Progress Report

Pratik Gajane

Montanuniversität Leoben

Experiments on Microgrid Simulator

Simple algorithm

Simple algorithm

If consumption $<$ EPV production
        action $=$ CC
else
        if total SOC $>$ additional demand
                action $=$ DD
        else
                action $=$ II

SOC	State of charge of a battery.
C	Charge a battery.
D	Discharge a battery.
I	Keep a battery idle.
additional demand	consumption - EPV production.

Performance of simple (no-lookahead) algorithm

Total reward for the year of 2014 = -91367

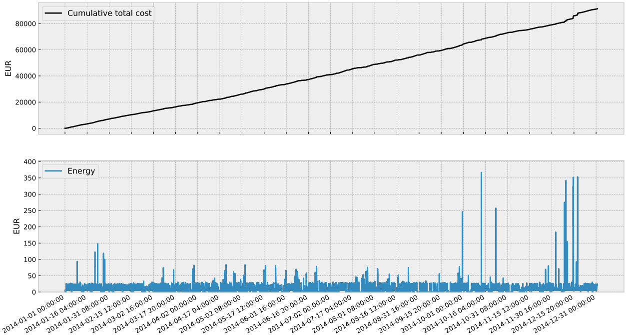

Figure 1: Cost and Energy profile

Why use future predictions?

- Batteries should
  1. have enough charge to satisfy future additional demands.
  2. have enough space to store future excess EPV production.
- In some cases, it is profitable to use batteries later. Why? Because, diesel generator has minimum stable generation (MSG).
- By using diesel generator,
  1. Energy wasted now = MSG - additional demand now.
  2. Energy wasted in future = MSG - additional demand in future.
- If Energy wasted now $<$ Energy wasted in future, then profitable to use diesel generator now (i.e. use batteries later)

Algorithm using future predictions

Lookahead Algorithm

If consumption < EPV production
    action = CC
else
    If it is profitable to use battery in future
        action = II
    else
        action = DD

Performance using future predictions

Total reward for the year of 2014 using 6-hour lookahead = -72506

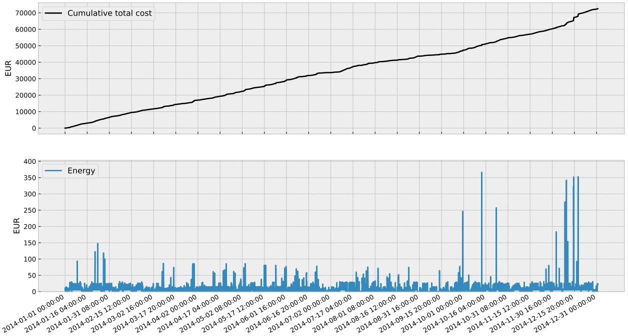

Figure 2: Cost and Energy profile

Performance using future predictions

Total reward for the year of 2014 using various values for lookahead

Lookahead	Cumulative reward
...	
2	-73668
4	-73498
6	-72506
8	-72834
10	-73193
...	

Predicting Consumption

- KNN with neighbours = consumption at the same time on the same day previous weeks.

  Neighbours for 2014-01-31 T 08-00-00 → 2014-01-24 T 08-00-00
  2014-01-17 T 08-00-00
  2014-01-10 T 08-00-00
  ...

- For the year 2014,

value of K	RMSE	Standard Deviation
...		
5	1.601	
6	1.589	
7	1.573	2.726
8	1.584	
9	1.597	
...		

Predicting EPV production

- KNN with neighbours = production at the same time on the previous days.

  Neighbours for 2014-01-31 T 08-00-00 $\rightarrow$ 2014-01-30 T 08-00-00
  2014-01-29 T 08-00-00
  2014-01-28 T 08-00-00
  ...

- For the year 2014,

value of K	RMSE	Standard Deviation
...		
14	1.336	
15	1.334	
16	1.329	2.14
17	1.330	
18	1.331	
...		

Making use of predictions with other algorithms

- Estimate state values of each state s as $\hat{V}(s)$ using DQN.
- Let $H =$ depth to which we can predict future EPV production and future consumption.
- Model-based Value Expansion (Feinberg et al. [2018])

$$\bar{Q}_t^H(s_t, a) = \sum_{\tau=t}^{t+H-1} \bar{r}_\tau + \hat{V}(\bar{s}_{t+H})$$

using imagined trajectory as follows:
At state s_t, take action $\bar{a}_t = a$, receive reward $\bar{r}_t$ and transition to state $\bar{s}_{t+1}$ and later

$$\{\bar{a}_{t+1}, \ldots, \bar{a}_{t+h-1}\} = \mathrm{argmax} \left[\sum_{\tau=t+1}^{t+H-1} \bar{r}_\tau + \hat{V}(\bar{s}_{t+H}) \right]$$

Insights and future work

- Better prediction for EPV production? Currently, we use KNN. Literature for weather prediction suggest neural networks might give improved results.
- Prediction only upto certain depth gives improved results.
- Better estimation of $\hat{V}$.

Variational Regret Bounds for RL

Variation in MDPs

- Regret bounds in RL literature depend on the number of changes l.
- For gradual changes, change could occur at every time step.
- Definition of variation for MDP:

$$
\begin{aligned}
V_T^r &:= \sum_{t=1}^{T-1} \max_{s,a} \left| \bar{r}_{t+1}(s,a) - \bar{r}_t(s,a) \right|, \\
V_T^p &:= \sum_{t=1}^{T-1} \max_{s,a} \left\| p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a) \right\|_1.
\end{aligned}
$$

  where $\bar{r}_t(s,a) :=$ mean reward of action a in state s at time t and $p_t(s'|s,a) :=$ prob. of transition to state s' from state s after taking action a at time t.
- Regret $R_T := \sum_{t=1}^{T} \left(\rho_T^G - r_t \right)$.
  where $r_t :=$ random reward at time t and $\rho_T^G :=$ average reward of the (global) non-stationary optimal policy which knows the reward distributions and transition probabilities up to time T.

UCRL with restarts and its regret bound

UCRL with restarts

- After every $\left\lceil \frac{T^{2/3}}{(V_T^r + D V_T^p)^{2/3}} \right\rceil$ steps, start a new phase.
- Only use history from the current phase to compute estimates.

Theorem (Regret Upper Bound)

The regret of UCRL with above restarting schedule is bounded with probability $1 - \delta$ *as,*

$$R_T \leq 34 DS\sqrt{A} T^{2/3} (V_T^r + D V_T^p)^{1/3} \sqrt{\log(8T^2/\delta)} + DSA \log_2\left(\frac{8T^2}{SA}\right)$$

when $T \geq SA$.

References

Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I. Jordan, Joseph E. Gonzalez, and Sergey Levine. Model-based value estimation for efficient model-free reinforcement learning. *CoRR*, abs/1803.00101, 2018. URL http://arxiv.org/abs/1803.00101.