
EPISODIC SPARSE COST EVALUATION FOR POLICY ANALYSIS IN STOCHASTIC SHORTEST PATH PROBLEMS

**Fahmi Izhari^{1*}, Rafi Septiawan Putra²,
Hasanal Fachri Satia Simbolon³, Ade Linhar P⁴**

^{1,2,3,4}UIN Syekh Ali Hasan Ahmad Addary, Padangsidempuan

⁵Universitas Mandiri Bina Prestasi, Medan

¹fahmi_izhari@uinsyahada.ac.id

Abstrak

Conventional evaluations of stochastic shortest path policies typically rely on dense reward or cost signals, which often obscure rare but behaviorally critical interactions. This paper introduces an episodic sparse-cost evaluation framework that assigns costs only to a small subset of state action pairs identified through a short probing phase, thereby decoupling cost accumulation from trajectory length. The objective of this study is to assess whether episodic sparse costs can provide a more interpretable and behavior-focused evaluation of policy execution compared to dense formulations. The framework is empirically validated through controlled navigation experiments under a fixed policy in a grid-based stochastic shortest path setting. In a representative episode, the agent successfully reached the terminal state in 95 steps, while incurring only two cost-triggering events drawn from a sparse support set of size five. This resulted in a total episodic cost of 2.0 and a hit rate of 0.021, indicating that more than 97% of agent environment interactions were cost-free. The temporal distribution of costs appeared as isolated impulses rather than continuous signals, enabling precise localization of critical decision points along the trajectory. These findings demonstrate that episodic sparse-cost evaluation yields bounded, event driven cost behavior that remains stable even for long trajectories. The proposed framework offers a transparent and scalable alternative for analyzing policy behavior in stochastic environments, particularly in

Episodic Sparse Cost Evaluation or Policy Analysis in Stochastic Shortest Path Problems

scenarios where rare violations, constraints, or risk sensitive interactions are of primary concern. Future research will extend this evaluation paradigm to multi-episode analysis, adaptive policies, and integration with constraint aware learning objectives.

Kata Kunci : episodic sparse cost; stochastic shortest path; policy evaluation; grid based navigation; rare event analysis; interpretable reinforcement learning; constraint aware decision making.

1. Pendahuluan

Sequential decision-making under uncertainty is a fundamental problem in artificial intelligence and reinforcement learning, particularly in domains where an agent must reach a target state while minimizing cumulative cost. One of the most established mathematical formulations for such problems is the stochastic shortest path (SSP) framework, which models goal directed behavior through a sequence of state transitions associated with costs until a terminal condition is achieved (Bertazzi, Mogre, & Trichakis, 2024; da Silva, Ramos, & Barbosa, 2023a, 2023b; Karia, Nayyar, & Srivastava, 2022). SSP models have been widely applied in navigation, planning, and safety-critical systems due to their ability to represent both uncertainty and long-term cost accumulation.

In most reinforcement learning studies, cost or reward signals are assumed to be dense, meaning that feedback is provided at nearly every interaction step. While this assumption simplifies learning and evaluation, it does not accurately capture many real-world conditions. In practical environments, penalties often arise only in exceptional or undesirable situations, such as collisions, unsafe actions, or constraint violations (Wang & Liang, 2025; Xu & Sankar, 2024). As a result, the underlying cost structure is inherently sparse, with only a limited number of state-action pairs producing non-zero costs.

Recent theoretical research has highlighted the importance of sparsity in SSP and online learning problems. Several studies

demonstrate that when costs are supported on a small subset of state-action pairs, the complexity of the decision-making problem can depend more strongly on the size of this support than on the total number of states and actions (Rolf et al., 2023; Wu, Han, Yan, Kuo, & Shen, 2025). This perspective introduces the concept of effective dimension, suggesting that sparse cost structures fundamentally alter learning dynamics and regret behavior.

