

A Appendix

A.1 Game-Theoretic Interpretation of CGD

In this section, we provide the game-theoretic interpretation of CGD and show why taking it into account is essential in multi-agent settings. Recall the CGD update rule:

$$\begin{aligned}\theta_{k+1} &= \theta_k + \alpha (I - \alpha\beta D_{\theta\phi} f D_{\phi\theta} g)^{-1} (\nabla_{\theta} f + \alpha D_{\theta\phi} f \nabla_{\phi} g) \\ \phi_{k+1} &= \phi_k + \beta (I - \beta\alpha D_{\phi\theta} g D_{\theta\phi} f)^{-1} (\nabla_{\phi} g + \beta D_{\phi\theta} g \nabla_{\theta} f)\end{aligned}\tag{25}$$

For a small enough $\alpha = \beta$, (25) can be equivalently written in block matrix form as:

$$\begin{pmatrix} \theta_{k+1} \\ \phi_{k+1} \end{pmatrix} = \begin{pmatrix} \theta_k \\ \phi_k \end{pmatrix} + \alpha \begin{pmatrix} I & -\alpha D_{\theta\phi} f \\ -\alpha D_{\phi\theta} g & I \end{pmatrix}^{-1} \begin{pmatrix} \nabla_{\theta} f \\ \nabla_{\phi} g \end{pmatrix}$$

, which can be re-written as:

$$\begin{pmatrix} \theta_{k+1} \\ \phi_{k+1} \end{pmatrix} = \begin{pmatrix} \theta_k \\ \phi_k \end{pmatrix} + \alpha (I - A)^{-1} \begin{pmatrix} \nabla_{\theta} f \\ \nabla_{\phi} g \end{pmatrix}, \text{ with } A = \begin{pmatrix} 0 & \alpha D_{\theta\phi} f \\ \alpha D_{\phi\theta} g & 0 \end{pmatrix}$$

If the spectral radius of $A < 1$, then we can use Neumann series to compute the inverse:

$$(I - A)^{-1} = \lim_{N \rightarrow \infty} \sum_{i=0}^N A^i = I + A + A^2 + A^3 + \dots$$

The partial sums yield update rules for different levels of reasoning in game theory. Observe that $N = 0$ corresponds to the SimGA update rule (4). This is equivalent to level 0 recursion reasoning in game theory, where each agent optimizes thinking that the other agent is constant (not changing its parameters). The second partial sum ($N = 1$) corresponds to level 1 reasoning, where each agent optimizes thinking that the other agent optimizes thinking that the agent is constant. A similar hierarchy is followed when going to higher levels. When $\lim N \rightarrow \infty$, we recover the Nash equilibrium given by CGD in the limit, which corresponds to infinite recursion reasoning. For practical implementation of CGD, the matrix inverse-vector products can be estimated efficiently via Krylov subspace methods like conjugate gradient (for zero-sum games) or GMRES (for general-sum games). Also, we never need to actually compute the mixed hessian since a much more efficient implementation can be done via hessian-vector products in deep learning libraries like PyTorch [Paszke et al., 2019] and TensorFlow [Abadi et al., 2016].

This game-theoretic interpretation answers why SimGA fails in the bilinear game. SimGA corresponds to level 0, and each agent updates its parameters thinking that the other agent does not update, which is obviously not true. Hence it fails to take into account the non-stationarity of other agent's parameters. On the contrary, CGD takes other agent's moves into account and updates accordingly. As a result, the bilinear game converges to the Nash equilibrium $(0, 0)$ using CGD (figure 1).

A.2 Interactive Policy Optimization Derivations

In this section, we provide the derivations of all the algorithms in section 4. In all the derivations below, we derive the expression for agent 1 ($i = 1$) and it will follow similarly for agent 2 ($i = 2$).

A.2.1 Vanilla Stochastic Policy Optimization

Here we provide the derivations of the gradients and mixed Hessians for vanilla stochastic policy gradients [Sutton et al., 1999].

Expression for gradient (12)

$$\nabla_{\theta^1} J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{p(\tau_{0:t})} [\gamma^t \nabla_{\theta^1} \log \pi^1(a_t^1 | s_t) Q^1(s_t, a_t^1, a_t^2)]$$

Proof. We start by considering relation between the state and action-value functions. To reduce notational complexity, let $\pi(a_t^i|s_t; \theta^i) = \pi^i(a_t^i|s_t)$, $Q_\theta^i(s_t, a_t^1, a_t^2) = Q^i(s_t, a_t^1, a_t^2)$, $V_\theta^i(s_t) = V^i(s_t)$.

$$V^1(s_0) = \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2$$

$$\begin{aligned} \nabla_{\theta^1} V^1(s_0) &= \int_{a_0^1} \int_{a_0^2} \nabla_{\theta^1} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \\ &\quad + \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) \nabla_{\theta^1} Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \end{aligned}$$

Unrolling using the definition of the action-value function:

$$\begin{aligned} \nabla_{\theta^1} V^1(s_0) &= \int_{a_0^1} \int_{a_0^2} \nabla_{\theta^1} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \\ &\quad + \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) \nabla_{\theta^1} \left[r^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) V^1(s_1) ds_1 \right] da_0^1 da_0^2 \\ \nabla_{\theta^1} V^1(s_0) &= \int_{a_0^1} \int_{a_0^2} \nabla_{\theta^1} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \\ &\quad + \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} V^1(s_1) ds_1 da_0^1 da_0^2 \end{aligned}$$

There is a clear recursion and unrolling $\nabla_{\theta^1} V^1(s_1)$ in the last term by one step we get:

$$\begin{aligned} \nabla_{\theta^1} V^1(s_0) &= \int_{a_0^1} \int_{a_0^2} \nabla_{\theta^1} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \\ &\quad + \gamma \int_{a_0^1} \int_{a_0^2} \int_{s_1} \int_{a_1^1} \int_{a_1^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} \pi^1(a_1^1|s_1) \pi^2(a_1^2|s_1) Q^1(s_1, a_1^1, a_1^2) da_1^1 da_1^2 ds_1 da_0^1 da_0^2 \\ &\quad + \gamma \int_{a_0^1} \int_{a_0^2} \int_{s_1} \int_{a_1^1} \int_{a_1^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) P(s_1|s_0, a_0^1, a_0^2) \pi^1(a_1^1|s_1) \pi^2(a_1^2|s_1) \gamma \int_{s_2} P(s_2|s_1, a_1^1, a_1^2) \nabla_{\theta^1} V^1(s_2) \\ &\quad \quad \quad ds_2 da_1^1 da_1^2 ds_1 da_0^1 da_0^2 \end{aligned}$$

Multiplying both sides by $\rho(s_0)$ (initial state distribution), integrating with respect to s_0 , and applying the log-derivative trick we get,

$$\begin{aligned} \nabla_{\theta^1} J^1 &= \int_{s_0} \rho(s_0) \nabla_{\theta^1} V^1(s_0) ds_0 \\ &= \int_{s_0} \int_{a_0^1} \int_{a_0^2} \rho(s_0) \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) \nabla_{\theta^1} \log \pi^1(a_0^1|s_0) Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 ds_0 \\ &\quad + \gamma \int_{s_0} \int_{a_0^1} \int_{a_0^2} \int_{s_1} \int_{a_1^1} \int_{a_1^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) P(s_1|s_0, a_0^1, a_0^2) \pi^1(a_1^1|s_1) \pi^2(a_1^2|s_1) \nabla_{\theta^1} \log \pi^1(a_1^1|s_1) Q^1(s_1, a_1^1, a_1^2) \\ &\quad \quad \quad da_1^1 da_1^2 ds_1 da_0^1 da_0^2 ds_0 + \dots \end{aligned}$$

