Appendix

VIII. Proof of Corollary 1

Proof: We initiate our analysis by considering Assumption 1, which concerns Lipschitz smoothness. For the risk-sensitive objective J_β with a Lipschitz smoothness constant L_β, we have that:

$$\begin{aligned} J_\beta(\theta_{t+1}) &\geq J_\beta(\theta_t) + \langle \nabla J_\beta(\theta_t), \theta_{t+1} - \theta_t \rangle - \frac{L_\beta}{2} \|\theta_{t+1} - \theta_t\|^2 \\ &= J_\beta(\theta_t) + \eta \left\langle \nabla J_\beta(\theta_t), \widehat{\nabla} J_\beta(\theta_t) \right\rangle - \frac{L_\beta \eta^2}{2} \left\| \widehat{\nabla} J_\beta(\theta_t) \right\|^2. \end{aligned} \tag{18}$$

Take expectations on both sides conditioned on θ_t and use Assumption 2, we get:

$$\begin{aligned} \mathbb{E}_t \left[J_\beta(\theta_{t+1}) \right] &\geq J_\beta(\theta_t) + \eta \langle \nabla J_\beta(\theta_t), \nabla J_\beta(\theta_t) \rangle - \frac{L_\beta \eta^2}{2} \mathbb{E}_t \left[\left\| \widehat{\nabla} J_\beta(\theta_t) \right\|^2 \right] \\ &\geq J_\beta(\theta_t) + \eta \|\nabla J_\beta(\theta_t)\|^2 - \frac{L_\beta \eta^2}{2} \left(2A(J_\beta^* - J_\beta(\theta_t)) + B \|\nabla J_\beta(\theta_t)\|^2 + C \right) \\ &= J_\beta(\theta_t) + \eta \left(1 - \frac{L_\beta B \eta}{2} \right) \|\nabla J_\beta(\theta_t)\|^2 - L_\beta \eta^2 A (J_\beta^* - J_\beta(\theta_t)) - \frac{L_\beta C \eta^2}{2}. \end{aligned} \tag{19}$$

Then we subtract J_β^* from both sides,

$$\mathbb{E}_t \left[J_\beta(\theta_{t+1}) \right] - J_\beta^* \geq -(1 + L_\beta \eta^2 A)(J_\beta^* - J_\beta(\theta_t)) + \eta \left(1 - \frac{L_\beta B \eta}{2} \right) \|\nabla J_\beta(\theta_t)\|^2 - \frac{L_\beta C \eta^2}{2}. \tag{20}$$

Take the expectation on both sides and rearrange the equation, we obtain:

$$\mathbb{E} \left[J_\beta^* - J_\beta(\theta_{t+1}) \right] + \eta \left(1 - \frac{L_\beta B \eta}{2} \right) \mathbb{E} \left[\|\nabla J_\beta(\theta_t)\|^2 \right] \leq (1 + L_\beta \eta^2 A) \mathbb{E} \left[J_\beta^* - J_\beta(\theta_t) \right] + \frac{L_\beta C \eta^2}{2}. \tag{21}$$

Define $\delta_t \stackrel{\text{def}}{=} \mathbb{E} \left[J_\beta^* - J_\beta(\theta_t) \right]$ and $r_t \stackrel{\text{def}}{=} \mathbb{E} \left[\|\nabla J_\beta(\theta_t)\|^2 \right]$, we can rewrite the above inequality as

$$\eta \left(1 - \frac{L_\beta B \eta}{2} \right) r_t \leq (1 + L_\beta \eta^2 A) \delta_t - \delta_{t+1} + \frac{L_\beta C \eta^2}{2}. \tag{22}$$

Now, we introduce a sequence of weights, denoted as $w_{-1}, w_0, w_1, \cdots, w_{T-1}$, based on a method used by [34, 38, 40]. We initialize w_{-1} with a positive value. We define w_t as $w_t =: \frac{w_{t-1}}{1 + L_\beta \eta^2 A}$ for all $t \geq 0$. It's important to note that when $A = 0$, all w_t are equal, i.e., $w_t = w_{t-1} = \cdots = w_{-1}$. By multiplying (22) by w_t / η, we can derive:

$$\begin{aligned} \left(1 - \frac{L_\beta B \eta}{2} \right) w_t r_t &\leq \frac{w_t (1 + L_\beta \eta^2 A)}{\eta} \delta_t - \frac{w_t}{\eta} \delta_{t+1} + \frac{L_\beta C \eta}{2} w_t \\ &= \frac{w_{t-1}}{\eta} \delta_t - \frac{w_t}{\eta} \delta_{t+1} + \frac{L_\beta C \eta}{2} w_t. \end{aligned} \tag{23}$$

