Learn to Bid: Deep Reinforcement Learning with Transformer for Energy Storage Bidding in Energy and Contingency Reserve Markets

Jinhao Li
Department of Data Science and AI
Monash University
stephlee175@gmail.com

Changlong Wang
Department of Civil Engineering
Monash University
chang.wang@monash.edu

Yanru Zhang
School of Computer Science & Engineering
UESTC
yanruzhang@uestc.edu.cn

Hao Wang*
Department of Data Science and AI
and Monash Energy Institute
Monash University
hao.wang2@monash.edu

Abstract

As part of efforts to tackle climate change, grid-scale battery energy storage systems (BESS) play an essential role in facilitating reliable and secure power system operation with variable renewable energy (VRE). BESS can balance time-varying electricity demand and supply in the spot market through energy arbitrage and in the frequency control ancillary services (FCAS) market through service enablement or delivery. Effective algorithms are needed for the optimal participation of BESS in multiple markets. Using deep reinforcement learning (DRL), we present a BESS bidding strategy in the joint spot and contingency FCAS markets, leveraging a transformer-based temporal feature extractor to exploit the temporal trends of volatile energy prices. We validate our strategy on real-world historical energy prices in the Australian National Electricity Market (NEM). We demonstrate that the novel DRL-based bidding strategy significantly outperforms benchmarks. The simulation also reveals that the joint bidding in both the spot and contingency FCAS markets can yield a much higher profit than in individual markets. Our work provides a viable use case for the BESS, contributing to the power system operation with high penetration of renewables.

1 Introduction and Background

Global warming will likely exceed 1.5 degrees Celsius in the 21st century, despite the nationally determined contributions (NDCs) committed before the 2021 United Nations Climate Change Conference (COP26) [1]. Mitigation efforts must be accelerated more urgently and rapidly [1]. As the main pillar for decarbonization, variable renewable energy (VRE) generation has been increasingly adopted in modern power systems [2]. This has also called for more energy storage to balance the increasing VRE generation for system reliability and security [2, 3]. Battery energy storage systems (BESS) can swiftly switch between two working modes, i.e., discharge and charge (storage) [4], in response

*Corresponding author: Hao Wang <hao.wang2@monash.edu>. Hao Wang's research has been supported in part by the FIT Startup Funding of Monash University and the Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) under Grant DE230100046.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

to the mismatch between VRE generation and electricity consumption for system reliability. Such demand-supply mismatches are reflected by price fluctuations in the real-time electricity spot market [5], which creates the financial incentive for the BESS to enter into the spot market for price arbitrage (i.e., buy low and sell high). Balancing demand-supply mismatch to minimize price volatility and market distortion is also in the consumer's interest. Apart from spot market participation, the BESS can also be financially rewarded for providing frequency control ancillary services (FCAS) [6] in the FCAS market for maintaining system security. Considering the multiple revenue streams the BESS is exposed to, optimal scheduling to participate in both spot and FCAS markets at the same time (i.e., joint bidding) is critical to unlocking the BESS' full potential in supporting a high VRE power system. Joint bidding is, however, challenging because the BESS has limited capacity, and energy prices are highly stochastic.

In addressing the concern above, the most explored approach in the literature is to derive real-time bidding strategies through stochastic optimization [7, 8], whose performance heavily relies on the quality of price forecasting. It is, however, notoriously difficult to forecast energy prices due to the high volatility of the spot and FCAS markets [9] and complex price drivers. Alternatively, deep reinforcement learning (DRL)-based methods have drawn increasing attention lately for their model-free paradigm and data-driven characteristics [10–12]. DRL can adaptively capture the dynamics of the electricity market in an online manner since it learns from historical data and past experiences. Such learned dynamics would enable the DRL to yield a fast and better response even facing frequent unexpected shifts in the underlying distribution of market prices [13].

However, it appears that there is a significant research gap in the literature. BESS joint bidding in multiple markets has not been adequately investigated [7, 8, 14, 15], particularly in contingency FCAS. The existing DRL-based methods [16, 17, 10–12, 18] tend to overlook the hidden temporal information inside the volatile streaming market prices. The novelty of our work shows that extracting the inherent temporal patterns in the underlying market prices will make the bidding strategy aware of and sensitive to recent changes in energy prices, thereby making better bidding decisions.

