
To Switch or Not to Switch?

Balanced Policy Switching in Offline Reinforcement Learning

Tao Ma¹ Xuzhi Yang¹ Zoltán Szabó¹

Abstract

Reinforcement learning (RL)—finding the optimal behaviour (also referred to as policy) maximizing the collected long-term cumulative reward—is among the most influential approaches in machine learning with a large number of successful applications. In several decision problems, however, one faces the possibility of policy switching—changing from the current policy to a new one—which incurs a non-negligible cost, and in the decision one is limited to using historical data without the availability for further online interaction. Despite the inevitable importance of this offline learning scenario, to our best knowledge, very little effort has been made to tackle the key problem of balancing between the gain and the cost of switching in a flexible and principled way. Leveraging ideas from the area of optimal transport, we initialize the systematic study of policy switching in offline RL. We establish fundamental properties and design a Net Actor-Critic algorithm for the proposed novel switching formulation. Numerical experiments demonstrate the efficiency of our approach on multiple robot control benchmarks of the Gymnasium and traffic light control from SUMO-RL.

1. Introduction

Reinforcement learning (RL, [Puterman, 2014](#)) is a fundamental tool in machine learning for advising agents to make sequential decisions, which has recently witnessed an unprecedented breakthrough from both theoretical and application perspective ([Sutton & Barto, 2018](#)). Successful applications of RL include for instance beating human expert players in games ([Mnih et al., 2013](#); [Silver et al., 2017](#)), dynamic treatment and automated medical diagno-

sis in healthcare ([Yu et al., 2021](#); [Sun et al., 2024](#)), robotics behaviour improvement ([Kober et al., 2013](#); [Tang et al., 2024](#)) and autonomous driving ([Kiran et al., 2021](#)). Due to its flexible design, RL is able to accommodate various important forms of optimal decision making.

In a broad sense, RL problems can be divided into two groups, online and offline RL, each of which has its distinct strengths and limitations. In the online setting, the agent can actively explore the unknown environment by executing actions according to the policies, and make use of the received rewards to adjust the behaviour for a higher future gain ([Arulkumaran et al., 2017](#)). However, in scenarios where random exploration may be impractical or even dangerous ([Singla et al., 2021](#)), gathering a static dataset is often a more adequate choice. Motivated by such constraints, offline RL has emerged as a promising approach ([Levine et al., 2020](#)). In the offline setting, some policies have already been applied in the environment and generated a large offline dataset. With such data, the agent cannot make further exploration, but is supposed to learn a better policy solely based on the available information ([Haarnoja et al., 2018](#); [Fujimoto et al., 2019](#); [Kumar et al., 2020](#); [Matsushima et al., 2021](#); [Fujimoto & Gu, 2021](#); [Kostrikov et al., 2021](#); [An et al., 2021](#); [Ma & Yang, 2024](#)). Due to the discrepancies between the policies that generated the offline data and the policy learned by an offline algorithm, solving decision problems offline is highly challenging, with expected sub-optimal performance ([Kumar et al., 2019](#)) compared to their online counterparts.

Despite the success of RL algorithms in the offline setting ([Kumar et al., 2020](#); [Kostrikov et al., 2021](#)), one key but moderately studied question is the cost of policy switching. Significant cost can occur when changing from an old policy to a new one. One common example is the adaptive policies in traffic light control ([Lopez et al., 2018](#)). Policy adjustment can lead to additional delays or even compromise traffic safety ([Han et al., 2023](#)). So a policy switch for such traffic control introduces significant costs (such as temporary traffic congestion, update of facilities). Other examples include the cost of updating hardware devices ([Mirhoseini et al., 2017](#)), the fees to employ

¹Department of Statistics, London School of Economics, Houghton Street, London, WC2A 2AE, UK. Correspondence to: Tao Ma <t.ma9@lse.ac.uk>.

human annotators for large models (He et al., 2023), the re-organization expenses of a company (LoPucki & Doherty, 2004), or the additional efforts to modify webpage designs (Theocharous et al., 2015). However, modelling such policy switching cost is a highly non-trivial task. For example, in the perspective of employees in a company, learning a new skill normally requires more efforts than relocating to a new team with similar tasks. Such scenario of strategy change and cost management is called organizational change management in the theory of business (By, 2005; Lauer, 2010). On the other hand, in the existing literature of RL to our best knowledge, the focus was only on somewhat simplistic schemes of costs, which include the global and the local switching cost (Bai et al., 2019; Gao et al., 2021; Wang et al., 2021; Qiao et al., 2022). Both definitions target to measure if two policies (or policies conditional on states) are the same or not, but ignore *how* the two (families of) policies are *different* from each other. In addition, these costs with limited forms of expressiveness were developed for the online setting.

In this work we focus on the offline RL setting. Our aim is to initialize the formulation and understanding of the key properties of policy switching in this scenario. Throughout the paper, we will consider the following prototype offline RL task: the agent has been relying on an old policy for a long term, with which rich offline data has been generated. Now in the beginning of a new episode, there is one chance for the agent to execute a policy with the possibility of switching to a new one, where the change can have a non-negligible cost. Our **goals** are three-fold:

1. How to rigorously formulate such offline policy switching problem, and balance between the potential gain and the cost?
2. Is there a way to construct a family of switching costs that are flexible and expressive?
3. How to design an algorithm to robustly find a better policy in the new problem formulation?

Given these three questions, our **contributions** can be summarized as follows.

1. We propose a new policy switching problem, by defining the novel net values and net Q-functions, and establish their fundamental properties which are in sharp contrast to their classic RL counterparts.
2. Motivated by mass transportation, we propose a flexible class of cost functions, which includes former definitions (local and global costs) as special cases.
3. An algorithm, named Net Actor-Critic (NAC), is proposed to find a new policy which improves the old policy towards the optimal in terms of net value, which is the first offline method for policy switching problem with costs to our best knowledge.

The paper is structured as follows. We begin with preliminaries on notations, classic RL settings and a review of former switching costs in Section 2. In Section 3 we introduce the notions of net value and net Q-function, with which the novel policy switching problem is formulated. A new family of cost functions are also provided relying on optimal transport (OT). We present our NAC algorithm to approximate the switch-optimal policy in Section 4; the numerical efficiency of the approach is demonstrated in Section 5. Further algorithmic details, extensions of the problem formulation, proofs and implementation details of experiments are provided in the Appendix.

2. Preliminaries

In this section we provide the necessary background for the manuscript. Notations are introduced in Section 2.1, and the classic RL settings and formerly proposed policy switching costs are elaborated in Section 2.2.

2.1. Notations

We introduce a few notations used throughout the paper. A σ -algebra on a set X is denoted by Σ_X . Given measurable spaces (X, Σ_X) and (Y, Σ_Y) , $(X \times Y, \Sigma_X \otimes \Sigma_Y)$ is the product space, where $\Sigma_X \otimes \Sigma_Y$ is the smallest σ -algebra generated by $\{A \times B : A \in \Sigma_X, B \in \Sigma_Y\}$. The set of all probability measure on (X, Σ_X) is denoted by $\mathcal{P}(X)$. Let \mathcal{F} be the collection of all real-valued functions on X , $\|f\|_\infty := \sup_{x \in X} |f(x)|$ ($f \in \mathcal{F}$), and define $\mathcal{G}(X, \|\cdot\|_\infty) := \{f \in \mathcal{F} : \|f\|_\infty < +\infty\}$; $\mathcal{G}(X, \|\cdot\|_\infty)$ is known to be complete. For any set $A \subseteq X$, $I_A : X \rightarrow \{0, 1\}$ is the indicator function of A : $I_A(x) = 1$ if $x \in A$, $I_A(x) = 0$ otherwise. For a set B , $|B|$ stands for its cardinality. The set of non-negative real numbers is denoted by $\mathbb{R}_{\geq 0}$; similarly, $\mathbb{R}_{> 0}$ stands for the set of positive reals. Let Id be the identity map. For any positive integer K , $[K] := \{1, \dots, K\}$. For any $a, b \in \mathbb{R}$, $a \wedge b := \min\{a, b\}$. A map T from a metric space (Z, ρ) into itself is called contraction if there exists a constant $c_T \in [0, 1)$ such that $\rho(T(z_1), T(z_2)) \leq c_T \rho(z_1, z_2)$ for all $z_1, z_2 \in Z$.

2.2. Classic RL Settings

In this subsection, we recall a few fundamental concepts of RL from the formulation of MDPs, alongside with the formerly proposed policy switching costs.

MDPs. We consider a time-homogeneous MDP, denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, with $(\mathcal{S}, \Sigma_{\mathcal{S}})$ and $(\mathcal{A}, \Sigma_{\mathcal{A}})$ being measurable state and action spaces, respectively. Given any pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, $P(\cdot | s, a) : \Sigma_{\mathcal{S}} \rightarrow [0, 1]$ is the transition kernel and $r(\cdot | s, a)$ encodes a stochastic reward with mean $R(s, a)$ and bounded support. Finally, $\gamma \in [0, 1)$ is the discount factor for future rewards. Given an MDP, a policy

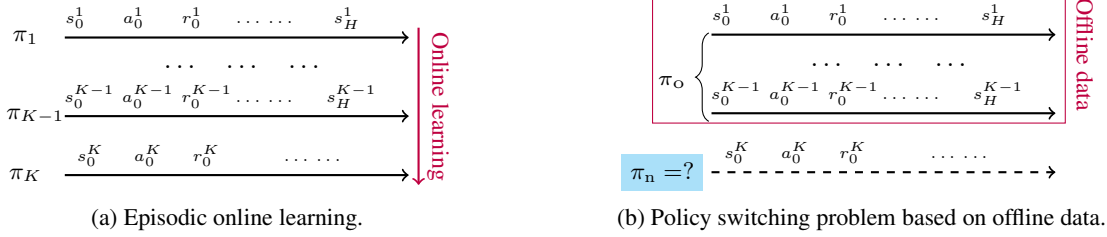


Figure 1: Comparison between previous setting of online learning and ours.

$\pi = \{\pi(\cdot|s), s \in \mathcal{S}\}$ of an agent is a collection of conditional distributions on $(\mathcal{A}, \Sigma_{\mathcal{A}})$, and Π is the collection of all policies. With these notations at hand, an MDP proceeds as follows. The agent starts at a fixed initial state $s_0 \in \mathcal{S}$. At any step $t \geq 0$, the agent is at state $s_t \in \mathcal{S}$, selects action $a_t \sim \pi(\cdot|s_t)$, receives a reward $r_t \sim r(\cdot|s_t, a_t)$, and is transitioned to $s_{t+1} \sim P(\cdot|s_t, a_t)$, the process of which creates one transition tuple $(s_t, a_t, r_t, s_{t+1}) \in \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$.

