Batch Q-learning for Energy Storage Management in Smart Grids

Mehnaz Mannan, Aditya Mahajan
mehnaz.mannan@mail.mcgill.ca, aditya.mahajan@mcgill.ca

McGill
Centre for Intelligent Machines

Introduction

We consider the energy management system (EMS) of a sustainable house that contains a renewable generation unit (e.g., rooftop solar generation) and an energy storage device (e.g., battery). The house is connected to the electricity grid; the EMS can purchase electricity from the grid, but cannot supply electricity back to the grid. The renewable generation, the energy demand in the house, and the electricity price vary in a stochastic manner. At each decision epoch, the EMS must meet the demand using the renewable generation, the electricity grid, and the storage device.

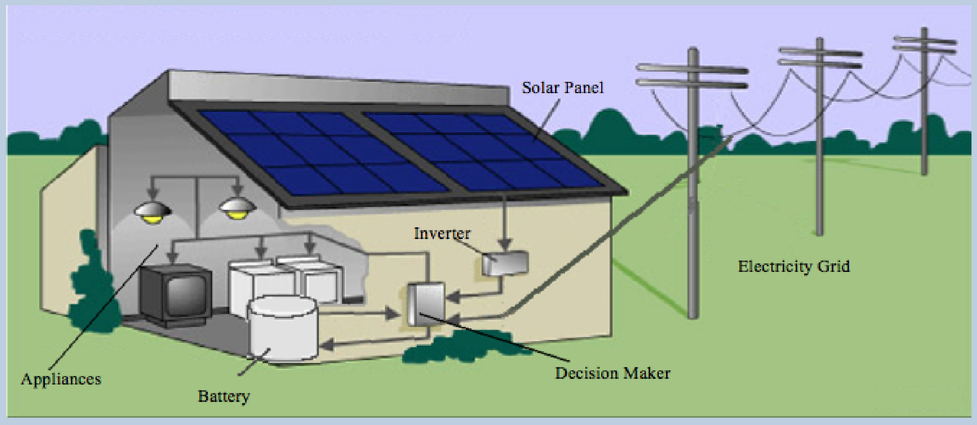

Figure 1: A sustainable house with an energy management system

We investigate optimal decision strategies for the EMS that determine when and how much energy is purchased from the grid and is stored in the storage device. We present a dynamic programming decomposition of the problem and use batch Q-learning to obtain an iterative algorithm that converges to the optimal decision strategy without any knowledge of the probability distributions of the renewable generation, the demand, and the electricity price. In the future, we plan to investigate the performance of the proposed batch Q-learning algorithm on real generation, demand, and price data.

Model

State: $(\mathbf{x}_t, \mathbf{y}_t, \mathbf{p}_t) \in [0, S] \times \mathbb{R} \times \mathbb{N}$

- ▷ S is the total capacity of the energy storage device.
- ▷ $\mathbf{x}_t$ denotes the level of useful energy in the storage at time t
- ▷ $\mathbf{y}_t$ is the net load i.e. demand minus renewable generation at time t
- ▷ $\mathbf{p}_t$ is the price of each unit of grid electricity at time t

Action: $\mathbf{x}_t \leq \mathbf{u}_t \leq S - \mathbf{x}_t$

- ▷ $\mathbf{u}_t$ denotes the amount of useful energy to be stored. Positive $\mathbf{u}_t$ means we are storing energy in the battery; negative $\mathbf{u}_t$ means we are extracting energy from the battery.

Dynamic Programming Decomposition:

$$V_t(x_t, y_t, p_t) = \min_{u_t} \; p_t[y_t + \frac{u_t}{\beta_{u_t}}]^+ + \gamma \mathbb{E}_{\mathbf{y}_{t+1}, \mathbf{p}_{t+1}}[V_{t+1}(x_t + u_t, \mathbf{y}_{t+1}, \mathbf{p}_{t+1})]$$

$$\text{where, } \beta_{u_t} = \begin{cases} \rho & \text{if } u_t \geq 0 \\ 1 & \text{otherwise} \end{cases} \quad \text{and } 0 < \gamma < 1$$

- ▷ β_{u_t} and γ are roundtrip efficiency and discount factor respectively.

