

What is MORL?

- Real-world building energy management tasks involve **conflicting objectives** (e.g., users' comfort vs. energy)
- MORL aims to obtain either a single policy readily adapted to various preferences or **a set of policies aligned with their respective preferences**
- A policy is **Pareto-optimal** if and only if it is not dominated by any other policies
- Our goal is then to find a multi-objective policy $\pi(a|s, \omega)$ such that the expected scalarized return $\omega^\top G^\pi$ is maximized, where $G^\pi = [G_1^\pi, G_2^\pi, \dots, G_n^\pi]^\top$, and the expected return of the i^{th} objective is given as $G_i^\pi = \mathbb{E}_{a_{t+1} \sim \pi(\cdot|s_t)} [\sum_t \gamma^t \mathcal{R}(s_t, a_t)_i]$ for some predefined time horizon

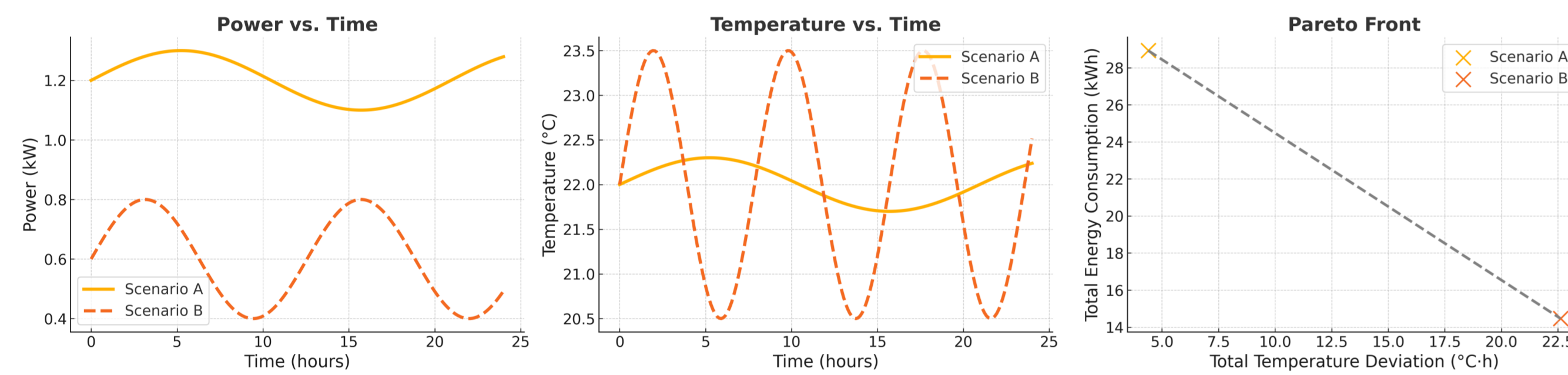


Figure 1. Illustration of Pareto optimal—Scenario A offers stable temperature but consumes more energy; Scenario B saves energy but sacrifices temperature stability.

Research Gaps and Our Solutions

- Gap 1:** low training efficiency
- Solution 1:** an efficient two-stage method
- Gap 2:** hard to cover the complete Pareto front
- Solution 2:** a novel constrained optimization formulation
- Gap 3:** inability to maximize utility for any given preference
- Solution 3:** assign a policy from the Pareto set that maximizes its utility for any given preference

Results Highlights

- +35% Hypervolume Gain and +9% Higher Expected Utility:** C-MORL achieves up to 35% higher hypervolume and 9% higher expected utility compared to baselines, indicating better Pareto front coverage
- Linear Time Complexity:** unlike traditional ϵ -constraint methods, C-MORL scales linearly with the number of objectives.
- Generalizes to up to 9 Objectives**