In particular, we develop a temporal-aware DRL-based bidding strategy for the BESS participating in both the energy spot (ES) and contingency FCAS markets. It leverages a transformer-based temporal feature extractor (TTFE) to fully unlock the value of multiple streaming energy prices in both markets. The new temporal-aware approach will help the BESS optimize charge/discharge for energy arbitrage and bid capacity for FCAS service enablement to maximize the overall revenue. We validate our method using the realistic electricity market data collected from the Australian National Electricity Market (NEM) [19], which supplies around 9 million customers with a trading value of 16.6 billion Australian dollars per annum. We present our DRL-based bidding method in Section 2.

2 Methodology

To optimize the joint bidding of the BESS in both spot and contingency FCAS markets, we develop a novel temporal-aware DRL-based bidding strategy with the help of TTFE as a feature extraction technique based on the transformer [20] to capture temporal patterns from time-series market prices. We introduce the TTFE in Section 2.1, followed by Section 2.2, where we formulate the continuous bidding problem as a Markov decision process and then introduce the soft actor-critic (SAC) algorithm to learn an optimal joint-bidding strategy. The framework of the developed joint-bidding strategy is illustrated in Fig. 1.

2.1 Transformer-based Temporal Feature Extractor

Through the extraction of temporal patterns by TTFE, the bidding strategy can be made more aware of and sensitive to changes in energy prices. In the spot market, detecting recent price fluctuations will assist the BESS in energy arbitrage, especially in discharging at higher prices. Whereas, being sensitive to recent price changes in the contingency FCAS market, the BESS can reserve enough capacity in advance in the event of a contingency.

We denote a price vector $\boldsymbol{\rho}_t$ at each bidding interval (i.e., 5 minutes) containing market prices in the spot and contingency FCAS markets. A temporal segment with length L is developed to store a series of price vectors, including the latest L price vectors, which can be formulated as

$$S_t = (\boldsymbol{\rho}_{t-L+1}, \boldsymbol{\rho}_{t-L+2}, \cdots, \boldsymbol{\rho}_t) \in \mathbb{R}^{L \times F}, \tag{1}$$

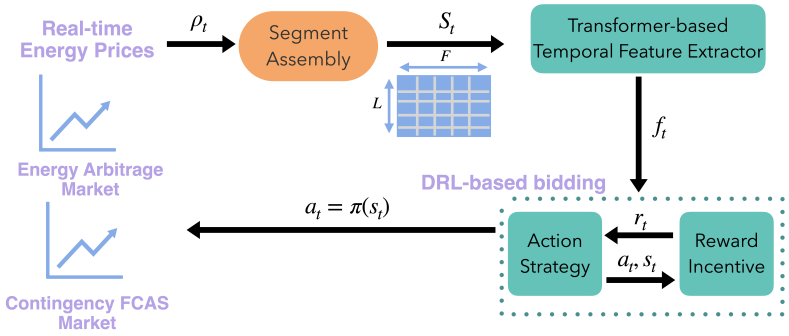


Figure 1: The framework of temporal-aware DRL-based joint bidding strategy.

where F represents the number of features, i.e., the number of participated markets.

To extract temporal patterns of market prices inside the temporal segment S_t, we introduce the transformer for its powerful capability in handling temporal sequences to explore mutual influence on market prices. Since the output of the transformer encoder is two-dimensional, i.e., the attention matrix, we apply one-dimensional global average pooling [21] to achieve feature extraction and aggregation. For each bidding interval, we prepare such a temporal segment defined in Eq. (1), feed it into the TTFE, and obtain the final extracted feature vector for the following DRL algorithm to bid, which can be formulated as

$$\begin{aligned} \boldsymbol{f}_t &= \text{TTFE}\left(S_t\right) \\ &= \text{Pooling}\left[\text{TransformerEncoder}\left(S_t\right)\right] \in \mathbb{R}^{1\times F'}, \end{aligned} \tag{2}$$

where F' is the extracted feature dimension.

2.2 Learning Joint Bidding Strategy via DRL

The optimal joint bidding problem requires consecutive decision-making. We model such a continuous process as a dynamic Markov decision process, consisting of four essential parts: state space $\mathbb{S}$, action space $\mathbb{A}$, probability space $\mathbb{P}$, and reward space $\mathbb{R}$ [22].

