MDPs and Value Iteration for Microgrid Trade Policies

Gianna Chien
ggchien

Santosh Mohan
santoshm

Priyanka Sekhar
psekhar

Abstract

This project investigates the strengths and limitations of using an MDP to model optimal trade policies between communities of microgrids. We use value iteration to learn optimal policies in different states, and use our model to observe likely trade behaviors in states of deficit (where one member has more money or energy than another). By defining and varying price, reward functions, production amounts, and consumption amounts, we are able to evaluate the effectiveness of this modeling strategy for use in smart city design. While the model does provide high level insight into behaviors in extremes, the requirement of manually setting constraints and rewards may cause modelers to miss nuances in real world behavior. However, despite these drawbacks, this model-based approach provides high level insight into macro behaviors that are useful in city prototyping and inventive structure iteration.

1 Background

Microgrids help to power islands, isolated areas, and increasingly serve as supplements to large centralized grids. However, the flexibility that comes with load balancing between multiple power sources also comes with a fair amount of uncertainty as different communities have different rates of energy production, rates of energy consumption, monetary resources, and views on energy prices. As global energy needs continue to change, accurate models for microgrids that capture these uncertainties can prove useful in many real-world scenarios such as optimizing power distribution after a disaster or in green city planning.

Markov decision processes (MDPs) provide a useful framework for modeling decision-making processes. By defining the states, actions, rewards, and transition probabilities of an MDP, the MDP can be solved to generate the optimal policy for each state in the state space [1]. MDPs have been used in a variety of fields of study, including human behavior; the Schelling Segregation Model, for example, shows that seemingly mild preferences among neighbors often results in the segregation of neighborhoods [2].

Past simulations of population dynamics include Conway's Game of Life, which encodes rules for the survival, multiplication, and death of the cells in a colony depending on the clustering of those cells within the environment [3]. Similar simulations concerning energy supply and demand include Bharadwaj, Reddy, and Bhatnagar's work examining energy sharing among microgrids using MDPs. This work used Q-learning to generate an optimal policy for maximizing profit earned by microgrids from selling excess energy while maintaining a low gap between demand and supply [4].

2 Introduction

In this project we investigate the versatility of MDPs and value iteration for modeling energy trade behavior within a community of microgrids. Because microgrids often have either excesses or deficits of energy, they must trade with other energy producers in real time.

2.1 Significance

Prior work on this topic has focused on learning optimal strategies in existing market conditions and microgrids, and Q-learning approaches have proved effective for this simulation of the status quo. However, the rise of "smart cities," self governance through blockchain, and new innovative civil structures around the world have made it possible for leaders to define their own constraints and rewards across a community. Furthermore, in places where the amount of real time data is sparse or inaccessible, either due to the lack of high qual-

ity data collection or legal reasons, accurate models may be difficult or impossible to learn.

Consequently, our work focuses on how to define constraints and parameters in microgrid MDPs to create desired trade behavior patterns. Rather than just learning what optimal policies are, our project investigates how we can shape behavior. This lens allows smart city planners to leverage our findings should they wish to prototype a set of reward structures, cost structures, or production targets in order observe how their plans will likely affect real world decisions. This project evaluates the strengths and constraints of value iteration on an MDP microgrid model for such prototyping.

2.2 Approach

To evaluate the expressive power of value iteration on an MDP of a microgrid community, we first define the states, actions, rewards and transitions to be used in our model-based approach. The details of these definitions can be found in section 3.

We then vary the initial state configuration (number of producers and consumers), the relative amount of energy production and consumption in a community, the reward functions, and the price of energy. The impact of these variations on trade behavior are shown in Section 5. Our implementation can be found on GitHub. [1]

3 Model

Our model consists of a 2x2 grid world. Each square on the grid has four attributes: a number of producers of energy, number of consumers of energy, available energy reserves, and a budget to buy energy. We define the state S of our grid world to be the values of each attribute in each square. We allow each square to trade with its adjacent square (in clockwise order). A square will consider the potential reward generated by either selling or buying energy with that adjacent square. In the offline case, we generate the state space and take the action that maximizes a reward function over all possible trades between any two squares. In the online case, we initialize a state S and only calculate policies for states reachable from S.