Evaluation & optimality. For the purpose of policy evaluation and optimization, the *value* at state s and the *Q-function* at state-action pair (s, a) of π are respectively defined as

$$V^\pi(s) := \mathbb{E}_\pi \left\{ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right\},$$

$$Q^\pi(s, a) := \mathbb{E}_\pi \left\{ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right\},$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expectation according to π . At a state s , the optimal value is $V^*(s) := \max_\pi V^\pi(s)$. With a state-action pair (s, a) , the optimal Q-function is $Q^*(s, a) := \max_\pi Q^\pi(s, a)$, which are both taken over all policies. With two policies $\pi_1, \pi_2 \in \Pi$, we say that π_1 is at least as good as π_2 if $V^{\pi_1}(s) \geq V^{\pi_2}(s)$ for all $s \in \mathcal{S}$. The optimal policy is then defined as one that is at least as good as any other policy. It is known that there always exists an *optimal policy*, and for any optimal policy π^* , $Q^{\pi^*}(s, a) = Q^*(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Online & offline RL. We now provide a description of online and offline RL for convenient comparison; see Fig. 1 for a visual illustration. In a given MDP, online learning is when the agent is allowed to switch policies throughout the process of learning. In practice, one specific example is when there are a total number of K episodes, and the agent can choose a new policy π_k for the k -th episode ($k \in [K]$). The data directly generated by policy π_k proposed by the agent can be collected in the episode k , which further helps learning the next policy π_{k+1} . On the other hand, offline learning is when a fixed dataset, containing transition tuples by following some policy π_o not proposed by the agent, is provided to the agent. And one needs to learn a better policy π_n , to apply in the following steps, only using this

dataset, without any further interaction with the environment.

Switching cost. The limited coverage of switching cost formulations in the literature (Bai et al., 2019; Gao et al., 2021; Wang et al., 2021; Qiao et al., 2022), to our best knowledge, all focus on the online learning setting. For a finite \mathcal{S} , the formerly proposed global and local switching costs (Bai et al., 2019) with K episodes are respectively

$$C^{\text{gl}}(\pi_1, \dots, \pi_K) = \sum_{k=1}^{K-1} I_{\{\pi_k \neq \pi_{k+1}\}}, \quad (1)$$

$$C^{\text{loc}}(\pi_1, \dots, \pi_K) = \sum_{k=1}^{K-1} \sum_{s \in \mathcal{S}} I_{\{\pi_k(\cdot|s) \neq \pi_{k+1}(\cdot|s)\}}. \quad (2)$$

As long as the policy is changed, global switching cost will increase by 1, while the increase in local switching cost is determined by how many states on which the conditional distributions are changed, which can be seen as a more fine-grained version of the global cost.

The **primary challenges** tackled in this paper are two-fold. First, our goal is to address the offline setting where an agent is only allowed to switch policy once, and this switch has a non-negligible cost. Second, the local switching cost is agnostic w.r.t. how different two distributions $\pi(\cdot|s)$ and $\pi'(\cdot|s)$ are (it increases by 1 as long as they are not identical); we aim to take into account that distributions far away are expected to incur higher costs than two similar ones.

3. Problem Formulation

In order to address the challenges outlined in Section 2, we introduce the net value and net Q-function, with which a novel policy switching problem in offline RL is proposed in Section 3.1. The considered switching cost family is detailed in Section 3.2, which significantly extends the existing switching costs.

3.1. The Policy Switching Problem

This section is dedicated to the formulation of our policy switching problem based on two new notions (net value and

net Q-function) introduced below, followed by establishing some of their fundamental theoretical properties.

The question. Enriched with the general setting of RL (Section 2.2), we consider the following scenario; see Fig. 1b for an illustration. There is a known old policy π_o , which has already been applied for $K - 1$ episodes and led to the forming of an offline dataset \mathcal{D} with size $n = (K - 1)H$. Now at $s_0 \in \mathcal{S}$, to begin the last episode, the agent needs to choose a policy. In addition, the agent is given a switching cost function C ; $C(\pi_o, \pi_n)$ measures the policy switching cost (from π_o to π_n) incurred in the beginning of the last episode. This gives rise to the switching extension captured by the augmented tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, s_0, \pi_o, C)$.

There are two fundamental questions to be addressed:

1. Is it profitable to switch to a different policy π_n from the old π_o ?
2. In the case of switching, which new policy π_n would better balance between the discounted total return in the last episode and the cost?

We use the following two new notions to address these questions.

Definition 3.1 (Net Value, Net Q-function). For $s \in \mathcal{S}$, $a \in \mathcal{A}$, define the net value function and the net Q-function as

$$\begin{aligned} V_N^{\pi_n}(s) &:= V^{\pi_n}(s) - C(\pi_o, \pi_n), \\ Q_N^{\pi_n}(s, a) &:= Q^{\pi_n}(s, a) - C(\pi_o, \pi_n). \end{aligned}$$

The value $V_N^{\pi_n}(s)$ measures after deducting the switching cost $C(\pi_o, \pi_n)$, the actual return in the last episode by adopting some new policy π_n and starting from state s . Notice that the net value function is defined for all possible initial states $s \in \mathcal{S}$ which will allow us to investigate optimality w.r.t. different initial states (Proposition 3.4(c)). Using the analogue of business strategies, the one-time switching cost $C(\pi_o, \pi_n)$ represents how much investment is needed to change to a new strategy π_n , the value $V^{\pi_n}(s)$ of a strategy is the total return in the future, while the net value $V_N^{\pi_n}(s)$ corresponds to the net income. The meaning of $Q_N^{\pi_n}(s, a)$ can be interpreted similarly. In the traffic light control problem, the cost is due to the temporary congestion of the intersection, and the value is the evaluation of the efficiency for vehicles to pass the intersection.

Having defined net values, we now formulate the notion of switch-optimal policy while fixing the initial state $s_0 \in \mathcal{S}$.

Definition 3.2 (Switch-optimal policy). Given an old policy π_o and a fixed initial state $s_0 \in \mathcal{S}$, a proposed policy π_n^* is said to be switch-optimal if, for any policy $\pi_c \in \Pi$,

$$V_N^{\pi_n^*}(s_0) \geq V_N^{\pi_c}(s_0). \quad (3)$$

Based on these definitions, our goal is to find a switch-optimal policy π_n^* or at least a policy π_n which improves upon the old policy π_o in terms of the net value function ($V_N^{\pi_n}(s_0) \geq V_N^{\pi_o}(s_0)$ where the r.h.s. equals to $V^{\pi_o}(s_0)$). If able to find such better π_n , the agent switches to this new policy; otherwise sticks with π_o in the last episode. It should be noted that, although we try to find some policy close to the switch-optimal one, in the offline setting this can be rather challenging; so a new policy with significant improvement often already suffices. It is important to note that π_n^* need not have a significantly large value close to the optimal value $V^*(s_0)$, as our goal is not to find a policy with the maximal value function. Instead, we aim to find a policy that best balances the future return and the cost.

Before moving on to our solution in the next section, we provide the following proposition for a deeper understanding of this new policy switching problem.

Assumption 3.3. The sets of values $\{V^\pi(s_0)\}_{\pi \in \Pi}$ and costs $\{C(\pi_o, \pi)\}_{\pi \in \Pi}$ are compact.

Beyond existence, the following result shows various distinct characteristics [see Proposition 3.4(b)-3.4(d)] specific to the switching setting.

Proposition 3.4. For any MDP, the followings hold.

- (a) If Assumption 3.3 is satisfied, then there always exists a switch-optimal policy.
- (b) There exists a cost function C , with which an optimal policy in value is not switch-optimal in net value.

If π_n^* is switch-optimal in a fixed initial state s_0 , then

- (c) if an alternative $s'_0 \in \mathcal{S}$ is fixed as initial state, then the switch-optimal policy may change.
- (d) it may not be the case that $Q_N^{\pi_n^*}(s_0, a) \geq Q_N^\pi(s_0, a)$ for all $a \in \mathcal{A}$ and all $\pi \in \Pi$.

Remarks:

- **Existence:** Under mild assumptions, Proposition 3.4(a) guarantees the existence of a switch-optimal policy, which ensures that the problem is well-posed. Proposition 3.4(b) distinguishes the policy switching problem from the classic policy learning problem, as the respective optimal policies are different with an appropriate choice of C . It should be noted that the optimal policies in the two problems are not always different.
- **Initial state dependence:** Proposition 3.4(c) and 3.4(d) indicate that the switch-optimal policy depends both on the initial state and the first action. This behaviour is in sharp contrast to the classic RL setting (Section 2.2) where an optimal policy achieves the highest value and Q-function *simultaneously* on all states/state-action pairs. Such different characteristic of the optimal policies in the switching problem calls for a new approach to im-

prove the candidate policy in the policy learning step of any proposed algorithm, as summing up returns over episodes with different initial states will be invalid in this case.

As a useful computation tool for policy evaluation (used later in Algorithm 1), we define the net Bellman operator and establish its contractive property.

Definition 3.5 (Net Bellman operator). Given any net Q-function $Q_N \in \mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$ and policy π , define the net Bellman operator $B^\pi : \mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty) \rightarrow \mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$ of net Q-function as

$$(B^\pi Q_N)(s, a) := R(s, a) - (1 - \gamma)C(\pi_o, \pi) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)}[V_N(s')],$$

with $V_N(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q_N(s, a)]$.

Proposition 3.6 (Policy evaluation with net Q-function). Given a net Bellman operator B^π with respect to a policy π , and any net Q-function $Q_N^0 \in \mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$, let

$$Q_N^{k+1} := B^\pi(Q_N^k) \text{ for } k = 0, 1, 2, \dots$$

Then B^π is a contraction with parameter $c_{B^\pi} = \gamma$ and

$$\lim_{k \rightarrow \infty} Q_N^k = Q_N^\pi,$$

where $Q_N^\pi \in \mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$ is the net Q-function of π .

Thanks to Proposition 3.6, one can use the net Bellman operator to evaluate a given policy π starting from an arbitrary net-Q function Q_N^0 . In this work, we represent net Q-functions by neural networks, replace all expectations with sampled data and tune the parameters so that the net Bellman backup error $\|(B^\pi Q_N) - Q_N\|_2$ is small enough.

3.2. The Family of Cost Functions

In this section, we first introduce two different components in the cost when switching from an old policy to a new one. Then we propose a general cost function family, which includes the reviewed local and global switching costs as specific cases. Finally we gradually zoom in to one specific choice of switching cost relying on optimal transport, which we also investigate numerically (Section 5).

Two components of switching cost. In various policy switching problems, the induced switching costs come from two different sources: *learning cost* and *transaction cost*. Learning cost is incurred when the new policy introduces unfamiliar jobs, which requires serious effort to absorb. Meanwhile, transaction corresponds to the adjustment cost on existing familiar jobs. Such separation of costs have been a longstanding subject of analysis in economics (Nilssen, 1992). Taking traffic control as an example, one wants to update the control policies of some intersection to see if vehicles can pass the intersection more efficiently. Such switch can involve the temporary congestion

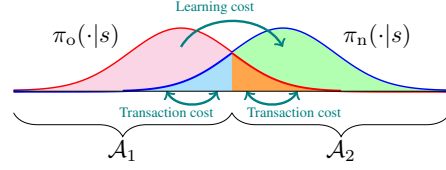


Figure 2: Transport switching cost.

due to configuring new equipments, which is the learning cost. While there is also influence to the traffic due to updating the software of the existing facilities or maintenance of the existing devices, which is the transaction cost. These analogues are reflected in the following cost family.

General cost family. We define a cost family

$$C(\pi_o, \pi_n) := \sigma \left(\int_{\mathcal{S}} f(s) F(\pi_o(\cdot|s), \pi_n(\cdot|s)) d\mu(s) \right), \quad (4)$$

$$F(\pi_o(\cdot|s), \pi_n(\cdot|s)) := c_l \mathcal{L}(\pi_o(\cdot|s), \pi_n(\cdot|s)) + c_t \mathcal{T}(\pi_o(\cdot|s), \pi_n(\cdot|s)), \quad (5)$$

with $\mathcal{L}, \mathcal{T} : \mathcal{P}(\mathcal{A}) \times \mathcal{P}(\mathcal{A}) \rightarrow \mathbb{R}$ capturing the learning cost and the transaction cost, with weights $c_l, c_t \in \mathbb{R}$, $f : \mathcal{S} \rightarrow \mathbb{R}$ measurable function representing the relative importance weighting of different states, μ a probability measure on \mathcal{S} , and activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$; see Fig. A1 for an illustration with finite state spaces ($|\mathcal{S}| < \infty$).