Using the probability distribution of a truncated trajectory from (7), we get:

$$\begin{aligned} \nabla_{\theta^1} J^1 &= \int_{\tau_{0:0}} p(\tau_{0:0}) \nabla_{\theta^1} \log \pi^1(a_0^1|s_0) Q^1(s_0, a_0^1, a_0^2) d\tau_{0:0} \\ &\quad + \gamma \int_{\tau_{0:1}} p(\tau_{0:1}) \nabla_{\theta^1} \log \pi^1(a_1^1|s_1) Q^1(s_1, a_1^1, a_1^2) d\tau_{0:1} \\ &\quad \quad \quad + \gamma^2 \int_{\tau_{0:2}} p(\tau_{0:2}) \nabla_{\theta^1} \log \pi^1(a_2^1|s_2) Q^1(s_2, a_2^1, a_2^2) d\tau_{0:2} + \dots \end{aligned}$$

$$\nabla_{\theta^1} J^1 = \sum_{t=0}^{T-1} \gamma^t \int_{\tau_{0:t}} p(\tau_{0:t}) \nabla_{\theta^1} \log \pi^1(a_t^1 | s_t) Q^1(s_t, a_t^1, a_t^2) d\tau_{0:t}$$

Finally, we can re-write the above equation as an expectation:

$$\nabla_{\theta^1} J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{p(\tau_{0:t})} [\gamma^t \nabla_{\theta^1} \log \pi^1(a_t^1 | s_t) Q^1(s_t, a_t^1, a_t^2)]$$

□

Expression for mixed hessian using monte-carlo rollouts (13)

Let $\pi(a_t^i | s_t; \theta^i) = \pi^i(a_t^i | s_t)$, $Q_\theta^i(s_t, a_t^1, a_t^2) = Q^i(s_t, a_t^1, a_t^2)$, $V_\theta^i(s_t) = V^i(s_t)$ to reduce notational complexity.

$$D_{\theta^1 \theta^2} J^1 = \mathbb{E}_{p(\tau)} \left[\sum_{t=0}^{T-1} \gamma^t r_t^1 \left(\sum_{k=0}^t \nabla_{\theta^1} \log \pi^1(a_k^1 | s_k) \right) \left(\sum_{k=0}^t \nabla_{\theta^2} \log \pi^2(a_k^2 | s_k) \right)^\top \right]$$

Proof. Let us consider the definition of the policy objective defined in (8):

$$J^1 = \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T-1} \gamma^t r_t^1(s_t, a_t^1, a_t^2) \right]$$

For simplicity of notation, $r_t^1(s_t, a_t^1, a_t^2) = r_t^1$. We can write the above equation as a sum of expectations:

$$J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{\tau \sim p(\tau)} [\gamma^t r_t^1]$$

Note that r_t^1 depends only up on the states and actions sampled from p till time step t . Hence we can re-write the expectation with respect to the truncated trajectory distribution (terms after time step t in the above equation integrate to 1):

$$J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{\tau_{0:t} \sim p(\tau_{0:t})} [\gamma^t r_t^1]$$

$$J^1 = \sum_{t=0}^{T-1} \mathbb{E}_M \left[\int_{a_0^1} \int_{a_0^2} \dots \int_{a_t^1} \int_{a_t^2} \gamma^t r_t^1 \left(\prod_{k=0}^t \pi^1(a_k^1 | s_k) \right) \left(\prod_{k=0}^t \pi^2(a_k^2 | s_k) \right) da_t^2 da_t^1 \dots da_0^2 da_0^1 \right]$$

$$D_{\theta^1 \theta^2} J^1 = \sum_{t=0}^{T-1} \mathbb{E}_M \left[\int_{a_0^1} \int_{a_0^2} \dots \int_{a_t^1} \int_{a_t^2} \gamma^t r_t^1 \nabla_{\theta^1} \left(\prod_{k=0}^t \pi^1(a_k^1 | s_k) \right) \nabla_{\theta^2} \left(\prod_{k=0}^t \pi^2(a_k^2 | s_k) \right)^\top da_t^2 da_t^1 \dots da_0^2 da_0^1 \right]$$

where M contains initial and state transition probability distributions. Applying the log-trick, we get:

$$D_{\theta^1 \theta^2} J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{\tau_{0:t} \sim p(\tau_{0:t})} \left[\gamma^t r_t^1 \nabla_{\theta^1} \log \left(\prod_{k=0}^t \pi^1(a_k^1 | s_k) \right) \nabla_{\theta^2} \log \left(\prod_{k=0}^t \pi^2(a_k^2 | s_k) \right)^\top \right]$$

$$D_{\theta^1 \theta^2} J^1 = \mathbb{E}_{p(\tau)} \left[\sum_{t=0}^{T-1} \gamma^t r_t^1 \left(\sum_{k=0}^t \nabla_{\theta^1} \log \pi^1(a_k^1 | s_k) \right) \left(\sum_{k=0}^t \nabla_{\theta^2} \log \pi^2(a_k^2 | s_k) \right)^\top \right] \quad (26)$$

□

Expression for mixed hessian using value functions (14)

(26) gives us an expression for mixed hessian. However, it would be of higher variance due to monte-carlo estimation. We can use the definition of the value functions to get a lower variance estimate. Let $\pi(a_t^i|s_t; \theta^i) = \pi^i(a_t^i|s_t)$, $Q_\theta^i(s_t, a_t^1, a_t^2) = Q^i(s_t, a_t^1, a_t^2)$, $V_\theta^i(s_t) = V^i(s_t)$ to reduce notational complexity.

$$\begin{aligned} D_{\theta^1 \theta^2} J^1 &= \sum_{t=0}^{T-1} \mathbb{E}_{p(\tau_{0:t})} [\gamma^t \nabla_{\theta^1} \log \pi^1(a_t^1|s_t) \nabla_{\theta^2} \log \pi^2(a_t^2|s_t)^\top Q^1(s_t, a_t^1, a_t^2)] \\ &+ \sum_{t=1}^{T-1} \mathbb{E}_{p(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \log \pi^1(a_t^1|s_t) \nabla_{\theta^2} \log \left(\prod_{k=0}^{t-1} \pi^2(a_k^2|s_k) \right)^\top Q^1(s_t, a_t^1, a_t^2) \right] \\ &+ \sum_{t=1}^{T-1} \mathbb{E}_{p(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \log \left(\prod_{k=0}^{t-1} \pi^1(a_k^1|s_k) \right) \nabla_{\theta^2} \log \pi^2(a_t^2|s_t)^\top Q^1(s_t, a_t^1, a_t^2) \right] \end{aligned}$$

Proof. Using the relation between state and action-value functions:

$$\begin{aligned} V^1(s_0) &= \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \\ \nabla_{\theta^2} V^1(s_0) &= \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \nabla_{\theta^2} \pi^2(a_0^2|s_0) Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \\ &+ \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) \nabla_{\theta^2} Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \\ D_{\theta^1 \theta^2} V^1(s_0) &= \int_{a_0^1} \int_{a_0^2} \nabla_{\theta^1} \pi^1(a_0^1|s_0) \nabla_{\theta^2} \pi^2(a_0^2|s_0)^\top Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \\ &+ \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \nabla_{\theta^1} Q^1(s_0, a_0^1, a_0^2) \nabla_{\theta^2} \pi^2(a_0^2|s_0)^\top da_0^1 da_0^2 \\ &+ \int_{a_0^1} \int_{a_0^2} \nabla_{\theta^1} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) \nabla_{\theta^2} Q^1(s_0, a_0^1, a_0^2)^\top da_0^1 da_0^2 \\ &+ \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) D_{\theta^1 \theta^2} Q^1(s_0, a_0^1, a_0^2) da_0^1 da_0^2 \quad (27) \end{aligned}$$