3.1 Constraints

To constrain the state space to something reasonable to enumerate, we only consider trades with adjacent squares. We also create reward functions that prioritize the buying of energy if the square at the next iteration will run out of energy due to local consumer demand.

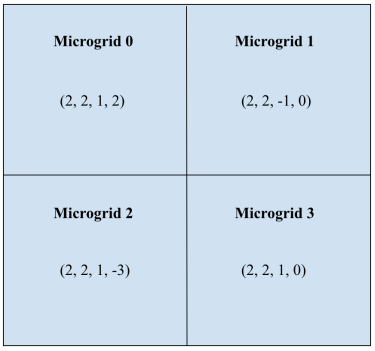

Figure 1: An example state of the microgrid municipality.

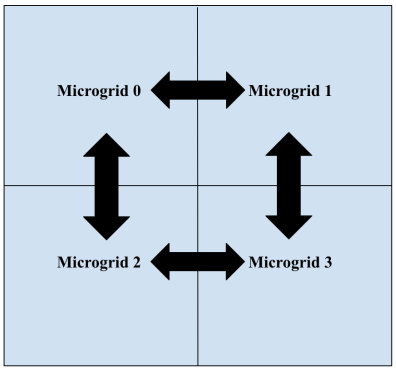

Figure 2: The possible trades that can be made at each timestep.

4 Value Iteration

To calculate the optimal policy, we utilize Gauss-Seidel value iteration with a discount factor γ of 0.5. We define:

- States: Data on each microgrid community is stored as a tuple of P energy producers, C energy consumers, E energy surplus, and M monetary surplus, where E and M change at each timestep. The overall state of the 2x2 municipality is a collection of the states of the individual microgrids. In order to generate the complete state space, as is necessary for offline value iteration, we constrain the possible states by defining constants MIN_UNIT and MAX_UNIT denoting

[1] https://github.com/ppsekhar/CS238

the minimum and maximum units of money and energy a given microgrid community can own and assuming that all money and energy amounts must be integers between MIN_UNIT and MAX_UNIT, inclusive.

- Transitions: Transitions are defined deterministically, where a given action taken from a specific state can only produce one possible next state. Therefore, only one transition with transition probability 1 exists for each action.

- Rewards: The reward of each microgrid community is based on its energy and monetary surplus. Specific reward functions are examined in more detail in section 5.3.

- Actions: Actions are defined as (buyer, seller, amount) tuples where the buyer is the ID number of the microgrid receiving energy, the seller is the ID number of the microgrid providing energy, and the amount is the number of energy units moving from buyer to seller. Amounts are constrained to one of [0, 1, 2, 3], where a trade of amount 0 indicates a non-action. Only one trade can be performed per timestep, and trades can only be performed between adjacent grids in a round-robin style. The constraints on trading are explained in more detail in section 3.1.

By iterating through the state space according to a pre-determined state ordering and repetitively updating the utility values and corresponding best action from each state according to Equation 1,

$$U_{k+1}(s) \leftarrow max_a[R(s,a)+\gamma \sum_{s'} T(s'|s,a)U_k(s')] \tag{1}$$

we converge to the optimal policy [1].

4.1 Online vs Offline

Because the state space grows exponentially with each increase in the MIN_UNIT to MAX_UNIT range, generating the complete state space can be intractable. Therefore, we implemented both an offline method, which generates a policy with an action for every state in the state space as previously described, and an online method, which generates a policy with an action for only the states reachable from the initial state. This is done by maintaining and iterating over a queue of viable next states that is continuously updated as more states are explored. We terminate updates when this queue is depleted, or after a predetermined number of timesteps.

4.2 Convergence

Since we want to understand the extent to which reward functions influence the outcomes of trade, it makes sense that we define a convergence condition that focuses on longer time horizons. We define a converged model to be that which has ran for at least N iterations and whose latest change in reward (delta) based on the action taken is less than the average of all previous deltas. We also ignore 0 deltas. This enforces the idea that the model should continue to evaluate new states that allow it to think longer-term.

5 Experiments

For all these experiments, the initial numbers of producers and consumers were set to 2 and 2 respectively. Future work would focus on extending these experiments to other initializations.

