# Multiagent Reinforcement Learning (MARL) frameworks for Peer-to-Peer Energy Trading with Voltage Control

Feng Chen and Andrew L. Liu

School of Industrial Engineering, Purdue University
West Lafayette, IN 47907
andrewliu@purdue.edu

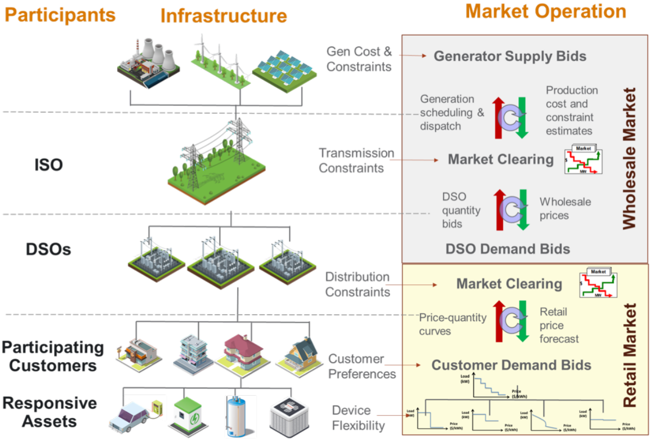
September 8th, 2023

# Outline

- Background and motivation: Issues with P2P energy trading
- Compare three MARL algorithms: PPO, MADDPG, EPG-Concensus
- Numerical results

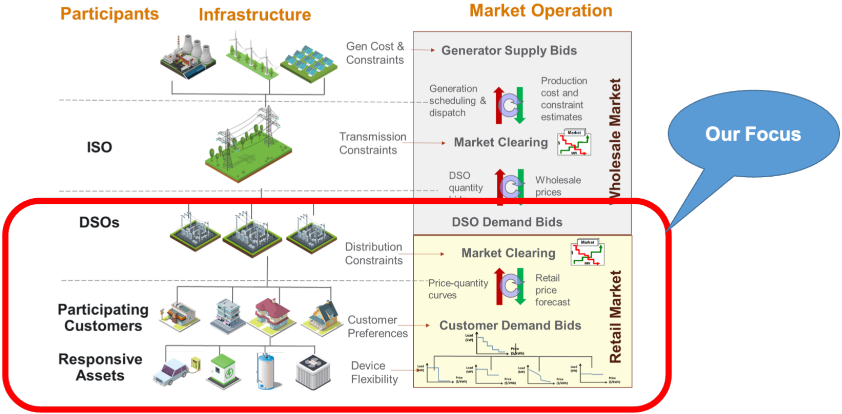
# Part I – Motivation

# Transactive Energy (PNNL's Vision)




Source: S. Widergren et al., DSO+T: Transactive Energy Coordination Framework Volume 3, PNNL-32170-3, January 2022.

# Transactive Energy (PNNL's Vision)




Source: S. Widergren et al., DSO+T: Transactive Energy Coordination Framework Volume 3, PNNL-32170-3, January 2022.

# Conceptual Models of TSO-DSO Coordination

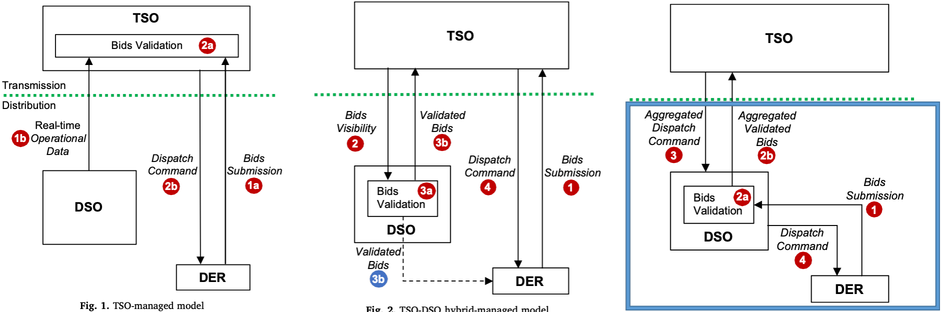



**Fig. 1.** TSO-managed model

**Fig. 2.** TSO-DSO hybrid-managed model

**Fig. 3.** DSO-managed model

Source: A. G. Givisez, K. Petrou and L. F. Ochoa, A Review on TSO-DSO Coordination Models and Solution Techniques. Electric Power Systems Research, 189 (2020) 106659