Q Learning Equation:

$$Q_t(x_t, y_t, p_t, u_t) = (1-\alpha)Q_{t-1}(x_t, y_t, p_t, u_t) + \alpha(p_t[y_t + \frac{u_t}{\beta_{u_t}}]^+ + \min_{u_{t+1}} \; Q_{t-1}(x_t + u_t, \mathbf{y}_{t+1}, \mathbf{p}_{t+1}, u_{t+1}))$$

Post-decision state: $(\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t, \tilde{\mathbf{p}}_t) \in [0, S] \times \mathbb{R} \times \mathbb{N}$

- ▷ $\tilde{\mathbf{x}}_t = \mathbf{x}_t + u_t, \tilde{\mathbf{y}}_t = \mathbf{y}_t, \tilde{\mathbf{p}}_t = \mathbf{p}_t$

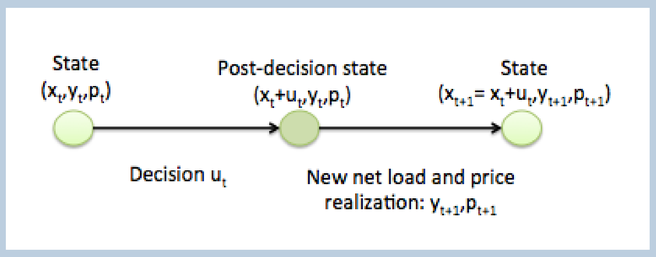

Figure 2: Illustration of the post-decision state

Batch Q Learning Equations:

$$V_t(x_t, y_t, p_t) = \min_{u_t} \; p_t[y_t + \frac{u_t}{\beta_{u_t}}]^+ + \gamma U_{t-1}(\tilde{x}_t, \tilde{y}_t, \tilde{p}_t)$$

$$U_t(\tilde{x}_t, \tilde{y}_t, \tilde{p}_t) = (1-\alpha)U_{t-1}(\tilde{x}_t, \tilde{y}_t, \tilde{p}_t) + \alpha V_t(\tilde{x}_t, \mathbf{y}_{t+1}, \mathbf{p}_{t+1}) \quad \forall \tilde{x}_t$$

Results

We consider the case where $x = [0, 1, 2, 3]$, $y = [-1, 0, 1]$, $p = [3, 15]$, $\gamma = 0.9$,$\rho = 1$. y and p evolve in an independent and identically distributed manner.
When $p = 3$

x \ y	-1	0	1
0	1	1	1
1	1	0	0
2	1	0	-1
3	0	0	-1

When $p = 15$

x \ y	-1	0	1
0	1	0	0
1	1	0	-1
2	1	0	-1
3	0	0	-1

When there is excess demand (**y** is positive), it is optimal to either extract from the storage, or to buy from the grid and keep in storage, or to do nothing. When $y = 1, x = 0, p = 3$, we buy and store some energy on top of buying energy to satisfy the excess demand. This is because it makes sense to buy energy at a lower price, so that the stored energy can be used to satisfy excess demand when price is high in future.

When there is excess generation, it is optimal to store all the excess and if necessary buy and from the grid and store as well.

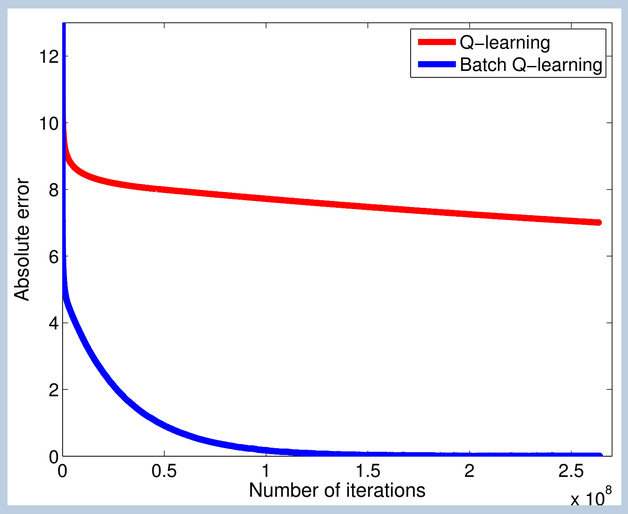

Figure 3: Convergence of Q learning and Batch Q learning value functions to DP value function

As can be seen from the graph, the Batch Q learning algorithm converges in about 1.3×10^8 iterations, while the regular Q learning algorithm does not converge even after 2.6×10^8 iterations. Hence the Batch Q learning algorithm is more effective for finding the optimal policy.

Future Work

The hourly electricity price, demand and wind generation output data for Ontario from 2006 to present, is available online through the Independent Electricity System Operator.We plan to show that Batch Q learning coverges more quickly to the DP solution than regular Q learning by applying these algorithms to 2006-2012 data.We will then apply the policy learned through Batch Q learning on 2013 data and illustrate the savings from storing according to this policy over having zero storage.