MSc Thesis Proposal

Stochastic Optimal Control of BESS

Dr. Xiaoyu Shen

Return Energy B.V.

x.shen@return.energy

2025.09.01

1 Background

Modern energy markets require decision tools that operate reliably under uncertainty and operational constraints. Reinforcement learning (RL) has recently gained traction as a practical, data-driven approach to sequential optimization, while established methods—such as dynamic programming (DP) and rolling intrinsic (RI)—remain widely used in energy storage operation and valuation. This thesis situates RL alongside these classical techniques within the context of energy storage contracts, emphasizing empirical performance, implementation robustness, and comparability under realistic stochastic price dynamics. The goal is to provide a clear, practice-oriented assessment of when and how RL offers tangible benefits relative to DP and RI in real-world settings.

2 Research Objectives

1. Study the literature to select a paper on reinforcement learning, replicate and improve it if needed.
2. Build two benchmarks: stochastic optimal control solver based on dynamic programming and one based on the Rolling Intrinsic algorithm.
3. Selection and calibration of the stochastic process needed for dynamic Programming.
4. Three-way comparison between Dynamic programming and stochastic control (DPSC) method, rollign intrinsic algorithm and reinforcement learning for BESS pricing.
5. Theoretical contribution: formalize the validity of the optimal storage pricing value as equivalent to optimal BESS policy.

Research Questions

1. How does a stochastic optimal control / dynamic programming (DPSC) approach perform, in terms of expected arbitrage revenue and policy structure, for a BESS trading in electricity spot markets, compared to reinforcement learning and rolling intrinsic algorithm?
2. How do different stochastic price models (e.g. regime-switching, jump-diffusion, Gaussian mixture models) affect the performance of DPSC?
3. Can a reinforcement learning method learn BESS trading policies that match or exceed the performance of the DPSC benchmark when only simulated or historical price trajectories are available, and without explicit knowledge of the underlying price process?

4. Can we combine insights from DPSC and rolling intrinsic methods to design improved RL algorithms tailored for BESS operation?

3 Methodology (Planned)

3.1 Battery Energy Storage System (BESS) Problem

We consider a battery with energy capacity $E_{\max}$, charging and discharging power limits

$$-P_{\max}^{\text{dis}} \leq u_t \leq P_{\max}^{\text{ch}},$$

and efficiency parameters $\eta_{\text{ch}}, \eta_{\text{dis}} \in (0, 1]$. The state of charge (SoC) $x_t \in [0, E_{\max}]$ evolves as

$$x_{t+1} = x_t + \Delta t \left(\eta_{\text{ch}} u_t^+ - \frac{1}{\eta_{\text{dis}}} u_t^- \right). \tag{1}$$

Let the electricity price be S_t. The instantaneous trading profit is

$$\Pi(u_t, S_t) = -S_t u_t, \tag{2}$$

where $u_t > 0$ represents buying energy and $u_t < 0$ represents selling.

The system state is defined as $s_t = (x_t, S_t)$.

A (possibly stochastic) policy π maps states to actions: $u_t = \pi(s_t)$.

The objective is to maximise expected discounted net revenue:

$$V^\pi(s_0) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \, \Pi(u_t, S_t) \,\middle|\, s_0 \right], \tag{3}$$

where $\gamma \approx 1$ is the discount factor. The optimal control problem is

$$V^*(s_0) = \sup_\pi V^\pi(s_0). \tag{4}$$

The electricity price follows a Markov process

$$S_{t+1} \sim \mathbb{P}(\cdot \mid S_t; \theta), \tag{5}$$

where θ denotes calibrated model parameters.

3.2 Electricity Price (Intraday Continuous Market) Modelling

Several candidate stochastic models can be considered for S_t.

3.2.1 Regime-Switching Mean-Reverting Process

$$dS_t = \kappa_{Z_t}(\mu_{Z_t} - S_t)dt + \sigma_{Z_t}\, dW_t, \tag{6}$$

with Z_t a finite-state Markov chain.

3.2.2 Jump-Diffusion Process

$$dS_t = \kappa(\mu - S_t)dt + \sigma\, dW_t + J_t\, dN_t, \tag{7}$$

where N_t is a Poisson process and J_t are i.i.d. jump sizes.