The family (4) subsumes various switching costs including the local and global ones with finite state space $|\mathcal{S}|$ [as in (1) & (2)]¹; see Table 1. This specialization also reveals that local/global costs can only measure the learning cost with simple indicator functions, thus are unable to capture the two different sources (learning and transaction) of the cost.

Proposed transport switching cost. We design a decomposition of the state-wise cost F specified in (5) into the sum of learning cost \mathcal{L} and transaction cost \mathcal{T} , relying on optimal transport (hence the name). We restrict our attention to a specific case of a more general construction (see the end of this section, and Section A for further details) (i) to keep the presentation simple, (ii) as it already conveys the key ideas, (iii) this specialization is easy-to-implement and already turns out to be beneficial as demonstrated by our numerical experiments on multiple RL benchmarks (Section 5).

In various decision problems the action space has a natural partition $\mathcal{A} = \cup_{i=1}^L \mathcal{A}_i$, like the different skill sets in a department of a company. For easier understanding, we focus on the case of $L = 2$; see Fig. 2 for an illustration with colors indicating the different cost terms defined below.

¹For finite state space, one can choose μ to be the uniform distribution on \mathcal{S} and get back (1) and (2).

Table 1: Choices of functions and parameters in the switching cost family.

Cost	$\sigma(x)$	$\mathcal{L}(\pi_o(\cdot s), \pi_n(\cdot s))$	$\mathcal{T}(\pi_o(\cdot s), \pi_n(\cdot s))$	f	c_l	c_t	μ
Local	$ \mathcal{S} x$	$I_{\{\pi_o(\cdot s) \neq \pi_n(\cdot s)\}}$	0	1	1	\mathbb{R}	Unif(\mathcal{S})
Global	$I_{\mathbb{R}_{>0}}(x)$	$I_{\{\pi_o(\cdot s) \neq \pi_n(\cdot s)\}}$	0	1	1	\mathbb{R}	Unif(\mathcal{S})
Transport	$\sigma(x)$	$ \pi_o(\mathcal{A}_1 s) - \pi_n(\mathcal{A}_1 s) $	$\pi_o(\mathcal{A}_1 s) \wedge \pi_n(\mathcal{A}_1 s) + \pi_o(\mathcal{A}_2 s) \wedge \pi_n(\mathcal{A}_2 s)$	f	\mathbb{R}	\mathbb{R}	μ

The construction consists of 2 steps:

Step 1: Mass moving. We move mass *across* \mathcal{A}_1 and \mathcal{A}_2 such that the mass in each component \mathcal{A}_i agree. The amount needed to be moved is defined as the learning cost

$$\mathcal{L}(\pi_o(\cdot|s), \pi_n(\cdot|s)) = |\pi_o(\mathcal{A}_1|s) - \pi_n(\mathcal{A}_1|s)|. \quad (6)$$

Step 2: Mass rearrangement. As some mass of $\pi_o(\cdot|s)$ remains in the same respective component during the first step—see the blue and orange areas in Fig. 2—this part of mass will incur a cost due to rearrangement *within* their own components, which gives rise to the transaction cost:

$$\begin{aligned} \mathcal{T}(\pi_o(\cdot|s), \pi_n(\cdot|s)) &= \pi_o(\mathcal{A}_1|s) \wedge \pi_n(\mathcal{A}_1|s) \\ &\quad + \pi_o(\mathcal{A}_2|s) \wedge \pi_n(\mathcal{A}_2|s). \end{aligned} \quad (7)$$

Connection to OT. The construction of (6) and (7) implicitly defines a near-optimal transport map between $\pi_o(\cdot|s)$ and $\pi_n(\cdot|s)$, and serves as a tight upper bound for the optimal transport in classic OT theories (see e.g. [Staudt & Hundrieser, 2023](#), Lemma 5.1). In addition, the definitions naturally extend to $L > 2$ by treating the mass transportation across components as learning cost, and rearrangement within components as transaction cost (definitions and details in Section A). And for any finite L we show, in a wide range of settings, that the proposed cost is optimal.

Proposition 3.7 (Optimality of the proposed cost). *Given old policy π_o , new policy π_n and any $\{\mathcal{A}_i\}_{i=1}^L$ as a partition of \mathcal{A} . Let the transport cost (of the OT problem) be $c(x, y) = c_l \sum_{i \neq j} I_{\mathcal{A}_i \times \mathcal{A}_j}(x, y) + c_t \sum_{i=1}^L I_{\mathcal{A}_i \times \mathcal{A}_i}(x, y)$ for $x, y \in \mathcal{A}$, with $c_l \geq c_t$. Then $F(\pi_o(\cdot|s), \pi_n(\cdot|s))$ solves the OT problem associated to c for any $s \in \mathcal{S}$.*

Due to limited space, formal definitions of the OT problem and its optimality are available in Section A. Note that the only assumption $c_l \geq c_t$ aligns with common scenarios in real life, when the learning cost due to unfamiliar jobs are expected to be higher than the transaction cost due to rearrangement. Such inequality also guides the choices of the (c_l, c_t) pair in our numerical experiments. With Proposition 3.7, our proposed definition not only captures the two sources of costs, but also is (near) optimal, which could be hard to simultaneously achieve by other discrepancies like

KL-divergence, total variation distance, or maximum mean discrepancy.

We briefly mention a further generalization of the transport switching cost. As justified by Proposition 3.7, in (6) and (7), we have implicitly used $I_{\mathcal{A}_1 \times \mathcal{A}_2}(a, a')$ as the similarity measurement of actions. We note that one can also employ different measurements like the L_2 distance (see Section A).

4. Net Actor-Critic

In this section, we propose the Net Actor-Critic algorithm (NAC; Algorithm 1) to approximate the switch-optimal policy. Note that with known cost function that depends only on policies, actor-critic approach would separate the calculation of induced costs by actor from the conservative Q-function estimation, preventing inaccurate cost computation due to pessimism. At high level, NAC starts from evaluating the old policy, then alternately improves and evaluates the new policy in each iteration, and finally compares the empirical net values of the resulting new policy with the old one for a switching decision.

Algorithm 1 Net Actor-Critic (NAC)

Input: Offline data \mathcal{D} , net Q-function parameters $\{\phi_i\}_{i \in [M]}$, target net Q-function parameters $\{\phi'_j\}_{j \in [M]}$, policy parameter θ , and learning rates $\rho_{nq}, \rho_\theta, \rho_{\text{stb}}$.

- 1: Apply Algorithm 2 to evaluate old policy π_o
- 2: **repeat**
- 3: Sample a mini-batch $B = \{(s, a, r, s')\}$ from \mathcal{D}
- 4: Generate $a' \sim \pi_\theta(\cdot|s')$, compute $y(r, s')$ by (8)
- 5: Update $Q_{N, \phi_i}: \phi_i \leftarrow \phi_i - \rho_{nq} \nabla_{\phi_i} J_{Q_N, i}, i \in [M]$
- 6: Improve policy with gradient ascent:

$$\theta \leftarrow \theta + \rho_\theta \nabla_\theta \mathbb{E}_{a \sim \pi_\theta(\cdot|s_0)} \left[\min_{i \in [M]} Q_{N, \phi_i}(s_0, a) \right]$$

- 7: Update $\phi'_j: \phi'_j \leftarrow \rho_{\text{stb}} \phi'_j + (1 - \rho_{\text{stb}}) \phi_j, j \in [M]$
- 8: **until** Stopping criterion met
- 9: (Optionally) apply Algorithm 2 to evaluate the resulting policy π_θ

Output: $\pi_{\text{out}} = w(\pi_o, \pi_\theta)$

Step 1: Old policy evaluation. As a preliminary step, we need to evaluate the value of π_o , as a reference for later new policy training. Since such algorithm is inspired by an offline fitted-Q evaluation (Munos & Szepesvári, 2008; Le et al., 2019), sharing similar structure as the evaluation part in Algorithm 1, due to limited space, we defer the presentation of Algorithm 2 to Section B.

Step 2: Off-policy evaluation. With offline data $\mathcal{D} = \{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^n$, we first evaluate the net Q-function of the current policy π_θ . Inspired by the pessimistic evaluation with clipped double Q-learning (Hasselt, 2010; Fujimoto et al., 2018), as well as the practical extension to multiple Q-evaluation (An et al., 2021), we train M net Q-functions, in the form of neural networks, in parallel, and take the minimum values to have a conservative estimation of the net Q-function. In addition, we also maintain separate target net Q-networks to improve the stability of evaluation process (Lillicrap et al., 2015). In each training iteration, we independently sample a mini-batch $B \subset \mathcal{D}$ instead of using the whole data. Hence, denoting the parameters of net Q-function by $\{\phi_i\}_{i \in [M]}$, that of the target net Q-functions by $\{\phi'_i\}_{i \in [M]}$, and that of the policy by θ , the target function for evaluation, calculated on $\{(s, a, r, s')\}$ is

$$y(r, s') := r + \gamma \min_{i \in [M]} Q_{N, \phi'_i}(s', a') - (1 - \gamma)C(\pi_o, \pi_\theta),$$

with $a' \sim \pi_\theta(\cdot | s')$. (8)

Then for each $i \in [M]$, we update the parameter values ϕ_i using the gradient of

$$J_{Q_{N,i}} := \mathbb{E}_{(s,a,r,s') \sim B} [Q_{N, \phi_i}(s, a) - y(r, s')]^2.$$

Step 3: Policy improvement. We improve the policy by applying stochastic policy gradient ascent with the objective

$$\max_{\theta} \mathbb{E}_{a \sim \pi_\theta(\cdot | s_0)} \left[\min_{i \in [M]} Q_{N, \phi_i}(s_0, a) \right],$$

which takes care of the state-dependent optimality, compared to former actor-critic methods. By alternatively running Step 2 and 3 the policy is expected to improve towards the underlying switch-optimal one. Especially with multiple net Q-functions and the existence of costs, the distribution shift issue commonly observed in offline RL is naturally handled. To save computational efforts and avoid over-fitting, a stopping criterion is applied; for further details the reader is referred to Section B. Note that when the search process for the switch-optimal policy finishes, we can optionally further evaluate the found policy π_θ by Algorithm 2 to have more accurate offline evaluation.

Step 4: Final decision. In the last step, the algorithm decides to switch to π_θ if its net value at s_0 exceeds the value

of the old policy, and such final decision criterion can be defined through a decision function w , where

$$w(\pi_o, \pi_\theta) := I_{\{V^{\pi_o}(s_0) \geq V_N^{\pi_\theta}(s_0)\}} \pi_o + I_{\{V^{\pi_o}(s_0) < V_N^{\pi_\theta}(s_0)\}} \pi_\theta.$$

5. Numerical Experiments

In this section we demonstrate the efficiency of the proposed NAC algorithm on various Gymnasium benchmarks (Towers et al., 2023) for robot control and SUMO-RL (Alegre, 2019) for traffic control.² The experiments were designed to answer the following **two questions** (in line with Section 3.1):

- Q1: If the old policy π_o is highly suboptimal in terms of its net value, can NAC find a new policy π_n to improve it (in terms of net value)?
- Q2: When the old policy π_o is already switch-optimal, will NAC advise the agent not to switch?