When we sum up both sides for $t = 0, 1, \cdots, T-1$, we get:

$$\begin{aligned} \left(1 - \frac{L_\beta B \eta}{2} \right) \sum_{t=0}^{T-1} w_t r_t &\leq \frac{w_{-1}}{\eta} \delta_0 - \frac{w_{T-1}}{\eta} \delta_T + \frac{L_\beta C \eta}{2} \sum_{t=0}^{T-1} w_t \\ &\leq \frac{w_{-1}}{\eta} \delta_0 + \frac{L_\beta C \eta}{2} \sum_{t=0}^{T-1} w_t. \end{aligned} \tag{24}$$

We can define W_T as $W_T =: \sum_{t=0}^{T-1} w_t$. By dividing both sides of the equation by W_T, we obtain:

$$\left(1 - \frac{L_\beta B \eta}{2} \right) \min_{0 \leq t \leq T-1} r_t \leq \frac{1}{W_T} \cdot \left(1 - \frac{L_\beta B \eta}{2} \right) \sum_{t=0}^{T-1} w_t r_t \leq \frac{w_{-1}}{W_T} \frac{\delta_0}{\eta} + \frac{L_\beta C \eta}{2}. \tag{25}$$

Note that,

$$W_T = \sum_{t=0}^{T-1} w_t \geq \sum_{t=0}^{T-1} \min_{0 \leq i \leq T-1} w_i = T w_{T-1} = \frac{T w_{-1}}{(1 + L_\beta \eta^2 A)^T}. \tag{26}$$

Use this in (25),

$$\left(1 - \frac{L_\beta B\eta}{2}\right) \min_{0\le t\le T-1} r_t \quad \le \quad \frac{(1+L\eta^2 A)^T}{\eta T}\delta_0 + \frac{LC\eta}{2}. \tag{27}$$

By substituting r_t in (27) with $\mathbb{E}\left[\|\nabla J_\beta(\theta_t)\|^2\right]$, we obtain:

$$\begin{aligned}\left(1 - \frac{L_\beta B\eta}{2}\right) \min_{0\le t\le T-1} \mathbb{E}\left[\|\nabla J_\beta(\theta_t)\|^2\right] \quad &\le \quad \frac{(1+L_\beta\eta^2 A)^T}{\eta T}\delta_0 + \frac{L_\beta C\eta}{2}, \\ \min_{0\le t\le T-1} \mathbb{E}\left[\|\nabla J_\beta(\theta_t)\|^2\right] \quad &\le \quad \frac{2\delta_0(1+L_\beta\eta^2 A)^T}{\eta T(2-L_\beta B\eta)} + \frac{L_\beta C\eta}{2-L_\beta B\eta}. \end{aligned} \tag{28}$$

The choice of our step size ensures that for both cases, whether $B > 0$ or $B = 0$, we have $1 - \frac{L_\beta B\eta}{2} > 0$.

■

IX. Proof of Theorem 1

Lemma 1: Subject to Assumption 4, for any non-negative integer t, and for any state-action pair $(s_t, a_t) \in \mathcal{S}\times\mathcal{A}$ at time t within a trajectory τ sampled under the parameterized policy π_θ, we have the following:

$$\mathbb{E}_{\tau\sim p(\cdot|\theta)}\left[\|\nabla_\theta \log\pi_\theta(a_t \mid s_t)\|^2\right] \quad \le \quad F_1^2, \tag{29}$$

$$\mathbb{E}_{\tau\sim p(\cdot|\theta)}\left[\left\|\nabla_\theta^2 \log\pi_\theta(a_t \mid s_t)\right\|\right] \quad \le \quad F_2. \tag{30}$$