# Utilizing DERs: Four Approaches

- Direct load control (DER aggregation)
- DSO-operated wholesale-style market – DLMP
- Price-based control (between DSO and DERs)

# Utilizing DERs: Four Approaches

- Direct load control (DER aggregation)
- DSO-operated wholesale-style market – DLMP
- Price-based control (between DSO and DERs)
- Peer-to-peer trading (among DERs and consumers) **over shared networks** – Our focus

# Utilizing DERs: Four Approaches

- Direct load control (DER aggregation)
- DSO-operated wholesale-style market – DLMP
- Price-based control (between DSO and DERs)
- Peer-to-peer trading (among DERs and consumers) **over shared networks** – Our focus
  - Continuous-time trading: continuous double-auction
  - Discrete-time trading (by rounds, $x$-hour ahead) – This work

# A Conceptual Peer-to-Peer Retail (Local) Energy Market


Source: `https://100percentrenewables.com.au/peer-to-peer-energy-trading/`

# Potential Issues of P2P Energy Trading

- Consumers/prosumers do not have the expertise, nor the time to bid, say, every hour

# Potential Issues of P2P Energy Trading

- Consumers/prosumers do not have the expertise, nor the time to bid, say, every hour – Solution: control automation

# Potential Issues of P2P Energy Trading

- Consumers/prosumers do not have the expertise, nor the time to bid, say, every hour – Solution: control automation

- A wholesale-market-like uniform price auction will NOT work:
  - All zero-marginal resources

# Potential Issues of P2P Energy Trading

- Consumers/prosumers do not have the expertise, nor the time to bid, say, every hour – Solution: control automation

- A wholesale-market-like uniform price auction will NOT work:
  - All zero-marginal resources
  - Consumers/prosumers do not know there own valuation of energy consumption/generation (due to zero marginal cost)

# Potential Issues of P2P Energy Trading

- Consumers/prosumers do not have the expertise, nor the time to bid, say, every hour – Solution: control automation

- A wholesale-market-like uniform price auction will NOT work:
  - All zero-marginal resources
  - Consumers/prosumers do not know there own valuation of energy consumption/generation (due to zero marginal cost)
  - Uncleared demand in a P2P market need to buy from utility/DSO at the utility rate (UR); uncleared energy from DERs need to sell to utility/DSO at feed-in tariff (FIT) (UR > FIT)

# Potential Issues of P2P Energy Trading

- Consumers/prosumers do not have the expertise, nor the time to bid, say, every hour – Solution: control automation

- A wholesale-market-like uniform price auction will NOT work:
  - All zero-marginal resources
  - Consumers/prosumers do not know there own valuation of energy consumption/generation (due to zero marginal cost)
  - Uncleared demand in a P2P market need to buy from utility/DSO at the utility rate (UR); uncleared energy from DERs need to sell to utility/DSO at feed-in tariff (FIT) (UR $>$ FIT) – UR and FIT are then de facto reserve prices of P2P trading, which are publicly known! $\implies$ Any double-auction design will lead to bang-bang outcomes (unless supply$_t$ $=$ demand$_t$). [Zhao et al., 2022]

# Potential Issues of P2P Energy Trading

- Consumers/prosumers do not have the expertise, nor the time to bid, say, every hour – Solution: control automation

- A wholesale-market-like uniform price auction will NOT work:
  - All zero-marginal resources
  - Consumers/prosumers do not know there own valuation of energy consumption/generation (due to zero marginal cost)
  - Uncleared demand in a P2P market need to buy from utility/DSO at the utility rate (UR); uncleared energy from DERs need to sell to utility/DSO at feed-in tariff (FIT) (UR $>$ FIT) – UR and FIT are then de facto reserve prices of P2P trading, which are publicly known! $\Longrightarrow$ Any double-auction design will lead to bang-bang outcomes (unless supply$_t$ $=$ demand$_t$). [Zhao et al., 2022]

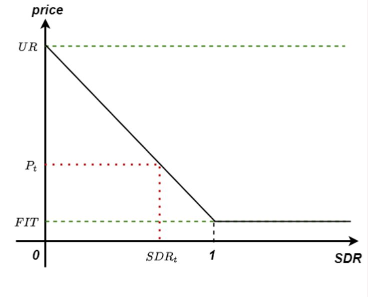
- P2P tradings only financial transactions; how to deal with shared network constraints – Solution: Add (fake) financial penalties for constraint violation in learning algorithms