The BESS's state is defined as the aggregation of the latest energy prices and the extracted temporal features defined in Eq. (2), along with the BESS's state of charge (SoC), denoted by $\boldsymbol{s}_t = (\text{SoC}_t, \boldsymbol{\rho}_t, \boldsymbol{f}_t)$. For the action space $\mathbb{A}$, the BESS takes action $\boldsymbol{a}_t$ to allocate the current battery capacity to bid in the ES and contingency FCAS markets. In DRL, an action strategy, expressed as $\pi : \mathbb{S} \to \pi(\mathbb{A})$, is commonly applied to learn how to process different states and estimated using function approximators [23], i.e., neural networks, which map states to a probabilistic distribution over actions. Designing an appropriate reward function $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$ plays a significant role in optimizing the proposed MDP for BESS revenue maximization in real-time bidding since the reward function assesses the quality of both the current state $\boldsymbol{s}_t$ and the selected action $\pi(\boldsymbol{s}_t)$. Details about the reward function are presented in Appendix A. We use SAC [24] to solve the developed MDP by maximizing the expected total rewards, which is formulated as

$$\max\ J(\pi) = \max \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{s}_t\sim\mathbb{S},\boldsymbol{a}_t\sim\pi}\left[r\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)\right], \tag{3}$$

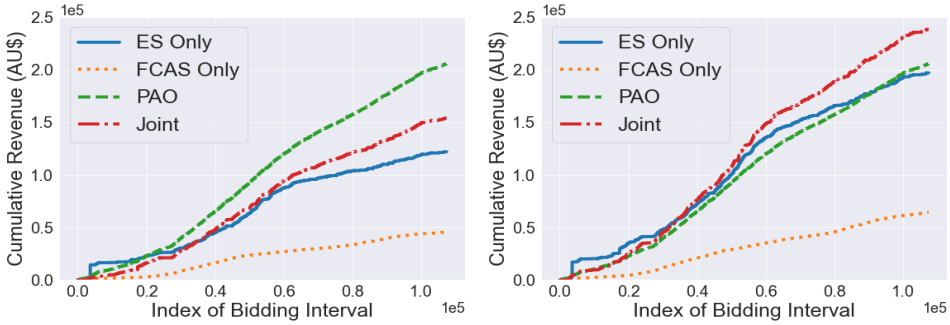
where T is the total number of bidding intervals. The associated algorithmic procedure is presented in Appendix B.

3 Experimental Results

The proposed temporal-aware DRL-based bidding strategy is trained and evaluated using real-world historical prices from 2016 to 2017 in the Victoria jurisdiction of the NEM. Specifically, energy prices in 2016 are used for training the bidding strategy via SAC; we then test the learned strategy on 2017 energy prices. The energy prices in the spot and contingency FCAS markets are collected every

Table 1: Cumulative revenue of bidding strategies trained with/without the TTFE.

Bid Scenario	Without TTFE	**With TTFE**	Boost
ES Market	AU$ $122,005$ (± 952)	AU$ $197,157$ (± 584)	62%
Contingency FCAS Market	AU$ $45,526$ (± 8)	AU$ $64,219$ (± 37)	41%
Joint Market	AU$ $\mathbf{153,952}$ (± 202)	AU$ $\mathbf{238,608}$ (± 349)	55%



(a) Strategy without TTFE.

(b) Strategy with TTFE.

Figure 2: Evaluation results of bidding strategies trained with/without the TTFE.

5 minutes, i.e., the time resolution for one bidding interval. We use 1 Nvidia TITAN RTX graphics process unit for training the DRL algorithms.

We compare three scenarios where the BESS can bid: 1) ES market only; 2) contingency FCAS market only; 3) a joint market. To examine the effectiveness of the proposed TTFE, we trained the DRL-based strategies without/with the TTFE. In addition, we developed a *predict-and-optimize* (PAO) method [25] to compare our proposed strategy with the benchmark. The PAO method relies on an LSTM network to forecast one-interval-ahead energy prices and a mixed integer linear programming solver (from the PuLP library [26]) to optimize the cumulative revenue.

The corresponding results in VIC are illustrated in Fig. 2a and 2b, respectively. The associated cumulative revenue is also presented in Table 1 for cross comparison. We trained DRL-based strategies 6 times to mitigate the randomness of the DRL algorithm with their corresponding averages and standard deviations shown in Table 1.

From Fig. 2a, joint-market bidding consistently generates higher revenue than individual market participation. This is because simultaneous joint bidding can maximize the full potential of the BESS and take advantage of the flexibility in both markets, which results in higher revenue.

