Relaxion of Assumption C.3 (Perfectly Pretrained LLMs)

As we discuss in Lines 395-405 of Section 5, the assumption of perfectly pertrained LLMs can be relaxed to accommodate a generalization error. Now we provide the proof as follows. We start with a concentration inequality for the maximum-likelihood estimate (MLE). Let $\mathcal{F}$ be a finite function class used to model a conditional distribution $p_{Y|X}(y|x)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Assume there is $f^\star \in \mathcal{F}$ such that $p(y|x) = f^\star(y|x)$ (realizable), and $f(\cdot|x) \in \Delta(\mathcal{Y})$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$ (proper). Let $D = \{x_i, y_i\}_{i\in[N]}$ denote a dataset of i.i.d samples where $x_i \sim p_X$ and $y_i \sim p_{Y|X}(\cdot|x_i)$. Let $\hat{f}$ be the MLE, which satisfies

$$\hat{f} = \operatorname*{argmax}_{f\in\mathcal{F}} \sum_{i\in[N]} \log f(y_i|x_i).$$

From the MLE analysis in [4], we know it holds that

$$\mathbb{E}_{x\sim p_X} d_{\mathrm{TV}}\left(\hat{f}(\cdot|x), p_{Y|X}(\cdot|x)\right) \le \frac{8\log(|\mathcal{F}|/\delta)}{N},$$

with probability at least $1-\delta$. Now, we use this result to analyze the generalization error in the pretraining phase of LLMs. In the following, we consider that the LLMs used in RAFA are not perfectly pretrained but the MLE in the pretraining dataset. Define $\tilde{P}_{\mathrm{LLM}(\mathcal{D})}$ as the *perfectly pretrained* LLM transition kernel estimator with the memory buffer $\mathcal{D}$ prompted as contexts. Define $\tilde{r}_{\mathrm{LLM}(\mathcal{D})}$ as the *perfectly pretrained* LLM reward estimator with the memory buffer $\mathcal{D}$ prompted as contexts. For the simplicity of later discussion, we denote by $P^{\mathrm{LLM}}((s',r)|s,a,\mathcal{D}))$ the state-reward (s',r) prediction probability of LLMs used in RAFA and the $\tilde{P}^{\mathrm{LLM}}((s',r)|s,a,\mathcal{D}))$ the state-reward (s',r) prediction probability of *perfectly pretrained* LLMs. Let $|\mathcal{F}_{\mathrm{LLM}}|$ be the cardinality of the function class of LLMs, N be the size of the pretraining dataset, ρ_{pre} be the prompt distribution in the pretraining dataset. Now we are ready to use the standard MLE analysis to show the generalization error bound of LLMs as follows. For any fixed distribution μ of $(s,a,\mathcal{D})$, it holds with at probability $1-\delta$ that

$$\begin{aligned}
\mathbb{E}_{(\mathcal{D},s,a)\sim\mu} d_{\mathrm{TV}}(\tilde{P}^{\mathrm{LLM}}(\cdot|s,a,\mathcal{D})\|P^{\mathrm{LLM}}(\cdot|s,a,\mathcal{D})) &\le \left\|\frac{\mathrm{d}\mu}{\mathrm{d}\rho_{pre}}\right\|_\infty \cdot \mathbb{E}_{(\mathcal{D},s,a)\sim\rho_{pre}} d_{\mathrm{TV}}(\tilde{P}^{\mathrm{LLM}}(\cdot|s,a,\mathcal{D})\|P^{\mathrm{LLM}}(\cdot|s,a,\mathcal{D})) \\
&\le \left\|\frac{\mathrm{d}\mu}{\mathrm{d}\rho_{pre}}\right\|_\infty \cdot \sqrt{\frac{8\log(|\mathcal{F}_{\mathrm{LLM}}|/\delta)}{N}},
\end{aligned}$$

where $\|\cdot\|_\infty$ denotes the infinity norm. Here, we assume that the pretraining dataset is large enough such that $\left\|\frac{\mathrm{d}\mu}{\mathrm{d}\rho_{pre}}\right\|_\infty$ is properly defined for any μ. We denote the Bellman operator induced by the perfectly pretrained LLM and $\mathcal{D}_{t_k}$ as $\tilde{B}_k$, which is defined as $(\tilde{B}_k V)(s,a) = \tilde{r}_{\mathrm{LLM}(\mathcal{D}_{t_k})}(s,a) + (\tilde{P}_{\mathrm{LLM}(\mathcal{D}_{t_k})}V)(s,a)$ for any s,a, and value function V. Then, by the definition of B_k, we have