# Alternative Market Clearing Mechanism SDR [Liu et al., 2017]

## Supply-Demand Ratio

Let $b_{i,t}$ be bid/ask of agent $i$ at time $t$: $b_{i,t} > 0$ (sell); $b_{i,t} < 0$ (buy). The supply-demand ratio (SDR):

$$SDR_t := \frac{\sum_{i \in \mathcal{S}_t} b_{i,t}}{-\sum_{i \in \mathcal{B}_t} b_{i,t}}.$$

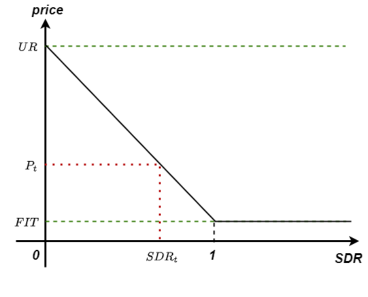


# Alternative Market Clearing Mechanism SDR [Liu et al., 2017]

## Supply-Demand Ratio

Let $b_{i,t}$ be bid/ask of agent $i$ at time $t$: $b_{i,t} > 0$ (sell); $b_{i,t} < 0$ (buy). The supply-demand ratio (SDR):

$$SDR_t := \frac{\sum_{i \in \mathcal{S}_t} b_{i,t}}{-\sum_{i \in \mathcal{B}_t} b_{i,t}}.$$




## Market Clearing Price under SDR

$$P_t := P(SDR_t) := \begin{cases} (FIT - UR) \cdot SDR_t + UR, & 0 \leq SDR_t \leq 1 \\ FIT, & SDR_t > 1. \end{cases}$$

# Part II – MARL Framework

# Single-agent (Agent $i$'s) RL Problem

**State Variables (in continuous space)**

$s_{i,t} := (d^p_{i,t}, d^q_{i,t}, v_{i,t}, e_{i,t}, PV_{i,t}) \in \mathcal{S}_\rangle$ – (baseload real power, baseload reactive power, voltage magnitude, battery state of charge, PV (real power) generation)

# Single-agent (Agent $i$'s) RL Problem

## State Variables (in continuous space)

$s_{i,t} := (d^p_{i,t}, d^q_{i,t}, v_{i,t}, e_{i,t}, PV_{i,t}) \in \mathcal{S}_\rangle$ – (baseload real power, baseload reactive power, voltage magnitude, battery state of charge, PV (real power) generation)

## Action (in continuous space)

$a_{i,t} := (a^q_{i,t}, a^e_{i,t}) \in \mathcal{A}_i = \mathcal{A}^q_i \times \mathcal{A}^e_i$ – (reactive power injection/withdraw, energy charge/discharge)

# Single-agent (Agent $i$'s) RL Problem

## State Variables (in continuous space)

$s_{i,t} := (d^p_{i,t}, d^q_{i,t}, v_{i,t}, e_{i,t}, PV_{i,t}) \in \mathcal{S}_\rangle$ – (baseload real power, baseload reactive power, voltage magnitude, battery state of charge, PV (real power) generation)

## Action (in continuous space)

$a_{i,t} := (a^q_{i,t}, a^e_{i,t}) \in \mathcal{A}_i = \mathcal{A}^q_i \times \mathcal{A}^e_i$ – (reactive power injection/withdraw, energy charge/discharge) (Underlying assumption: PV/battery connected to a smart inverter: can set reactive power setpoints within a range)

# Single-agent (Agent $i$'s) RL Problem

## State Variables (in continuous space)

$s_{i,t} := (d^p_{i,t}, d^q_{i,t}, v_{i,t}, e_{i,t}, PV_{i,t}) \in \mathcal{S}_{\rangle}$ – (baseload real power, baseload reactive power, voltage magnitude, battery state of charge, PV (real power) generation)

## Action (in continuous space)

$a_{i,t} := (a^q_{i,t}, a^e_{i,t}) \in \mathcal{A}_i = \mathcal{A}^q_i \times \mathcal{A}^e_i$ – (reactive power injection/withdraw, energy charge/discharge) (Underlying assumption: PV/battery connected to a smart inverter: can set reactive power setpoints within a range)

The actual bids = net energy of PV generation minus baseload demand (of real power) and charge/discharge to the battery:

$$b_{i,t} = \begin{cases} PV_{i,t} - d^p_{i,t} - \min(a^e_{i,t}, \frac{\overline{e}_i - e_{i,t}}{\eta^c_i}), & \text{if } a^e_{i,t} \geq 0, \\ PV_{i,t} - d^p_{i,t} - \max(a^e_{i,t}, -e_{i,t} \cdot \eta^d_i), & \text{otherwise,} \end{cases}$$

where $\eta^c_i$ and $\eta^d_i$ are the charging and discharging efficiency of agent $i$'s battery, resp., and $\overline{e}_i$ is the battery capacity.