We define the following notation:

$$\begin{aligned} A_t &= \int_{a_t^1} \int_{a_t^2} \nabla_{\theta^1} \pi^1(a_t^1|s_t) \nabla_{\theta^2} \pi^2(a_t^2|s_t)^\top Q^1(s_t, a_t^1, a_t^2) da_t^1 da_t^2 \\ B_t &= \int_{a_t^1} \int_{a_t^2} \pi^1(a_t^1|s_t) \nabla_{\theta^1} Q^1(s_t, a_t^1, a_t^2) \nabla_{\theta^2} \pi^2(a_t^2|s_t)^\top da_t^1 da_t^2 \\ C_t &= \int_{a_t^1} \int_{a_t^2} \nabla_{\theta^1} \pi^1(a_t^1|s_t) \pi^2(a_t^2|s_t) \nabla_{\theta^2} Q^1(s_t, a_t^1, a_t^2)^\top da_t^1 da_t^2 \end{aligned}$$

Multiplying (27) by $\rho(s_0)$ (initial state distribution) both sides, integrating, and unrolling using recursion, we get:

$$\begin{aligned} D_{\theta^1 \theta^2} J^1 &= \int_{s_0} \rho(s_0) D_{\theta^1 \theta^2} V^1(s_0) ds_0 = \int_{s_0} \rho(s_0) (A_0 + B_0 + C_0) ds_0 \\ &+ \gamma \int_{s_0} \int_{a_0^1} \int_{a_0^2} \int_{s_1} \rho(s_0) \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) P(s_1|s_0, a_0^1, a_0^2) (A_1 + B_1 + C_1) ds_1 da_0^2 da_0^1 ds_0 + \dots \end{aligned} \quad (28)$$

Let us consider each of the three recursions (**A**, **B**, **C**) separately. First recursion:

$$\mathbf{A} = \int_{s_0} \rho(s_0) A_0 ds_0 + \gamma \int_{s_0} \int_{a_0^1} \int_{a_0^2} \int_{s_1} \rho(s_0) \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) P(s_1|s_0, a_0^1, a_0^2) A_1 ds_1 da_0^2 da_0^1 ds_0 + \dots$$

Applying the log-derivative trick for each of the terms and using the definition of the probability distribution of a trajectory from (7), we get:

$$\begin{aligned} \mathbf{A} &= \int_{\tau_{0:0}} p(\tau_{0:0}) \nabla_{\theta^1} \log \pi^1(a_0^1|s_0) \nabla_{\theta^2} \log \pi^2(a_0^2|s_0)^\top Q^1(s_0, a_0^1, a_0^2) d\tau_{0:0} \\ &\quad + \gamma \int_{\tau_{0:1}} p(\tau_{0:1}) \nabla_{\theta^1} \log \pi^1(a_1^1|s_1) \nabla_{\theta^2} \log \pi^2(a_1^2|s_1)^\top Q^1(s_1, a_1^1, a_1^2) d\tau_{0:1} \\ &\quad + \gamma^2 \int_{\tau_{0:2}} p(\tau_{0:2}) \nabla_{\theta^1} \log \pi^1(a_2^1|s_2) \nabla_{\theta^2} \log \pi^2(a_2^2|s_2)^\top Q^1(s_2, a_2^1, a_2^2) d\tau_{0:2} + \dots \\ \mathbf{A} &= \sum_{t=0}^{T-1} \gamma^t \int_{\tau_{0:t}} p(\tau_{0:t}) \nabla_{\theta^1} \log \pi^1(a_t^1|s_t) \nabla_{\theta^2} \log \pi^2(a_t^2|s_t)^\top Q^1(s_t, a_t^1, a_t^2) d\tau_{0:t} \end{aligned}$$

Writing as an expectation:

$$\mathbf{A} = \sum_{t=0}^{T-1} \mathbb{E}_{p(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \log \pi^1(a_t^1|s_t) \nabla_{\theta^2} \log \pi^2(a_t^2|s_t)^\top Q^1(s_t, a_t^1, a_t^2) \right]$$

Second recursion:

$$\mathbf{B} = \int_{s_0} \rho(s_0) B_0 ds_0 + \gamma \int_{s_0} \int_{a_0^1} \int_{a_0^2} \int_{s_1} \rho(s_0) \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) P(s_1|s_0, a_0^1, a_0^2) B_1 ds_1 da_0^2 da_0^1 ds_0 + \dots$$

Let us consider the first element with B_0 :

$$B_0 = \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \nabla_{\theta^1} Q^1(s_0, a_0^1, a_0^2) \nabla_{\theta^2} \pi^2(a_0^2|s_0)^\top da_0^1 da_0^2$$

Using the definition of the action-value function:

$$\begin{aligned} B_0 &= \int_{a_0^1} \int_{a_0^2} \pi^1(a_0^1|s_0) \nabla_{\theta^1} \left[r^1(s_0, a_0) + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) V^1(s_1) ds_1 \right] \nabla_{\theta^2} \pi^2(a_0^2|s_0)^\top da_0^1 da_0^2 \\ \int_{s_0} \rho(s_0) B_0 ds_0 &= \gamma \int_{s_0} \int_{a_0^1} \int_{a_0^2} \int_{s_1} \rho(s_0) \pi^1(a_0^1|s_0) P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} V^1(s_1) \nabla_{\theta^2} \pi^2(a_0^2|s_0)^\top ds_1 da_0^1 da_0^2 ds_0 \end{aligned} \quad (29)$$

We have the following:

$$\begin{aligned} \nabla_{\theta^1} V^1(s_1) &= \int_{a_1^1} \int_{a_1^2} \nabla_{\theta^1} \pi^1(a_1^1|s_1) \pi^2(a_1^2|s_1) Q^1(s_1, a_1^1, a_1^2) da_1^1 da_1^2 \\ &\quad + \int_{a_1^1} \int_{a_1^2} \pi^1(a_1^1|s_1) \pi^2(a_1^2|s_1) \nabla_{\theta^1} Q^1(s_1, a_1^1, a_1^2) da_1^1 da_1^2 \end{aligned}$$

Substituting $\nabla_{\theta^1} V^1(s_1)$ from the above equation into (29) and using the definition of action-value function again, we will get a recursion, solving which gives us:

$$\begin{aligned} \mathbf{X}_0 &= \int_{s_0} \rho(s_0) B_0 ds_0 = \gamma \int_{\tau_{0:1}} p(\tau_{0:1}) \nabla_{\theta^1} \log \pi^1(a_1^1|s_1) \nabla_{\theta^2} \log \pi^2(a_0^2|s_0)^\top Q^1(s_1, a_1^1, a_1^2) d\tau_{0:1} \\ &\quad + \gamma^2 \int_{\tau_{0:2}} p(\tau_{0:2}) \nabla_{\theta^1} \log \pi^1(a_2^1|s_2) \nabla_{\theta^2} \log \pi^2(a_0^2|s_0)^\top Q^1(s_2, a_2^1, a_2^2) d\tau_{0:2} \\ &\quad + \gamma^3 \int_{\tau_{0:3}} p(\tau_{0:3}) \nabla_{\theta^1} \log \pi^1(a_3^1|s_3) \nabla_{\theta^2} \log \pi^2(a_0^2|s_0)^\top Q^1(s_3, a_3^1, a_3^2) d\tau_{0:3} + \dots \end{aligned}$$

Similarly, for the second term of \mathbf{B} , we get:

$$\begin{aligned}
\mathbf{X}_1 &= \gamma \int_{s_0} \int_{a_0^1} \int_{a_0^2} \int_{s_1} \rho(s_0) \pi^1(a_0^1|s_0) \pi^2(a_0^2|s_0) P(s_1|s_0, a_0^1, a_0^2) B_1 ds_1 da_0^2 da_0^1 ds_0 \\
&= \gamma^2 \int_{\tau_{0:2}} p(\tau_{0:2}) \nabla_{\theta^1} \log \pi^1(a_2^1|s_2) \nabla_{\theta^2} \log \pi^2(a_1^2|s_1)^\top Q^1(s_2, a_2^1, a_2^2) d\tau_{0:2} \\
&\quad + \gamma^3 \int_{\tau_{0:3}} p(\tau_{0:3}) \nabla_{\theta^1} \log \pi^1(a_3^1|s_3) \nabla_{\theta^2} \log \pi^2(a_1^2|s_1)^\top Q^1(s_3, a_3^1, a_3^2) d\tau_{0:3} \\
&\quad + \gamma^4 \int_{\tau_{0:4}} p(\tau_{0:4}) \nabla_{\theta^1} \log \pi^1(a_4^1|s_4) \nabla_{\theta^2} \log \pi^2(a_1^2|s_1)^\top Q^1(s_4, a_4^1, a_4^2) d\tau_{0:4} + \dots
\end{aligned}$$