5.1 Production and Consumption

We examined the effect of differing production and consumption values on the optimal policies generated. Three conditions were considered:

- Equal consumption and production: The number of units of energy consumed by each consumer at each timestep is equal to the number of units of energy produced by each producer at each timestep. This was simulated by making each consumer consume 1 unit of energy and each producer produce 1 unit of energy.

- Higher consumption than production: The number of units of energy consumed by each consumer at each timestep is higher than the number of units of energy produced by each producer at each timestep. This was simulated by making each consumer consume 2 units of energy and each producer produce 1 unit of energy.

- Lower consumption than production: The number of units of energy consumed by each consumer at each timestep is lower than the number of units of energy produced by each producer at each timestep. This was simulated by making each consumer consume 1

unit of energy and each producer produce 2 units of energy.

All simulations were performed using the online method with MIN_UNITS as -5 and MAX_UNITS as 5. Each unit of energy costed one unit of money.

In the following graphs, a state with an "energy disparity" is defined as a state where the difference in energy between some pair of communities is at least MAX_UNIT - MIN_UNIT, while a state with a "money disparity" is defined as a state where the difference in money between some pair of communities is at least MAX_UNIT - MIN_UNIT.

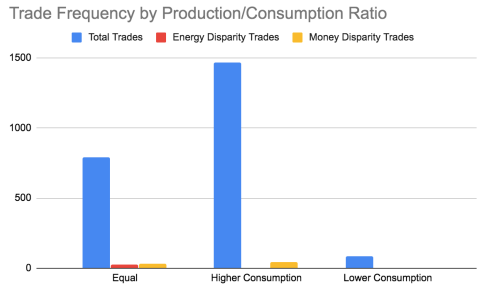

Figure 3: The number of states where the optimal action is to trade with a neighbor, split by producer/consumer ratio.

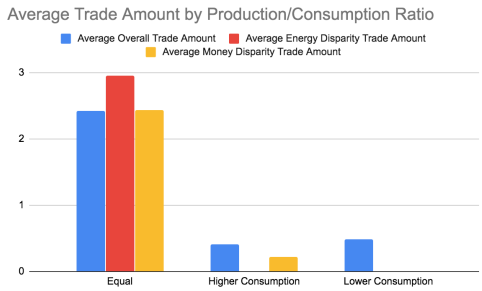

Figure 4: The average amount of energy bought or sold in states where the optimal action is to trade with a neighbor.

As shown in Figure 3, higher consumption encourages a higher volume of trading, as more energy is required to support each microgrid community. Similarly, lower consumption discourages trading, as less energy is required for the community. Figure 4 demonstrates that on average much smaller amounts of energy are traded in the higher and lower consumption conditions as opposed to the equal production/consumption value condition. This is likely because in the higher consumption condition, microgrid communities require more resources and thus are more hesitant to sell large amounts of energy, and in the lower consumption condition microgrid communities require less resources and thus gain no benefit from hoarding large amounts of energy. This explanation could also account for the lack of energy disparity trades, as communities did not trade sufficient amounts of energy to create such disparities in our online simulation.

5.2 Rewards

We tested different rewards functions to see how incentives can be used to influence behaviors. These reward functions map to broad categories of real world behaviors, but a city planner could use a wide range or combination of different rewards to prototype their city's incentive structure. In this project we used the following rewards:

- *Greedy* - If the designer wishes to assume that money is more important to the community than energy, money can have a higher weight.
- *Regulated* - If the city designer wishes to institute a tax, this reward function can be used to penalize members with too much money.
- *Rich* - If the city designer assumes society benefits when wealth is concentrated, for example by having a rich district where schools and hospitals are, then this reward function can be used to offer a bonus based on the amount of money the wealthiest community has.
- *Decay* - If the designer assumes that the more money one has, the less they value energy and money, the relative reward of an individual community compared to neighbors can be discounted proportionally to its wealth.
- *Energy Deficit* - In all other scenarios the reward function assumes the community cannot run an energy deficit and penalizes them steeply. However, if the designer wishes to allow communities to consume more energy than they produce (perhaps to simulate the scenario where communities can purchase energy from someone outside the modeled trade partners), then a reward function that instead penalizes a financial deficit could be used.