Despite these theoretical advances, empirical validation of sparse SSP models remains limited. Many existing works focus on abstract or synthetic environments that obscure how sparse costs are encountered during actual agent trajectories (Ding et al., 2025). Consequently, it is often unclear whether sparsity-aware theoretical guarantees accurately reflect agent behavior in interpretable, spatially structured environments.

Grid-based environments have long been used as experimental platforms for reinforcement learning due to their transparency and controllability. The MiniGrid environment, in particular, has been extensively employed to study exploration strategies, generalization, and representation learning in navigation tasks (Chevalier-Boisvert et al., 2023; de Tinguy, Van de Maele, Verbelen, & Dhoedt, 2024). However, prior MiniGrid-based studies primarily emphasize reward maximization and policy learning, with cost functions typically modeled as dense negative rewards or stationary penalties. The empirical behavior of agents under episodic sparse cost conditions has received comparatively little attention.

A small number of studies have explored sparse penalties in navigation and control problems, yet these works often treat sparsity implicitly rather than as a primary experimental variable (Balogun et al., 2024). Moreover, existing empirical evaluations rarely provide step-level trajectory data that enable detailed analysis of when, where, and how sparse costs are activated during an episode. This lack of fine-grained empirical evidence limits the ability to bridge theoretical SSP analyses

Episodic Sparse Cost Evaluation or Policy Analysis in Stochastic Shortest Path Problems

with practical reinforcement learning behavior (Zou, Zeng, & Zhang, 2024).

Based on these observations, a clear research gap can be identified: there is a lack of empirical studies that explicitly construct episodic sparse cost structures and systematically analyze their interaction with agent trajectories in SSP settings. Addressing this gap is essential for understanding how sparsity-aware theoretical results manifest in practice.

Therefore, the purpose of this study is to conduct an empirical analysis of sparse episodic cost behavior in a stochastic shortest path framework using the MiniGrid environment as a controlled testbed. Rather than proposing a new learning algorithm, this research focuses on observing and characterizing how agents encounter sparse costs during navigation under known transition dynamics. Specifically, this study seeks to answer the following research questions :

(1) How frequently do agents encounter sparse cost events along their trajectories?; (2) How are sparse cost activations distributed over time within an episode?; (3) How do repeated visits to cost-bearing state-action pairs contribute to cumulative episodic cost?

By addressing these questions, this work aims to provide empirical evidence that complements existing theoretical studies on sparse SSP models. The findings are expected to enhance understanding of sparsity-driven behavior in sequential decision-making and to serve as a foundation for future research on sparsity-aware reinforcement learning methods.

2. Metode Penelitian

This study employs an episodic evaluation method to analyze sparse cost behavior in a stochastic shortest path (SSP) environment. The environment is modeled as an SSP with a designated initial state, goal state, and known transition dynamics. The primary objective of the

method is to observe how costs accumulate along agent trajectories when penalties are restricted to a limited number of state–action configurations within each episode.

Algorithm 1 Episodic Sparse-Cost SSP Evaluation

Input: SSP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, s_0, s_g)$; policy π ; sparsity M ; probe length T_p ; horizon H ; unit cost $\lambda > 0$