Proof: For $t > 0$ and $(s_t, a_t) \in \mathcal{S}\times\mathcal{A}$, we have

$$\mathbb{E}_\tau\left[\|\nabla_\theta \log\pi_\theta(a_t \mid s_t)\|^2\right] = \mathbb{E}_{s_t}\left[\mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\left[\|\nabla_\theta \log\pi_\theta(a_t \mid s_t)\|^2 |s_t\right]\right] \overset{14}{\le} F_1^2, \tag{31}$$

where the first equality is obtained by the Markov property. Similarly, we have

$$\mathbb{E}_\tau\left[\left\|\nabla_\theta^2 \log\pi_\theta(a_t \mid s_t)\right\|\right] = \mathbb{E}_{s_t}\left[\mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\left[\left\|\nabla_\theta^2 \log\pi_\theta(a_t \mid s_t)\right\| |s_t\right]\right] \overset{15}{\le} F_2. \tag{32}$$

■

Then Lemma 1 is then used for the derivation of L and L_β.

Assumption 1 is equivalent to $\left\|\nabla^2 J(\theta)\right\| \le L$ for the risk-neutral REINFORCE and $\left\|\nabla^2 J_\beta(\theta)\right\| \le L_\beta$ for the risk-sensitive REINFORCE. We first take the second order derivative of the risk-neutral objective w.r.t. θ, in order to derive the L-Lipschitz smooth constant.

$$\begin{aligned}
\nabla^2 J(\theta) \quad &\overset{3}{=} \quad \nabla_\theta \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t \mid s_t)R(t)\right] \\
&= \quad \nabla_\theta\left[\int p(\tau\mid\theta)\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t \mid s_t)R(t)d\tau\right] \\
&= \quad \int \nabla_\theta p(\tau\mid\theta)\left(\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t \mid s_t)R(t)\right)^\top d\tau + \int p(\tau\mid\theta)\sum_{t=0}^{\infty}\nabla_\theta^2 \log\pi_\theta(a_t \mid s_t)R(t)d\tau \\
&= \quad \int p(\tau\mid\theta)\nabla_\theta\log p(\tau\mid\theta)\left(\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t \mid s_t)R(t)\right)^\top d\tau + \int p(\tau\mid\theta)\sum_{t=0}^{\infty}\nabla_\theta^2 \log\pi_\theta(a_t \mid s_t)R(t)d\tau \\
&= \quad \mathbb{E}_\tau\left[\nabla_\theta\log p(\tau\mid\theta)\left(\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t \mid s_t)R(t)\right)^\top\right] + \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta^2 \log\pi_\theta(a_t \mid s_t)R(t)\right] \\
&= \quad \underbrace{\mathbb{E}_\tau\left[\sum_{k=0}^{\infty}\nabla_\theta\log\pi_\theta(a_k\mid\theta_k)\left(\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t \mid s_t)R(t)\right)^\top\right]}_{①} + \underbrace{\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta^2 \log\pi_\theta(a_t \mid s_t)R(t)\right]}_{②}.
\end{aligned} \tag{33}$$

We individually bound the aforementioned two terms for the risk-neutral REINFORCE.

For the term ①,

$$
\begin{aligned}
\|①\| &= \left\|\mathbb{E}_\tau\left[\sum_{k=0}^{\infty}\nabla_\theta\log\pi_\theta(a_k\mid\theta_k)\left(\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t\mid s_t)R(t)\right)^\top\right]\right\| \\
&= \left\|\mathbb{E}_\tau\left[\sum_{k=0}^{\infty}\nabla_\theta\log\pi_\theta(a_k\mid\theta_k)\left(\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t\mid s_t)\sum_{t'=t}^{\infty}\gamma^{t'}r(s_{t'},a_{t'})\right)^\top\right]\right\| \\
&= \left\|\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\left(\sum_{t'=0}^{t}\nabla_\theta\log\pi_\theta(a_{t'}\mid\theta_{t'})\right)\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid\theta_k)\right)^\top\right]\right\| \\
&\leq \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t\left|r(s_t,a_t)\right|\left\|\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k\mid\theta_k)\right\|^2\right] \\
&\leq r_{max}\sum_{t=0}^{\infty}\gamma^t\sum_{k=0}^{t}\mathbb{E}_\tau\left[\|\nabla_\theta\log\pi_\theta(a_k\mid\theta_k)\|^2\right] \\
&\overset{29}{\leq} r_{max}F_1^2\sum_{t=0}^{\infty}\gamma^t(t+1) \\
&= \frac{r_{max}F_1^2}{(1-\gamma)^2},
\end{aligned}
\tag{34}
$$

where the third line is due to the fact that the future actions do not depend on the past rewards.