3.2.3 Gaussian Mixture Innovations

$$S_{t+1} = \mu + \phi S_t + \varepsilon_{t+1}, \qquad \varepsilon_{t+1} \sim \sum_{k=1}^{K} w_k \mathcal{N}(0, \sigma_k^2). \tag{8}$$

Calibration relies on maximum likelihood estimation or EM-type algorithms.

3.3 Trading Models

3.3.1 Rolling Intrinsic Model (Deterministic Benchmark)

At each time t, given the deterministic forward curve $F_t(\tau)$, the Rolling Intrinsic method solves the deterministic linear programme:

$$\begin{aligned} \max_{u_{t:T-1}} \quad & \sum_{\tau=t}^{T-1} -F_t(\tau)u_\tau \\ \text{s.t.} \quad & x_{\tau+1} = x_\tau + \Delta t \left(\eta_{\text{ch}} u_\tau^+ - \frac{1}{\eta_{\text{dis}}} u_\tau^- \right), \\ & 0 \le x_\tau \le E_{\max}, \\ & -P_{\max}^{\text{dis}} \le u_\tau \le P_{\max}^{\text{ch}}. \end{aligned} \tag{9}$$

Only the first control u_t is executed; at $t+1$ the forward curve updates and problem (9) is re-solved.

3.3.2 Dynamic Programming / Stochastic Optimal Control

When S_t is stochastic, the Bellman equation becomes

$$V_t(x_t, S_t) = \max_{u_t} \left\{ -S_t u_t + \gamma\, \mathbb{E}\left[V_{t+1}(x_{t+1}, S_{t+1}) \mid S_t \right] \right\}. \tag{10}$$

The terminal condition is $V_T(x_T, S_T) = 0$.

This can be solved by combining grid-based backward induction with numerical integration methods to evaluate the conditional expectation. Candidate techniques include:

- COS (Fourier-cosine) numerical integration techniques;
- Least-Squares Monte Carlo (LSMC) where continuation values are estimated via regression:

$$\hat{V}_{t+1}(s) \approx \sum_k \beta_k \phi_k(s).$$

The associated optimal policy is

$$\pi^*(s_t) = \arg\max_{u_t} Q^*(s_t, u_t),$$

with

$$Q^*(s_t, u_t) = -S_t u_t + \gamma \mathbb{E}[V_{t+1}(x_{t+1}, S_{t+1})].$$

3.3.3 Reinforcement Learning Benchmark

RL approximates the optimal policy or value function from sampled trajectories without requiring a transition model. A parametrised action-value function Q_θ satisfies

$$Q_\theta(s_t, u_t) \approx Q^*(s_t, u_t), \tag{11}$$

with stochastic gradient updates

$$\theta \leftarrow \theta + \alpha \nabla_\theta \left(-S_t u_t + \gamma \max_{u'} Q_\theta(s_{t+1}, u') - Q_\theta(s_t, u_t) \right)^2. \tag{12}$$

Candidate methods include Deep Q-learning, Soft Actor Critic (SAC), or policy gradient methods.

4 Timeline (Indicative)

Task	Duration
(Benchmark 1 - Dynamic Programming: both COS and LSMC numerical solutions)	Month 1-2
Modelling and Calibration of Underlying Process (regime switching, Jump-diffusion, Mixture Gaussian, etc.):	Month 2-3
Benchmark 2 - Rolling Intrinsic	Month 3-4
Literature Review and replication of Reinforcement Learning Method	Month 4-5
Initial Tests, Comparison and Analysis of the Results	Month 5-6
Model Improvements and Innovations	Month 6-8
Finalization	Month 8-9

5 References

- Boris C. Boonstra and CornelisW. Oosterlee, Valuation of electricity storage contracts using the COS method, Applied Mathematics and Computation 410 (2021) 126416.
- Rick van Voorbergen, Msc Thesis Utrecht University, Comparing Least Squares Monte Carlo with the Rolling Intrinsic strategy, for batteries trading on the intraday electricity market, 2025-09.
- Shengren Hou, PhD Thesis TU Delft, Deep reinforcement learning based optimal distribution networks operation, 2025-08.