Supplement to the paper: Distributed Reinforcement Learning via Aggregative Actor-Critic

Andrea Camisa, Lorenzo Sforni, Giuseppe Notarstefano

APPENDIX

A. Proof of Lemma 3.1

We first reformulate problem (11) as a standard linear quadratic optimal control problem. By denoting as $A = \text{blkdiag}(A_1, \ldots, A_N)$ and $B = \text{blkdiag}(B_1, \ldots, B_N)$ the overall system matrices (here blkdiag is the block diagonal operator) and by $w_k = (w_{1,k}, \ldots, w_{N,k})$ the overall disturbance vector, the system dynamics can be rewritten as

$$x_{k+1} = Ax_k + Bu_k + w_k.$$

Let us define $H = [H_1, \ldots, H_N] \in \mathbb{R}^{s \times n}$ and let us define the following matrices and vectors

$$\begin{aligned}
&\mathbb{R}^{n\times n} \ni Q = \text{blkdiag}(Q_1, \ldots, Q_N) + \frac{1}{N^2}\sum_{i=1}^{N} H^\top F_i H,\\
&\mathbb{R}^{m\times m} \ni R = \text{blkdiag}(R_1, \ldots, R_N),\\
&\mathbb{R}^{n} \ni q = [q_1^\top, \ldots, q_N^\top]^\top + \frac{1}{N}\sum_{i=1}^{N} H^\top f,\\
&\mathbb{R}^{m} \ni r = [r_1^\top, \ldots, r_N^\top]^\top.
\end{aligned}$$

Moreover, let us define the augmented state $\tilde{x} = \left[\begin{smallmatrix} 1 \\ x \end{smallmatrix}\right] \in \mathbb{R}^{n+1}$ with system matrices $\tilde{A} = \text{blkdiag}(1, A)$, $\tilde{B} = \left[\begin{smallmatrix} 0 \\ B \end{smallmatrix}\right]$, and the augmented cost matrices $\tilde{Q} = \left[\begin{smallmatrix} 0 & q^\top \\ q & Q \end{smallmatrix}\right]$ and $\tilde{S} = \left[\begin{smallmatrix} r^\top \\ 0 \end{smallmatrix}\right]$. With these positions, problem (11) is seen to be equivalent to the optimal control problem

$$\begin{aligned}
\min_{u}\ & \mathbb{E}\bigg[\frac{1}{2}\sum_{k=0}^{\infty} \alpha^k \big(\tilde{x}^\top \tilde{Q}\tilde{x} + u^\top R u + 2\tilde{x}^\top \tilde{S} u\big)\bigg] \\
\text{subj. to }\ & \tilde{x}_{k+1} = \tilde{A}\tilde{x}_k + \tilde{B}u_k + \tilde{w}_k,
\end{aligned} \tag{36}$$

where $\tilde{w}_k = \left[\begin{smallmatrix} 0 \\ w_k \end{smallmatrix}\right]$. By using standard dynamic programming arguments (see, e.g., [18]), it can be seen that the optimal solution of problem (36) is a linear feedback $u = \tilde{K}\tilde{x}$, where $\tilde{K} = -(R + \alpha \tilde{B}^\top \tilde{P}\tilde{B})^{-1}(\tilde{S} + \alpha\tilde{B}^\top \tilde{P}\tilde{A})\tilde{x}$ and $\tilde{P}$ is the solution of a suitable Algebraic Riccati Equation. Let us write $\tilde{K} \in \mathbb{R}^{(n+1)\times m}$ as $\tilde{K} = [v^\star \ \ K^\star]$, with $v^\star \in \mathbb{R}^m$ and $K^\star \in \mathbb{R}^{n\times m}$. Then, the linear feedback becomes

$$u = \tilde{K}\tilde{x} = \begin{bmatrix} v^\star & K^\star \end{bmatrix}\begin{bmatrix} 1 \\ x \end{bmatrix} = K^\star x + v,$$

and the proof follows. $\square$

B. Proof of Proposition 3.2

Fix the initial states to $\bar{x}_i$ and the policy parameters of π_i to some K_i, v_i for all $i \in \mathbb{I}$. Since each system follows the policy π_i such that $u_{i,k} = K_i x_{i,k} + v_i + \eta_{i,k}$, it holds

$$\begin{aligned}
x_{i,k+1} &= A_i x_{i,k} + B_i(K_i x_{i,k} + v_i + \eta_{i,k}) + w_{i,k} \\
&= (A_i + B_i K_i)x_{i,k} + B_i v_i + B\eta_{i,k} + w_{i,k}.
\end{aligned} \tag{37}$$

The evolution of each system i can be thus written in closed form as

$$x_{i,k} = \underbrace{(A_i + B_i K_i)^k}_{:=\Phi_{i,k}}\bar{x}_i + \underbrace{\sum_{\tau=0}^{k-1} A_i^{k-\tau-1}(B_i v_i + B\eta_{i,\tau} + w_{i,\tau})}_{:=\xi_{i,k}}$$

Similarly we can also express the input as a function of the initial state and of the noise realizations:

$$u_{i,k} = K_i x_{i,k} + v_i + \eta_{i,k} = K_i\Phi_{i,k}\bar{x}_i + \underbrace{K_i\xi_{i,k} + \eta_{i,k}}_{:=\psi_{i,k}}$$