# State Transition and Reward Function

**Battery state of charge ($e_{i,t}$)**

$$e_{i,t+1} := E_i(e_{i,t}, a^e_{i,t}) := \max\left\{ \min\left[ e_{i,t} + \eta_i^c \max(a^e_{i,t}, 0) + \frac{1}{\eta_i^d} \min(a^e_{i,t}, 0), \overline{e}_i \right], 0 \right\},$$

# State Transition and Reward Function

## Battery state of charge ($e_{i,t}$)

$$e_{i,t+1} := E_i(e_{i,t}, a^e_{i,t}) := \max\left\{ \min\left[ e_{i,t} + \eta_i^c \max(a^e_{i,t}, 0) + \frac{1}{\eta_i^d} \min(a^e_{i,t}, 0), \overline{e}_i \right], 0 \right\},$$

## Reward function

$$r_{i,t} = R^m_{i,t}(a^e_{i,t}; a^e_{-i,t}, s_t) + R^v(a_{i,t}; a_{-i,t}, s_t)/I.$$

# State Transition and Reward Function

## Battery state of charge ($e_{i,t}$)

$$e_{i,t+1} := E_i(e_{i,t}, a^e_{i,t}) := \max\left\{\min\left[e_{i,t} + \eta^c_i \max(a^e_{i,t}, 0) + \frac{1}{\eta^d_i}\min(a^e_{i,t}, 0), \overline{e}_i\right], 0\right\},$$

## Reward function

$$r_{i,t} = R^m_{i,t}(a^e_{i,t}; a^e_{-i,t}, s_t) + R^v(a_{i,t}; a_{-i,t}, s_t)/I.$$

$$R^m_{i,t} := \begin{cases} \mathbb{I}_{i\in\mathcal{B}_t} \times \left[SDR_t \cdot P_t \cdot b_{i,t} + (1 - SDR_t) \cdot UR \cdot b_{i,t}\right] & \\ + \mathbb{I}_{i\in\mathcal{S}_t} \times \left(P_t \cdot b_{i,t}\right), & 0 \leq SDR_t \leq 1 \\ FIT \cdot b_{i,t}, & SDR_t > 1, \end{cases}$$

# Reward Function (cont.) Constraint Violation Penalty

$$R_t^v/I = -\lambda \sum_{j:Bus} \left[ \left( \max(0, |V_{j,t}| - \overline{V}_j) + \max(0, \underline{V}_j - |V_{j,t}|) \right] \Big/ \Big( I, \right.$$

- $I$ – the no. of agents, $\lambda$ – an arbitrary large number (the fake penalty for voltage violation)
- Assumption – The voltage violation is equally shared among all agents (again, this is NOT real, only for training)

# Reward Function (cont.) Constraint Violation Penalty

$$R_t^v/I = -\lambda \sum_{j:Bus} \left[ \left( \max(0, |V_{j,t}| - \overline{V}_j) + \max(0, \underline{V}_j - |V_{j,t}|) \right] \middle/ \left( I, \right.\right.$$

- $I$ – the no. of agents, $\lambda$ – an arbitrary large number (the fake penalty for voltage violation)
- Assumption – The voltage violation is equally shared among all agents (again, this is NOT real, only for training)
- If voltage violation $> 0$, all bids are rejected; agents resubmit bids

# Reward Function (cont.) Constraint Violation Penalty

$$R_t^v/I = -\lambda \sum_{j:Bus} \left[ \left( \max(0, |V_{j,t}| - \overline{V}_j) + \max(0, \underline{V}_j - |V_{j,t}|) \right] \right/ I,$$

