

## A PROOF OF THEOREM 1

The set of bijections  $\{o^i\}_{i \in \mathcal{N}}$  induces a one-to-one mapping between  $\Pi$  and  $\Pi_o$ , and therefore the first equality holds. For the second equality, consider an arbitrary state  $s = (s^1, \dots, s^N) \in \mathcal{S}$  and the permutation  $M$  that swaps a pair of agents  $(i, j)$ , such that  $Ms = M(\dots, s^i, \dots, s^j, \dots) = (\dots, (Ms)^i = s^j, \dots, (Ms)^j = s^i, \dots)$ . Due to the permutation invariance of the transition and reward functions by condition (ii) of Definition 1, there exists an optimal state-based joint policy  $\pi_* \in \Pi$  such that  $\pi_*^i(\cdot|s) = \pi_*^j(\cdot|Ms)$ . Consider the corresponding optimal observation-based joint policy  $\pi_{*o} \in \Pi_o$  that is the bijective mapping of  $\pi_*$ , such that  $\pi_{*o}^i(\cdot|o^i(s)) = \pi_*^i(\cdot|s)$  and  $\pi_{*o}^j(\cdot|o^j(Ms)) = \pi_*^j(\cdot|Ms)$ . We therefore have

$$\pi_{*o}^i(\cdot|o^i(s)) = \pi_{*o}^j(\cdot|o^j(Ms)). \quad (3)$$

Further, since  $\{o^i\}_{i \in \mathcal{N}}$  are permutation preserving by condition (iii) of Definition 1, we have  $o^i(s) = o^j(Ms) \in \mathcal{O}$  in Equation (3). Since Equation (3) holds for arbitrary  $s \in \mathcal{S}$  and  $i, j \in \mathcal{N}$ , and thus it follows that  $\pi_{*o}^i(\cdot|o) = \pi_{*o}^j(\cdot|o) \forall o \in \mathcal{O}$ , i.e., the second equality holds.

## B PROOF OF THEOREM 2

### B.1 ASSUMPTIONS

We make the following assumptions that are necessary to establish the convergence.

**Assumption 1.** The Markov game has finite state and action spaces and bounded rewards. Further, for any joint policy, the induced Markov chain is irreducible and aperiodic.

**Assumption 2.** The critic class is linear, i.e.,  $Q(s, a; \omega) = \phi(s, a)^\top \omega$ , where  $\phi(s, a) \in \mathbb{R}^K$  is the feature of  $(s, a)$ . Further, the feature vectors  $\phi(s, a) \in \mathbb{R}^K$  are uniformly bounded by any  $(s, a)$ . The feature matrix  $\Phi \in \mathbb{R}^{|S||A| \times K}$  has full column rank.

**Assumption 3.** The stepsizes  $\beta_{\omega,t}$  and  $\beta_{\theta,t}$  satisfy

$$\sum_t \beta_{\omega,t} = \sum_t \beta_{\theta,t} = \infty, \quad \sum_t \beta_{\omega,t}^2 = \sum_t \beta_{\theta,t}^2 < \infty, \quad \beta_{\theta,t} = o(\beta_{\omega,t}).$$

In addition,  $\lim_t \beta_{\omega,t+1} \beta_{\omega,t}^{-1} = 1$ .

**Assumption 4.** We assume the nonnegative matrices  $C_t \in \{C_{\omega,t}, C_{\theta,t}\}$  satisfy the following conditions: (i)  $C_t$  is row stochastic (i.e.,  $C_t \mathbf{1} = \mathbf{1}$ ) and  $\mathbb{E}[C_t]$  is column stochastic (i.e.,  $\mathbf{1}^\top \mathbb{E}[C_t] = \mathbf{1}^\top$ ) for all  $t > 0$ ; (ii) The spectral norm of  $\mathbb{E}[(W_t^\top (I - \frac{1}{N} \mathbf{1}\mathbf{1}^\top) W_t)]$  is strictly smaller than one; (iii)  $W_t$  and  $(s_t, \{r_t^i\})$  are conditionally independent given the  $\sigma$ -algebra generated by the random variables before time  $t$ .