The third term of \mathbf{B} is equal to:

$$\begin{aligned}
\mathbf{X}_2 &= \gamma^3 \int_{\tau_{0:3}} p(\tau_{0:3}) \nabla_{\theta^1} \log \pi^1(a_3^1|s_3) \nabla_{\theta^2} \log \pi^2(a_2^2|s_2)^\top Q^1(s_3, a_3^1, a_3^2) d\tau_{0:3} \\
&\quad + \gamma^4 \int_{\tau_{0:4}} p(\tau_{0:4}) \nabla_{\theta^1} \log \pi^1(a_4^1|s_4) \nabla_{\theta^2} \log \pi^2(a_2^2|s_2)^\top Q^1(s_4, a_4^1, a_4^2) d\tau_{0:4} \\
&\quad + \gamma^5 \int_{\tau_{0:5}} p(\tau_{0:5}) \nabla_{\theta^1} \log \pi^1(a_5^1|s_5) \nabla_{\theta^2} \log \pi^2(a_2^2|s_2)^\top Q^1(s_5, a_5^1, a_5^2) d\tau_{0:5} + \dots
\end{aligned}$$

We see a clear pattern here. We evaluate $\mathbf{B} = \mathbf{X}_0 + \mathbf{X}_1 + \mathbf{X}_2 + \dots$ to get the contribution of \mathbf{B} -terms to the mixed hessian $D_{\theta^1 \theta^2} J^1$ (28). For clarity, we define the following short-hand notation:

$$\nabla_{\theta^1} \log \pi^1(a_t^1|s_t) = f_t^1, \quad \nabla_{\theta^2} \log \pi^2(a_t^2|s_t) = g_t^1$$

With some rearrangement,

$$\begin{aligned}
\mathbf{B} &= \mathbf{X}_0 + \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \dots = \gamma \int_{\tau_{0:1}} p(\tau_{0:1}) f_1^1 [g_0^2]^\top Q^1(s_1, a_1^1, a_1^2) d\tau_{0:1} \\
&\quad + \gamma^2 \int_{\tau_{0:2}} p(\tau_{0:2}) f_2^1 [g_0^2 + g_1^2]^\top Q^1(s_2, a_2^1, a_2^2) d\tau_{0:2} \\
&\quad + \gamma^3 \int_{\tau_{0:3}} p(\tau_{0:3}) f_3^1 [g_0^2 + g_1^2 + g_2^2]^\top Q^1(s_3, a_3^1, a_3^2) d\tau_{0:3} \\
&\quad + \gamma^4 \int_{\tau_{0:4}} p(\tau_{0:4}) f_4^1 [g_0^2 + g_1^2 + g_2^2 + g_3^2]^\top Q^1(s_4, a_4^1, a_4^2) d\tau_{0:4} + \dots
\end{aligned}$$

Substituting the actual values back:

$$\mathbf{B} = \sum_{t=1}^{T-1} \gamma^t \int_{\tau_{0:t}} p(\tau_{0:t}) \nabla_{\theta^1} \log \pi^1(a_t^1|s_t) \left[\sum_{k=0}^{t-1} \nabla_{\theta^2} \log \pi^2(a_k^2|s_k) \right]^\top Q^1(s_t, a_t^1, a_t^2) d\tau_{0:t}$$

Writing as an expectation:

$$\mathbf{B} = \sum_{t=1}^{T-1} \mathbb{E}_{p(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \log \pi^1(a_t^1|s_t) \nabla_{\theta^2} \log \left(\prod_{k=0}^{t-1} \pi^2(a_k^2|s_k) \right)^\top Q^1(s_t, a_t^1, a_t^2) \right]$$

With a similar process, we can derive the contribution of \mathbf{C} -terms to the mixed hessian $D_{\theta^1 \theta^2} J$ (28):

$$\mathbf{C} = \sum_{t=1}^{T-1} \mathbb{E}_{p(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \log \left(\prod_{k=0}^{t-1} \pi^1(a_k^1|s_k) \right) \nabla_{\theta^2} \log \pi^2(a_t^2|s_t)^\top Q^1(s_t, a_t^1, a_t^2) \right]$$

Combining \mathbf{A} , \mathbf{B} , \mathbf{C} , we get the required final result. \square

A.2.2 Natural and Trust Region Policy Optimization

Here we provide the derivations pertaining to natural [Kakade, 2001] and trust region policy gradients [Schulman et al., 2015].

Relation between policy performance at θ^1, θ^2 and θ_k^1, θ_k^2 (15)

$$J^1(\theta^1, \theta^2) = J^1(\theta_k^1, \theta_k^2) + \mathbb{E}_{\tau \sim p(\tau_\theta)} \left[\sum_{t=0}^{T-1} \gamma^t A_{\theta_k}^1(s_t, a_t^1, a_t^2) \right]$$

Proof. Let us start with R.H.S

$$\begin{aligned} & \mathbb{E}_{\tau \sim p(\tau_\theta)} \left[\sum_{t=0}^{T-1} \gamma^t A_{\theta_k}^1(s_t, a_t^1, a_t^2) \right] \\ &= \mathbb{E}_{\tau \sim p(\tau_\theta)} \left[\sum_{t=0}^{T-1} \gamma^t Q_{\theta_k}^1(s_t, a_t^1, a_t^2) - \gamma^t V_{\theta_k}^1(s_t) \right] \\ &= \mathbb{E}_{\tau \sim p(\tau_\theta)} \left[\sum_{t=0}^{T-1} \gamma^t r^1(s_t, a_t^1, a_t^2) + \gamma^{t+1} V_{\theta_k}^1(s_{t+1}) - \gamma^t V_{\theta_k}^1(s_t) \right] \\ &= \mathbb{E}_{\tau \sim p(\tau_\theta)} \left[\sum_{t=0}^{T-1} \gamma^t r^1(s_t, a_t^1, a_t^2) \right] + \mathbb{E}_{\tau \sim p(\tau_\theta)} [\gamma^T V_{\theta_k}^1(s_T) - V_{\theta_k}^1(s_0)] \\ &= J^1(\theta^1, \theta^2) + 0 - \mathbb{E}_{\rho(s_0)} [V_{\theta_k}^1(s_0)] = J^1(\theta^1, \theta^2) - J^1(\theta_k^1, \theta_k^2) \end{aligned}$$

We get the above expression since value of terminal state $V_{\theta_k}^1(s_T) = 0$. \square

Policy improvement bound (17)

$$J^1(\theta^1, \theta^2) \geq L^1(\theta^1, \theta^2) - \epsilon \sqrt{2D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))}$$

where $\epsilon = \max_s \left| \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s) \right] \right|$

Proof. Let us re-write the policy performance at $\theta = (\theta^1, \theta^2)$ and the surrogate objective:

$$\begin{aligned} J^1(\theta^1, \theta^2) &= J^1(\theta_k^1, \theta_k^2) + \mathbb{E}_{\tau \sim p(\tau_\theta)} \left[\sum_{t=0}^{T-1} \gamma^t A_{\theta_k}^1(s_t, a_t^1, a_t^2) \right] \\ L^1(\theta^1, \theta^2) &= J^1(\theta_k^1, \theta_k^2) + \mathbb{E}_{\tau \sim p(\tau_{\theta_k})} \left[\sum_{t=0}^{T-1} \gamma^t \frac{\pi(a_t^1 | s_t; \theta^1)}{\pi(a_t^1 | s_t; \theta_k^1)} \frac{\pi(a_t^2 | s_t; \theta^2)}{\pi(a_t^2 | s_t; \theta_k^2)} A_{\theta_k}^1(s_t, a_t^1, a_t^2) \right] \end{aligned}$$