- $I$ – the no. of agents, $\lambda$ – an arbitrary large number (the fake penalty for voltage violation)
- Assumption – The voltage violation is equally shared among all agents (again, this is NOT real, only for training)
- If voltage violation $> 0$, all bids are rejected; agents resubmit bids
- $\overline{V}^j / \underline{V}^j$: upper/lower voltage limit of Bus $j$
- $V_{j,t}$: voltage magnitude at Bus $j$ **after** each agent makes the decision, calculated by solving a bus injection model – Bids validation (done by DSO or Blockchain)

$$p_k = \sum_{j=1}^{N} \left( |V_k||V_j|(G_{kj}\cos(\alpha_k - \alpha_j) + B_{kj}\sin(\alpha_k - \alpha_j)), \right.$$

$$q_k = \sum_{j=1}^{N} \left( |V_k||V_j|(G_{kj}\sin(\alpha_k - \alpha_j) - B_{kj}\cos(\alpha_k - \alpha_j)), \right.$$

$$\text{for } k = 1, 2, \ldots, N,$$

# MARL with Continuous State & Action Spaces

It's all about policy gradient!

For a generic policy $\pi(a|s,\theta)$ and a performance measure $J(\theta)$,

$$\theta_{t+1} = \theta_t + \alpha \widehat{\nabla J(\theta_t)}.$$

# Three MARL Frameworks

**Completely <u>decentralized</u> learning/execution**

-- no communication among peers

**Middle Ground: Consensus-based, decentralized actor-critic MARL**

– Each peer maintains an estimate of the centralized critic function

-- Update the estimates through neighbors to reach a consensus

-- Decentralized actor (policy) update

**<u>Centralized</u> Learning/Decentralized Execution**

- Centralized critic (action-value) function estimation (need other agents' policies)
- Decentralized actor (policy) update

# Three MARL Frameworks The Details

## Performance measure $J$

- Pure decentralized and MADDPG $J_i(\theta_i) = \mathbb{E}_{\pi_{\theta_i}}[\sum_{t=0}^{T} \gamma_i^t r_{i,t}]$

- Consensus: $J(\theta) = \mathbb{E}_{\pi_\theta}\left[\left(\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T}\left(\frac{1}{I}\sum_{i=1}^{I} r_{i,t}\right)\right]\right.$

# Three MARL Frameworks The Details

## Performance measure $J$

- Pure decentralized and MADDPG $J_i(\theta_i) = \mathbb{E}_{\pi_{\theta_i}}[\sum_{t=0}^{T} \gamma_i^t r_{i,t}]$

- Consensus: $J(\theta) = \mathbb{E}_{\pi_\theta}\left[\left(\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T}\left(\frac{1}{I}\sum_{i=1}^{I} r_{i,t}\right)\right)\right]$

## Policy Gradient

- Purely decentralized: $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim\rho^\theta, a_i\sim\pi_{\theta_i}}\left[\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s_i)Q_i^\pi(s_i;a_i)\right]$ (PPO implementation: [Feng et al., 2023])
- MADDGP: $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim\rho^\theta, a_i\sim\pi_{\theta_i}}\left[\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s_i)Q_i^\pi(s;a_1,\ldots,a_I)\right]$

# Three MARL Frameworks The Details

## Performance measure $J$

- Pure decentralized and MADDPG $J_i(\theta_i) = \mathbb{E}_{\pi_{\theta_i}}[\sum_{t=0}^{T} \gamma_i^t r_{i,t}]$

- Consensus: $J(\theta) = \mathbb{E}_{\pi_\theta}\left[\left(\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T}\left(\frac{1}{I}\sum_{i=1}^{I} r_{i,t}\right)\right]\right($

## Policy Gradient

- Purely decentralized: $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim\rho^\theta, a_i\sim\pi_{\theta_i}}\left[\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s_i)Q_i^\pi(s_i; a_i)\right]$ (PPO implementation: [Feng et al., 2023])

- MADDGP: $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim\rho^\theta, a_i\sim\pi_{\theta_i}}\left[\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s_i)Q_i^\pi(s; a_1, \ldots, a_I)\right]$

- Consensus: Expected policy gradient (EPG) $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim\rho^\theta, a_{-i}\sim\pi_{\theta_{-i}}} I_{\theta_i}^Q(s, a_{-i})$, where $I_{\theta_i}^Q(s, a_{-i}) = \mathbb{E}_{a_i\sim\pi_{\theta_i}}\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s)Q_i^\pi(s; a_1, \ldots, a_I)$.