**Assumption 5.** The critic update is stable, i.e.,  $\sup_t \|\omega_t^i\| < \infty$ , for all  $i$ . For the actor update,  $\{\theta_t^i\}$  belongs to a compact set for all  $i$  and  $t$ .

### B.2 CRITIC CONVERGENCE

In this subsection, we establish critic convergence under a fixed joint policy in Lemma 3. Specifically, given a fixed joint policy  $\pi = (\pi^1, \dots, \pi^N)$ , we aim to show that the critic update converges to  $\omega_\pi$ , which is the unique solution to the Mean Square Projected Bellman Error (MSPBE):

$$\omega_\pi = \arg \min_{\omega} \|\Phi \omega - \Pi T_\pi(\Phi \omega)\|_{D_\pi}^2,$$

which also satisfies

$$\Phi^\top D_\pi [T_\pi(\Phi \omega_\pi) - \Phi \omega_\pi] = 0,$$

where  $T_\pi$  is the Bellman operator for  $\pi$ ,  $\Pi$  is the projection operator for the column space of  $\Phi$ , and  $D_\pi = \text{diag}[d_\pi(s, a) : s \in \mathcal{S}, a \in \mathcal{A}]$  for the stationary distribution  $d_\pi$  induced by  $\pi$ .

**Lemma 3.** Under the assumptions, for any give joint policy  $\pi$ , with distributed critic parameters  $\omega_t^i$  generated from Equation 1 using on-policy transitions  $(s_t, a_t, r_t, s_{t+1}, a_{t+1}) \sim \pi$ , we have  $\lim_t \omega_t^i = \omega_\pi$  almost surely (a.s.) for any  $i \in \mathcal{N}$ , where  $\omega_\pi$  is the MSPBE minimizer for joint policy  $\pi$ .

*Proof.* We use the same proof techniques as Zhang et al. (2018).

Let  $\phi_t = \phi(s_t, a_t)$ ,  $\delta_t = [\delta_t^1, \dots, \delta_t^N]^\top$ , and  $\omega_t = [\omega_t^1, \dots, \omega_t^N]^\top$ . The update of  $\omega_t$  in Equation 1 can be rewritten in a compact form of  $\omega_{t+1} = (C_{\omega,t} \otimes I)(\omega_t + \beta_{\omega,t} y_t)$  where  $\otimes$  is the Kronecker product,  $I$  is the  $K \times K$  identity matrix, and  $y_t = [\delta_t^1 \phi_t^\top, \dots, \delta_t^N \phi_t^\top]^\top \in \mathbb{R}^{KN}$ . Define operator  $\langle \cdot \rangle : \mathbb{R}^{KN} \rightarrow \mathbb{R}^K$  as

$$\langle \omega \rangle = \frac{1}{N} (\mathbf{1}^\top \otimes I) \omega = \frac{1}{N} \sum_{i \in \mathcal{N}} \omega^i$$

for any  $\omega = [(\omega^1)^\top, \dots, (\omega^N)^\top]^\top \in \mathbb{R}^{KN}$  with  $\omega^i \in \mathbb{R}^K$  for any  $i \in \mathcal{N}$ . We decompose  $\omega_t$  into its agreement component  $\mathbf{1} \otimes \langle \omega_t \rangle$  and its disagreement component  $\omega_{\perp,t} := \omega_t - \mathbf{1} \otimes \langle \omega_t \rangle$ . To prove  $\omega_t = \omega_{\perp,t} + \mathbf{1} \otimes \langle \omega_t \rangle \xrightarrow{\text{a.s.}} \mathbf{1} \otimes \omega_\pi$ , we next show  $\omega_{\perp,t} \xrightarrow{\text{a.s.}} 0$  and  $\langle \omega_t \rangle \xrightarrow{\text{a.s.}} \omega_\pi$  respectively.