Most importantly, introducing the TTFE can substantially improve the bidding performance and revenue creation in all the three bidding scenarios. What stands out in Fig. 2b is the significant revenue boost after introducing TTFE in joint markets (shown in the red dash-dotted line). This considerable improvement has surpassed the PAO method (shown in the green dashed line), where our DRL-based strategy excels by approximately 16% (AU$$32,848$ in total).

4 Conclusion

In this paper, we developed a model-free DRL-based strategy for bidding in energy spot and contingency FCAS markets to maximize revenue in real time. The TTFE captures temporal patterns of energy prices in both markets, allowing DRL-based bidding strategies to be aware of and sensitive to price changes. Based on the modelling results, we can draw two major conclusions: 1) bidding in joint markets can dramatically improve the viability of the BESS; and 2) the proposed TTFE empowers the DRL-based bidding strategy to make better decisions, with outcomes significantly outperforming the PAO benchmark. Future work will factor in battery degradation in joint bidding and study bidding in other four jurisdictions of the NEM.

References

[1] M. Meinshausen, J. Lewis, C. McGlade, J. Gütschow, Z. Nicholls, R. Burdon, L. Cozzi, and B. Hackmann, "Realization of Paris Agreement pledges may limit warming just below 2 °C," *Nature*, vol. 604, no. 7905, pp. 304–309, apr 2022. [Online]. Available: https://www.nature.com/articles/s41586-022-04553-z

[2] IPCC, *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, P. Shukla, J. Skea, R. Slade, A. A. Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley, Eds. Cambridge, UK and New York, NY, USA: Cambridge University Press, 2022.

[3] ARENA, "Large-scale battery storage knowledge sharing report," 2019. [Online]. Available: https://arena.gov.au/knowledge-bank/large-scale-battery-storage-knowledge-sharing-report/

[4] S. R. Sinsel, R. L. Riemke, and V. H. Hoffmann, "Challenges and solution technologies for the integration of variable renewable energy sources—a review," *Renewable Energy*, vol. 145, pp. 2271–2285, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148119309875

[5] Q. Wang, C. Zhang, Y. Ding, G. Xydis, J. Wang, and J. Østergaard, "Review of real-time electricity markets for integrating distributed energy resources and demand response," *Applied Energy*, vol. 138, pp. 695–706, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261914010988

[6] AEMO, "Guide to ancillary services in the national electricity market," 2015. [Online]. Available: https://aemo.com.au/energy-systems/electricity/national-electricity-market-nem/system-operations/ancillary-services

[7] K. Abdulla, J. de Hoog, V. Muenzel, F. Suits, K. Steer, A. Wirth, and S. Halgamuge, "Optimal operation of energy storage systems considering forecasts and battery degradation," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2086–2096, 2018.

[8] D. Krishnamurthy, C. Uckun, Z. Zhou, P. R. Thimmapuram, and A. Botterud, "Energy storage arbitrage under day-ahead and real-time price uncertainty," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 84–93, 2018.

[9] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *International Journal of Forecasting*, vol. 30, no. 4, pp. 1030–1081, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207014001083

[10] J. Cao, D. Harrold, Z. Fan, T. Morstyn, D. Healey, and K. Li, "Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4513–4521, 2020.

[11] B. Huang and J. Wang, "Deep-reinforcement-learning-based capacity scheduling for pv-battery storage system," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2272–2283, 2021.

[12] X. Wei, Y. Xiang, J. Li, and X. Zhang, "Self-dispatch of wind-storage integrated system: A deep reinforcement learning approach," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 3, pp. 1861–1864, 2022.

[13] M. Anwar, C. Wang, F. de Nijs, and H. Wang, "Proximal policy optimization based reinforcement learning for joint bidding in energy and frequency regulation markets," *IEEE Power & Energy Society General Meeting (PESGM)*, 2022.

[14] A. A. Mohamed, R. J. Best, X. Liu, and D. J. Morrow, "Single electricity market forecasting and energy arbitrage maximization framework," *IET Renewable Power Generation*, vol. 16, no. 1, pp. 105–124, 2022. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/rpg2.12345

[15] D. R. Jiang and W. B. Powell, "Optimal hour-ahead bidding in the real-time electricity market with battery storage using approximate dynamic programming," *INFORMS J. on Computing*, vol. 27, no. 3, p. 525–543, aug 2015.