Gymnasium. We selected three environments of Gymnasium (version 0.29.1) to test these hypotheses and the performance of NAC: Ant-v4, HalfCheetah-v4 and Hopper-v4. Common characteristics of the environments are that

- their state and action spaces are continuous ($\mathcal{S} \subseteq \mathbb{R}^{d_S}$, $\mathcal{A} \subseteq \mathbb{R}^{d_A}$),
- the environments are challenging (due to their large dimensional state/action spaces; see Table 2),
- the aim of different 3D robots as agents in the environments is to fast move forward and remain healthy.

To simulate an already switch-optimal old policy π_o (to Q2), we relied on the online version of the NAC algorithm. To obtain a highly sub-optimal old policy π_o (to Q1), we initialized π_o randomly for the HalfCheetah-v4 and the Hopper-v4 environment. For Ant-v4, most random policies were so weak that the agent could hardly learn anything useful from it, not to say improve. So we instead used a policy π_o that was trained online for a few steps; this ensured that the agent could receive some positive rewards but π_o was still far from optimal. For each environment and question (Q1 and Q2), we performed 10 Monte Carlo experiments to assess the performance of NAC. In our experiments, we set $c_l = 5$ and $c_t = 0$ in the cost³, and all the hyperparameters of the algorithms and parameters of the cost are provided in Section B. Note that we discard implementing the former methods with global/local costs, because 1) they are only for online setting; 2) those costs can only handle finite state space but we deal with more complex continuous spaces; 3) even in finite space case, any switch will lead to constant

²All the code replicating our experiments is available at <https://github.com/xiaobaobaochifan/NAC>.

³The $c_t = 0$ choice was made as it is the simplest setting which already goes beyond the local/global switching costs. Due to limited space, further results on $c_t \in \{0.1, 1\}$ are in Section C.

Table 2: Performance of NAC on Gymnasium benchmarks. 1st column: environment considered. 2nd column: $\dim(\mathcal{S})$. 3rd column: $\dim(\mathcal{A})$. 4th column: (sub)optimality of the old policy. 5-7th columns: performance measures, for ‘‘Improvement’’ as mean \pm std. The average net values of old policies are -14.2 (Ant-v4), -60.5 (HalfCheetah-v4), 15.6 (Hopper-v4).

Environment \mathcal{M}	d_S	d_A	Old policy π_o	Improvement	Switch proportion	Responsible rate
Ant-v4	27	8	suboptimal	58.2 ± 23.7	100.0%	90.0%
HalfCheetah-v4	17	6	suboptimal	18.5 ± 19.5	80.0%	70.0%
Hopper-v4	11	3	suboptimal	27.7 ± 16.9	100.0%	100.0%
Ant-v4	27	8	optimal	/	0.0%	100.0%
HalfCheetah-v4	17	6	optimal	/	0.0%	100.0%
Hopper-v4	11	3	optimal	/	20.0%	80.0%

Table 3: Performance of NAC on SUMO-RL benchmarks. 1st column: c_l . 2nd column: c_t . 3rd column: (sub)optimality of the old policy. 4-6th columns: performance measures, for ‘‘Improvement’’ as mean \pm std. The average net values of suboptimal old policies are 21.9, 19.5, 21.9.

c_l	c_t	Old policy π_o	Improvement	Switch proportion	Responsible rate
0.5	0.01	suboptimal	7.0 ± 0.7	100.0%	100.0%
5.0	0.1	suboptimal	6.2 ± 0.4	100.0%	100.0%
10.0	1.0	suboptimal	6.4 ± 0.7	100.0%	100.0%
0.5	0.01	optimal	/	0.0%	100.0%
5.0	0.1	optimal	/	0.0%	100.0%
10.0	1.0	optimal	/	0.0%	100.0%

costs, which provides no information for policy learning.

Our performance measures reported (Table 2, with additional ablation study in Section C) were as follows. With optimal old policies (Q2), we counted the proportion of repetitions over all random seeds when the algorithm advised the agent to switch; the perfect value is 0%. For suboptimal old policies (Q1), we calculated the same proportion (but the perfect value is 100% instead). Such ratio is reported in the column with label ‘‘Switch proportion’’. For all suboptimal cases, we report the mean \pm std of the improvement in net value, with label ‘‘Improvement’’. In addition, we also considered the performance measure ‘‘Responsible rate’’. Recall that the NAC approach makes its decision by comparing the offline-estimated values of the old and newly-obtained policies. We also evaluated the two compared policies (π_o and π_n) in an online fashion, providing a more accurate ‘‘ground truth’’. The performance measure ‘‘Responsible rate’’ counts the proportion that the decision made by NAC agrees with the one provided by the online evaluator.

Table 2 shows that for suboptimal old policies (Q1), in all environments NAC could significantly improve the net values (by relying on the offline data generated by such weak policies); the highest increase was 58.2 in Ant-v4, noting

that the average net value of old policy was only -14.2 . In terms of switch decisions, in at least 80% of the cases NAC advised the agent to switch to a new policy π_n ; these results show that NAC encouraged the agent to explore better policies with high probability. For already optimal old policies (Q2), only in Hopper-v4 there were as low as 20% of the cases where the algorithm advised to switch, while in other environments the decision was always to stick with the old one. Such high probabilities to keep old policies made sure that the agent did not switch to a less profitable policy. We can see that NAC provides responsible decisions in most cases: if due to randomness, the learned policy is not good enough, NAC will likely advise not to switch.

Traffic data. To further showcase the applicability of NAC, we implemented the algorithm on the data from SUMO-RL (Alegre, 2019), an environment designed for developing and assessing traffic control algorithms in realistic urban scenarios. The specific scenario concerned is an intersection with stochastically arriving vehicles. By observing the current light conditions, densities of vehicles and the numbers of queued ones, the agent can adaptively change the phases of lights in each direction to increase the speed for vehicles to pass and ease the pressure of the intersection. The instant reward is the efficiency of vehicles to cross the intersection. Switching to a new policy negatively influ-

ences such efficiency by delays, which is the cost.

According to Table 3, the efficiency of vehicles near the intersection has been significantly increased, showcasing the effectiveness of our proposed method. Particularly, starting with sub-optimal old policies with net-value ~ 20 , the improvement on average is $6.2 - 7.0$ (with std $0.4 - 0.7$). In addition, the switch proportion is 100% when starting from sub-optimal old policies and 0% when starting from an optimal one. In all cases, the responsible rate is 100%.

Further discussions of experiment settings, implementations and limitations are available in Section B and C.

Impact Statement

We do not see any direct negative societal impact arising from the proposed problem formulation of policy switching or the Net Actor-Critic algorithm.

References

- Achiam, J. Spinning Up in Deep Reinforcement Learning. 2018.
- Alegre, L. N. SUMO-RL. <https://github.com/LucasAlegre/sumo-rl>, 2019.
- An, G., Moon, S., Kim, J.-H., and Song, H. O. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, pp. 7436–7447, 2021.
- Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.
- Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. Provably efficient Q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pp. 8004–8013, 2019.
- By, R. T. Organisational change management: A critical review. *Journal of Change Management*, 5(4):369–380, 2005.
- Folland, G. B. *Real analysis: modern techniques and their applications, 2nd edition*. John Wiley & Sons, 1999.
- Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.
- Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 20132–20145, 2021.
- Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pp. 1587–1596, 2018.
- Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062, 2019.
- Gao, M., Xie, T., Du, S. S., and Yang, L. F. A provably efficient algorithm for linear Markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.
- Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018.
- Han, Y., Wang, M., and Leclercq, L. Leveraging reinforcement learning for dynamic traffic control: A survey and challenges for field implementation. *Communications in Transportation Research*, 3:100104, 2023.
- Hasselt, H. Double Q-learning. In *Advances in Neural Information Processing Systems*, pp. 2613–2621, 2010.
- He, X., Lin, Z., Gong, Y., Jin, A., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., and Chen, W. AnnoLLM: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*, 2023.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Salhab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.
- Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.
- Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit Q-learning. *arXiv preprint arXiv:2110.06169*, 2021.
- Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11784–11794, 2019.

- Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1179–1191, 2020.
- Lauer, T. *Change Management*. Springer, 2010.
- Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712, 2019.
- Lei, J. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.
- Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., and Wießner, E. Microscopic traffic simulation using SUMO. In *International Conference on Intelligent Transportation Systems*, pp. 2575–2582, 2018.
- LoPucki, L. M. and Doherty, J. W. The determinants of professional fees in large bankruptcy reorganization cases. *Journal of Empirical Legal Studies*, 1(1):111–141, 2004.
- Ma, T. and Yang, X. A framework for policy evaluation enhancement by diffusion models. In *The Second Tiny Papers Track at ICLR*, 2024.
- Ma, T., Zhu, J., Cai, H., Qi, Z., Chen, Y., Shi, C., and Laber, E. B. Sequential knockoffs for variable selection in reinforcement learning. *arXiv preprint arXiv:2303.14281*, 2023.
- Matsushima, T., Furuta, H., Matsuo, Y., Nachum, O., and Gu, S. Deployment-efficient reinforcement learning via model-based offline optimization. In *International Conference on Learning Representations*, 2021.
- Mirhoseini, A., Pham, H., Le, Q. V., Steiner, B., Larsen, R., Zhou, Y., Kumar, N., Norouzi, M., Bengio, S., and Dean, J. Device placement optimization with reinforcement learning. In *International Conference on Machine Learning*, pp. 2430–2439, 2017.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing Atari with deep reinforcement learning. *arXiv:1312.5602*, 2013.
- Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5):815–857, 2008.
- Nilssen, T. Two kinds of consumer switching costs. *The RAND Journal of Economics*, pp. 579–589, 1992.
- Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- Qiao, D., Yin, M., Min, M., and Wang, Y.-X. Sample-efficient reinforcement learning with $\log \log(t)$ switching cost. In *International Conference on Machine Learning*, pp. 18031–18061, 2022.
- Riedmiller, M. Neural fitted Q iteration—first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, pp. 317–328, 2005.
- Silver, D., Schrittwieser, et al. Mastering the game of Go without human knowledge. *Nature*, 550:354–359, 2017.
- Singla, A., Rafferty, A. N., Radanovic, G., and Heffernan, N. T. Reinforcement learning for education: Opportunities and challenges. *arXiv preprint arXiv:2107.08828*, 2021.
- Staudt, T. and Hundrieser, S. Convergence of empirical optimal transport in unbounded settings. *arXiv preprint arXiv:2306.11499*, 2023.
- Sun, H., Hüyük, A., Jarrett, D., and van der Schaar, M. Accountability in offline reinforcement learning: Explaining decisions with a corpus of examples. *Advances in Neural Information Processing Systems*, 36, 2024.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.
- Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., and Stone, P. Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8, 2024.
- Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. Ad recommendation systems for life-time value optimization. In *International Conference on World Wide Web*, pp. 1305–1310, 2015.
- Towers, M., Terry, J. K., Kwiatkowski, A., Balis, J. U., Cola, G. d., Deleu, T., Goulão, M., Kallinteris, A., KG, A., Krimmel, M., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Shen, A. T. J., and Younis, O. G. Gymnasium, March 2023. URL <https://zenodo.org/record/8127025>.