**Convergence of  $\omega_{\perp,t} \xrightarrow{a.s.} 0$ .** We first establish that, for any  $M > 0$ , we have

$$\sup_t \mathbb{E} \left[ \left\| \beta_{Q,t}^{-1} \omega_{\perp,t} \right\|^2 \cdot \mathbf{1}_{\{\sup_t \|\omega_t\| \leq M\}} \right] < \infty. \quad (4)$$

To show Equation 4 let  $\{\mathcal{F}_t\}$  be the filtration of  $\mathcal{F}_t = \sigma(r_{\tau-1}s_\tau, a_\tau, \omega_\tau, C_{\omega,\tau-1}; \tau \leq t)$ ,  $J = \frac{1}{N}(\mathbb{1}\mathbb{1}^\top \otimes I)$  such that  $J\omega_t = \mathbb{1} \otimes \langle \omega_t \rangle$ ,  $(I - J)\omega_t = \omega_{\perp,t}$ . The following facts about  $\otimes$  will be useful:

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \quad (5)$$

This enables us to write  $\omega_{\perp,t+1}$  as

$$\begin{aligned} \omega_{\perp,t+1} &= (I - J)\omega_{t+1} \\ &= (I - J) [(C_{\omega,t} \otimes I)(\omega_t + \beta_{\omega,t}y_t)] \\ &= (I - J) [(C_{\omega,t} \otimes I)(\mathbb{1} \otimes \langle \omega_t \rangle + \omega_{\perp,t} + \beta_{\omega,t}y_t)] \\ &\quad (\text{By Equation 5 and Assumption 4 we have } (C_{\omega,t} \otimes I)(\mathbb{1} \otimes \langle \omega_t \rangle) = (C_{\omega,t}\mathbb{1}) \otimes (I\langle \omega_t \rangle) = \mathbb{1} \otimes \langle \omega_t \rangle) \\ &= (I - J) [\mathbb{1} \otimes \langle \omega_t \rangle + (C_{\omega,t} \otimes I)(\omega_{\perp,t} + \beta_{\omega,t}y_t)] \\ &= (I - J) [(C_{\omega,t} \otimes I)(\omega_{\perp,t} + \beta_{\omega,t}y_t)] \quad ((I - J)(\mathbb{1} \otimes \langle \omega_t \rangle) = 0) \\ &= [(I - \mathbb{1}\mathbb{1}^\top/N) \otimes I] [(C_{\omega,t} \otimes I)(\omega_{\perp,t} + \beta_{\omega,t}y_t)] \quad (I - J = (I - \mathbb{1}\mathbb{1}^\top/N) \otimes I) \\ &= [(I - \mathbb{1}\mathbb{1}^\top/N)C_{\omega,t} \otimes I](\omega_{\perp,t} + \beta_{\omega,t}y_t) \quad (\text{By Equation 5}). \end{aligned}$$