[16] H. Wang and B. Zhang, "Energy storage arbitrage in real-time markets via reinforcement learning," *IEEE Power & Energy Society General Meeting (PESGM)*, 2018.

[17] H. Xu, X. Li, X. Zhang, and J. Zhang, "Arbitrage of energy storage in electricity markets with deep reinforcement learning," *CoRR*, vol. abs/1904.12232, 2019. [Online]. Available: http://arxiv.org/abs/1904.12232

[18] V.-H. Bui, A. Hussain, and H.-M. Kim, "Double deep q -learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 457–469, 2020.

[19] AEMO, "About the national electricity market (nem)," 2020. [Online]. Available: https://aemo.com.au/energy-systems/electricity/national-electricity-market-nem/about-the-national-electricity-market-nem

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[21] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013. [Online]. Available: http://arxiv.org/abs/1312.4400

[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[24] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: https://proceedings.mlr.press/v80/haarnoja18b.html

[25] A. N. Elmachtoub and P. Grigas, "Smart "predict, then optimize"," 2017. [Online]. Available: https://arxiv.org/abs/1710.08005

[26] S. Mitchell, M. J. O'Sullivan, and I. Dunning, "Pulp : A linear programming toolkit for python," 2011.

Appendix

A: Reward Function Design in Optimal Bidding Strategy

$$r\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) = r_t^{\mathrm{ES}} + b_t^{\mathrm{dch}}\left(r_t^{\mathrm{FR}} + r_t^{\mathrm{SR}} + r_t^{\mathrm{DR}}\right) + b_t^{\mathrm{ch}}\left(r_t^{\mathrm{FL}} + r_t^{\mathrm{SL}} + r_t^{\mathrm{DL}}\right), \tag{4}$$

$$\begin{aligned} r_t^{\mathrm{ES}} &= p_t^{\mathrm{ES}}\rho_t^{\mathrm{ES}}\left(b_t^{\mathrm{dch}}\eta^{\mathrm{dch}} - b_t^{\mathrm{ch}}\frac{1}{\eta^{\mathrm{ch}}}\right), \\ &\quad + p_t^{\mathrm{ES}}|\rho_t^{\mathrm{ES}} - \bar{\rho}_t^{\mathrm{ES}}|\left(b_t^{\mathrm{ch}}\mathbb{I}^{\mathrm{ch}}\frac{1}{\eta^{\mathrm{ch}}} + b_t^{\mathrm{dch}}\mathbb{I}^{\mathrm{dch}}\eta^{\mathrm{dch}}\right), \end{aligned} \tag{5}$$

$$r_t^{\mathrm{FR}} = b_t^{\mathrm{dch}}\eta^{\mathrm{dch}}p_t^{\mathrm{FR}}\rho_t^{\mathrm{FR}}, \tag{6}$$

$$r_t^{\mathrm{FL}} = b_t^{\mathrm{ch}}\frac{1}{\eta^{\mathrm{ch}}}p_t^{\mathrm{FL}}\rho_t^{\mathrm{FL}}, \tag{7}$$

$$r_t^{\mathrm{SR}} = b_t^{\mathrm{dch}}\eta^{\mathrm{dch}}p_t^{\mathrm{SR}}\rho_t^{\mathrm{SR}}, \tag{8}$$

$$r_t^{\mathrm{SL}} = b_t^{\mathrm{ch}}\frac{1}{\eta^{\mathrm{ch}}}p_t^{\mathrm{SL}}\rho_t^{\mathrm{SL}}, \tag{9}$$

$$r_t^{\mathrm{DR}} = b_t^{\mathrm{dch}}\eta^{\mathrm{dch}}p_t^{\mathrm{DR}}\rho_t^{\mathrm{DR}}, \tag{10}$$

$$r_t^{\mathrm{DL}} = b_t^{\mathrm{ch}}\frac{1}{\eta^{\mathrm{ch}}}p_t^{\mathrm{DL}}\rho_t^{\mathrm{DL}}. \tag{11}$$

