}$$

Continuing the analysis for $T \sim \Theta(1/\gamma)$ steps and performing a low-order approximation for ReLU output, we are able to obtain an important sub-result ($m \sim O(2^k)$ needed):

$$f_{\mathcal{T}}^{(T)}(x) \approx \sum_{i \in S} c_i x_i + \sum_{U \subseteq S, |U| \geq 2} c_U \chi_U(x)$$

We show that it is possible to upper bound the second, higher order sum term by $\Omega(\gamma^2)$ and $\forall i \in S, c_i = \Theta((1-c)\gamma)$.

Lemma 4. By definition of degree-1 leaky, the optimal Bayes classifier for the population has an ℓ_2-margin $\geq \gamma$

Using the four Lemmas above as pieces, we can retain the remaining proof structure for Theorem B.1. from Panigrahi et al. to obtain the upper-bound as stated. This completes the proof sketch for our key result. Essentially, we show that Phase 1 (i.e. a checkpoint learning S through degree-1 correlations) is possible. Phase 2 remains largely the same for the analysis. □