We then have

$$\begin{aligned} &\mathbb{E} \left[ \left\| \beta_{Q,t+1}^{-1} \omega_{\perp,t+1} \right\|^2 \mid \mathcal{F}_t \right] \\ &= (\|x\|^2 = x^\top x, A = (I - \mathbb{1}\mathbb{1}^\top/N)C_{\omega,t} \otimes I, A^\top A = C_{\omega,t}^\top (I - \mathbb{1}\mathbb{1}^\top/N)C_{\omega,t} \otimes I) \\ &= \frac{\beta_{\omega,t}^2}{\beta_{Q,t+1}^2} \mathbb{E} \left[ (\beta_{\omega,t}^{-1} \omega_{\perp,t} + y_t)^\top (C_{\omega,t}^\top (I - \mathbb{1}\mathbb{1}^\top/N)C_{\omega,t} \otimes I) (\beta_{\omega,t}^{-1} \omega_{\perp,t} + y_t) \mid \mathcal{F}_t \right] \\ &\quad (A = C_{\omega,t}^\top (I - \mathbb{1}\mathbb{1}^\top/N)C_{\omega,t}, B = I, \|A \otimes B\| = \|A\| \|B\|) \\ &\quad (x^\top Ax = \|x^\top Ax\| \leq \|x^\top\| \|A\| \|x\| = \|A\| x^\top x) \\ &\quad (C_{\omega,t} \text{ and } (r_t, y_t) \text{ are independent conditioning on } \mathcal{F}_t) \\ &\leq \frac{\beta_{\omega,t}^2}{\beta_{Q,t+1}^2} \rho \mathbb{E} \left[ (\beta_{\omega,t}^{-1} \omega_{\perp,t} + y_t)^\top (\beta_{\omega,t}^{-1} \omega_{\perp,t} + y_t) \mid \mathcal{F}_t \right] \\ &\quad (\text{where } \rho \text{ is the spectral norm of } \mathbb{E}[C_{\omega,t}^\top (I - \mathbb{1}\mathbb{1}^\top/N)C_{\omega,t}]) \\ &= \frac{\beta_{\omega,t}^2}{\beta_{Q,t+1}^2} \rho \left( \mathbb{E} \left[ \left\| \beta_{\omega,t}^{-1} \omega_{\perp,t} \right\|^2 \mid \mathcal{F}_t \right] + 2\mathbb{E} [\langle \beta_{\omega,t}^{-1} \omega_{\perp,t}, y_t \rangle \mid \mathcal{F}_t] + \mathbb{E} [\|y_t\|^2 \mid \mathcal{F}_t] \right) \\ &\quad (\text{By Cauchy-Schwarz } |\langle u, v \rangle| \leq \|u\| \|v\|) \\ &\leq \frac{\beta_{\omega,t}^2}{\beta_{Q,t+1}^2} \rho \left( \mathbb{E} \left[ \left\| \beta_{\omega,t}^{-1} \omega_{\perp,t} \right\|^2 \mid \mathcal{F}_t \right] + 2\mathbb{E} [\left\| \beta_{\omega,t}^{-1} \omega_{\perp,t} \right\| \|y_t\| \mid \mathcal{F}_t] + \mathbb{E} [\|y_t\|^2 \mid \mathcal{F}_t] \right) \\ &\quad (\text{Quantities are deterministic given } \mathcal{F}_t) \\ &= \frac{\beta_{\omega,t}^2}{\beta_{Q,t+1}^2} \rho \left( \left\| \beta_{\omega,t}^{-1} \omega_{\perp,t} \right\|^2 + 2 \left\| \beta_{\omega,t}^{-1} \omega_{\perp,t} \right\| \|y_t\| + \|y_t\|^2 \right). \quad (6) \end{aligned}$$

Since  $\mathbb{E}[\|y_t\|^2 \mid \mathcal{F}_t] = \mathbb{E}[\sum_{i \in \mathcal{N}} \|\delta_t^i \phi_t\|^2 \mid \mathcal{F}_t]$  with  $\delta_t^i = r_t + \gamma \phi_t^\top \omega_t^i - \phi_{t+1}^\top \omega_t^i$ , by Assumptions 1 and 2 the rewards  $r_t$  and the features  $\phi_t$  are bounded, and thus we have that  $\mathbb{E}[\|y_t\|^2 \mid \mathcal{F}_t]$  is bounded on set  $\{\sup_{\tau \leq t} \|\omega_\tau\| \leq M\}$  for any given  $M > 0$ . We can then following the proof of Lemma 5.3 in Zhang et al. (2018) and its sequel to show Equation 4 and conclude the step of  $\omega_{\perp,t} \xrightarrow{a.s.} 0$ .

