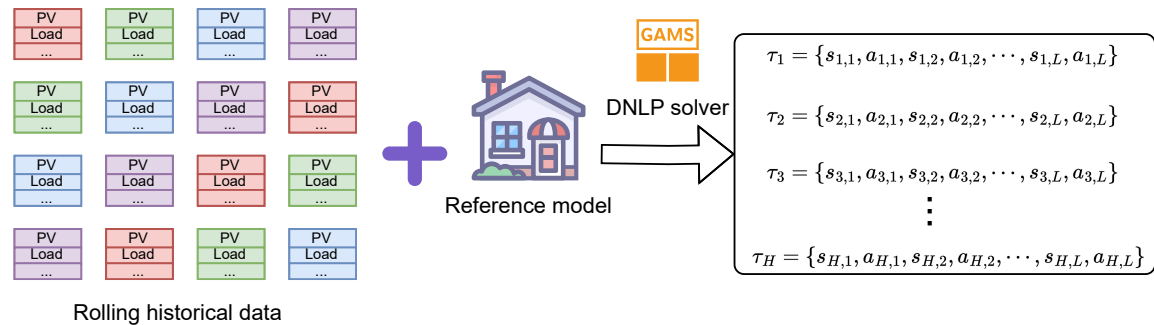
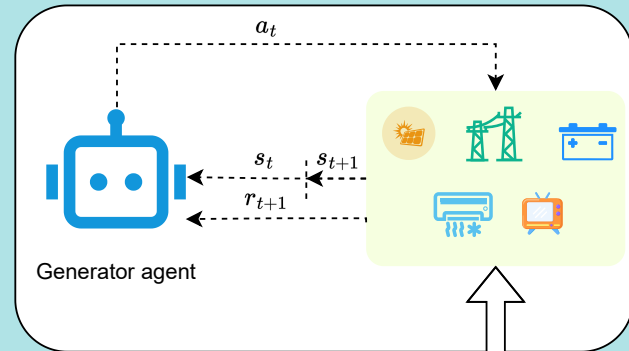


Part I Generate multiple expert trajectories based on optimization solver



Part III Learn operation strategy leveraging reward function learned



Sample expert trajectories

Sample generator trajectories

$$s_1^e, a_1^e, s_1^{e'}, s_2^e, a_2^e, s_2^{e'}, \dots, s_N^e, a_N^e, s_N^{e'} \quad s_1^g, a_1^g, s_1^{g'}, s_2^g, a_2^g, s_2^{g'}, \dots, s_N^g, a_N^g, s_N^{g'}$$

$$\text{Discriminator: } \hat{d}_i = \frac{\exp r_\phi(s_i, a_i, s_i')}{\exp r_\phi(s_i, a_i, s_i') + \pi(a_i | s_i)}$$

$$\hat{d}_1^e, \hat{d}_2^e, \dots, \hat{d}_N^e$$

$$1, 1, \dots, 1$$

$$\hat{d}_1^g, \hat{d}_2^g, \dots, \hat{d}_N^g$$

$$0, 0, \dots, 0$$

Calculate binary cross-entropy reward loss and optimize:

$$\mathcal{L}_D = -\frac{1}{N} \sum_{i=1}^N [d_i \log \hat{d}_i + (1 - d_i) \log(1 - \hat{d}_i)]$$

Part II Learn reward function from expert trajectories rather than manually design