Similar to Schulman et al. [2015], let us define $\bar{A}_{\theta_k}^1(s)$ as the expected advantage of $(\pi(\theta^1), \pi(\theta^2))$ over $(\pi(\theta_k^1), \pi(\theta_k^2))$ in state s :

$$\bar{A}_{\theta_k}^1(s) = \mathbb{E}_{a^1 \sim \pi(\cdot | s; \theta^1), a^2 \sim \pi(\cdot | s; \theta^2)} [A_{\theta_k}^1(s, a^1, a^2)]$$

Using this, the policy performance at (θ^1, θ^2) can be written as:

$$J^1(\theta^1, \theta^2) = J^1(\theta_k^1, \theta_k^2) + \mathbb{E}_{\tau \sim p(\tau_\theta)} \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s_t) \right] \quad (30)$$

The surrogate objective can be written as:

$$L^1(\theta^1, \theta^2) = J^1(\theta_k^1, \theta_k^2) + \mathbb{E}_{\tau \sim p(\tau_{\theta_k})} \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s_t) \right] \quad (31)$$

The difference between these equations is that the states are sampled using $(\pi(\theta^1), \pi(\theta^2))$ in (30) and using $(\pi(\theta_k^1), \pi(\theta_k^2))$ in (31). Now we can write:

$$\begin{aligned}
|J^1(\theta^1, \theta^2) - L^1(\theta^1, \theta^2)| &= \left| \mathbb{E}_{\tau \sim p(\tau_\theta)} \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s_t) \right] - \mathbb{E}_{\tau \sim p(\tau_{\theta_k})} \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s_t) \right] \right| \\
&= \left| \int_{\tau} p(\tau_\theta) - p(\tau_{\theta_k}) \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s_t) \right] d\tau \right| \\
&\leq \int_{\tau} \left| p(\tau_\theta) - p(\tau_{\theta_k}) \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s_t) \right] \right| d\tau \\
&\leq \int_{\tau} |p(\tau_\theta) - p(\tau_{\theta_k})| \left| \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s_t) \right] \right| d\tau \\
&\leq \int_{\tau} |p(\tau_\theta) - p(\tau_{\theta_k})| \epsilon d\tau = \int_{\tau} |p(\tau_{\theta_k}) - p(\tau_\theta)| \epsilon d\tau \\
&= 2\epsilon D_{TV}(p(\tau_{\theta_k}), p(\tau_\theta)) \leq \epsilon \sqrt{2D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))}
\end{aligned}$$

where $\epsilon = \max_s \left| \left[\sum_{t=0}^{T-1} \gamma^t \bar{A}_{\theta_k}^1(s) \right] \right|$, $D_{TV}(p(\tau_{\theta_k}), p(\tau_\theta)) = \frac{1}{2} \int_{\tau} |p(\tau_{\theta_k}) - p(\tau_\theta)| d\tau$,

and $D_{TV}(p(\tau_{\theta_k}), p(\tau_\theta)) \leq \sqrt{\frac{1}{2} D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))}$ via Pinsker's inequality.

From this, we get:

$$J^1(\theta^1, \theta^2) \geq L^1(\theta^1, \theta^2) - \epsilon \sqrt{2D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))}$$

□

Derivation of natural gradient update rule (19)

$$\begin{bmatrix} \theta^1 - \theta_k^1 \\ \theta^2 - \theta_k^2 \end{bmatrix} = \begin{bmatrix} \lambda_1 F_1 & -D_{\theta^1 \theta^2} L^1 \\ -D_{\theta^2 \theta^1} L^2 & \lambda_2 F_2 \end{bmatrix}^{-1} \begin{bmatrix} \nabla_{\theta^1} L^1 \\ \nabla_{\theta^2} L^2 \end{bmatrix}$$

, where $F_1 = D_{\theta^1 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))|_{\theta=\theta_k}$, $F_2 = D_{\theta^2 \theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))|_{\theta=\theta_k}$

Proof. Differentiating (16), for $i, j \in \{1, 2\}, i \neq j$, the gradients and mixed Hessians of L^i (at $\theta = \theta_k$) are given by:

$$\nabla_{\theta^i} L^i(\theta^1, \theta^2) = \mathbb{E}_{\tau \sim p(\tau_{\theta_k})} \left[\sum_{t=0}^{T-1} \gamma^t \nabla_{\theta^i} \log \pi(a_t^i | s_t; \theta^i) A_{\theta_k}^i(s_t, a_t^1, a_t^2) \right]_{\theta=\theta_k} \quad (32)$$

$$D_{\theta^i \theta^j} L^i(\theta^1, \theta^2) = \mathbb{E}_{\tau \sim p(\tau_{\theta_k})} \left[\sum_{t=0}^{T-1} \gamma^t \nabla_{\theta^i} \log \pi(a_t^i | s_t; \theta^i) \nabla_{\theta^j} \log \pi(a_t^j | s_t; \theta^j)^\top A_{\theta_k}^i(s_t, a_t^1, a_t^2) \right]_{\theta=\theta_k} \quad (33)$$

Note that $\nabla_{\theta^i} L^i = \nabla_{\theta^i} J^i$ at $\theta = \theta_k$. This is another advantage of using the surrogate objective that it matches J^i to first order. The bilinear approximation of the surrogate objective in (18) is given by,

$$\begin{aligned}
&\max_{\theta^1} (\theta^1 - \theta_k^1)^\top \nabla_{\theta^1} L^1 + (\theta^1 - \theta_k^1)^\top D_{\theta^1 \theta^2} L^1 (\theta^2 - \theta_k^2), \\
&\max_{\theta^2} (\theta^2 - \theta_k^2)^\top \nabla_{\theta^2} L^2 + (\theta^2 - \theta_k^2)^\top D_{\theta^2 \theta^1} L^2 (\theta^1 - \theta_k^1) \\
&\text{subject to } D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta)) \leq \delta
\end{aligned} \quad (34)$$

We approximate the KL-divergence constraint with a quadratic approximation. For $\theta = (\theta^1, \theta^2), \theta_k = (\theta_k^1, \theta_k^2)$, Taylor expansion for D_{KL} around θ_k :

$$\begin{aligned}
D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta)) &= D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))|_{\theta=\theta_k} + \begin{bmatrix} \theta^1 - \theta_k^1 & \theta^2 - \theta_k^2 \end{bmatrix} \nabla_{\theta} D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))|_{\theta=\theta_k} \\
&\quad + \frac{1}{2} \begin{bmatrix} (\theta^1 - \theta_k^1)^\top & (\theta^2 - \theta_k^2)^\top \end{bmatrix} D_{\theta}^2 D_{KL}(p(\tau_{\theta_k}), p(\tau_\theta))|_{\theta=\theta_k} \begin{bmatrix} \theta^1 - \theta_k^1 \\ \theta^2 - \theta_k^2 \end{bmatrix} + \text{H.O.T}
\end{aligned}$$

where H.O.T. stands for higher order terms (that we would ignore in our approximation) and

$$\begin{aligned}\nabla_{\theta} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} &= \left[\begin{array}{c} \nabla_{\theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) \\ \nabla_{\theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) \end{array} \right]_{\theta=\theta_k} \\ D_{\theta}^2 D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} &= \left[\begin{array}{cc} D_{\theta^1 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) & D_{\theta^1 \theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) \\ D_{\theta^2 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) & D_{\theta^2 \theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) \end{array} \right]_{\theta=\theta_k}\end{aligned}$$

$D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} = 0$ since KL-divergence between two same distributions is 0.

$$\begin{aligned}\nabla_{\theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} &= \left(\nabla_{\theta^1} \int_{\tau} p(\tau_{\theta_k}) \log \frac{p(\tau_{\theta_k})}{p(\tau_{\theta})} d\tau \right)_{\theta=\theta_k} \\ &= - \int_{\tau} p(\tau_{\theta_k}) \frac{\nabla_{\theta^1} p(\tau_{\theta})|_{\theta=\theta_k}}{p(\tau_{\theta_k})} d\tau = 0\end{aligned}$$

Similarly, $\nabla_{\theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} = 0$.