**Convergence of  $\langle \omega_t \rangle \xrightarrow{a.s.} \omega_\pi$ .** We write the update of  $\langle \omega_t \rangle$  as

$$\begin{aligned}\langle \omega_{t+1} \rangle &= \frac{1}{N} (\mathbb{1}^\top \otimes I) \omega_{t+1} \\ &= \frac{1}{N} (\mathbb{1}^\top \otimes I) [(C_{\omega,t} \otimes I)(\mathbb{1} \otimes \langle \omega_t \rangle + \omega_{\perp,t} + \beta_{\omega,t} y_t)] \\ &= \langle \omega_t \rangle + \beta_{\omega,t} \langle (C_{\omega,t} \otimes I)(y_t + \beta_{\omega,t}^{-1} \omega_{\perp,t}) \rangle \quad (\text{By Equation 5}).\end{aligned}$$

We rewrite the above update as

$$\begin{aligned}\langle \omega_{t+1} \rangle &= \langle \omega_t \rangle + \beta_{\omega,t} \mathbb{E}[\langle \delta_t \rangle \phi_t \mid \mathcal{F}_t] + \beta_{\omega,t} \xi_t \\ \text{where} \quad \xi_t &= \langle (C_{\omega,t} \otimes I)(y_t + \beta_{\omega,t}^{-1} \omega_{\perp,t}) \rangle - \mathbb{E}[\langle \delta_t \rangle \phi_t \mid \mathcal{F}_t]\end{aligned} \quad (7)$$

We can verify that the following conditions hold (with probability 1) regarding the update of  $\langle \omega_t \rangle$  in Equation 7:

1.  $\mathbb{E}[\langle \delta_t \rangle \phi_t \mid \mathcal{F}_t]$  is Lipschitz continuous in  $\langle \omega_t \rangle$ ,
2.  $\xi_t$  is a martingale difference sequence and satisfies  $\mathbb{E}[\|\xi_{t+1}\|^2 \mid \mathcal{F}_t] \leq K(1 + \|\omega_t\|^2)$  for some constant  $K$ ,

such that the conditions in Assumption B.1 of Zhang et al. (2018) are satisfied (with probability 1) and the behavior of Equation 7 is related to its corresponding ODE (see Theorem B.2 in Zhang et al. (2018)):

$$\begin{aligned}\dot{\langle \omega \rangle} &= \sum_{s,a} d_\pi(s,a) \mathbb{E}[\langle \delta \rangle \phi \mid s,a] \\ &= \sum_{s,a} d_\pi(s,a) \mathbb{E}_{s',a'} [(r(s,a) + \gamma \phi^\top(s,a) \langle \omega \rangle - \phi^\top(s,a) \langle \omega \rangle) \phi(s,a) \mid s,a] \\ &= \Phi^\top D_\pi (\gamma P^\pi - I) \Phi \langle \omega \rangle + \Phi^\top D_\pi R\end{aligned}$$

Note that  $(\gamma P^\pi - I)$  has all eigenvalues with negative real parts, so does  $(\Phi^\top D_\pi (\gamma P^\pi - I) \Phi)$  since  $\Phi$  is assumed to be full column rank. Hence, the ODE is globally asymptotically stable, with its equilibrium satisfying

$$\Phi^\top D_\pi [R + (\gamma P^\pi - I) \Phi \langle \omega \rangle] = 0,$$

which is the MSPBE minimizer, i.e.,  $\langle \omega \rangle = \omega_\pi$ . This concludes the step of  $\langle \omega_t \rangle \xrightarrow{a.s.} \omega_\pi$  and the proof of Lemma 3.  $\square$

### B.3 ACTOR CONVERGENCE

In this subsection, we establish the convergence of actor update with critic parameters  $\omega_t^i$  in Equation 2 replaced with the critic convergence point established in Lemma 3. Then, by the two-timescale nature of the algorithm, we establish the convergence of  $\{\omega_t^i\}$  and  $\{\theta_t^i\}$  generated by Equation 1 Equation 2.