The reward functions for each of the above scenarios are defined in Equation 2, with α and β

values defined in Table 1. E_c and M_c represent the amount of Energy and Money for a community c in a given state.

$$R_{total} = \sum_c \alpha * E_c + \beta * M_c \quad (2)$$

The exception is the RICH reward function, which is defined in Equation 3.

$$R_{total} = \beta^3 * max(M_c) + \sum_c \alpha * E_c \quad (3)$$

Policy	α	β
Baseline	1	1 if $E_c > 0$; 0 otherwise
Greedy	1	3 if $E_c > 0$; 0 otherwise
Regulated	1	0 if $M_c > MAX$; 0 if $E_c <= 0$; 3 Otherwise
Rich*	1	3 if $E_c > 0$; 0 otherwise
Decay	$\frac{1}{M_c^2}$	$\frac{1}{M_c^2}$
Energy Deficit	1 if $M_c > 0$; 0 Otherwise	1

Table 1: α and β values for each reward function. MAX was hardcoded as MAX UNIT for this experiment.

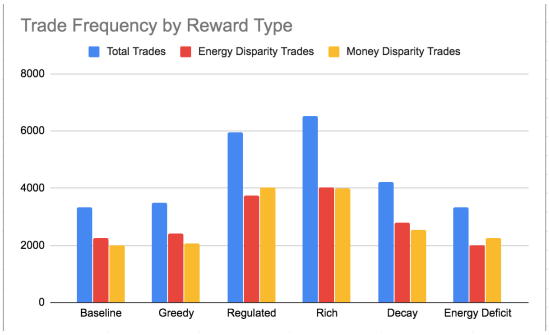

Figure 5: The number of states where the optimal action is to trade with a neighbor for each reward function.

Energy disparity and money disparity states are defined the same way as in section 5.1. In this particular experiment, MAX_UNIT was set to 1 and MIN_UNIT was set to -1.

The *Rich* and *Regulated* policies both encourage a higher volume of trading as shown in Figure 5. However, reward functions that assume having a rich person is beneficial to society as a whole tend to also encourage greater amounts of energy traded in states of deficit, as shown in Figure 6. This is likely because the community with the most money will seek to minimize energy surplus, while other communities may be willing to buy excess energy to boost the richest member.

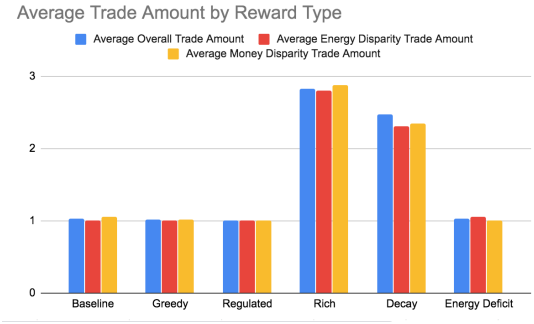

Figure 6: The average amount of energy bought or sold in states where the optimal action is to trade with a neighbor.

More nuanced reward mechanisms are possible, however it is difficult to fine tune every possible set of behavioral tendencies into an explicitly selected function. Thus, much important information may be missing for those prototyping incentive schemes.

We visualized some examples of "extreme" actions by the reward function. The following are the energy and budget states resulting from an action and corresponding reward value. The simulation was run from the same initial states in the online learning regime in Figure 7.

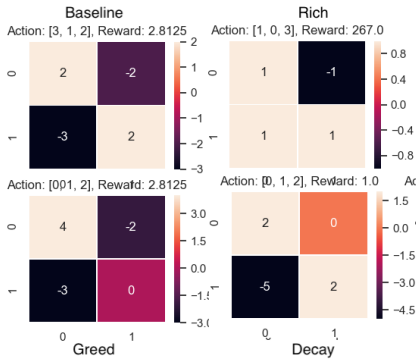

Figure 7: Budget resulting from the action taken in title: (Buyer, seller, amount)

We observe that, against the baseline, the *Rich* reward function "spreads the wealth" better. This is potentially a byproduct of more wealthy squares

trading more often with squares in a deficit. In contrast, *Greedy* does little to change the baseline results and *Decay* seems to concentrate the deficit in a single state.

5.3 Price

We examined the effect of varying prices of energy on the optimal policies generated. Three conditions were considered:

- Equal value of money and energy: 1 unit of energy costs 1 unit of money.
- Higher value of money than energy: 1 unit of energy costs less than 1 unit of money (0.5 money per unit energy in simulations).
- Lower value of money than energy: 1 unit of energy costs more than 1 unit of money (2 money per unit energy in simulations).