- Villani, C. *Optimal transport: old and new*. Springer, 2009.
- Wang, T., Zhou, D., and Gu, Q. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. In *Advances in Neural Information Processing Systems*, pp. 13524–13536, 2021.
- Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25:2620–2648, 2019.
- Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Appendix

In Section A, we elaborate the general optimal transport based switching cost, which we specialized in the main body of the paper. Algorithmic details are provided in Section B. Additional experimental results are given in Section C. Section D is dedicated to proofs.

A. General Formula for Transport Switching Cost

In this section, we aim to provide generalizations for the transport switching cost in two directions: a) employ general measurements of similarity between two actions instead of indicator functions; b) consider the partition with multiple components, i.e. $L > 2$. The construction is inspired by a technique that are widely used to obtain the convergence rate of empirical Wasserstein distance (Fournier & Guillin, 2015; Weed & Bach, 2019; Lei, 2020; Staudt & Hundrieser, 2023).

A.1. The Classic OT Theory

Before delving into the details of the cost construction, we introduce two concepts from the optimal transport theory.

Definition A.1 (Feasible transport plan). Given any measure spaces (X, Σ_X, μ) and (Y, Σ_Y, ν) . Then for any measure σ on $(X \times Y, \Sigma_X \otimes \Sigma_Y)$, we say σ is a feasible transport plan between μ and ν if for any $A \in \Sigma_X$ and any $B \in \Sigma_Y$ we have

$$\sigma(A \times Y) = \mu(A), \quad \text{and} \quad \sigma(X \times B) = \nu(B),$$

and we write as $\sigma \in \mathcal{C}(\mu, \nu)$.

Definition A.2 (Optimal transport plan). Given measure spaces (X, Σ_X, μ) and (Y, Σ_Y, ν) and any nonnegative measurable function $c : X \times Y \rightarrow \mathbb{R}_{\geq 0}$ satisfies some continuity conditions (see e.g. Villani, 2009, Theorem 4.1). Then we say $\sigma^* \in \mathcal{C}(\mu, \nu)$ is an optimal transport plan if and only if

$$\sigma^* \in \arg \min_{\sigma \in \mathcal{C}(\mu, \nu)} \left\{ \int_{X \times Y} c(x, y) d\sigma(x, y) \right\}.$$

A.2. The Generalized Costs

Now we are ready to introduce our construction. For every practical problem, the action space could be naturally divided into several groups, which then forms a partition of \mathcal{A} , denoted by $\{\mathcal{A}_\ell\}_{\ell=1}^L$. Therefore, for each fixed state s , when switching from $\pi_o(\cdot|s)$ to $\pi_n(\cdot|s)$, the learning cost is to consider the probability mass that is transported *between* different components of $\{\mathcal{A}_\ell\}_{\ell=1}^L$. While the transaction cost focuses on the probability mass that moves *within* each component of $\{\mathcal{A}_\ell\}_{\ell=1}^L$. We elaborate the intuition in the coming paragraphs

Learning cost. For any $s \in \mathcal{S}$, let $a_\ell^s := \pi_o(\mathcal{A}_\ell|s)$, $b_\ell^s := \pi_n(\mathcal{A}_\ell|s)$, then we immediately have the following decomposition

$$\pi_o(\cdot|s) = \sum_{\ell=1}^L a_\ell^s \pi_{o,\ell}(\cdot|s) \quad \text{and} \quad \pi_n(\cdot|s) = \sum_{\ell=1}^L b_\ell^s \pi_{n,\ell}(\cdot|s)$$

where $\pi_{o,\ell}(\cdot|s) := \pi_o(\cdot|s)I_{\mathcal{A}_\ell}/a_\ell^s$ and $\pi_{n,\ell}(\cdot|s) := \pi_n(\cdot|s)I_{\mathcal{A}_\ell}/b_\ell^s$ are conditional distributions on \mathcal{A}_ℓ . Then if $a_\ell^s \neq b_\ell^s$, we need to transport $|a_\ell^s - b_\ell^s|$ amount of mass in or out of \mathcal{A}_ℓ , which is captured by the following two measures on \mathcal{A} :

$$\rho^s := \sum_{\ell=1}^L (a_\ell^s - b_\ell^s)_+ \pi_{o,\ell}(\cdot|s) \quad \text{and} \quad \eta^s := \sum_{\ell=1}^L (b_\ell^s - a_\ell^s)_+ \pi_{n,\ell}(\cdot|s), \quad (\text{A1})$$

with $(a_\ell^s - b_\ell^s)_+ := (a_\ell^s - b_\ell^s)I_{\{a_\ell^s - b_\ell^s \geq 0\}}$ and $(b_\ell^s - a_\ell^s)_+ := (b_\ell^s - a_\ell^s)I_{\{b_\ell^s - a_\ell^s \geq 0\}}$. If we further define $\tau_\ell^s := a_\ell^s \wedge b_\ell^s$. Then ρ^s determines how much mass above τ_ℓ^s should be moved from $\pi_o(\cdot|s)$ in each \mathcal{A}_ℓ . Similar intuition applies to η^s . Thus, any feasible transport plan between ρ^s and η^s , i.e.

$$\gamma^s \in \mathcal{C}(\rho^s, \eta^s), \quad (\text{A2})$$

would lead to the first step of transportation between π_o^s and π_n^s , i.e. mass moving (while the second step would be shape matching in each component), and the induced cost during this cross-component transportation models the learning cost. Specifically, we define the learning cost as

$$\mathcal{L}_{c_1}(\pi_o(\cdot|s), \pi_n(\cdot|s)) := \int_{\mathcal{A} \times \mathcal{A}} c_1^s(x, y) d\gamma^s(x, y), \quad (\text{A3})$$

where $c_1^s : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}_{\geq 0}$ is the cost function measures the similarity/distance between two actions.

Transaction cost. The above transport plan γ^s guarantees that $\pi_o(\cdot|s)$ has the same amount of mass as $\pi_n(\cdot|s)$ by moving across different components. Then inside each \mathcal{A}_ℓ with $a_\ell^s \leq b_\ell^s$, we also need to properly rearrange the mass within \mathcal{A}_ℓ , such that the mass has same distribution as $\pi_n(\cdot|s)$, as the second step of a plan, and the cost incurred by this within-component rearrangement is transaction cost. Such rearrangement can be described by

$$\lambda^s := \sum_{\ell=1}^L \tau_\ell^s \lambda_\ell^s, \quad (\text{A4})$$

where each $\lambda_\ell^s \in \mathcal{C}(\pi_{o,\ell}(\cdot|s), \pi_{n,\ell}(\cdot|s))$. With another cost function $c_2^s : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}_{\geq 0}$, the transaction cost is defined as the induced cost during this within-component rearrangement:

$$\mathcal{T}_{c_2}(\pi_o(\cdot|s), \pi_n(\cdot|s)) := \int_{\mathcal{A} \times \mathcal{A}} c_2^s(x, y) d\lambda^s(x, y). \quad (\text{A5})$$

Moreover, the following proposition justifies that the combination of the two steps produces a feasible transportation between π_o^s and π_n^s .

Proposition A.3 (Feasibility of the proposed transport plan). *Given old policy π_o and an candidate new policy π_n . For each fixed $s \in \mathcal{S}$, suppose γ^s and λ^s are defined as (A2) and (A4), we have $\gamma^s + \lambda^s \in \mathcal{C}(\pi_o^s, \pi_n^s)$.*

Proposition A.3 assures that any feasible transport plan γ^s and λ^s will lead to a feasible transport plan between $\pi_o(\cdot|s)$ and $\pi_n(\cdot|s)$ and lead to a transport switching cost via (A3) and (A5). In fact, it is possible to be more ambitious by choosing γ^s and λ^s to be the optimal/near-optimal transport plan. In the following proposition, we demonstrate that the formulation of transport switching cost we defined in (6) and (7) can be seen as the optimal value of (A3) and (A5) for specific choice of cost functions c_1^s and c_2^s .

Proposition A.4 (Respective optimality of the cost terms). *Let $L = 2$, i.e. $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$ and $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$. For each fixed $s \in \mathcal{S}$, we define $c_1^s(x, y) = \sum_{i,j=1}^2 I_{\mathcal{A}_i \times \mathcal{A}_j}(x, y)$ and $c_2^s(x, y) \equiv 1$, for any $(x, y) \in \mathcal{A} \times \mathcal{A}$. In this case, we have*

$$\begin{aligned} (6) &= \min \left\{ \int_{\mathcal{A} \times \mathcal{A}} c_1^s(x, y) d\gamma^s(x, y) : \gamma^s \in \mathcal{C}(\rho^s, \eta^s) \right\}, \\ (7) &= \sum_{\ell=1}^2 \tau_\ell^s \min \left\{ \int_{\mathcal{A} \times \mathcal{A}} c_2^s(x, y) d\lambda_\ell^s(x, y) : \lambda_\ell^s \in \mathcal{C}(\pi_{o,\ell}(\cdot|s), \pi_{n,\ell}(\cdot|s)) \right\}. \end{aligned}$$

Before discussing the main theoretical result (i.e. Proposition 3.7), we need to explicitly emphasize that, although the above generalized costs simultaneously incorporate two directions of extensions compared to (6) and (7), what Proposition 3.7 is concerned with is only the extension of (6) and (7) by allowing $L > 2$, which is formally defined by

$$\begin{aligned} F(\pi_o(\cdot|s), \pi_n(\cdot|s)) &:= c_l \mathcal{L}(\pi_o(\cdot|s), \pi_n(\cdot|s)) + c_t \mathcal{T}(\pi_o(\cdot|s), \pi_n(\cdot|s)), \quad \text{where} \\ \mathcal{L}(\pi_o(\cdot|s), \pi_n(\cdot|s)) &:= \sum_{\ell=1}^L (a_\ell^s - b_\ell^s)_+, \quad \text{and} \\ \mathcal{T}(\pi_o(\cdot|s), \pi_n(\cdot|s)) &:= \sum_{\ell=1}^L a_\ell^s \wedge b_\ell^s. \end{aligned}$$

In such a way, for any $L \geq 2$, $\mathcal{L}(\pi_o(\cdot|s), \pi_n(\cdot|s))$ captures the total amount of mass to move across different elements of the partition $\{\mathcal{A}_\ell\}_{\ell=1}^L$, and $\mathcal{T}(\pi_o(\cdot|s), \pi_n(\cdot|s))$ depicts the total amount of mass remaining in each regions.

And for completeness, in the following we can re-state the Proposition 3.7 with formal OT terminologies:

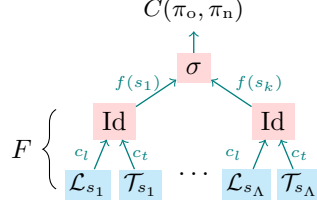


Figure A1: Proposed cost function family. Here we use the shorthands $\mathcal{L}_{s_i} := \mathcal{L}(\pi_o(\cdot|s_i), \pi_n(\cdot|s_i))$ and $\mathcal{T}_{s_i} := \mathcal{T}(\pi_o(\cdot|s_i), \pi_n(\cdot|s_i))$.

Algorithm 2 Offline Net Value Evaluation

Input: Offline data \mathcal{D} , initial values in target net Q-function parameters $\{\phi'_j\}_{j \in [M]}$, net Q-function parameters $\{\phi_i\}_{i \in [M]}$, policy π , learning rates ρ_{nq}, ρ_{stb} .

