Supplementary Materials

Wei-Fang Sun[1,2] Cheng-Kuang Lee[2] Chun-Yi Lee[1]
[1]Department of Computer Science, National Tsing Hua University, Taiwan
[2]NVIDIA AI Technology Center, NVIDIA Corporation

S1 Theorems and Proofs

In this section, we elaborate on the definitions, and provide the proofs of the theorems discussed in the main manuscript.

Proposition 1. *Monotonicity for utility distributions:*

$$\begin{aligned} Z_{\text{jt}}(\mathbf{h}, \mathbf{u}) &= \Psi(Z_1(h_1, u_1), ..., Z_{\text{K}}(h_{\text{K}}, u_{\text{K}})|s) \\ &= M(Z_1(h_1, u_1), ..., Z_{\text{K}}(h_{\text{K}}, u_{\text{K}})|s), \end{aligned}$$

where M is a monotonic transformation that satisfies $\frac{\partial M}{\partial Q_k} \geq 0, \forall k \in \mathbb{K}$, is not a sufficient condition for DIGM, although the equality may hold for special cases of M and $[Z_k(h_k, u_k)]_{k \in \mathbb{K}}$.

Proof. We consider a degenerated case and prove the theorem by contradiction. Consider a case where there is only a single agent ($\text{K} = 1$), with a single fully observable state and an exponential transformation $M(Z_1(h_1, u_1)|s) = \exp(Z_1(h_1, u_1))$. The (joint) action space of this case consists of two (joint) actions: $\mathbb{U}_{\text{jt}} = \mathbb{U}_1 = \{u_1^*, u_1'\}$, where u_1^* is the optimal action (with expected return 2) and u_1' is the suboptimal action (with expected return 1.5). We define the probability mass function (PMF) of $Z_1(h_1, u_1^*)$ to be:

$$p(z) = \begin{cases} 1 & \text{if } z = 2 \\ 0 & \text{otherwise,} \end{cases}$$

and the PMF of $Z_1(h_1, u_1')$ to be:

$$p(z) = \begin{cases} 0.5 & \text{if } z = 0 \\ 0.5 & \text{if } z = 3 \\ 0 & \text{otherwise.} \end{cases}$$

By the definition above, we can calculate the followings:

$$\begin{cases} \mathbb{E}[Z_1(h_1, u_1^*)] = 1 \cdot 2 = 2 \\ \mathbb{E}[Z_1(h_1, u_1')] = 0.5 \cdot 0 + 0.5 \cdot 3 = 1.5 \\ \arg\max_{u_1} \mathbb{E}[Z_1(h_1, u_1)] = u_1^* \end{cases}$$

$$\begin{cases} \mathbb{E}[M(Z_1(h_1, u_1^*)|s)] = \mathbb{E}[\exp(Z_1(h_1, u_1^*))] = e^2 \approx 7.39 \\ \mathbb{E}[M(Z_1(h_1, u_1')|s)] = \mathbb{E}[\exp(Z_1(h_1, u_1'))] = 0.5 \cdot e^0 + 0.5 \cdot e^3 \approx 10.54 \\ \arg\max_{u_1} \mathbb{E}[M(Z_1(h_1, u_1)|s)] = u_1' \end{cases}$$

Assume, to the contrary, that *Monotonicity* for utility distributions is a sufficient condition for DIGM. By the definition of DIGM:

$$\begin{aligned}
&\arg\max_{\mathbf{u}} \mathbb{E}[Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u})] = \left(\arg\max_{u_1} \mathbb{E}[Z_1(h_1,u_1)]\right) \\
\Rightarrow\ &\arg\max_{u_1} \mathbb{E}[M(Z_1(h_1,u_1)|s)] = \arg\max_{u_1} \mathbb{E}[Z_1(h_1,u_1)] \\
\Rightarrow\ & u_1' = u_1^* \ (\Rightarrow\Leftarrow \text{contradiction}).
\end{aligned}$$

A contradiction occurs since $u_1' \neq u_1^*$, showing that *Monotonicity* is not a sufficient condition for DIGM. Since there exist a case where DIGM does not hold for $\mathrm{K}=1$, it certainly does not hold for all $\mathrm{K} \in \mathbb{Z}$.