For $D_{\theta^1 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k}$:

$$\begin{aligned}D_{\theta^1 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} &= \left(D_{\theta^1 \theta^1} \int_{\tau} p(\tau_{\theta_k}) \log \frac{p(\tau_{\theta_k})}{p(\tau_{\theta})} d\tau \right)_{\theta=\theta_k} \\ &= - \left(\int_{\tau} p(\tau_{\theta_k}) D_{\theta^1 \theta^1} \log p(\tau_{\theta}) d\tau \right)_{\theta=\theta_k}\end{aligned}$$

Using definition of probability distribution of full trajectory,

$p(\tau_{\theta}) = \rho(s_0) \prod_{t=0}^{T-1} \pi(a_t^1 | s_t; \theta^1) \pi(a_t^2 | s_t; \theta^2) P(s_{t+1} | s_t, a_t^1, a_t^2)$. Taking log and differentiating:

$$\log p(\tau_{\theta}) = \log \rho(s_0) + \sum_{t=0}^{T-1} \log \pi(a_t^1 | s_t; \theta^1) + \sum_{t=0}^{T-1} \log \pi(a_t^2 | s_t; \theta^2) + \sum_{t=0}^{T-1} \log P(s_{t+1} | s_t, a_t^1, a_t^2)$$

$$\nabla_{\theta^1} \log p(\tau_{\theta}) = \sum_{t=0}^{T-1} \nabla_{\theta^1} \log \pi(a_t^1 | s_t; \theta^1), \quad D_{\theta^1 \theta^1} \log p(\tau_{\theta}) = \sum_{t=0}^{T-1} D_{\theta^1 \theta^1} \log \pi(a_t^1 | s_t; \theta^1)$$

$$D_{\theta^1 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} = -\mathbb{E}_{\tau \sim p(\tau_{\theta_k})} \left[\sum_{t=0}^{T-1} D_{\theta^1 \theta^1} \log \pi(a_t^1 | s_t; \theta^1) \right]_{\theta=\theta_k}$$

Similarly, it follows for $D_{\theta^2 \theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k}$.

$D_{\theta^1 \theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} = D_{\theta^2 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} = 0$ as $D_{\theta^1 \theta^2} \log p(\tau_{\theta}) = D_{\theta^2 \theta^1} \log p(\tau_{\theta}) = 0$.

Alternatively, we can also get the hessian of the KL-divergence using only first-order derivatives by using Fisher information:

$$\begin{aligned}D_{\theta^1 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} &= \left(D_{\theta^1 \theta^1} \int_{\tau} p(\tau_{\theta_k}) \log \frac{p(\tau_{\theta_k})}{p(\tau_{\theta})} d\tau \right)_{\theta=\theta_k} \\ &= - \left(\nabla_{\theta^1} \int_{\tau} p(\tau_{\theta_k}) \frac{\nabla_{\theta^1} p(\tau_{\theta})}{p(\tau_{\theta})} d\tau \right)_{\theta=\theta_k} \\ &= - \left(\int_{\tau} \frac{p(\tau_{\theta_k})}{p(\tau_{\theta})} D_{\theta^1 \theta^1} p(\tau_{\theta}) d\tau \right)_{\theta=\theta_k} + \left(\int_{\tau} p(\tau_{\theta_k}) \frac{\nabla_{\theta^1} p(\tau_{\theta})}{p(\tau_{\theta})} \frac{\nabla_{\theta^1} p(\tau_{\theta})^{\top}}{p(\tau_{\theta})} d\tau \right)_{\theta=\theta_k} \\ &= - \int_{\tau} D_{\theta^1 \theta^1} p(\tau_{\theta})|_{\theta=\theta_k} d\tau + \mathbb{E}_{\tau \sim p(\tau_{\theta_k})} \left[\nabla_{\theta^1} \log p(\tau_{\theta}) \nabla_{\theta^1} \log p(\tau_{\theta})^{\top} \right]_{\theta=\theta_k} \\ &= 0 + \mathbb{E}_{\tau \sim p(\tau_{\theta_k})} \left[\left(\sum_{t=0}^{T-1} \nabla_{\theta^1} \log \pi(a_t^1 | s_t; \theta^1) \right) \left(\sum_{t=0}^{T-1} \nabla_{\theta^1} \log \pi(a_t^1 | s_t; \theta^1)^{\top} \right) \right]_{\theta=\theta_k}\end{aligned}$$

We can see that this is the Fisher Information Matrix.

Similarly, we can derive for $D_{\theta^2 \theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k}$. Finally,

$$\begin{aligned} D_{\theta}^2 D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta}))|_{\theta=\theta_k} &= \begin{bmatrix} D_{\theta^1 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) & D_{\theta^1 \theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) \\ D_{\theta^2 \theta^1} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) & D_{\theta^2 \theta^2} D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) \end{bmatrix}_{\theta=\theta_k} \\ &= \begin{bmatrix} F_1 & 0 \\ 0 & F_2 \end{bmatrix} \end{aligned}$$

Finally, for the quadratic approximation of the KL-divergence, we have:

$$D_{KL}(p(\tau_{\theta_k}), p(\tau_{\theta})) \approx \frac{1}{2} [(\theta^1 - \theta_k^1)^\top \quad (\theta^2 - \theta_k^2)^\top] \begin{bmatrix} F_1 & 0 \\ 0 & F_2 \end{bmatrix} \begin{bmatrix} \theta^1 - \theta_k^1 \\ \theta^2 - \theta_k^2 \end{bmatrix} \quad (35)$$

We use the quadratic approximation in (35) instead of the KL-divergence constraint in (34). Using Lagrangian duality, we can write the final optimization problem with a bilinear approximation of the surrogate objective and a quadratic approximation of the KL-divergence:

$$\begin{aligned} \max_{\theta^1} \Delta \theta^1^\top \nabla_{\theta^1} L^1 + \Delta \theta^1^\top D_{\theta^1 \theta^2} L^1 \Delta \theta^2 - \frac{\lambda_1}{2} (\Delta \theta^1^\top F_1 \Delta \theta^1 + \Delta \theta^2^\top F_2 \Delta \theta^2 - 2\delta) \\ \max_{\theta^2} \Delta \theta^2^\top \nabla_{\theta^2} L^2 + \Delta \theta^2^\top D_{\theta^2 \theta^1} L^2 \Delta \theta^1 - \frac{\lambda_2}{2} (\Delta \theta^1^\top F_1 \Delta \theta^1 + \Delta \theta^2^\top F_2 \Delta \theta^2 - 2\delta) \end{aligned} \quad (36)$$

, where $\Delta \theta^1 = \theta^1 - \theta_k^1$, $\Delta \theta^2 = \theta^2 - \theta_k^2$. Note that this natural gradient formulation uses KL-penalty in contrast to Euclidean penalty in (5), indicating that it takes care of the information geometry and that the change in the policy is within the trust region δ . Differentiating, we get:

$$\begin{aligned} \nabla_{\theta^1} L^1 + D_{\theta^1 \theta^2} L^1 (\theta^2 - \theta_k^2) - \lambda_1 F_1 (\theta^1 - \theta_k^1) &= 0 \\ \nabla_{\theta^2} L^2 + D_{\theta^2 \theta^1} L^2 (\theta^1 - \theta_k^1) - \lambda_2 F_2 (\theta^2 - \theta_k^2) &= 0 \end{aligned} \quad (37)$$