$$\begin{aligned}
\left|\big((\tilde{B}_k - B_k)V_t\big)(s,a)\right| &= \left|\mathbb{E}_{(s',r)\sim\tilde{P}_{\mathrm{LLM}}(\cdot|s,a,\mathcal{D}_{t_k})}[r+\gamma\cdot V(s')] - \mathbb{E}_{(s',r)\sim P_{\mathrm{LLM}}(\cdot|s,a,\mathcal{D}_{t_k})}[r+\gamma\cdot V(s')]\right| \\
&\le 2L\cdot d_{\mathrm{TV}}(\tilde{P}^{\mathrm{LLM}}(\cdot|s,a,\mathcal{D}_{t_k})\|P^{\mathrm{LLM}}(\cdot|s,a,\mathcal{D}_{t_k})),
\end{aligned}$$

where the last inequality uses the definition of L (recall that L is the bound of $|r+V(s)|$ for any reward r, state s, and value V) and Hölder's inequality. In the proof of Theorem C.7 (the analysis of the regret of RAFA), we need modify (D.12) and (D.20) with the following equality:

$$\begin{aligned}
\mathbb{E}_{\pi^k}\left[\big((B_k - B_{\theta^\star})V_t\big)(s_t,\pi^k(s_t))\right] &= \mathbb{E}_{\pi^k}[(\tilde{B}_k - B_{\theta^\star})V_t)(s_t,\pi^k(s_t))] \\
&\quad + \mathbb{E}_{\pi^k}\left[\big((\tilde{B}_k - B_k)V_t\big)(s_t,\pi^k(s_t))\right] \\
&\le \mathbb{E}_{\pi^k}[(\tilde{B}_k - B_{\theta^\star})V_t)(s_t,\pi^k(s_t))] \\
&\quad + 2L\cdot\mathbb{E}_{\pi^k}[d_{\mathrm{TV}}(\tilde{P}^{\mathrm{LLM}}(\cdot|s,a,\mathcal{D}_{t_k})\|P^{\mathrm{LLM}}(\cdot|s,a,\mathcal{D}_{t_k}))].
\end{aligned}$$

By Proposition 5.1 (perfectly pretrained LLM perform BMA) and the fact that $\tilde{B}_k$ is the Bellman operator induced by the perfectly pretrained LLM and $\mathcal{D}_{t_k}$, we can analyze $\mathbb{E}_{\pi^k}[(\tilde{B}_k - B_{\theta^\star})V_t)(s_t,\pi^k(s_t))]$ in the

same way as in the previous proof of Theorem C.7. Hence, the additional regret without Assumption C.3 (perfectly pretrained LLMs) is less than

$$(1-\gamma)^{-1} \cdot \mathbb{E}\Big[\sum_{k=0}^{K-1} \mathbb{E}_{\pi^k}\Big[\sum_{t=t_k}^{t_{k+1}-1} 4L \cdot d_{\mathrm{TV}}(\tilde{P}^{\mathrm{LLM}}(\cdot \mid s,a,\mathcal{D}_{t_k}) \| P^{\mathrm{LLM}}(\cdot \mid s,a,\mathcal{D}_{t_k}))\Big]\Big]$$
$$\leq 2 \cdot \sqrt{\frac{4 \cdot \log(\mathcal{F}_{\mathrm{LLM}}/\delta)}{N}} \cdot \sup_{t<T} \left\| \frac{\mathrm{d}\mu_t}{\mathrm{d}\rho_{pre}} \right\|_{\infty} \cdot T,$$

with probability at least $1-\delta$. Here, $|\mathcal{F}_{\mathrm{LLM}}|$ is the cardinality of the function class of LLMs, N is the size of the pretraining dataset, ρ_{pre} is the prompt distribution in the pretraining dataset, and μ_t is the marginal distribution of $(s_t, \pi_t(s_t), \mathcal{D}_t)$ for RAFA. When the size N of pretraining dataset tends to infinity, the additional regret decays to zero. Hence, we justify that Assumption C.3 (Perfectly Pretrained LLMs) holds approximately if the pretraining dataset is very large.