For the term ②,

$$
\begin{aligned}
\|②\| &= \left\|\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta^2\log\pi_\theta(a_t\mid s_t)R(t)\right]\right\| \\
&= \left\|\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta^2\log\pi_\theta(a_t\mid s_t)\sum_{t'=t}^{\infty}\gamma^{t'}r(s_{t'},a_{t'})\right]\right\| \\
&= \left\|\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\left(\sum_{k=0}^{t}\nabla_\theta^2\log\pi_\theta(a_k\mid s_k)\right)\right]\right\| \\
&\leq \mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\gamma^t\left|r(s_t,a_t)\right|\left(\sum_{k=0}^{t}\left\|\nabla_\theta^2\log\pi_\theta(a_k\mid s_k)\right\|\right)\right] \\
&\leq r_{\max}\sum_{t=0}^{\infty}\gamma^t\left(\sum_{k=0}^{t}\mathbb{E}_\tau\left[\left\|\nabla_\theta^2\log\pi_\theta(a_k\mid s_k)\right\|\right]\right) \\
&\overset{30}{\leq} r_{\max}F_2\sum_{t=0}^{\infty}\gamma^t(t+1) \\
&= \frac{r_{\max}F_2}{(1-\gamma)^2},
\end{aligned}
\tag{35}
$$

where the third line is also due to the fact that the future actions do not depend on the past rewards.

Finally,

$$
\left\|\nabla^2 J(\theta)\right\| \leq \frac{r_{\max}}{(1-\gamma)^2}(F_1^2+F_2), \tag{36}
$$

so $L = \frac{r_{\max}}{(1-\gamma)^2}(F_1^2+F_2)$.

Then we take the second order derivative of the risk-sensitive objective w.r.t. θ, in order to derive the L_β-Lipschitz smoothness constant.

$$
\begin{aligned}
\nabla^2 J_\beta(\theta) &\overset{8}{=} \nabla_\theta \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)}\right] \\
&= \nabla_\theta \left[\int p(\tau \mid \theta) \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)} d\tau\right] \\
&= \int \nabla_\theta p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)}\right)^\top d\tau + \int p(\tau \mid \theta) \sum_{t=0}^{\infty} \nabla_\theta^2 \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)} d\tau \\
&= \int p(\tau \mid \theta) \nabla_\theta \log p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)}\right)^\top d\tau \\
&\quad + \int p(\tau \mid \theta) \sum_{t=0}^{\infty} \nabla_\theta^2 \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)} d\tau \\
&= \mathbb{E}_\tau \left[\nabla_\theta \log p(\tau \mid \theta) \left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)}\right)^\top\right] + \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \nabla_\theta^2 \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)}\right] \\
&= \underbrace{\mathbb{E}_\tau \left[\sum_{k=0}^{\infty} \nabla_\theta \log \pi_\theta(a_k \mid \theta_k) \left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)}\right)^\top\right]}_{③} \\
&\quad \underbrace{+ \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \nabla_\theta^2 \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)}\right]}_{④}.
\end{aligned} \tag{37}
$$