- 1: **repeat**
- 2: Sample a mini-batch $B = \{(s, a, r, s')\}$ from \mathcal{D}
- 3: Generate $a' \sim \pi(\cdot|s')$, compute update target by

$$y(r, s') = r - (1 - \gamma)C(\pi_o, \pi) + \gamma \min_{j \in [M]} Q_{N, \phi'_j}(s', a')$$
- 4: For each $i \in [M]$, update net Q-function $\{Q_{N, \phi_i}\}_{i \in [M]}$ by

$$\phi_i \leftarrow \phi_i - \rho_{nq} \nabla_{\phi_i} \sum_{(s, a, r, s') \in B} [\{y(r, s') - Q_{N, \phi_i}(s, a)\}^2]$$
- 5: Update target net Q-function by $\phi'_j \leftarrow \rho_{stb} \phi'_j + (1 - \rho_{stb}) \phi_j$
- 6: **until** Convergence criterion met
- 7: $Q_N^\pi(s_0, a) := \min_{i \in [M]} Q_{N, \phi_i}(s_0, a)$, for all $a \in \mathcal{A}$

Output: $V_N^\pi(s_0) = \mathbb{E}_{a \sim \pi(\cdot|s_0)}[Q_N^\pi(s_0, a)]$

Proposition A.5 (Optimality of the proposed cost). *Given old policy π_o , candidate policy π_n and any $\{\mathcal{A}_\ell\}_{\ell=1}^L$ as a partition of \mathcal{A} , i.e. $\mathcal{A} = \cup_{\ell=1}^L \mathcal{A}_\ell$ and $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for all $i \neq j$. Let $c(x, y) = c_l \sum_{i \neq j} I_{\mathcal{A}_i \times \mathcal{A}_j}(x, y) + c_t \sum_{\ell=1}^L I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y)$ with constants $c_l \geq c_t$. Then for any $s \in \mathcal{S}$, we have*

$$F(\pi_o(\cdot|s), \pi_n(\cdot|s)) = \min \left\{ \int_{\mathcal{A} \times \mathcal{A}} c(x, y) d\sigma^s(x, y) : \sigma^s \in \mathcal{C}(\pi_o(\cdot|s), \pi_n(\cdot|s)) \right\}.$$

In particular, when $L = 2$, the left-hand side of the above recovers exactly the cost in (5), with learning cost and transaction costs chosen as (6) and (7), respectively.

A.3. The Graphical Interpretation of the Cost Family

Finally we provide a visual explanation of the cost family defined in (4) by the Figure A1.

B. Further Details on Numerical Experiments

In this section we provide additional details in the algorithm of NAC as well as implementation techniques in related experiments.

B.1. Offline Evaluation

The offline evaluation method used in Algorithm 1 is presented here as Algorithm 2, which is mainly inspired by the widely used offline fitted-Q evaluation (Riedmiller, 2005; Munos & Szepesvári, 2008; Le et al., 2019; Ma et al., 2023).

B.2. Stopping of the Algorithms

For evaluation purpose in Algorithm 2, since it is used either for a fine evaluation of the given old policy or the finally found new policy, and such numerical results are directly used for comparisons between such two policies, we need

Algorithm 3 Stopping criterion

Input: The list of average estimated net values of new policies in the last 2 epochs $[v_1, v_2]$. Net value of the old policy v_o . Net value increase rate upper bound $\alpha > 0$, increase upper bound $b_u > 0$, decrease lower bound $b_d > 0$, stopping flag $\beta = 0$.

```

1: if  $v_0 > 0$  then
2:   if  $v_1 > (1 + \alpha)v_0$  and  $v_2 > (1 + \alpha)v_0$  then
3:      $\beta = 1$ 
4:   end if
5: else
6:   if  $v_1 > 0$  and  $v_2 > 0$  then
7:      $\beta = 1$ 
8:   end if
9: end if
10: if  $v_1 \geq v_0 + b_u$  and  $v_2 > v_0 + b_u$  then
11:    $\beta = 1$ 
12: end if
13: if  $v_1 \leq v_0 - b_d$  and  $v_2 > v_0 - b_d$  then
14:    $\beta = 1$ 
15: end if

```

Output: Stop the training when $\beta = 1$.

both evaluation process to nearly converge, which only needs the total number of epochs (each epoch contains 1000 evaluation/training steps) to be large.

On the other hand, the case of the policy learning process in Algorithm 1, i.e. line 2-8, is more complicated. First, due to offline settings, especially with quite weak old policy as the teacher, the sample distribution of transition tuples in a given offline dataset can be very different from one generated by an optimal policy. If the total number of epochs are too high, not only the later training epochs are possibly not contributing to improving the policy, but also the loss in either net values or net Q-networks may diverge due to over-fitting. Motivated by such observations, we introduce a set of stopping criterion, which contains the following several requirements:

First, we set a threshold named “epochs_stop”, which is the least number of epochs for policy learning, and we never stop the training before the epoch number reaches “epochs_stop”. Second, as presented in Algorithm 3, we terminate when the current new policy either significantly improves over the old policy or has been even worse for consecutive 2 epochs. This ensures that the NAC training part will be appropriately stopped even when the old policy is optimal or highly suboptimal. All hyper-parameters will be explicitly provided in Section B.

B.3. Training Stability

As a important universal observation in offline RL, the Q-networks during the training process will be over-optimistic on state-action pairs that do not appear in the offline dataset. As in half of the cases in experiments, we deal with very weak old policies, some of which even have negative net values. The offline data and the policy cloning effect due to the existence of switching costs will further enhance such over-optimistic behaviour. As a result, to prevent highly volatile updates in each training step, we perform gradient clipping on both the gradient w.r.t. policy parameters and Q-network parameters. Those values are also reported in the next subsection.

B.4. Hyper-Parameters

In this subsection we provide hyper-parameters for each distinct experiment settings. Note that the implementation is also inspired by the Spinningup project (Achiam, 2018).

B.5. Compute Resources

Our experiments ran on a single Precision 7875 Tower workstation, with AMD Ryzen Threadripper PRO 7945WX CPU (64 MB cache, 12 cores, 24 threads, 4.7GHz to 5.3GHz), NVIDIA RTX 6000 Ada GPU. In the training process, the

Table 4: Shared hyper-parameters.

Parameter	value
Number of repetitions	10
Offline data size	1000000
Batch sample size	256
Train from scratch	False
c_l	5
Random seeds	{4, 5, ..., 13}
Number of Monte Carlo state samples for cost function in training	10
Number of Monte Carlo state samples for cost function in evaluation	10000
Steps per epoch	1000
Number of epochs in training	100
Number of epochs in evaluation	50
Discount rate γ	0.99
Learning rate	0.0003
M	2
Number of Monte Carlo action samples for net value estimate in evaluation	10000
Maximum length of one episode	1000
Maximum 2-norm for gradient in Q-networks	1
Epochs_stop	20
Net value increase upper bound b_u	50
Net value decrease bound b_d	10
Optimizer	Adam (Kingma & Ba, 2015)

Table 5: Hyper-parameters for Ant-v4, (sub)optimal old policy.

Parameter	value
Maximum 2-norm for gradients in net values	1
Number of Monte Carlo action samples for net value estimate in training	1000
Net value increase rate upper bound α	1
c_t	{0, 1}

Table 6: Hyper-parameters for HalfCheetah-v4, (sub)optimal old policy.

Parameter	value
Maximum 2-norm for gradients in net values	5
Number of Monte Carlo action samples for net value estimate in training	2000
Net value increase rate upper bound α	0.15
c_t	{0, 0.1}

Table 7: Hyper-parameters for Hopper-v4, (sub)optimal old policy.

Parameter	value
Maximum 2-norm for gradients in net values	1
Number of Monte Carlo action samples for net value estimate in training	1000
Net value increase rate upper bound α	1
c_t	{0, 0.1}

Table 8: Hyper-parameters for SUMO-RL that are different from Gymnasium, (sub)optimal old policy.

Parameter	value
Environment name	sumo-rl-v0
Intersection type	Simple intersection
Offline data size	100000
Maximum 2-norm for gradients in net values	None
Maximum 2-norm for gradients in net Q-functions	None
Number of epochs in training	150
Maximum length of one episode	100
Steps per epoch	400
Net value increase rate upper bound α	0.5
Net value increase upper bound b_u	5.0
Net value decrease bound b_d	5.0
(c_l, c_t)	$\{(0.5, 0.01), (5.0, 0.1), (10.0, 1.0)\}$

memory needed was around 7.2GB. The time to get all results was within one week.

B.6. The SUMO-RL Environment

As we implement our algorithm in a novel real problem, the traffic control problem, we first introduce the general problem setting, and then describe the concerned technical details of the environment we used in the experiments.

B.6.1. THE TRAFFIC CONTROL PROBLEM

Efficient traffic signal control is crucial for urban areas, as it directly impacts congestion levels, travel times, fuel consumption, and overall quality of life for residents. Frequent policy switches due to online algorithms can potentially influence traffic safety (Han et al., 2023). In contrast, our approach focuses on training a traffic signal control policy using only offline data, striking a balance between maximizing future expected rewards (speeds of vehicles) and minimizing potential switching cost (such as temporary traffic congestion, update of facilities). To this end, we employ the SUMO-RL environment (Alegre, 2019) designed for developing and assessing traffic control algorithms in realistic urban scenarios.

B.6.2. ENVIRONMENT DETAILS

Simulation of Urban MObility (SUMO) is a widely used simulation system of transport and large road networks (Lopez et al., 2018). And it has been adopted to an RL-friendly interface by SUMO-RL (Alegre, 2019) for both single agent and multi-agent scenarios. While we refer readers to the documentations of both open-source projects for complete set of details, here we introduce core technical settings of SUMO-RL interface.

State. Given a complex traffic network, there are several/single intersection(s). The agent(s) is allowed to observe the traffic information and adaptively change the traffic lights phases. Especially, the observation, i.e. the state variables, is a long vector with both discrete and continuous entries. The state variables include the phase of the current traffic lights, a binary variable ‘min.green’ indicating whether minimum green light time has passed, the density of vehicles in each lane of each directions (North, South, East, West, each containing multiple lanes) as well as the respective number of queuing vehicles in each lane.

Action. Actions correspond to distinct traffic light configurations in each intersection, i.e. the ways one can change the lights. And to model the real life cases, there is another constant ‘yellow.time’ so that after a new action being executed, the light will be turned yellow for such amount of seconds before turning green/red.

Reward. For different goals of a problem setting, one can define different instant rewards, for example, the total delay around intersections (the summation of all queuing times of all nearby vehicles). While in our experiment, we focus on the average speed of all nearby vehicles.

In our implementation, by default we only provide textual output. However, it is ready to generate graphical user interface

Table 9: Additional performance of the NAC algorithm on various Gymnasium benchmarks. 1st column: environment considered. 2nd column: $\dim(\mathcal{S})$. 3rd column: $\dim(\mathcal{A})$. 4th column: (sub)optimality of the old policy π_o . 5-7th columns: performance measures. The performance measures are meant as mean \pm std. The average net values of old policies are -14.2 (Ant-v4), -52.8 (HalfCheetah-v4), 17.0 (Hopper-v4).