Output: Trajectory τ_k , cumulative cost C_k , hit count Hits_k , hit rate HR_k

```

1:  $s \leftarrow s_0$ 
2:  $\mathcal{D}_k \leftarrow \emptyset$ 
3: for  $t = 0$  to  $T_p - 1$  do
4:    $a_t \sim \pi(\cdot | s)$ 
5:    $\mathcal{D}_k \leftarrow \mathcal{D}_k \cup \{(s, a_t)\}$ 
6:    $s \sim P(\cdot | s, a_t)$ 
7: end for
8:  $P_k \leftarrow \text{Sample}_M(\mathcal{D}_k) \{|P_k| = M\}$ 
9:  $s \leftarrow s_0$ 
10:  $C_k \leftarrow 0$ 
11:  $\text{Hits}_k \leftarrow 0$ 
12:  $\tau_k \leftarrow \emptyset$ 
13: for  $t = 0$  to  $H - 1$  do
14:    $a_t \sim \pi(\cdot | s)$ 
15:    $c_t \leftarrow \lambda \cdot \mathbb{I}[(s, a_t) \in P_k]$ 
16:    $C_k \leftarrow C_k + c_t$ 
17:    $\text{Hits}_k \leftarrow \text{Hits}_k + \mathbb{I}[(s, a_t) \in P_k]$ 
18:    $\tau_k \leftarrow \tau_k \cup \{(s, a_t, c_t)\}$ 
19:    $s' \sim P(\cdot | s, a_t)$ 
20:   if  $s' = s_g$  then
21:     break
22:   end if
23:    $s \leftarrow s'$ 
24: end for
25:  $\text{HR}_k \leftarrow \text{Hits}_k / \max\{1, |\tau_k|\}$ 
26: return  $(\tau_k, C_k, \text{Hits}_k, \text{HR}_k)$ 

```

Algorithm 1. Episodic Sparse-Cost SSP Evaluation

The experimental procedure follows two sequential phases, as formally specified in Algorithm 1. In the first phase, a short probe rollout is performed from the initial state using a fixed policy. During this probe, reachable state action pairs are collected to form an empirical basis for defining cost locations. From these collected pairs, a fixed size episodic sparse support set is sampled, ensuring that costs are assigned only to configurations that can be encountered by the agent within the environment.

Episodic Sparse Cost Evaluation or Policy Analysis in Stochastic Shortest Path Problems

In the second phase, the agent is reset to the initial state and interacts with the environment using the same policy until either the goal state is reached or a predefined horizon is exceeded. At each step, a unit cost is incurred only when the executed state-action pair belongs to the previously constructed sparse support set. Throughout the episode, the interaction trajectory, cumulative cost, and cost activation events are recorded. The episode-level outputs consist of the full trajectory, total accumulated cost, number of cost hits, and hit rate, which together characterize the impact of sparse episodic costs on agent behavior in the SSP setting.

3. Hasil dan Pembahasan

The proposed episodic sparse cost evaluation framework was empirically validated on a grid based stochastic shortest path navigation task using a fixed forward-biased policy. The evaluation was conducted on an empty grid environment, where the agent navigates from a predefined initial state toward a terminal goal state under deterministic transitions. The sparse cost mechanism was instantiated by sampling a support set of size from a short probing rollout prior to the main episode.

The agent successfully completed the task in 95 steps, starting from the initial pose and terminating at . Visual snapshots of the environment at the beginning and termination of the episode are shown in Figure 1, confirming that the agent reached the intended goal state without premature termination.