Let  $\theta = [(\theta^1)^\top, \dots, (\theta^N)^\top]^\top$  and  $\omega_\theta$  be the critic convergence point for joint policy parameterized by  $\theta$  as established in Lemma 3. Define

$$A_{t,\theta}^i = Q(s_t, a_t; \omega_\theta) \quad \psi_{t,\theta}^i = \nabla_{\theta^i} \log \pi^i(a_t^i | o^i(s_t); \theta^i)$$

for an arbitrary  $\theta$ . We study the variant of Equation 2 where  $\omega_t^i$  is replaced by  $\omega_{\theta_t}$ :

$$\begin{aligned}A_{t,\theta_t}^i &= Q(s_t, a_t; \omega_{\theta_t}) \quad \psi_{t,\theta_t}^i = \nabla_{\theta^i} \log \pi^i(a_t^i | o^i(s_t); \theta_t^i) \\ \tilde{\theta}_{t+1}^i &= \theta_t^i + \beta_{\theta,t} \cdot A_{t,\theta_t}^i \cdot \psi_{t,\theta_t}^i \\ \theta_{t+1}^i &= \sum_{j \in \mathcal{N}} c_{\theta,t}(i,j) \cdot \tilde{\theta}_t^j\end{aligned} \quad (8)$$

which can be rewritten as

$$\theta_{t+1} = (C_{\theta,t} \otimes I)(\theta_t + \beta_{\theta,t} y_{t,\theta_t})$$

where  $y_{t,\theta_t} = [(A_{t,\theta_t}^1 \cdot \psi_{t,\theta_t}^1)^\top, \dots, (A_{t,\theta_t}^N \cdot \psi_{t,\theta_t}^N)^\top]^\top$ .

Similar to the critic convergence, we make the decomposition  $\theta_t = \theta_{\perp,t} + \mathbb{1} \otimes \langle \theta_t \rangle$  and then show  $\theta_{\perp,t} \xrightarrow{a.s.} 0$  and convergence of  $\langle \theta_t \rangle$  respectively.

**Convergence of  $\theta_{\perp,t} \xrightarrow{a.s.} 0$ .** In light of the argument for  $\omega_{\perp,t} \xrightarrow{a.s.} 0$  in the proof of Lemma 3, it suffices to show that the boundedness of  $y_{t,\theta_t}$ . Here,  $y_{t,\theta_t} = [(A_{t,\theta_t}^1 \cdot \psi_{t,\theta_t}^1)^\top, \dots, (A_{t,\theta_t}^N \cdot \psi_{t,\theta_t}^N)^\top]^\top$  is bounded because 1)  $A_{t,\theta_t}^i = Q(s_t, a_t; \omega_{\theta_t})$  is bounded since  $\omega_{\theta_t}$  is the MSPBE minimizer; (2)  $\psi_{t,\theta_t}^i$  is bounded since by Assumption 5 it is a continuous function over a compact set.

**Convergence of  $\langle \theta_t \rangle$ .** We write the update of  $\langle \theta_t \rangle$  in Equation 8 as

$$\begin{aligned} \langle \theta_{t+1} \rangle &= \frac{1}{N} (\mathbb{1}^\top \otimes I) \theta_{t+1} \\ &= \frac{1}{N} (\mathbb{1}^\top \otimes I) [(C_{\theta,t} \otimes I)(\mathbb{1} \otimes \langle \theta_t \rangle + \theta_{\perp,t} + \beta_{\theta,t} y_{t,\theta_t})] \\ &= \langle \theta_t \rangle + \beta_{\theta,t} \langle (C_{\theta,t} \otimes I)(y_{t,\theta_t} + \beta_{\theta,t}^{-1} \theta_{\perp,t}) \rangle \quad (\text{By Equation 5}). \end{aligned}$$

We rewrite the above update as

$$\begin{aligned} \langle \theta_{t+1} \rangle &= \langle \theta_t \rangle + \beta_{\theta,t} \mathbb{E}_{s_t \sim d_{\langle \theta_t \rangle}, a_t \sim \pi_{\langle \theta_t \rangle}} [\langle y_{t,\theta_t} \rangle | \mathcal{F}_t] + \beta_{\theta,t} \xi_t \\ \text{where} \quad \xi_t &= \langle (C_{\theta,t} \otimes I)(y_{t,\theta_t} + \beta_{\theta,t}^{-1} \theta_{\perp,t}) \rangle - \mathbb{E}_{s_t \sim d_{\langle \theta_t \rangle}, a_t \sim \pi_{\langle \theta_t \rangle}} [\langle y_{t,\theta_t} \rangle | \mathcal{F}_t] \end{aligned}$$