Environment \mathcal{M}	d_S	d_A	Old policy π_o	c_t	Improvement	Switch proportion	Responsible rate
Ant-v4	27	8	suboptimal	1	55.2 ± 23.2	90.0%	100.0%
HalfCheetah-v4	17	6	suboptimal	0.1	24.1 ± 8.9	100.0%	100.0%
Hopper-v4	11	3	suboptimal	0.1	52.0 ± 19.8	100.0%	100.0%
Ant-v4	27	8	optimal	1	/	0.0%	100.0%
HalfCheetah-v4	17	6	optimal	0.1	/	0.0%	100.0%
Hopper-v4	11	3	optimal	0.1	/	0.0%	100.0%

to visualize results. Further details can be found in the aforementioned documentations and our source code.

B.7. Limitations

Throughout the paper we considered a general cost formulation relying on optimal transport (OT). We paid specific attention to costs within this class, specified by (6) and (7). One can always consider more general costs, but this instantiation of the costs is theoretically justified, probably the simplest to explain, and already provides a more fine-grained quantification for the switching cost compared to existing approach (local and global switching cost), conveying the key ideas.

C. Further Experimental Results

In this section we provide additional results in Gymnasium environments, when the coefficient of transaction cost c_t takes nonzero values, followed by ablation study.

C.1. Additional Results

We implement NAC when $c_t \in \{0.1, 1\}$ to provide comparisons to the above results when $c_t = 0$, the results of which are presented in Table 9.

C.2. Ablation Study.

Here we mainly want to understand if the scale of cost functions influence the policy learning performance in different environments. As seen in Table 2 and 9, we increased c_t from 0 to 0.1 or 1. Especially note that, to guarantee fair comparisons, in each environment, apart from c_t , all hyperparameters are kept exactly the same, independent of c_t values or the (sub)optimality of the given old policies, which can be checked according to Table 5, 6 and 7. Finally, by comparing the results environment-wise, we can see that, when increasing c_t by an appropriate value, the performance in both optimal and suboptimal old policy cases are similar or slightly better than when $c_t = 0$. Such observation is important, as it shows that NAC training process is robust to different scaling of cost functions, which makes it applicable to different scenarios.

D. Proofs

This section is about our proofs.

D.1. Proof of Lemma 3.4(a)

Proof. By Assumption 3.3, both $\{V^\pi(s_0)\}_{\pi \in \Pi}$ and $\{C(\pi_o, \pi)\}_{\pi \in \Pi}$ are compact in the topology generated by open sets in \mathbb{R} , which then implies that both sets are sequentially compact. Then consider the set of resulting net values $\{V_N^\pi(s_0)\}_{\pi \in \Pi}$. For any sequence $(z_i)_{i \geq 1} \subseteq \{V_N^\pi(s_0)\}_{\pi \in \Pi}$, by definition of net values, we must have sequences $(x_i)_{i \geq 1} \subseteq \{V^\pi(s_0)\}_{\pi \in \Pi}$ and $(y_i)_{i \geq 1} \subseteq \{C(\pi_o, \pi)\}_{\pi \in \Pi}$, such that $z_i = x_i - y_i$ for all i (and specifically, for each given i , x_i, y_i corresponds to value and cost of the same policy). Since $\{V^\pi(s_0)\}_{\pi \in \Pi}$ is sequentially compact, there exists a subsequence $(x_{i_j})_{j \geq 1}$

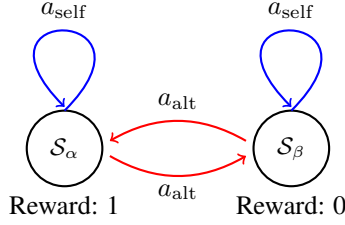


Figure D2: Illustration of the constructed MDP.

of $(x_i)_{i \geq 1}$ such that for some $x \in \{V^\pi(s_0)\}_{\pi \in \Pi}$, $x_{i_j} \rightarrow x$ as $j \rightarrow +\infty$. On top of such sequence of indices $(i_j)_{j \geq 1}$, Since $\{\bar{C}(\pi_0, \pi)\}_{\pi \in \Pi}$ is sequentially compact, there exists a further subsequence of $(i_j)_{j \geq 1}$, denoted as $(i_{j_k})_{k \geq 1}$ such that $y_{i_{j_k}} \rightarrow y$, as $k \rightarrow +\infty$, for some $y \in \{C(\pi_0, \pi)\}_{\pi \in \Pi}$. So we immediately know $(z_{i_{j_k}})_{k \geq 1}$, as a subsequence of $(z_i)_{i \geq 1}$, converges to $x - y$. Then $\{V_N^\pi(s_0)\}_{\pi \in \Pi}$ is sequentially compact, and especially attains its supremum by some policy in Π . \square

D.2. Proof of Lemma 3.4(b)

Proof. For an arbitrarily given MDP, all we need is to construct a counter-example, so that the optimal policy is not switch-optimal. So we just discuss how to design such a counter-example. Given current behaviour policy π_0 and some fixed initial state s_0 , let's consider an arbitrary policy π and the optimal policy in value function π^* . By definition of optimality, we know $V^{\pi^*}(s_0) \geq V^\pi(s_0)$, and especially we denote the gap by M , i.e. $M := V^{\pi^*}(s_0) - V^\pi(s_0)$. Now as long as in some problem settings, the cost function C is larger in π^* , i.e. $C(\pi_0, \pi^*) > C(\pi_0, \pi)$, it could then be the case that V_N^π dominates that of π^* . To be more specific, whenever $C(\pi_0, \pi^*) - C(\pi_0, \pi) > M$, we would have $V_N^\pi > V_N^{\pi^*}$, making π^* not switch-optimal. \square

D.3. Proof of Proposition 3.4(c)

Proof. Given an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, with a fixed initial state s_0 , let's consider the following example, which is also depicted in Figure D2. To begin with, let $\mathcal{S} = \{s_\alpha, s_\beta\}$, and $\mathcal{A} = \{a_{\text{self}}, a_{\text{alt}}\}$. At any state, taking action a_{self} means trying to stay in the same state, guaranteed by letting $P(s|s, a_{\text{self}}) = 1$ for any $s \in \mathcal{S}$. Meanwhile, we also let $P(s_\alpha|s_\beta, a_{\text{alt}}) = P(s_\beta|s_\alpha, a_{\text{alt}}) = 1$, so that whenever the action a_{alt} is taken at any state, the environment would transit the agent to the other state. Then we define the rewards as $r(s_\alpha, a_{\text{self}}) = r(s_\beta, a_{\text{alt}}) = 1$, while $r(s_\alpha, a_{\text{alt}}) = r(s_\beta, a_{\text{self}}) = 0$. Together with the definition of the transition dynamics, it just means that the reward is 1 if and only if the state to arrive at is s_α , and vanishes otherwise. Finally, choose $\gamma = 0.99$ for simplicity.

With such environment, any policy π from the pool of feasible policies Π takes the form of $\pi = \{\pi(\cdot|s_\alpha), \pi(\cdot|s_\beta)\}$. Now let's construct a specific example, where the current behaviour policy is π_0 , with $\pi_0(a|s) = 1/2$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. We focus on the case when there is no transaction cost. In addition, for any candidate policy π to switch to from π_0 , the learning cost is high whenever in one state π is a stochastic policy. And actually let such high learning cost to be 500. On the other hand, when π is always deterministic at any state, the learning cost is low. Especially, if in all states, the action to execute is the same, cost is 25, while 50 if actions to take are different in different states.

By the above settings, all the candidate policies (excluding π_0) can be divided into 2 groups: $\{\pi_{n1}, \pi_{n2}, \pi_{n3}, \pi_{n4}\}$, and $\Pi \setminus \{\pi_0, \pi_{n1}, \pi_{n2}, \pi_{n3}, \pi_{n4}\}$, where $\pi_{n1}(a_{\text{self}}|s) = 1$ for any $s \in \mathcal{S}$; $\pi_{n2}(a_{\text{alt}}|s) = 1$ for any $s \in \mathcal{S}$; $\pi_{n3}(a_{\text{self}}|s_\alpha) = 1$ and $\pi_{n3}(a_{\text{alt}}|s_\beta) = 1$; $\pi_{n4}(a_{\text{alt}}|s_\alpha) = 1$ and $\pi_{n4}(a_{\text{self}}|s_\beta) = 1$.

Now let's calculate the values and net values for all policies in the first group. First we consider π_{n1} . If initial state is s_α , since one would always take a_{self} , the agent would stay in s_α , to earn reward 1 recursively for all rounds, leading to

$$V^{\pi_{n1}}(s_\alpha) = \sum_{i=0}^{\infty} 0.99^i \cdot 1 = \frac{1}{1-0.99} = 100.$$

On the other hand, if $s_0 = s_\beta$, then the agent would stay in s_β , receiving zero rewards, so that $V^{\pi_{n1}}(s_\beta) = 0$. By such way, we can continue to know $V^{\pi_{n2}}(s_\alpha) = 49.75$, $V^{\pi_{n2}}(s_\beta) = 50.25$ for π_{n2} ; $V^{\pi_{n3}}(s_\alpha) = V^{\pi_{n3}}(s_\beta) = 100$ for π_{n3} ; and

$V^{\pi_{n4}}(s_\alpha) = V^{\pi_{n4}}(s_\beta) = 0$ for π_{n4} . Recall there is low costs of switch 25 for first 2 policies and 50 for the next 2, we finally know $V_N^{\pi_{n1}}(s_\alpha) = 75$, $V_N^{\pi_{n1}}(s_\beta) = -25$ for π_{n1} ; $V_N^{\pi_{n2}}(s_\alpha) = 24.75$, $V_N^{\pi_{n2}}(s_\beta) = 25.25$ for π_{n2} ; $V_N^{\pi_{n3}}(s_\alpha) = V_N^{\pi_{n3}}(s_\beta) = 50$ for π_{n3} ; and $V_N^{\pi_{n4}}(s_\alpha) = V_N^{\pi_{n4}}(s_\beta) = -50$ for π_{n4} .

Let's consider the second group. Due to the problem settings with horizon $\gamma = 0.99$ and the maximal immediate reward of 1, the highest possible return from any initial state is bounded by 100, then for all policies in the second group, the net value in any state is bounded by -400 due to the high cost, and could never be switch-optimal in any state, given the results in the first group.

For complete comparisons, we could know that for the current policy π_o , net values share the same numbers as its values, which are $V^{\pi_o}(s_\alpha) = V^{\pi_o}(s_\beta) = 50$.

By comparing the net values among all policies, we observe that, if $s = s_\alpha$, then the switch-optimal is π_{n1} ; while, if $s = s_\beta$, then either the switch-optimal is π_{n3} or that we just don't make a switch. □

D.4. Proof of Proposition 3.4(d)

Proof. The proof would be quite straight-forward if we follow the example settings in the Proof in Appendix D.3 in the above. As discussed there, we know that π_{n1} is switch-optimal at initial state $s = s_\alpha$. Then to compute the corresponding net Q-functions, first consider the case $Q_N^{\pi_{n1}}(s_\alpha, a_{\text{self}})$. If a_{self} is executed at state s_α , according to the transition dynamics, the agent would remain in such state until termination, leading to $Q^{\pi_{n1}}(s_\alpha, a_{\text{self}}) = 100$, and then $Q_N^{\pi_{n1}}(s_\alpha, a_{\text{self}}) = 75$. On the other hand, if a_{alt} is executed at state s_α , it would arrive at s_β and remain there, having $Q^{\pi_{n1}}(s_\alpha, a_{\text{alt}}) = 0$, and $Q_N^{\pi_{n1}}(s_\alpha, a_{\text{alt}}) = -25$. Following the same idea, we can immediately know $Q^{\pi_{n3}}(s_\alpha, a_{\text{alt}}) = 99$ and $Q_N^{\pi_{n3}}(s_\alpha, a_{\text{alt}}) = 49 > Q_N^{\pi_{n1}}(s_\alpha, a_{\text{alt}}) = -25$, showing that π_{n1} is not switch-optimal for every action $a \in \mathcal{A}$. □

D.5. Proof of Proposition 3.6

Proof. The proof consists of two parts. First we want to show that such net Bellman operator B^π is a contraction map on $\mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$ under $\|\cdot\|_\infty$ -norm. Then we show such repeated implementations of the contraction lead to the unique evaluation. Without loss of generality, we focus on the proof when state and action spaces are finite, while we note that the proof can be readily extended to the continuous case.