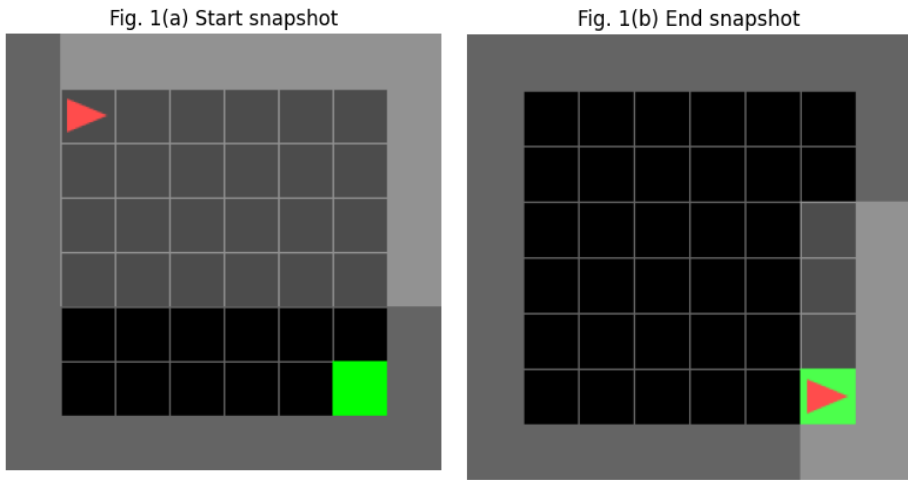


Fig. 1. MiniGrid environment snapshots at (a) episode start and (b) termination

The episode level quantitative results are summarized in Table 1. Despite the relatively long trajectory, only two cost-triggering events were observed, resulting in a total episodic cost of 2.0 and a hit rate of approximately 0.021. This observation highlights a key property of the proposed formulation: the magnitude of accumulated cost does not scale with episode length, but is instead governed by rare encounters with a small subset of state–action pairs.

Table 1. Episodic Sparse Cost Evaluation Results

Episode	Steps	Hits	Total Cost	Hit Rate	Hit Steps
1	95	5	2.0	0.021	13, 77

Formally, the episodic cumulative cost is defined as where the instantaneous cost at step t follows the sparse episodic definition with denoting the sampled sparse support and in this experiment. Under this formulation, costs are incurred exclusively when the agent revisits

Episodic Sparse Cost Evaluation or Policy Analysis in Stochastic Shortest Path Problems

specific state–action pairs identified during the probe phase, while all other interactions remain cost-free.

The spatial behavior of the agent throughout the episode is illustrated in Figure 2, which depicts the trajectory in grid coordinates. The trajectory exhibits a monotonic progression toward the goal region, with extended traversal along the lower boundary of the grid. Importantly, the two cost-triggering events occur at intermediate locations along the trajectory rather than near the start or termination, indicating that sparse costs are not boundary artifacts but arise from selective interactions between the policy and the environment.

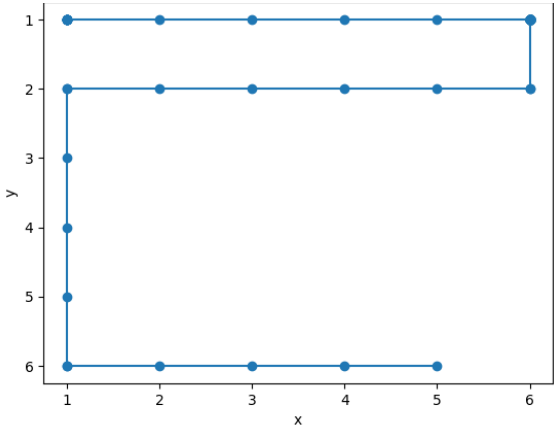


Fig. 2. Agent trajectory in grid coordinates (x,y) over one episode.

The exact cost-triggering events are detailed in Table 2, showing that penalties occurred only at steps 13 and 77. These steps correspond to state–action configurations belonging to the sampled support, while the remaining 93 steps incurred zero cost. This selective activation underscores the episodic and non-dense nature of the cost signal.

Step	x	y	Dir	Action	Sparse Cost
13	5	1	0	2	1.0
77	6	1	1	2	1.0

The temporal distribution of sparse costs is visualized in Figure 3, where the cost signal appears as isolated impulses rather than a continuous stream. This behavior contrasts sharply with dense reward or penalty formulations commonly used in reinforcement learning, and provides a clearer interpretation of when and where critical interactions occur.

