Proof. (Fang & Xie, 2022) By the Taylor expansion on θ, we have for any θ in $\mathcal{B}$ that

$$\sum_{j=2}^{k} \mathrm{KL}_i(\theta^\star||\theta) \geq \frac{1}{2}\sum_{j=2}^{k}(\theta-\theta^\star)^\top I_{j,\theta^*}(\theta-\theta^\star) + (k-1)O(\|\theta-\theta^\star\|^3) \tag{10}$$

$$\geq \frac{1}{2}(k-1)\lambda_{\min}(I_{j,\theta^*})\|\theta-\theta^\star\|^2 \tag{11}$$

$$\implies \|\theta-\theta^\star\|^2 \leq \frac{(\epsilon_\theta+\xi_\theta(r))}{\frac{1}{2}(k-1)(\min_j\ \lambda_{\min}(I_{j,\theta^*}))}. \tag{12}$$

We can bound the last KL term (k-th token) with the above term:

$$\mathrm{KL}_k(\theta^\star||\theta) = \frac{1}{2}(\theta-\theta^\star)^\top I_{k,\theta^*}(\theta-\theta^\star) + O(\|\theta-\theta^\star\|^3) \tag{13}$$

$$\leq \frac{1}{2}(\max_j\ \lambda_{\max}(I_{j,\theta^*}))\|\theta-\theta^\star\|^2 + O(\|\theta-\theta^\star\|^2) \tag{14}$$

$$\leq \frac{(\epsilon_\theta+\xi_\theta(r))(\max_j\ \lambda_{\max}(I_{j,\theta^*})+O(1))}{(k-1)\min_j\ \lambda_{\min}(I_{j,\theta^*})}. \tag{15}$$

Rearranging above equation, and with $\mathrm{KL}_k(\theta^\star||\theta) = \mathbb{E}_{x_{\text{test}}\sim p_{\text{prompt}}}[KL(p_{\text{prompt}}(y_{\text{test}}|x_{\text{test}})\|p(y_{\text{test}}|x_{\text{test}},\theta))]$, there is

$$\mathbb{E}_{x_{\text{test}}\sim p_{\text{prompt}}}[KL(p_{\text{prompt}}(y_{\text{test}}|x_{\text{test}})\|p(y_{\text{test}}|x_{\text{test}},\theta))] \leq \frac{(\epsilon_\theta+\xi_\theta(r))(\max_j\ \lambda_{\max}(I_{j,\theta^*})+O(1))}{(k-1)\min_j\ \lambda_{\min}(I_{j,\theta^*})} \tag{16}$$

Combining Equation 16 with Equation 6 into Lemma 1 completes the proof. □

KV Cache Sparsification. Revisiting the Equation 4 in Condition 2, the $\xi_\theta(r)$ is enlarged with the sparsity ratio r. The higher compression ratio r (means that more KV cache are discarded), the more noise $\xi_\theta(r)$. Then it leads to the higher bound of the $\lim_{n\to\infty} L_{0-1}(f_n)$ in Equation 5 in Lemma 1. Next, we discuss how KV cache compression influences the Equation 4.

Token-level Importance Measure. The token-level KV cache methods usually calculate the importance of different tokens. Then, the KV cache with indexes that have higher importance will be preserved. Such indexes are normaly choosed as the attention score. Considering that the case in Figure 1, where each token in the i-th training [2] example sequence ($O_i = [x_i, y_i]$) might be compressed, and tokens are sparsified concretely without out dependency to other tokens. However, in the generation process of the i-th training example, $O_i = [x_i, y_i]$ is sampled from $p(O_i|h_i^{\text{start}}, \theta^\star)$ and $p_\theta^j(o) := p(O[j] = o|O[1:j-1], \theta)$ of the j-th token with previous tokens and the analogous distribution $p_{prompt}^j := p_{prompt}(O[j] = o|O[1:j-1])$. And the KL divergence is defined as $\mathrm{KL}_j(\theta^\star||\theta) := \mathbb{E}_{O[1:j-1]\sim p_{prompt}}[\mathrm{KL}(p_{prompt}^j||p_\theta^j)]$, which means that in a training example $O_i = [x_i, y_i] = O_i[1:k]$, each token $O_i[j]$ has strong dependency with $O_i[1:j-1]$, noises on previous any j-th token will influence the distinguishability on the following tokens, i.e. requiring larger $\{\mathrm{KL}_u(\theta^\star||\theta)\}_{u>j}$.

On the other hand, the token-level sparsity enlarges the requirement on the distinguishability uniformly for each example O_i (the case in Figure 1), which uniformly loses the bound of $L_{0\text{-}1}(f_n)$ as in Equation 9.

Chunk-level Importance Measure. Different from token-level importance measure, ChunkKV regards tokens in a continuous window as a basic unit that should be left or discarded as a whole. The preserved window can be regarded as saving the complete $O_i = [x_i, y_i]$ without noise. Thus, ChunkKV reduces the noise $\xi_\theta(r)$ for the preserved O_i, which lowers the bound of $L_{0\text{-}1}(f_n)$.

More intuitively, ChunkKV focus on reducing the noise on some complete training examples, but some other examples overall with low importance will be discarded. Then, the model identifies the x_{test} from those clean and more related training examples O_i and neglect those O_i with less importance.

Note that here, we do not provide the rigorous proof on how KV cache sparsity enhances the requirement of the distinguishability and how different $\mathrm{KL}_j(\theta^\star||\theta)$ on $O_i = [x_i, y_i]$ influences the bound $L_{0\text{-}1}(f_n)$. We left this as the future work to analyze how KV cache sparsity influences the in-context learning.

[2] Here, training means prompt learning (Fang & Xie, 2022).