# Three MARL Frameworks The Details

## Performance measure $J$

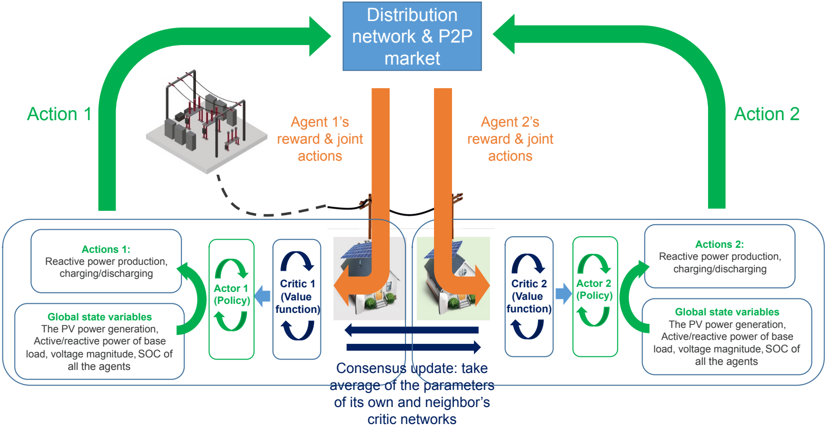
- Pure decentralized and MADDPG $J_i(\theta_i) = \mathbb{E}_{\pi_{\theta_i}}[\sum_{t=0}^{T} \gamma_i^t r_{i,t}]$
- Consensus: $J(\theta) = \mathbb{E}_{\pi_\theta}\left[\left(\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T}\left(\frac{1}{I}\sum_{i=1}^{I} r_{i,t}\right)\right)\right]$

## Policy Gradient

- Purely decentralized: $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim\rho^\theta, a_i\sim\pi_{\theta_i}}\left[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i|s_i) Q_i^\pi(s_i; a_i)\right]$ (PPO implementation: [Feng et al., 2023])
- MADDGP: $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim\rho^\theta, a_i\sim\pi_{\theta_i}}\left[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i|s_i) Q_i^\pi(s; a_1, \ldots, a_I)\right]$
- Consensus: Expected policy gradient (EPG) $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim\rho^\theta, a_{-i}\sim\pi_{\theta_{-i}}} I_{\theta_i}^Q(s, a_{-i})$, where $I_{\theta_i}^Q(s, a_{-i}) = \mathbb{E}_{a_i\sim\pi_{\theta_i}} \nabla_{\theta_i} \log \pi_{\theta_i}(a_i|s) Q_i^\pi(s; a_1, \ldots, a_I)$.

To deal with the centralized critic function, each agent $i$ use $\tilde{Q}(a_i, a_{-i}; w_{i,t})$ to approximate $Q_i^\pi(s; a_1, \ldots, a_I)$. Agent $i$ use weighted average of $w_t^j$, all $j$'s in $i$'s neighbor, to obtain $w_{i,t+1}$.

# Illustration of the Consensus MARL lgorithm



# Part III – Numerical Results

# Simulation Inputs

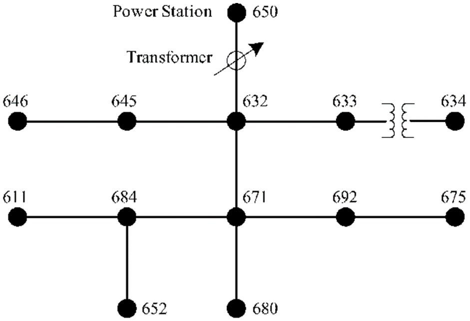



Figure: Test case: IEEE 13-bus feeder

- UR and FIT: $P_{UR} = 14$ ¢/KWh, $P_{FIT} = 5$ ¢/KWh.
- Agents: 12 prosumers, one at each bus (except the substation)
- PV and storage per agent: PV: 30KW, storage: 50KWh, charging/discharging efficiency: 0.95/0.9