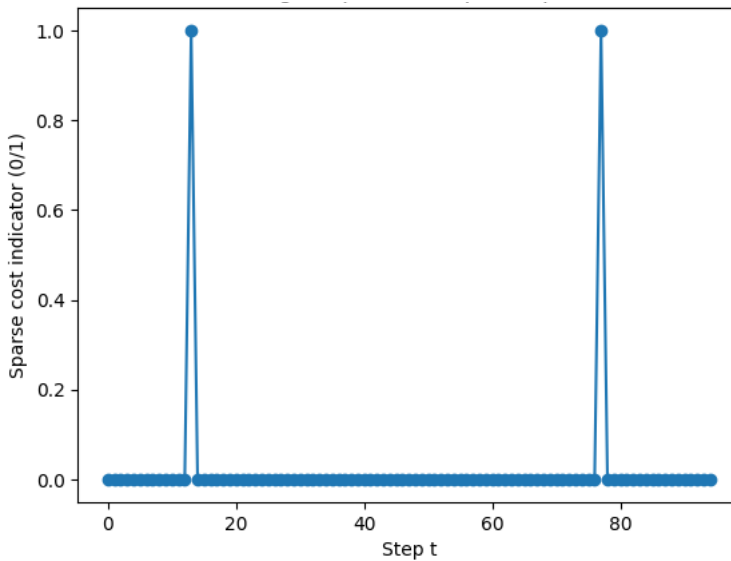


Fig. 3. Step-wise sparse cost indicator showing cost-triggering events

The cumulative effect of the sparse cost mechanism over the episode is further illustrated in Figure 4, which presents the cumulative sparse cost as a function of time steps. The plot exhibits a piecewise-constant profile with two discrete upward jumps occurring precisely at

Episodic Sparse Cost Evaluation or Policy Analysis in Stochastic Shortest Path Problems

steps 13 and 77, corresponding to the cost-triggering events reported in Table 2. Outside these points, the cumulative cost remains flat, indicating the absence of penalization during the majority of the episode.

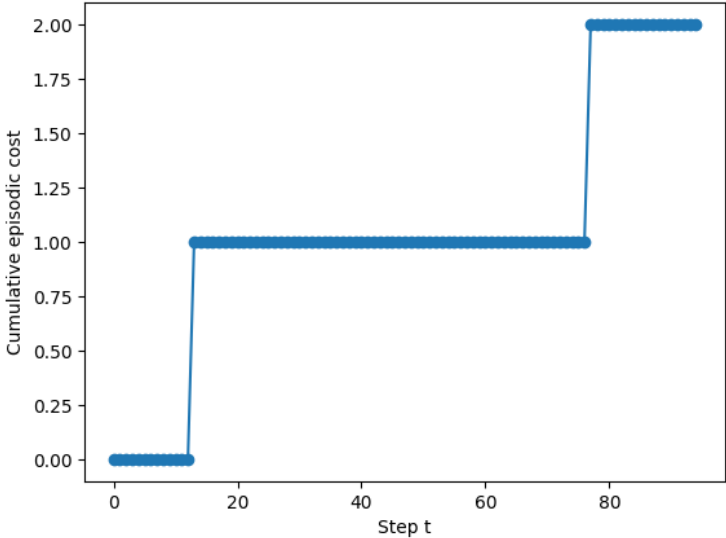


Fig 4. Cumulative sparse cost over one episode

This cumulative behavior provides additional empirical evidence that the proposed sparse-cost formulation does not introduce progressive or time-dependent penalization. Instead, cost accumulation is strictly event-driven and directly attributable to specific state-action encounters contained in the sparse support set. Consequently, the cumulative cost remains interpretable and bounded, reinforcing the claim that episodic sparse costs effectively decouple trajectory length from cost magnitude. The proportion of cost-triggering steps relative to the episode length is quantified by the hit rate yielding a value of 0.021 for the evaluated episode. This result indicates that more than 97% of agent-environment interactions were cost-free, demonstrating that the

proposed framework effectively isolates rare but significant events without perturbing overall task solvability.

Overall, the results empirically validate the effectiveness of episodic sparse-cost evaluation for policy analysis in stochastic shortest path environments. By decoupling trajectory length from cost accumulation and enabling explicit identification of critical state-action encounters, the framework provides an interpretable and behavior-focused alternative to dense cost formulations. This makes it particularly suitable for analyzing risk-sensitive or constraint-aware decision-making scenarios where rare events, rather than average performance, are of primary interest.

4. Kesimpulan dan Saran

This work presented an episodic sparse-cost evaluation framework for stochastic shortest path problems, aimed at analyzing policy behavior under rare and selective cost signals. The proposed formulation assigns costs only to a small subset of state action pairs identified through a preliminary probing phase, thereby decoupling cost accumulation from episode length and eliminating dense or time-dependent penalties.