where  $\mathcal{F}_t = \sigma(\theta_\tau, \tau \leq t)$ ,  $\pi_{\langle \theta_t \rangle}$  is the *joint* policy where each individual policy is parameterized by  $\langle \theta_t \rangle$ . Note that  $\xi_t$  is a martingale difference sequence. By Assumption 5  $\xi_t$  is bounded and further by Assumption 3 we have  $\sum_t \|\beta_{\theta,t} \xi_t\|^2 < \infty$ . By arguments in the proof of Theorem 4.7 in Zhang et al. (2018), we can apply Kushner-Clark lemma and conclude that  $\langle \theta_t \rangle$  converges almost sure to a point in the set of asymptotically stable equilibria of

$$\langle \dot{\theta} \rangle = \mathbb{E}_{s_t \sim d_{\langle \theta \rangle}, a_t \sim \pi_{\langle \theta \rangle}} [\langle y_{t,\langle \theta \rangle} \rangle] = \mathbb{E}_{s_t \sim d_{\langle \theta \rangle}, a_t \sim \pi_{\langle \theta \rangle}} \left[ \sum_i A_{t,\langle \theta \rangle}^i \cdot \psi_{t,\langle \theta \rangle}^i \right].$$

## C VISUALIZATION OF THE LEARNED COMMUNICATION RULE

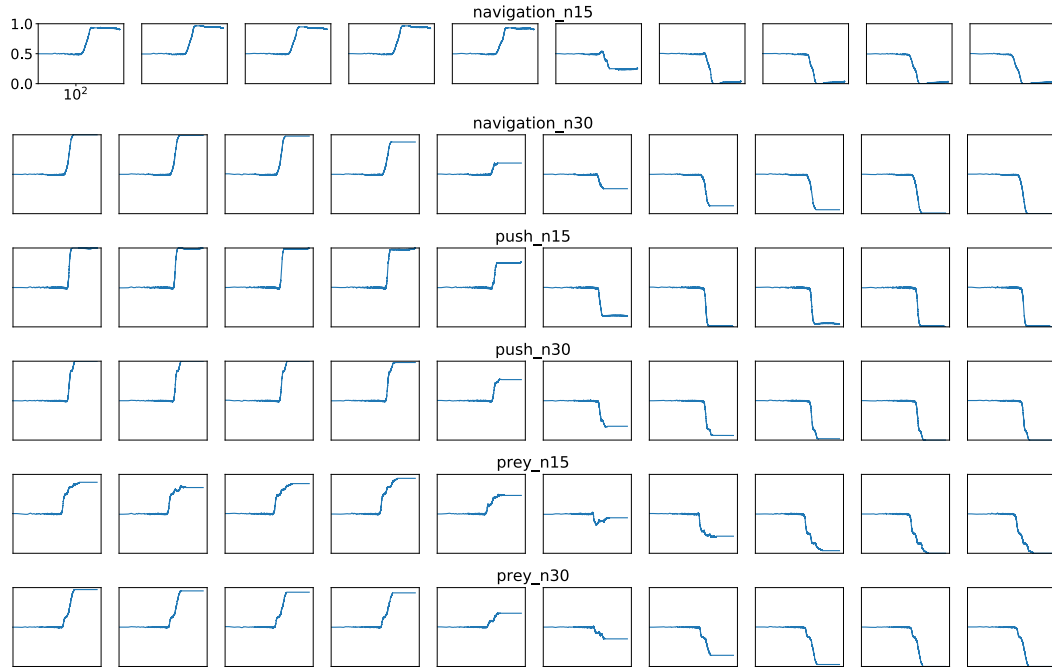


Figure 7: Y-axis: Avg communication rate. X-axis: training step in log scale. The average communication rate for detectable 10 nearby agents, with the order increasing in distance from left to right.