□

Theorem 1. *Given a deterministic joint action-value function* Q_{jt}*, a stochastic joint action-value function* Z_{jt}*, and a factorization function* Ψ *for deterministic utilities:*

$$Q_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) = \Psi(Q_1(h_1,u_1),...,Q_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}})|s),$$

such that $[Q_k]_{k\in\mathbb{K}}$ *satisfy IGM for* Q_{jt} *under* $\mathbf{h}$*, the following distributional factorization:*

$$Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) = \Psi(Z_1(h_1,u_1),...,Z_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}})|s).$$

is insufficient to guarantee that $[Z_k]_{k\in\mathbb{K}}$ *satisfy DIGM for* Z_{jt} *under* $\mathbf{h}$*.*

Proof. A contradiction is provided by Proposition 1.

□

Theorem 2 (DFAC Theorem)**.** *Given a deterministic joint action-value function* Q_{jt}*, a stochastic joint action-value function* Z_{jt}*, and a factorization function* Ψ *for deterministic utilities:*

$$Q_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) = \Psi(Q_1(h_1,u_1),...,Q_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}})|s),$$

such that $[Q_k]_{k\in\mathbb{K}}$ *satisfy IGM for* Q_{jt} *under* $\mathbf{h}$*, by Mean-Shape Decomposition, the following distributional factorization:*

$$\begin{aligned}
Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) &= \mathbb{E}[Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u})] + (Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) - \mathbb{E}[Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u})]) \\
&= Z_{\mathrm{mean}}(\mathbf{h},\mathbf{u}) + Z_{\mathrm{shape}}(\mathbf{h},\mathbf{u}) \\
&= \Psi(Q_1(h_1,u_1),...,Q_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}})|s) \\
&+ \Phi(Z_1(h_1,u_1),...,Z_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}})|s).
\end{aligned}$$

is sufficient to guarantee that $[Z_k]_{k\in\mathbb{K}}$ *satisfy DIGM for* Z_{jt} *under* $\mathbf{h}$*, where* $\mathrm{Var}(\Psi)=0$ *and* $\mathbb{E}[\Phi]=0$*.*

Proof. By mean-shape decomposition:

$$
\begin{aligned}
& \arg\max_{\mathbf{u}} \{\mathbb{E}[Z_{\text{jt}}(\mathbf{h}, \mathbf{u})]\} \\
= & \arg\max_{\mathbf{u}} \{\mathbb{E}[Z_{\text{mean}}(\mathbf{h}, \mathbf{u}) + Z_{\text{shape}}(\mathbf{h}, \mathbf{u})]\} \\
= & \arg\max_{\mathbf{u}} \{\mathbb{E}[Z_{\text{mean}}(\mathbf{h}, \mathbf{u})] + \mathbb{E}[Z_{\text{shape}}(\mathbf{h}, \mathbf{u})]\} \\
= & \arg\max_{\mathbf{u}} \{\mathbb{E}[\Psi(Q_1(h_1, u_1), ..., Q_{\text{K}}(h_{\text{K}}, u_{\text{K}})|s)] \\
& + \mathbb{E}[\Phi(Z_1(h_1, u_1), ..., Z_{\text{K}}(h_{\text{K}}, u_{\text{K}})|s)]\} \\
= & \arg\max_{\mathbf{u}} \{\Psi(Q_1(h_1, u_1), ..., Q_{\text{K}}(h_{\text{K}}, u_{\text{K}})|s) + 0\} \\
= & \arg\max_{\mathbf{u}} \{\Psi(Q_1(h_1, u_1), ..., Q_{\text{K}}(h_{\text{K}}, u_{\text{K}})|s)\} \\
= & \begin{pmatrix} \arg\max_{u_1} Q_1(h_1, u_1) \\ \vdots \\ \arg\max_{u_{\text{K}}} Q_{\text{K}}(h_{\text{K}}, u_{\text{K}}) \end{pmatrix} \\
\Rightarrow & \arg\max_{\mathbf{u}} \mathbb{E}[Z_{\text{jt}}(\mathbf{h}, \mathbf{u})] = \begin{pmatrix} \arg\max_{u_1} \mathbb{E}[Z_1(h_1, u_1)] \\ \vdots \\ \arg\max_{u_{\text{K}}} \mathbb{E}[Z_{\text{K}}(h_{\text{K}}, u_{\text{K}})] \end{pmatrix}.
\end{aligned}
$$