Code



Paper

Contact: ruohong.liu@eng.ox.ac.uk

How C-MORL Works

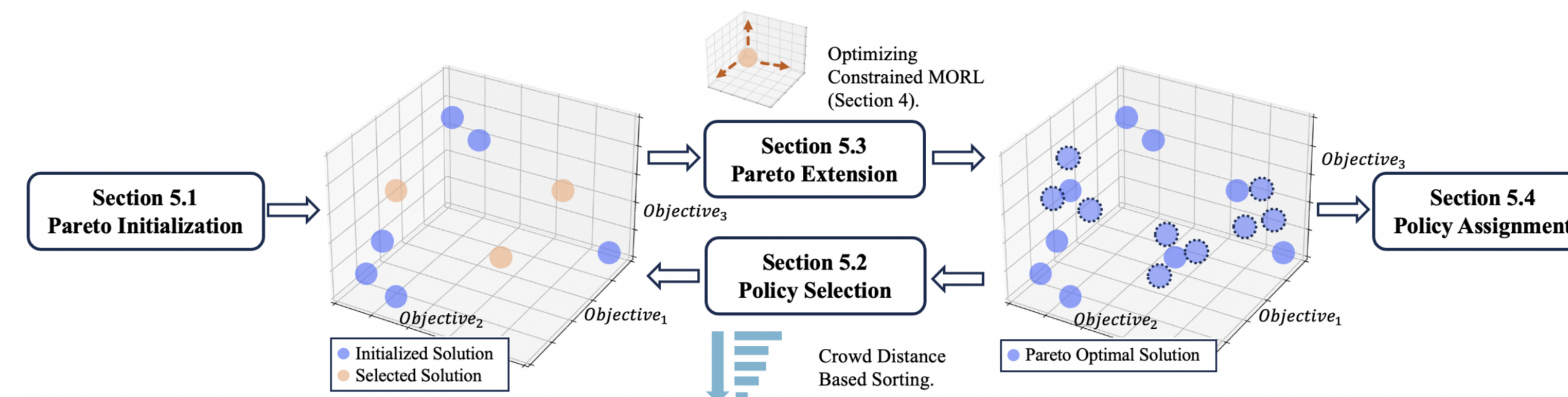


Figure 2. C-MORL workflow.

Pareto initialization: train several initial policies to derive the initial solution set \mathcal{X}_{init} .

Policy selection: the Pareto-optimal policies are selected based on their crowd distance:

$$D(P(j)) = \sum_{i=1}^n \frac{\tilde{G}_i(k+1) - \tilde{G}_i(k-1)}{\tilde{G}_{i,max} - \tilde{G}_{i,min}}. \quad (1)$$

Policies with larger crowd distance represent under-explored regions of the Pareto front and are prioritized for expansion.

Pareto extension: fill the Pareto front towards different directions by solving constrained optimization on the selected policies.

- Convert the MORL problem as a constrained MDP:

$$\max_{\pi} G_l^\pi \quad \text{s.t.} \quad G_i^\pi \geq d_i \quad i = 1, \dots, n, i \neq l. \quad (2)$$