Writing in block matrix form:

$$\begin{bmatrix} \nabla_{\theta^1} L^1 \\ \nabla_{\theta^2} L^2 \end{bmatrix} + \begin{bmatrix} 0 & D_{\theta^1 \theta^2} L^1 \\ D_{\theta^2 \theta^1} L^2 & 0 \end{bmatrix} \begin{bmatrix} \theta^1 - \theta_k^1 \\ \theta^2 - \theta_k^2 \end{bmatrix} - \begin{bmatrix} \lambda_1 F_1 & 0 \\ 0 & \lambda_2 F_2 \end{bmatrix} \begin{bmatrix} \theta^1 - \theta_k^1 \\ \theta^2 - \theta_k^2 \end{bmatrix} = 0 \quad (38)$$

After rearrangement, the final update rule for both agents is:

$$\begin{bmatrix} \theta^1 - \theta_k^1 \\ \theta^2 - \theta_k^2 \end{bmatrix} = \begin{bmatrix} \lambda_1 F_1 & -D_{\theta^1 \theta^2} L^1 \\ -D_{\theta^2 \theta^1} L^2 & \lambda_2 F_2 \end{bmatrix}^{-1} \begin{bmatrix} \nabla_{\theta^1} L^1 \\ \nabla_{\theta^2} L^2 \end{bmatrix}$$

□

A.2.3 Deterministic Policy Optimization

Here we provide the derivations of the gradients and mixed Hessians for deterministic policy gradients [Silver et al., 2014, Lillicrap et al., 2019].

Derivation of gradient (22)

$$\nabla_{\theta^1} J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{q(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \mu(s_t; \theta^1) \nabla_{a_t^1} Q_{\theta^1}^1(s_t, a_t^1, a_t^2) \Big|_{\substack{a_t^1 = \mu(s_t; \theta^1) \\ a_t^2 = \mu(s_t; \theta^2)}} \right]$$

Proof. The return, state and action-value functions are defined as:

$$\begin{aligned} R^1(\tau_t) &= \sum_{k=t}^{T-1} \gamma^{k-t} r^1(s_k, \mu(s_k; \theta^1), \mu(s_k; \theta^2)), \\ V_{\theta}^1(s_t) &= \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{k=t}^{T-1} \gamma^{k-t} r^1(s_k, \mu(s_k; \theta^1), \mu(s_k; \theta^2)) \mid s_t \right], \\ Q_{\theta}^1(s_t, \mu(s_t; \theta^1), \mu(s_t; \theta^2)) &= \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{k=t}^{T-1} \gamma^{k-t} r^1(s_k, \mu(s_k; \theta^1), \mu(s_k; \theta^2)) \mid s_t, \mu(s_t; \theta^1), \mu(s_t; \theta^2) \right] \end{aligned} \quad (39)$$

Using the relation between state and action-value function:

$$V_\theta^1(s_0) = Q_\theta^1(s_0, \mu(s_0; \theta^1), \mu(s_0; \theta^2))$$

$$\nabla_{\theta^1} V_\theta^1(s_0) = \nabla_{\theta^1} Q_\theta^1(s_0, \mu(s_0; \theta^1), \mu(s_0; \theta^2))$$

$$\nabla_{\theta^1} V_\theta^1(s_0) = \nabla_{\theta^1} \left[r^1(s_0, \mu(s_0; \theta^1), \mu(s_0; \theta^2)) + \gamma \int_{s_1} P(s_1|s_0, \mu(s_0; \theta^1), \mu(s_0; \theta^2)) V_\theta^1(s_1) ds_1 \right]$$

$$\begin{aligned} \nabla_{\theta^1} V_\theta^1(s_0) &= \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} r^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} P(s_1|s_0, a_0^1, a_0^2) V_\theta^1(s_1) ds_1 \\ &\quad + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} V_\theta^1(s_1) ds_1, \text{ where } a_0^1 = \mu(s_0; \theta^1), a_0^2 = \mu(s_0; \theta^2) \end{aligned}$$

Henceforth, we will use $a_t^1 = \mu(s_t; \theta^1)$, $a_t^2 = \mu(s_t; \theta^2)$.

$$\begin{aligned} \nabla_{\theta^1} V_\theta^1(s_0) &= \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} \left[r^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) V_\theta^1(s_1) ds_1 \right] \\ &\quad + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} V_\theta^1(s_1) ds_1 \end{aligned}$$

$$\nabla_{\theta^1} V_\theta^1(s_0) = \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} Q_\theta^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} V_\theta^1(s_1) ds_1$$

There is a clear recursion and unrolling $\nabla_{\theta^1} V_\theta^1(s_1)$ in the last term by one step we get:

$$\begin{aligned} \nabla_{\theta^1} V_\theta^1(s_0) &= \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} Q_\theta^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} \mu(s_1; \theta^1) \nabla_{a_1^1} Q_\theta^1(s_1, a_1^1, a_1^2) ds_1 \\ &\quad + \gamma^2 \int_{s_1} \int_{s_2} P(s_1|s_0, a_0^1, a_0^2) P(s_2|s_1, a_1^1, a_1^2) \nabla_{\theta^1} V_\theta^1(s_2) ds_2 ds_1 \end{aligned}$$

Multiplying both sides by $\rho(s_0)$ (initial state distribution) and integrating with respect to s_0 we get,

$$\begin{aligned} \nabla_{\theta^1} J^1 &= \int_{s_0} \rho(s_0) \nabla_{\theta^1} V_\theta^1(s_0) ds_0 = \int_{s_0} \rho(s_0) \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} Q_\theta^1(s_0, a_0^1, a_0^2) ds_0 \\ &\quad + \gamma \int_{s_0} \int_{s_1} \rho(s_0) P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} \mu(s_1; \theta^1) \nabla_{a_1^1} Q_\theta^1(s_1, a_1^1, a_1^2) ds_1 ds_0 \\ &\quad + \gamma^2 \int_{s_0} \int_{s_1} \int_{s_2} \rho(s_0) P(s_1|s_0, a_0^1, a_0^2) P(s_2|s_1, a_1^1, a_1^2) \nabla_{\theta^1} \mu(s_2; \theta^1) \nabla_{a_2^1} Q_\theta^1(s_2, a_2^1, a_2^2) ds_2 ds_1 ds_0 + \dots \end{aligned}$$

Using the definition of the probability distribution of a truncated trajectory from (21):

$$\begin{aligned} \nabla_{\theta^1} J^1 &= \int_{\tau_{0:0}} q(\tau_{0:0}) \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} Q_\theta^1(s_0, a_0^1, a_0^2) d\tau_{0:0} \\ &\quad + \gamma \int_{\tau_{0:1}} q(\tau_{0:1}) \nabla_{\theta^1} \mu(s_1; \theta^1) \nabla_{a_1^1} Q_\theta^1(s_1, a_1^1, a_1^2) d\tau_{0:1} \\ &\quad + \gamma^2 \int_{\tau_{0:2}} q(\tau_{0:2}) \nabla_{\theta^1} \mu(s_2; \theta^1) \nabla_{a_2^1} Q_\theta^1(s_2, a_2^1, a_2^2) d\tau_{0:2} + \dots \\ \nabla_{\theta^1} J^1 &= \sum_{t=0}^{T-1} \gamma^t \int_{\tau_{0:t}} q(\tau_{0:t}) \nabla_{\theta^1} \mu(s_t; \theta^1) \nabla_{a_t^1} Q_\theta^1(s_t, a_t^1, a_t^2) d\tau_{0:t} \end{aligned}$$

Writing as an expectation:

$$\nabla_{\theta^1} J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{q(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \mu(s_t; \theta^1) \nabla_{a_t^1} Q_{\theta}^1(s_t, a_t^1, a_t^2) \Big|_{\substack{a_t^1 = \mu(s_t; \theta^1), \\ a_t^2 = \mu(s_t; \theta^2)}} \right]$$

□

Derivation of mixed hessian (23)

$$D_{\theta^1 \theta^2} J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{q(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \mu(s_t; \theta^1) D_{a_t^1 a_t^2} Q_{\theta}^1(s_t, a_t^1, a_t^2) \Big|_{\substack{a_t^1 = \mu(s_t; \theta^1), \\ a_t^2 = \mu(s_t; \theta^2)}} \nabla_{\theta^2} \mu(s_t; \theta^2)^\top \right]$$