# Input Data (cont.)

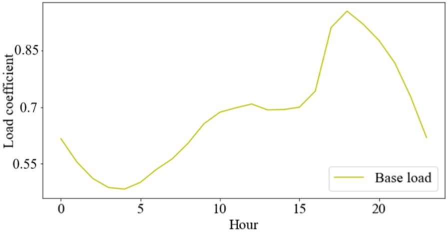



Figure: Average daily baseload shape

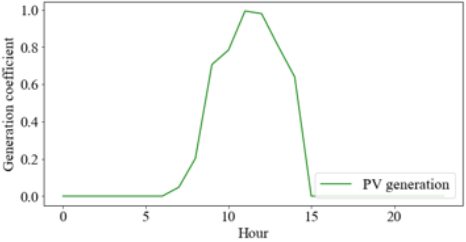



Figure: Daily PV output shape

# Numerical Results Rewards and Voltage Violation

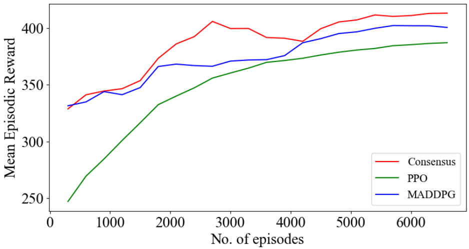



Figure: 30-epi. moving avg. of episodic total reward

# Numerical Results Rewards and Voltage Violation

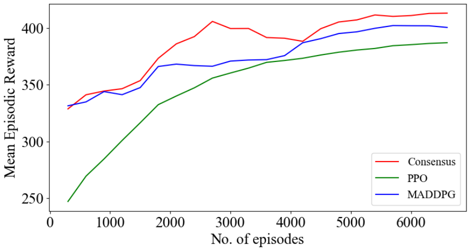



Figure: 30-epi. moving avg. of episodic total reward

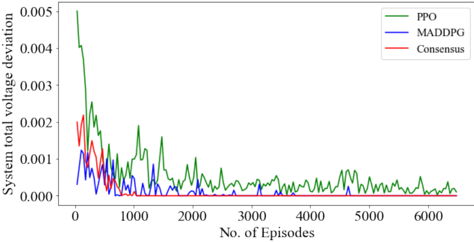



Figure: Voltage violation $[0.96pu, 1.04pu]$

# Market Clearing Price (under SDR)

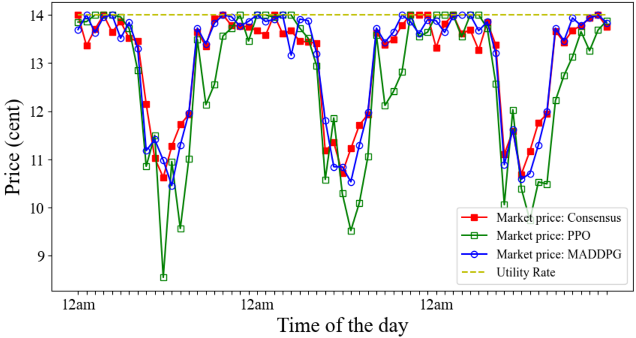



Figure: Hourly clearing prices (the last 3 days)

# Summary and Future Research

## Summay

- MARL is promising in P2P energy trading
  - Can realize control-automation
  - Decentralized learning among networked agents can learn to avoid constraint violation

# Summary and Future Research

## Summay

- MARL is promising in P2P energy trading
  - Can realize control-automation
  - Decentralized learning among networked agents can learn to avoid constraint violation
- But, the devil is in the details!

# Summary and Future Research

## Summay

- MARL is promising in P2P energy trading
  - Can realize control-automation
  - Decentralized learning among networked agents can learn to avoid constraint violation
- But, the devil is in the details!

## Future Research

- Scalability
- Cybersecurity: Byzantine agents [Figura et al., 2021]
- Real-time implementation (need to couple with demand and solar prediction)

# Thank you!

Acknowledgment: This research is partially supported by National Science Foundation grant ECCS-2129631 and the U.S. Department of Energy, Office of Electricity, under Award Number DE-OE0000921.

# References:


Feng, C., Liu, A. L., and Chen, Y. (2023).
Decentralized voltage control with peer-to-peer energy trading in a distribution network.
In *Proceedings of the 56th Hawaii International Conference on System Sciences*, pages 2600–42609.

Figura, M., Lin, Y., Liu, J., and Gupta, V. (2021).
Resilient consensus-based multi-agent reinforcement learning with function approximation.
*arXiv preprint arXiv:2111.06776*.

Liu, N., Yu, X., Wang, C., Li, C., Ma, L., and Lei, J. (2017).
Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers.
*IEEE Transactions on Power Systems*, 32(5):3569–3583.

Zhao, Z., Feng, C., and Liu, A. L. (2022).
Comparisons of auction designs through multiagent learning in peer-to-peer energy trading.
*IEEE Transactions on Smart Grid*, 14(1):593–605.