All simulations were performed using the online method with MIN_UNITS as -5 and MAX_UNITS as 5. Each producer produced 1 unit of energy and each consumer consumed 1 unit of energy per timestep.

The following graphs use the same definitions of "energy disparity" and "money disparity" as in previous sections.

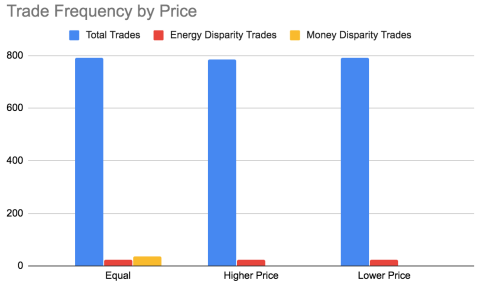

Figure 8: The number of states where the optimal action is to trade with a neighbor, split by relative price.

The price of each unit of energy did not seem to have any significant effect on the total trade frequency, as shown by Figure 8. However, it is notable that in the higher price condition, there were no money disparity trades. This is most likely because the high price of energy made communities unable or unwilling to trade in situations with money disparity. Additionally, although there were fewer money disparity trades in the lower price condition than there were in the equal price condition (2 as opposed to 35), the average trade amount was higher for money disparity trades in the lower price condition; in other words, trades during money disparities were made for as many units of energy as possible. This is likely because in the lower price condition, a community must trade more energy in order to obtain the same amount of money as in the other conditions, resulting in higher trade amounts in the case of money disparity.

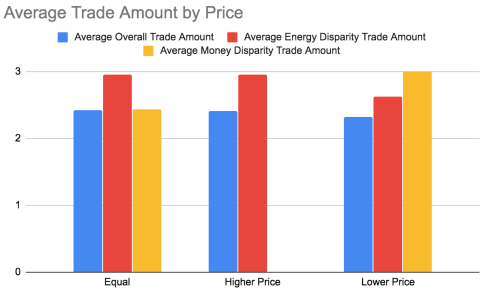

Figure 9: The average amount of energy bought or sold in states where the optimal action is to trade with a neighbor.

6 Discussion and Conclusions

MDPs and value iteration provide a lot of power for those designing new trade systems. However, the primary limitations are speed of iteration through the massive state space, difficulties with hand-defining expressive representations of human behavior in reward and transition functions, and the impact of factors not captured in this model, such as external factors that can influence price and energy production such as weather. However, this model does seem to be sufficient to gain high level insight into key elements necessary to tweak behaviors, such as appropriate taxes and programmatic incentives.

7 Future Work

Future iterations of this project would relax assumptions and introduce more complexity into the transition and reward functions. For example, we would extend to more than just 4 communities with adjacent trade partners and relax the restriction of only allowing neighbors to trade. Transition functions would be non-deterministic and reward functions would be based on well-established behavioral economic reward functions. Actions would be more diverse than trading 0, 1, 2, or 3 units of energy.

8 Contributions

- Priyanka - Helped implement initial offline value iteration and MDP definition, implemented an online alternative for large state spaces, implemented reward functions, structured and contributed to final writeup.
- Gianna - Helped implement initial offline value iteration and MDP definition, implemented price and production/consumption variation and analyzed effects, contributed to final writeup.
- Santosh - Helped implement the visualization of actions correlated with extrem rewards. Contributed to final writeup, specifically the effects of reward functions on actions taken and a description of the model.

References

[1] Mykel J. Kochenderfer. *Decision Making Under Uncertainty*. Massachusetts Institute of Technology, 2015.

[2] Erez Hatna and Itzhak Benenson. The schelling model of ethnic residential dynamics: Beyond the integrated - segregated dichotomy of patterns. *Journal of Artificial Societies and Social Simulation*, Jan 2012.

[3] Lorena Caballero, Bob Hodge, and Sergio Hernandez. Conway's "game of life" and the epigenetic principle. *Frontiers in Cellular and Infection Microbiology*, June 2016.

[4] D. Raghuram Bharadwaj, D. Sai Koti Reddy, and Shalabh Bhatnagar. A unified decision making framework for supply and demand management in microgrid networks. *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, Oct 2018.