The experimental results demonstrate that the cumulative episodic cost remains bounded and interpretable, even for long trajectories. Costs are incurred only at isolated steps corresponding to revisits of the sampled support set, while the majority of agent environment interactions remain cost free. This confirms that the framework effectively captures event driven interactions rather than averaging behavior over time.

From an analytical standpoint, the proposed approach provides a transparent mechanism for identifying critical decision events along a trajectory. Unlike conventional dense cost formulations, which may obscure the influence of individual actions, episodic sparse costs enable explicit attribution of penalties to specific state-action encounters. This

Episodic Sparse Cost Evaluation or Policy Analysis in Stochastic Shortest Path Problems

property is particularly valuable for evaluating policies in settings where safety violations, constraint breaches, or rare risks are of primary concern.

The current study focuses on single-episode evaluation under a fixed policy, which constitutes a limitation of the present analysis. Future research directions include extending the framework to multi-episode statistical evaluation, adaptive or learned policies, and more complex stochastic environments. Integrating episodic sparse costs directly into policy optimization objectives also represents a promising avenue for future work.

In conclusion, episodic sparse-cost evaluation offers a principled, interpretable, and behavior focused alternative to dense cost schemes, making it well-suited for diagnostic analysis and risk aware assessment in sequential decision-making systems.

Daftar Pustaka

- Balogun, A., Olajube, A., Awelewa, A., Agoro, S., Okafor, F., Sanni, T., ... Ajilore, A. (2024). Control strategies in enhanced stand-alone mini-grid operations for the NESI—an overview. *Frontiers in Energy Research*, 12, 1397482.
- Bertazzi, L., Mogre, R., & Trichakis, N. (2024). Dynamic project expediting: a stochastic shortest-path approach. *Management Science*, 70(6), 3748–3768.
- Chevalier-Boisvert, M., Dai, B., Towers, M., Perez-Vicente, R., Willems, L., Lahlou, S., ... Terry, J. (2023). Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, 36, 73383–73394.

-
- da Silva, J. M., Ramos, G. de O., & Barbosa, J. L. V. (2023). Multi-Objective Decision-Making Meets Dynamic Shortest Path: Challenges and Prospects. *Algorithms*, 16(3), 162.
- de Tinguy, D., Van de Maele, T., Verbelen, T., & Dhoedt, B. (2024). Spatial and temporal hierarchy for autonomous navigation using active inference in minigrid environment. *Entropy*, 26(1).
- Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., ... Sukiennik, N. (2025). Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3), 1–38.
- Karia, R., Nayyar, R. K., & Srivastava, S. (2022). Learning generalized policy automata for relational stochastic shortest path problems. *Advances in Neural Information Processing Systems*, 35, 30625–30637.
- Rolf, B., Jackson, I., Müller, M., Lang, S., Reggelin, T., & Ivanov, D. (2023). A review on reinforcement learning algorithms and applications in supply chain management. *International Journal of Production Research*, 61(20), 7151–7179.
- Wang, H., & Liang, G. (2025). Association Rules Between Urban Road Traffic Accidents and Violations Considering Temporal and Spatial Constraints: A Case Study of Beijing. *Sustainability*, 17(4), 1680.
- Wu, Q., Han, J., Yan, Y., Kuo, Y.-H., & Shen, Z.-J. M. (2025). Reinforcement learning for healthcare operations management: methodological framework, recent developments, and future research directions. *Health Care Management Science*, 28(2), 298.
- Xu, C., & Sankar, R. (2024). A comprehensive review of autonomous driving algorithms: Tackling adverse weather conditions, unpredictable traffic violations, blind spot monitoring, and emergency maneuvers. *Algorithms*, 17(11), 526.
- Zou, M., Zeng, Q., & Zhang, X. (2024). Weakly-supervised action learning in procedural task videos via process knowledge decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7), 5575–5588.