ϵ_θ resulting from the distribution mismatch between the prompt and the pertaining distributions for each example. Letting $p^i_\theta(o)$ and p^i_{prompt} correspond to the concept θ and $\theta^\star$.

Condition 1 (distinguishability (Fang & Xie, 2022))**.** *The $\theta^\star$ is distinguishable if for all $\theta \in \Omega$, $\theta \neq \theta^\star$,*

$$\sum_{i=1}^{k} KL_i(\theta^\star || \theta) > \epsilon_\theta, \tag{3}$$

where the $KL_i(\theta^\star||\theta) := \mathbb{E}_{O[1:i-1]\sim p_{prompt}}[KL(p^i_{prompt}||p^i_\theta)]$.

Noises from KV Cache Compression. Naturally, because of the sparsified KV cache, some history tokens in $o_{1:t-1}$ at different layers lost its attention score calculation with respect to the next word prediction o_t. We can regard this as the noise added onto the $o_{1:t-1}$. Thus, distincting $\theta^\star$ from θ requires larger KL divergence. Following (Zhou et al., 2024), we provide the following second condition about the distinguishability with the KV cache sparsity.

Condition 2 (distinguishability under sparsified KV cache)**.** *With the noise introduced by the sparsified KV cache of the sparse ratio r, the distribution mismatch between the prompt and the pretraining distribution that is approximated by LLM is enlarged, resulting in a varied requirement with error term $\xi_\theta(r)$ for θ^* being distinguishable if for all $\theta \in \Theta$, $\theta \neq \theta^*$,*

$$\sum_{i=1}^{k} KL_i(\theta^* || \theta) > \epsilon_\theta + \xi_\theta(r), \quad \text{where} \quad \xi_\theta(r) \propto r. \tag{4}$$

Lemma 1 (noisy-relaxed bound in (Fang & Xie, 2022; Zhou et al., 2024))**.** *let $\mathcal{B}$ denotes the set of θ which does not satisfy Condition 1. We assume that $KL(p_{prompt}(y_{test}|x_{test}))||p(y_{test}|x_{test},\theta)$ is bounded for all θ and that $\theta^\star$ minimizes the multi-class logistic risk as,*

$$L_{CE}(\theta) = -\mathbb{E}_{x_{test}\sim p_{prompt}}[p_{prompt}(y_{test}|x_{test}) \cdot \log p(y_{test}|x_{test},\theta)]. \tag{5}$$

If

$$\mathbb{E}_{x_{test}\sim p_{prompt}}[KL(p_{prompt}(y_{test}|x_{test}))||p(y_{test}|x_{test},\theta))] \leq (\epsilon_\theta + \xi_\theta(r)), \quad \forall \quad \theta \in \mathcal{B}, \tag{6}$$

then

$$\lim_{n\to\infty} L_{0-1}(f_n) \leq \inf_f L_{0-1}(f) + g^{-1}\left(\sup_{\theta\in\mathcal{B}}(\epsilon_\theta)\right), \tag{7}$$

where $g(\nu) = \frac{1}{2}\big((1-\nu)\log(1-\nu) + (1+\nu)\log(1+\nu)\big)$ is the calibration function (Steinwart, 2007; Pires & Szepesvári, 2016) for the multiclass logistic loss for $\nu \in [0, 1]$.

Following (Kleijn & der Vaart, 2012; Fang & Xie, 2022), KL divergence is assumed to haver the 2nd-order Taylor expansion with the concept θ. Then, we have the following theorem and proof.

Theorem 1. *(Fang & Xie, 2022; Zhou et al., 2024) Let the set of θ which does not satisfy Equation 3 in Condition 1 to be $\mathcal{B}$. Assume that KL divergences have a 2nd-order Taylor expansion around $\theta^\star$:*

$$\forall j > 1, \;\; KL_i(\theta^\star||\theta) = \frac{1}{2}(\theta - \theta^\star)^\top I_{j,\theta^*}(\theta - \theta^\star) + O(\|\theta - \theta^\star\|^3) \tag{8}$$

where I_{j,θ^*} is the Fisher information matrix of the j-th token distribution with respect to $\theta^\star$. Let $\gamma_{\theta^*} = \frac{\max_j \lambda_{max}(I_{j,\theta^*})}{\min j \lambda_{min}(I_{j,\theta^*})}$ where $\lambda_{max}, \lambda_{min}$ return the largest and smallest eigenvalues. Then for $k \geq 2$ and as $n \to \infty$, the 0-1 risk of the in-context learning predictor f_n is bounded as

$$\lim_{n\to\infty} L_{0\text{-}1}(f_n) \leq \inf_f L_{0\text{-}1}(f) + g^{-1}\left(O\left(\frac{\gamma_{\theta^*}\sup_{\theta\in\mathbb{B}}(\epsilon_\theta + \xi_\theta(r))}{k-1}\right)\right) \tag{9}$$