The equations above show that $[Z_k]_{k \in \mathbb{K}}$ satisfy ***DIGM*** for Z_{jt} under $\mathbf{h}$.

□

Theorem 3. *Given a quantile mixture:*

$$F^{-1}(\omega) = \sum_{k=1}^{\text{K}} \beta_k F_k^{-1}(\omega)$$

with K *components* $[F_k^{-1}]_{k \in \mathbb{K}}$ *and non-negative model parameters* $[\beta_k]_{k \in \mathbb{K}}$*. There exist a set of random variables* $Z = F^{-1}(\tau)$ *and* $[Z_k = F_k^{-1}(\tau)]_{k \in \mathbb{K}}$ *corresponding to the quantile functions* F^{-1} *and* $[F_k^{-1}]_{k \in \mathbb{K}}$*, respectively, where* τ *is a random variable uniformly distributed on* $[0, 1]$*, with the following relationship:*

$$Z \overset{D}{=} \sum_{k \in \mathbb{K}} \beta_k Z_k.$$

Proof. We first prove the case for a quantile mixture with $\text{K} = 2$ components, and then generalize it to all $\text{K} \in \mathbb{Z}$. For $\text{K} = 2$, the quantile mixture is simplified as follows:

$$F^{-1}(\tau) = \beta_1 F_1^{-1}(\tau) + \beta_2 F_2^{-1}(\tau)$$

For notational simplicity, let $X = \beta_1 Z_1$, $Y = \beta_2 Z_2$, and τ is a latent variable shared among the random variables X, Y, and Z. The corresponding CDFs of the random variables X, Y, and Z are F_X, F_Y, and F_Z, respectively, with $X(\tau) = F_X^{-1}(\tau)$, $Y(\tau) = F_Y^{-1}(\tau)$, and $Z(\tau) = F_Z^{-1}(\tau)$. Under this notation, the above equation can be re-written as:

$$F_Z^{-1}(\tau) = F_X^{-1}(\tau) + F_Y^{-1}(\tau).$$

The goal is to prove that there exist random variables (X, Y, Z) such that the following holds:

$$Z \stackrel{D}{=} X + Y$$

By the definition of the CDF of $X + Y$, the following holds:

$$\begin{aligned}
& F_{X+Y}(z), \forall z \in \mathbb{R} \\
&= \Pr(X + Y \leq z), \forall z \in \mathbb{R} \\
&= \Pr(\{\tau \in [0,1] : X(\tau) + Y(\tau) \leq z\}), \forall z \in \mathbb{R} \\
&= \Pr(\{\tau \in [0,1] : F_X^{-1}(\tau) + F_Y^{-1}(\tau) \leq z\}), \forall z \in \mathbb{R} \\
&= \sup\{\tau \in [0,1] : F_X^{-1}(\tau) + F_Y^{-1}(\tau) \leq z\}, \forall z \in \mathbb{R} \\
&= \inf\{\tau \in [0,1] : z \leq F_X^{-1}(\tau) + F_Y^{-1}(\tau)\}, \forall z \in \mathbb{R} \\
&= \inf\{\tau \in [0,1] : z \leq F_Z^{-1}(\tau)\}, \forall z \in \mathbb{R} \\
&= F_Z(z), \forall z \in \mathbb{R}. \\
\Rightarrow & Z \stackrel{D}{=} X + Y.
\end{aligned}$$

The proof for quantile mixtures with two components can be iteratively applied to quantile mixtures with $\mathrm{K} \in \mathbb{Z}$ components.