ES, FR, FL, SR, SL, DR, and DL are abbreviations of energy arbitrage market, fast-raise contingency FCAS market, fast-lower contingency FCAS market, slow-raise contingency FCAS market, slow-lower contingency FCAS market, delay-raise contingency FCAS market, and delay-lower contingency FCAS market. b_t^{dch}/b_t^{ch} are binary variables to determine discharging/charging operations of the BESS. ρ_t indicates the market clearing price (MCP). p_t is the allocated capacity for bidding in the corresponding market. η^{dch} and η^{ch} are denoted as discharging/charging efficiencies. $\bar{\rho}_t^{\mathrm{ES}}$ represents the exponential moving average energy prices, formulated as

$$\bar{\rho}_t^{\mathrm{ES}} = \lambda\bar{\rho}_{t-1}^{\mathrm{ES}} + (1-\lambda)\,\bar{\rho}_t^{\mathrm{ES}}, \tag{12}$$

where λ is the smoothing parameter.

$\mathbb{I}^{\mathrm{ch}}$ and $\mathbb{I}^{\mathrm{dch}}$ in Eq. (5) are indicators for the BESS to learn when to charge/discharge, formulated as

$$\mathbb{I}^{\mathrm{ch}} = \begin{cases} -1, & \rho_t^{\mathrm{ES}} > \bar{\rho}_t^{\mathrm{ES}}, \\ 0, & \rho_t^{\mathrm{ES}} = \bar{\rho}_t^{\mathrm{ES}}, \\ 1, & \rho_t^{\mathrm{ES}} < \bar{\rho}_t^{\mathrm{ES}}, \end{cases} \tag{13}$$

$$\mathbb{I}^{\mathrm{dch}} = \begin{cases} -1, & \rho_t^{\mathrm{ES}} < \bar{\rho}_t^{\mathrm{ES}}, \\ 0, & \rho_t^{\mathrm{ES}} = \bar{\rho}_t^{\mathrm{ES}}, \\ 1, & \rho_t^{\mathrm{ES}} > \bar{\rho}_t^{\mathrm{ES}}, \end{cases} \tag{14}$$

B: The Algorithmic Procedure of the Temporal-aware DRL-based Bidding Strategy

The TTFE can be considered as a preprocessing unit for its sequential SAC algorithm. Combining the output of TTFE $\boldsymbol{f}_t$, i.e., extracted temporal features of market prices, with market prices $\boldsymbol{\rho}_t$ and the BESS's SoC, we formulate the state of the proposed MDP in Section 2.2 as

$$\boldsymbol{s}_t = (\text{SoC}_t, \boldsymbol{\rho}_t, \boldsymbol{f}_t)\,. \tag{15}$$

The SAC algorithm defines an action strategy network π_ϕ to allocate the BESS's capacity for bidding in the energy spot market and contingency FCAS market. The value network V_ψ and Q network Q_θ are proposed to assess the quality of bidding decisions and current states. We use gradient descent to update the above three networks. The detailed algorithmic procedure of our DRL-based bidding strategy is presented in Algorithm 1.

Algorithm 1 The Temporal-aware DRL-based bidding strategy

Initialise parameters of the TTFE.
Initialise parameters of the action strategy network ϕ, value network ψ, and Q network θ.
Initialise target value network $\hat{\psi}$ with ψ: $\hat{\psi} \leftarrow \psi$.
Initialise the replay buffer $\mathbb{B}$.
for $t = 1, \cdots, T$ **do**
  Feed the temporal segment S_t into TTFE, and obtain the extracted feature vector $\boldsymbol{f}$.
  Prepare the current state $\boldsymbol{s}_t$.
  Get action $\boldsymbol{a}_t = \pi_\theta(\boldsymbol{s}_t)$ and reward r_t.
  if action violates capacity constraint **then**
    $\boldsymbol{a}_t \leftarrow \mathbf{0}$.
  end if
  Transit into the next state $\boldsymbol{s}_{t+1}$ via $\mathbb{P}$.
  Store transition $\{\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1}\}$ into replay buffer $\mathbb{B}$.
  if collect sufficient transitions **then**
    Update π_ϕ, V_ψ, and Q_θ using gradient descent
    $\phi \leftarrow \phi - \eta_\pi \nabla_\phi J_\pi(\phi)$,
    $\psi \leftarrow \psi - \eta_V \nabla_V J_V(\psi)$,
    $\theta \leftarrow \theta - \eta_Q \nabla_Q J_Q(\theta)$.
    Update target value network $V_{\hat{\psi}}$
    $\hat{\psi} \leftarrow \tau\psi + (1-\tau)\hat{\psi}$.
  end if
end for