Proof. Using the relation between state and action-value functions:

$$V_{\theta}^1(s_0) = Q_{\theta}^1(s_0, \mu(s_0; \theta^1), \mu(s_0; \theta^2))$$

$$\nabla_{\theta^2} V_{\theta}^1(s_0) = \nabla_{\theta^2} Q_{\theta}^1(s_0, \mu(s_0; \theta^1), \mu(s_0; \theta^2))$$

$$\nabla_{\theta^2} V_{\theta}^1(s_0) = \nabla_{\theta^2} \left[r^1(s_0, \mu(s_0; \theta^1), \mu(s_0; \theta^2)) + \gamma \int_{s_1} P(s_1|s_0, \mu(s_0; \theta^1), \mu(s_0; \theta^2)) V_{\theta}^1(s_1) ds_1 \right]$$

$$\begin{aligned} \nabla_{\theta^2} V_{\theta}^1(s_0) &= \nabla_{\theta^2} \mu(s_0; \theta^2) \nabla_{a_0^2} r^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} \nabla_{\theta^2} \mu(s_0; \theta^2) \nabla_{a_0^2} P(s_1|s_0, a_0^1, a_0^2) V_{\theta}^1(s_1) ds_1 \\ &\quad + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^2} V_{\theta}^1(s_1) ds_1, \text{ where } a_0^1 = \mu(s_0; \theta^1), a_0^2 = \mu(s_0; \theta^2) \end{aligned}$$

$$\begin{aligned} \nabla_{\theta^2} V_{\theta}^1(s_0) &= \nabla_{\theta^2} \mu(s_0; \theta^2) \nabla_{a_0^2} \left[r^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) V_{\theta}^1(s_1) ds_1 \right] \\ &\quad + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^2} V_{\theta}^1(s_1) ds_1, \text{ where } a_0^1 = \mu(s_0; \theta^1), a_0^2 = \mu(s_0; \theta^2) \end{aligned}$$

$$\nabla_{\theta^2} V_{\theta}^1(s_0) = \nabla_{\theta^2} \mu(s_0; \theta^2) \nabla_{a_0^2} Q_{\theta}^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^2} V_{\theta}^1(s_1) ds_1 \quad (40)$$

Henceforth, we will use $a_t^1 = \mu(s_t; \theta^1)$, $a_t^2 = \mu(s_t; \theta^2)$. Taking derivative with respect to θ^1 and applying chain rule we get:

$$\begin{aligned} D_{\theta^1 \theta^2} V_{\theta}^1(s_0) &= \nabla_{\theta^1} \mu(s_0; \theta^1) D_{a_0^1 a_0^2} Q_{\theta}^1(s_0, a_0^1, a_0^2) \nabla_{\theta^2} \mu(s_0; \theta^2)^\top \\ &\quad + \gamma \int_{s_1} \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^2} V_{\theta}^1(s_1)^\top ds_1 + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) D_{\theta^1 \theta^2} V_{\theta}^1(s_1) ds_1 \end{aligned}$$

$$\begin{aligned} D_{\theta^1 \theta^2} V_{\theta}^1(s_0) &= \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} \left[\nabla_{\theta^2} \mu(s_0; \theta^2) \nabla_{a_0^2} Q_{\theta}^1(s_0, a_0^1, a_0^2) + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^2} V_{\theta}^1(s_1) ds_1 \right] \\ &\quad + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) D_{\theta^1 \theta^2} V_{\theta}^1(s_1) ds_1 \end{aligned}$$

Using (40) and $\nabla_{\theta^2} V_{\theta}^1(s_0) = \nabla_{\theta^2} Q_{\theta}^1(s_0, a_0^1, a_0^2)$,

$$D_{\theta^1 \theta^2} V_{\theta}^1(s_0) = \nabla_{\theta^1} \mu(s_0; \theta^1) \nabla_{a_0^1} [\nabla_{\theta^2} Q_{\theta}^1(s_0, a_0^1, a_0^2)] + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) D_{\theta^1 \theta^2} V_{\theta}^1(s_1) ds_1$$

$$D_{\theta^1 \theta^2} V_{\theta}^1(s_0) = \nabla_{\theta^1} \mu(s_0; \theta^1) D_{a_0^1 a_0^2} Q_{\theta}^1(s_0, a_0^1, a_0^2) \nabla_{\theta^2} \mu(s_0; \theta^2)^{\top} + \gamma \int_{s_1} P(s_1|s_0, a_0^1, a_0^2) D_{\theta^1 \theta^2} V_{\theta}^1(s_1) ds_1$$

Multiplying both sides by $\rho(s_0)$ (initial state distribution) and integrating with respect to s_0 we get,

$$\begin{aligned} D_{\theta^1 \theta^2} J^1 &= \int_{s_0} \rho(s_0) D_{\theta^1 \theta^2} V_{\theta}^1(s_0) ds_0 = \int_{s_0} \rho(s_0) \nabla_{\theta^1} \mu(s_0; \theta^1) D_{a_0^1 a_0^2} Q_{\theta}^1(s_0, a_0^1, a_0^2) \nabla_{\theta^2} \mu(s_0; \theta^2)^{\top} ds_0 \\ &+ \gamma \int_{s_0} \int_{s_1} \rho(s_0) P(s_1|s_0, a_0^1, a_0^2) \nabla_{\theta^1} \mu(s_1; \theta^1) D_{a_1^1 a_1^2} Q_{\theta}^1(s_1, a_1^1, a_1^2) \nabla_{\theta^2} \mu(s_1; \theta^2)^{\top} ds_1 ds_0 + \dots \end{aligned}$$

Using the definition of the probability distribution of a truncated trajectory from (21):

$$\begin{aligned} D_{\theta^1 \theta^2} J^1 &= \int_{\tau_{0:0}} q(\tau_{0:0}) \nabla_{\theta^1} \mu(s_0; \theta^1) D_{a_0^1 a_0^2} Q_{\theta}^1(s_0, a_0^1, a_0^2) \nabla_{\theta^2} \mu(s_0; \theta^2)^{\top} d\tau_{0:0} \\ &+ \gamma \int_{\tau_{0:1}} q(\tau_{0:1}) \nabla_{\theta^1} \mu(s_1; \theta^1) D_{a_1^1 a_1^2} Q_{\theta}^1(s_1, a_1^1, a_1^2) \nabla_{\theta^2} \mu(s_1; \theta^2)^{\top} d\tau_{0:1} \\ &+ \gamma^2 \int_{\tau_{0:2}} q(\tau_{0:2}) \nabla_{\theta^1} \mu(s_2; \theta^1) D_{a_2^1 a_2^2} Q_{\theta}^1(s_2, a_2^1, a_2^2) \nabla_{\theta^2} \mu(s_2; \theta^2)^{\top} d\tau_{0:2} + \dots \end{aligned}$$

$$D_{\theta^1 \theta^2} J^1 = \sum_{t=0}^{T-1} \gamma^t \int_{\tau_{0:t}} q(\tau_{0:t}) \nabla_{\theta^1} \mu(s_t; \theta^1) D_{a_t^1 a_t^2} Q_{\theta}^1(s_t, a_t^1, a_t^2) \nabla_{\theta^2} \mu(s_t; \theta^2)^{\top} d\tau_{0:t}$$

Writing as an expectation:

$$D_{\theta^1 \theta^2} J^1 = \sum_{t=0}^{T-1} \mathbb{E}_{q(\tau_{0:t})} \left[\gamma^t \nabla_{\theta^1} \mu(s_t; \theta^1) D_{a_t^1 a_t^2} Q_{\theta}^1(s_t, a_t^1, a_t^2) \Big|_{\substack{a_t^1 = \mu(s_t; \theta^1) \\ a_t^2 = \mu(s_t; \theta^2)}} \nabla_{\theta^2} \mu(s_t; \theta^2)^{\top} \right]$$

□