□

S2 Hyperparameters and Settings

S2.1 Stochastic Two Step Game

In the stochastic two step game described in Section 4, each agent is implemented as an IQN with two hidden layers comprised of 64 units and 512 units, respectively, with a ReLU nonlinearity at the end of each layer. We optimize the IQNs with $\mathrm{N} = \mathrm{N}' = 32$ quantile samples, where each of them is encoded into a 64-dimensional intermediate embedding and projected to a 512-dimensional quantile embedding by a single hidden layer. Each agent performs independent ϵ-greedy action selection, with full exploration (i.e., $\epsilon = 1$). We set the discount factor γ to 0.99. The replay buffer contains the state-action pairs of the latest $2k$ episodes, from which we uniformly sample a batch of size 512 for training. The target network is updated every 100 episodes. The optimizer is set to Adam, in which its learning rate is set to 1×10^{-4}. We train for $20k$ timesteps ($10k$ episodes). All agent networks share parameters, and the one-hot encoded agent id ($[1\ 0]^T$ for agent 1 and $[0\ 1]^T$ for agent 2) is concatenated to each agent's observation. We do not pass the previous actions taken by the agents as their inputs. Each agent receives the full state as its input. For *DMIX*, we use a mixing network with 8 units.

Each state is one-hot encoded. The starting state for the first timestep is State 1 (one-hot: $[1\ 0\ 0]^T$). At State 1, if Agent 1 selects Action A, the agents transit to State 2A (one-hot: $[0\ 1\ 0]^T$). On the other hand, if agent 1 selects Action B, the agents transit to State 2B (one-hot: $[0\ 0\ 1]^T$).

S2.2 SMAC

We tuned the hyperparameters of both the baselines and their distributional variants by selecting their hidden layer sizes from $\{32, 64, 128, 256, 512\}$ and choose the best ones. The quantile samples of *DIQL* and *DDN*

Table S1: A summary of the optimal hidden state sizes of the baseline methods and their distributional variants.

Maps	IQL	VDN	QMIX	QR-MIX	DIQL	DDN	DMIX
3s5z_vs_3s6z	512	128	128	128	256	512	256
6h_vs_8z	128	128	256	256	512	512	256
MMM2	256	64	64	128	512	512	256
27m_vs_30m	256	64	64	64	512	128	128
corridor	256	128	256	64	512	128	64

Table S2: The detailed settings of the *Super Hard* maps.

Difficulty	Map	Player's Team	Enemy's Team
Super Hard	6h_vs_8z	6 Hydralisks	8 Zealots
	3s5z_vs_3s6z	3 Stalkers & 5 Zealots	3 Stalkers & 6 Zealots
	MMM2	7 Marines, 2 Marauders & 1 Medivac	8 Marines, 3 Marauders & 1 Medivac
	27m_vs_30m	27 Marines	30 Marines
	corridor	6 Zealots	24 Zerglings

are simply set to $\mathrm{N} = \mathrm{N}' = 1$, since they do not require the calculation of the expected value during the optimization process. As for *DMIX*, the numbers of quantile samples are set to $\mathrm{N} = \mathrm{N}' = 8$ as in [1]. The optimizers follow those used in DQN and IQN. All of the other hyperparameters follow those used in SMAC. Table S1 lists the hyperparameters adopted for the baselines and their distributional variants. The StarCraft version we used is 4.10.

References

[1] W. Dabney, G. Ostrovski, D. Silver, and R. Munos. Implicit quantile networks for distributional reinforcement learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 1096–1105, Jul. 2018.