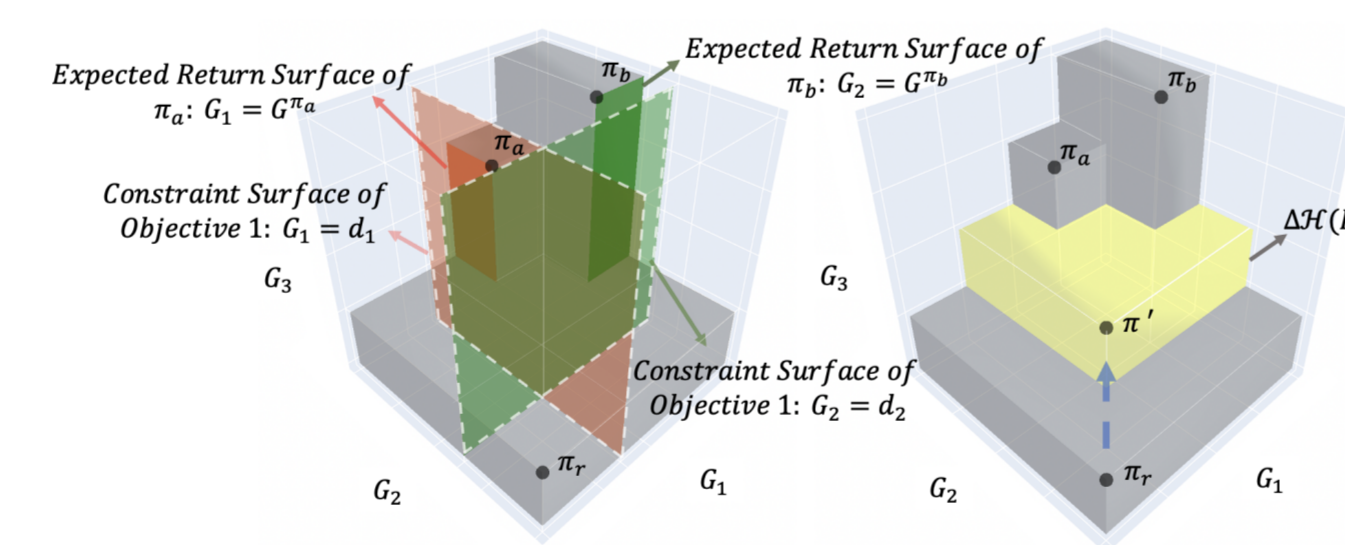


Figure 3. Visualization of criteria for specifying constraint values. π_r denotes initial point. The expected return $G^{\pi_a}(G_1^a)$ of solution $P_a(P_b)$ in objective 1(2) is the $(k-1)^{th}$ value in list $\tilde{G}_1(\tilde{G}_2)$, respectively. Therefore, specifying constraints values $d_1 \geq G^{\pi_a}$ and $d_2 \geq G^{\pi_b}$ is sufficient for the feasible solution of corresponding Eq. 2 to be Pareto-optimal solution.

- Specify appropriate constraint values: utilize only the expected return of the policy from the most recent step and an adjustable parameter β :

$$\pi_{r+1,i} = \arg \max_{\pi \in \Pi_\theta} \{G_l^\pi : G_i^\pi \geq \beta G_i^{\pi_r}, i = 1, \dots, n, i \neq l\}. \quad (3)$$

- Solve constrained policy optimization: solve the problem above via the interior point method (IPO):

$$\max_{\pi} G_l^\pi + \sum_{i=1}^n \phi(G_i^\pi) \quad (4)$$

Policy assignment: given preference ω , the surrogate execution policy selected to maximize its utility:

$$\pi_\omega^{SMP} = \max_{\pi \in \Pi_P} f_\omega(G^\pi). \quad (5)$$

Experiments

- Building-3d:** A building thermal control environment that controls the temperature of a commercial building with 23 zones. The conflicting objectives are minimizing energy cost, reducing the temperature difference, and managing the ramping rate of power consumption:

$$\mathcal{R}_1 = M - 0.05 * \sum_i |T_i[t] - T_{i,user}[t]|;$$

$$\mathcal{R}_2 = M - 0.05 * \sum_i c[t]|P_i[t]|;$$

$$\mathcal{R}_3 = M - |(\sum_i |P_i[t]| - \sum_i |P_i[t-1]|)|.$$

- Building-9d:** A modified version of Building-3d. Instead of calculating the reward based on all zones collectively, this version evaluates the reward for each of the three floors of the commercial building separately. Consequently, this results in a total of $3 \times 3 = 9$ objectives.

Table: Evaluation of HV, EU, and SP for MORL tasks. T/O indicates that the training time exceeded the maximum limit of 100 hours.

Environments	Metrics	CAPQL	Q-Pensieve	PG-MORL	GPI-LS	MORL/D	C-MORL
Building-3d	HV(10^{12})	0.33±0.18	1.00±0.02	0.83±0.02	0.26±0.04	0.87±0.38	1.15±0.00
	EU(10^4)	0.75±0.09	0.96±0.00	0.93±0.01	0.74±0.01	0.95±0.00	1.02±0.00
	SP(10^5)	0.18±0.08	0.92±0.78	0.04±0.02	0.07±0.09	7.31±2.20	0.07±0.06
Building-9d	HV(10^{31})	4.29±0.73	7.28±0.57	T/O	T/O	T/O	7.93±0.07
	EU(10^3)	3.31±0.06	3.46±0.03	T/O	T/O	T/O	3.50±0.00
	SP(10^3)	4.34±3.72	1.04±0.38	T/O	T/O	T/O	2.79±0.40

Future Research Directions

- Extend C-MORL to real-world sustainability challenges in power systems
- Develop generalizable MORL methods across diverse environments

References

- Lucas Nunes Alegre, Ana Bazzan, and Bruno C Da Silva. Optimistic linear support and successor features as a basis for optimal policy transfer. In International Conference on Machine Learning, pp. 394–413. PMLR, 2022.
- Florian Felten, Lucas N. Alegre, Ann Now´e, Ana L. C. Bazzan, El Ghazali Talbi, Gr´egoire Danoy, and Bruno Castro da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), 2023a.
- Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pp. 4940–4947, 2020.
- Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. Prediction-guided multi-objective reinforcement learning for continuous robot control. In International conference on machine learning, pp. 10607–10616. PMLR, 2020.