Notice that, since we suppose $\mathbb{E}[w_{i,k}] = 0, \mathbb{E}[\eta_{i,k}] = 0$ and both of them i.i.d. we have

$$\mathbb{E}\left[x_{i,k}\right] = \mathbb{E}\left[\Phi_{i,k}\bar{x}_i\right] + \mathbb{E}\left[\xi_{i,k}\right] = \Phi_{i,k}\bar{x}_i \tag{38}$$

and, similarly,

$$\mathbb{E}\left[u_{i,t}\right] = \mathbb{E}\left[K_i\Phi_{i,k}\bar{x}_i\right] + \mathbb{E}\left[\psi_{i,k}\right] = K_i\Phi_{i,k}\bar{x}_i. \tag{39}$$

For ease of exposition, let us assume that the linear terms in the cost are zero, i.e., that $q_i = 0, r_i = 0, f_i = 0$ (the derivations that follow are similar for the case in which the linear terms are nonzero). Thus we must consider

$$\begin{aligned}
J_\pi(x) = \sum_{i=1}^{N}\mathbb{E}\bigg[\frac{1}{2}\sum_{k=0}^{\infty}\alpha^k\big(&x_{i,k}^\top Q_i x_{i,k} + u_{i,k}^\top R_i u_{i,k} \\
&+ \sigma(x_k)^\top F_i \sigma(x_k)\big)\bigg].
\end{aligned} \tag{40}$$

Considering the closed form evolution of each system, exploiting the linearity of the expected value and using the definition of $\sigma(x)$, we obtain

$$\begin{aligned}
J_\pi(x) = \sum_{i=1}^{N}\frac{1}{2}\sum_{k=0}^{\infty}\Bigg\{&\mathbb{E}\bigg[\bar{x}_i^\top\alpha^k\bigg(\Phi_{i,k}^\top Q_i\Phi_{i,k} \\
&\qquad + \Phi_{i,k}^\top K_i^\top R_i K_i\Phi_{i,k}\bigg)\bar{x}_i\bigg] \\
&+ \mathbb{E}\bigg[2\alpha^k\bigg(\xi_{i,k}^\top Q_i\Phi_{i,k} + \psi_{i,k}^\top R_i K_i\Phi_{i,k}\bigg)\bar{x}_i\bigg] \\
&+ \mathbb{E}\bigg[\alpha^k\bigg(\xi_{i,k}^\top Q_i\xi_{i,k} + \psi_{i,k}^\top R_i\psi_{i,k}\bigg)\bigg] \\
&+ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{\ell=1}^{N}\Bigg(\mathbb{E}\bigg[\bar{x}_\ell^\top\alpha^k\bigg(\Phi_{\ell,k}^\top H_\ell^\top F_i H_j\Phi_{j,k}\bigg)\bar{x}_\ell\bigg] \\
&\quad + \mathbb{E}\bigg[2\alpha^k\bigg(\xi_{\ell,k}^\top H_\ell^\top F_i H_j\Phi_{j,k}\bigg)\bar{x}_j\bigg] \\
&\quad + \mathbb{E}\bigg[\alpha^k\bigg(\xi_{\ell,k}^\top H_\ell^\top F_i H_j\xi_{j,k}\bigg)\bigg]\Bigg)\Bigg\}.
\end{aligned}$$

Then, in light of (38) and (39) and defining

$$\begin{aligned}
\tilde{P}_i &:= \sum_{k=0}^{\infty} \alpha^k \left(\tilde{\Phi}_{i,k}^\top Q_i \tilde{\Phi}_{i,k} + \tilde{\Phi}_{i,k}^\top K_i^\top R_i K_i \tilde{\Phi}_{i,k} \right) \\
\tilde{S}_i &:= F_i \\
\tilde{\sigma}(\bar{x}) &:= \frac{1}{N} \sum_{i=0}^{N} \underbrace{\sum_{k=0}^{\infty} \sqrt{\alpha^k} H_i \Phi_{i,k}}_{:=\tilde{H}_i} \bar{x}_i \\
\zeta_i &:= \sum_{k=0}^{\infty} \alpha^k \left(\xi_{i,k}^\top Q_i \xi_{i,k} + \psi_{i,k}^\top R_i \psi_{i,k} \right) \\
\varsigma &:= \frac{1}{N} \sum_{i=0}^{N} \sum_{k=0}^{\infty} \sqrt{\alpha^k} H_i \xi_{i,k},
\end{aligned}$$

we can finally write:

$$J_\pi(x) = \frac{1}{2} \sum_{i=1}^{N} \left(\bar{x}_i^\top \tilde{P}_i \bar{x}_i + \tilde{\sigma}(\bar{x})^\top \tilde{S}_i \tilde{\sigma}(\bar{x}) + \tilde{\rho}_i \right), \tag{41}$$

with $\tilde{\rho}_i = \mathbb{E}[\zeta_i] + \mathbb{E}[\varsigma^\top \tilde{S}_i \varsigma]$. For the case in which the linear terms are nonzero, there will be additional linear terms in (41). The proof follows. $\square$