We also bound the above two terms separately for the risk-sensitive REINFORCE algorithm. For the term ③,

$$
\begin{aligned}
\|③\| &= \left\|\mathbb{E}_\tau \left[\sum_{k=0}^{\infty} \nabla_\theta \log \pi_\theta(a_k \mid \theta_k) \left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta R(t)}\right)^\top\right]\right\| \\
&= \left\|\mathbb{E}_\tau \left[\sum_{k=0}^{\infty} \nabla_\theta \log \pi_\theta(a_k \mid \theta_k) \left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \frac{1}{\beta} e^{\beta \sum_{t'=t}^{\infty} \gamma^{t'} r(s_{t'}, a_{t'})}\right)^\top\right]\right\| \\
&= \left\|\mathbb{E}_\tau \left[\frac{1}{\beta} e^{\beta \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)} \left(\sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'} \mid \theta_{t'})\right) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid \theta_k)\right)^\top\right]\right\| \\
&\leq \mathbb{E}_\tau \left[\frac{1}{|\beta|} e^{|\beta| \sum_{t=0}^{\infty} \gamma^t |r(s_t, a_t)|} \left\|\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid \theta_k)\right\|^2\right] \\
&\overset{13}{=} \mathbb{E}_\tau \left[\alpha \sum_{t=0}^{\infty} \gamma^t |r(s_t, a_t)| \left\|\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid \theta_k)\right\|^2\right] \\
&\leq \alpha \cdot r_{max} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\tau \left[\left\|\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k \mid \theta_k)\right\|^2\right] \\
&= \alpha \cdot r_{max} \sum_{t=0}^{\infty} \gamma^t \sum_{k=0}^{t} \mathbb{E}_\tau \left[\|\nabla_\theta \log \pi_\theta(a_k \mid \theta_k)\|^2\right] \\
&\leq \alpha \cdot r_{max} F_1^2 \sum_{t=0}^{\infty} \gamma^t (t+1) \\
&= \alpha \cdot \frac{r_{max} F_1^2}{(1-\gamma)^2},
\end{aligned} \tag{38}
$$

where in the third line, we use the fact that the future actions do not depend on the past rewards. In the fifth line, we use Assumption 3.

For the term ④,

$$
\begin{aligned}
\|④\| &= \left\|\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta^2\log\pi_\theta(a_t \mid s_t)\cdot\beta e^{\beta R(t)}\right]\right\| \\
&= \left\|\mathbb{E}_\tau\left[\sum_{t=0}^{\infty}\nabla_\theta^2\log\pi_\theta(a_t \mid s_t)\cdot\frac{1}{\beta}e^{\beta\sum_{t'=t}^{\infty}\gamma^{t'}r(s_{t'},a_{t'})}\right]\right\| \\
&= \left\|\mathbb{E}_\tau\left[\frac{1}{\beta}e^{\beta\sum_{t=0}^{\infty}\gamma^{t}r(s_t,a_t)}\left(\sum_{k=0}^{t}\nabla_\theta^2\log\pi_\theta(a_k \mid s_k)\right)\right]\right\| \\
&\leq \mathbb{E}_\tau\left[\frac{1}{|\beta|}e^{|\beta|\sum_{t=0}^{\infty}\gamma^{t}|r(s_t,a_t)|}\left(\sum_{k=0}^{t}\left\|\nabla_\theta^2\log\pi_\theta(a_k \mid s_k)\right\|\right)\right] \\
&\overset{13}{=} \mathbb{E}_\tau\left[\alpha\cdot\sum_{t=0}^{\infty}\gamma^{t}|r(s_t,a_t)|\left(\sum_{k=0}^{t}\left\|\nabla_\theta^2\log\pi_\theta(a_k \mid s_k)\right\|\right)\right] \\
&\leq \alpha\cdot r_{\max}\sum_{t=0}^{\infty}\gamma^{t}\left(\sum_{k=0}^{t}\mathbb{E}_\tau\left[\left\|\nabla_\theta^2\log\pi_\theta(a_k \mid s_k)\right\|\right]\right) \\
&\overset{30}{\leq} \alpha\cdot r_{\max}F_2\sum_{t=0}^{\infty}\gamma^{t}(t+1) \\
&= \alpha\cdot\frac{r_{\max}F_2}{(1-\gamma)^2}
\end{aligned}
\tag{39}
$$

Finally,

$$
\left\|\nabla^2 J_\beta(\theta)\right\| \leq \alpha\cdot\frac{r_{\max}}{(1-\gamma)^2}(F_1^2+F_2), \tag{40}
$$

so $L_\beta = \alpha\cdot\frac{r_{\max}}{(1-\gamma)^2}(F_1^2+F_2)$, where $0<\alpha<1$.