Now, take arbitrarily two net Q-functions $Q_N^{\pi_1}, Q_N^{\pi_2} \in \mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$, we observe that

$$\begin{aligned} & \|B^\pi Q_N^{\pi_1} - B^\pi Q_N^{\pi_2}\|_\infty \\ &= \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |(R(s,a) - (1-\gamma)C(\pi_0, \pi) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')} Q_N^{\pi_1}(s', a')) \\ & \quad - (R(s,a) - (1-\gamma)C(\pi_0, \pi) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')} Q_N^{\pi_2}(s', a'))| \\ &= \gamma \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |[\mathbb{E}_{s' \sim P(\cdot|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')} Q_N^{\pi_1}(s', a')] - [\mathbb{E}_{s' \sim P(\cdot|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')} Q_N^{\pi_2}(s', a')]| \\ &= \gamma \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\mathbb{E}_{s' \sim P(\cdot|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')} [Q_N^{\pi_1}(s', a') - Q_N^{\pi_2}(s', a')]| \\ &\leq \gamma \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |Q_N^{\pi_1}(s, a) - Q_N^{\pi_2}(s, a)| = \gamma \|Q_N^{\pi_1} - Q_N^{\pi_2}\|_\infty. \end{aligned}$$

By such we know B^π is a contraction mapping, whenever $\gamma \in [0, 1)$. After that we review the following theorem.

Theorem D.1 (Banach Fixed-point Theorem). *For a non-empty complete metric space (X, d) with contraction $\mathcal{T} : X \rightarrow X$, \mathcal{T} has a unique fixed point $\mathbf{x}^* \in X$. In addition, starting from arbitrary point $\mathbf{x}_0 \in X$, and define a new sequence as $\{\mathbf{x}_n\} = \{\mathcal{T}\mathbf{x}_{n-1}\}$, then we have $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{x}^*$.*

Since the concerned net Q-functions are defined based a finite horizon MDP with bounded reward function, they are contained in $\mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$. Given that $\mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$ is complete when metrized by $\|\cdot\|_\infty$ -norm (see e.g. Folland, 1999, pp.121), starting with a $Q_N^0 \in \mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$, Q_N^k converges to a fixed point $Q_N^* \in \mathcal{G}(\mathcal{S} \times \mathcal{A}, \|\cdot\|_\infty)$, which, by the definition of fixed point, satisfies the net Bellman equation:

$$Q_N^*(s, a) = R(s, a) - (1-\gamma)C(\pi_0, \pi) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')} Q_N^*(s', a')$$

Still due to the Banach fixed-point theorem, we know the corresponding fixed point Q_N^* is unique, which means that $Q_N^* = Q_N^\pi$. Then such iterations of backup would converge, i.e. $\lim_{k \rightarrow \infty} Q_N^k = Q_N^\pi$. \square

D.6. Proof of Proposition A.3

Proof. Writing $\sigma^s := \gamma^s + \lambda^s$, we only need to show that for any measurable set $G \subseteq \mathcal{A}$, we have $\sigma^s(G \times \mathcal{A}) = \pi_o(G|s)$ and $\sigma^s(\mathcal{A} \times G) = \pi_n(G|s)$. Recall that $\gamma^s \in \mathcal{C}(\rho^s, \eta^s)$ and $\lambda^s = \sum_{\ell=1}^L \tau_\ell^s \lambda_\ell^s$ with each $\lambda_\ell^s \in \mathcal{C}(\pi_o(\cdot|s), \pi_n(\cdot|s))$, thus we have

$$\sigma^s(G \times \mathcal{A}) = \gamma^s(G \times \mathcal{A}) + \sum_{\ell=1}^L (a_\ell \wedge b_\ell) \lambda_\ell(G \times \mathcal{A}) = \rho^s(G) + \sum_{\ell=1}^L (a_\ell \wedge b_\ell) \pi_{o,\ell}(G|s).$$

Then we take into the explicit form of ρ^s defined in (A1), it yields that

$$\begin{aligned} \sigma^s(G \times \mathcal{A}) &= \sum_{\ell=1}^L (a_\ell - b_\ell)_+ \pi_{o,\ell}(G|s) + \sum_{\ell=1}^L (a_\ell \wedge b_\ell) \pi_{o,\ell}(G|s) \\ &= \sum_{\ell=1}^L a_\ell \pi_{o,\ell}(G|s) = \pi_o(G|s). \end{aligned}$$

A similar calculation can be carried out to obtain $\sigma^s(\mathcal{A} \times G) = \pi_n(G|s)$. Hence the claim is verified. \square

D.7. Proof of Proposition A.4

Proof. When $L = 2$, recall the definition of ρ^s and η^s in (A1), we have

$$\begin{aligned} \rho^s &= (a_1^s - b_1^s)_+ \pi_{o,1}(\cdot|s) + (a_2^s - b_2^s)_+ \pi_{o,2}(\cdot|s), \\ \eta^s &= (b_1^s - a_1^s)_+ \pi_{n,1}(\cdot|s) + (b_2^s - a_2^s)_+ \pi_{n,2}(\cdot|s). \end{aligned}$$

Since $\pi_o(\cdot|s)$ and $\pi_n(\cdot|s)$ are probability measures, exactly one of two cases must hold: either (1) $a_1^s \geq b_1^s$ and $b_2^s \geq a_2^s$, or (2) $a_1^s \leq b_1^s$ and $b_2^s \leq a_2^s$. Consequently, ρ^s and η^s take one of two forms: either $\rho^s = (a_1^s - b_1^s)_+ \pi_{o,1}(\cdot|s)$ and $\eta^s = (b_2^s - a_2^s)_+ \pi_{n,2}(\cdot|s)$, or $\rho^s = (a_2^s - b_2^s)_+ \pi_{o,2}(\cdot|s)$ and $\eta^s = (b_1^s - a_1^s)_+ \pi_{n,1}(\cdot|s)$. Without loss of generality, we assume the former case. Given that γ^s is a feasible transport plan between ρ^s and η^s , it must concentrate on $\mathcal{A}_1 \times \mathcal{A}_2 \subseteq \mathcal{A}_1 \times \mathcal{A}$, thus we have

$$\int_{\mathcal{A} \times \mathcal{A}} c_1^s(x, y) d\gamma^s(x, y) = \gamma^s(\mathcal{A}_1 \times \mathcal{A}) \stackrel{(a)}{=} (a_1^s - b_1^s)_+ \pi_{o,1}(\mathcal{A}_1) = a_1^s - b_1^s = (6).$$

where equality (a) is by the definition of feasible transport plan in Definition A.1. Note this holds for all feasible transport plan, it naturally becomes the optimal transport cost.

As for the transaction cost, since that λ_ℓ^s represents a feasible transport plan between the distributions $\pi_{o,\ell}(\cdot|s)$ and $\pi_{n,\ell}(\cdot|s)$ and both distributions are concentrated on \mathcal{A}_ℓ , we have λ_ℓ^s must concentrate on $\mathcal{A}_\ell \times \mathcal{A}_\ell$. Consequently, by the form of λ^s defined (A4), it follows that

$$\int_{\mathcal{A} \times \mathcal{A}} c_2^s(x, y) d\lambda^s(x, y) = \tau_1^s \lambda_1^s(\mathcal{A}_1 \times \mathcal{A}_1) + \tau_2^s \lambda_2^s(\mathcal{A}_2 \times \mathcal{A}_2) = \tau_1^s + \tau_2^s = (7).$$

Again, since the calculation above does not depend on the specific choice of each λ_ℓ^s , it coincides with the case when we choosing each λ_ℓ^s as the optimal one. \square

D.8. Proof of Proposition 3.7

Proof. We show the statement by going through several steps. First we provide a novel perspective to decompose the cost of any feasible transport plan; then we calculate the cost of one specific transport plan, which corresponds to our proposal in the transport cost by (6) and (7) up to any L (in the sense that the plan exactly attains a cost of $F(\pi_o(\cdot|s), \pi_n(\cdot|s))$); finally we verifies that any other feasible transport plan can never achieve a cost strictly lower than the above case.

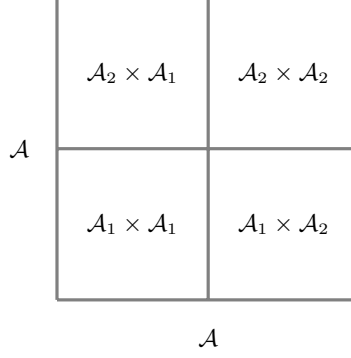


Figure D3: The product action space $\mathcal{A} \times \mathcal{A}$ is partitioned into four components.

Step 1 (Decomposition of the cost). For any feasible transport plan $\sigma^s = \gamma^s + \lambda^s$, by definition the cost is

$$\begin{aligned}
 & \int_{\mathcal{A} \times \mathcal{A}} c(x, y) d\sigma^s(x, y) \\
 &= \int_{\mathcal{A} \times \mathcal{A}} \sum_{l=1}^L [I_{\mathcal{A}_l \times \mathcal{A}}(x, y) c(x, y)] d\sigma^s(x, y) \\
 &= \sum_{l=1}^L \int_{\mathcal{A} \times \mathcal{A}} I_{\mathcal{A}_l \times \mathcal{A}}(x, y) c(x, y) d\sigma^s(x, y) \\
 &= \sum_{l=1}^L \int_{\mathcal{A}_l \times \mathcal{A}} c(x, y) d\sigma^s(x, y),
 \end{aligned}$$

where the first inequality is due to the fact that, for any valid partition $\{\mathcal{A}_l\}_{l=1}^L$ of the set \mathcal{A} , and any $x, y \in \mathcal{A}$, one always has

$$\sum_{l=1}^L I_{\mathcal{A}_l \times \mathcal{A}}(x, y) \equiv 1.$$

Now for convenience, we introduce a new notation by

$$d_l^s := \int_{\mathcal{A}_l \times \mathcal{A}} c(x, y) d\sigma^s(x, y),$$

which immediately leads to the following simplified decomposition:

$$\int_{\mathcal{A} \times \mathcal{A}} c(x, y) d\sigma^s(x, y) = \sum_{l=1}^L d_l^s.$$

To begin with, for each $l \in [L]$, d_l^s just represents the total cost due to any transport which has a starting location inside \mathcal{A}_l , which provides a novel and straightforward view of the total cost due to any feasible transport plan σ . In addition, as $\{\mathcal{A}_l\}_{l=1}^L$ is a partition, the decomposition of total cost is naturally constructed, allowing a clear comparison between any two different transport plans, which will be showcased in the following steps.

Step 2 (The case of the proposed transport cost). Now we specify one transport plan $\sigma^s = \gamma^s + \lambda^s$, where γ^s is any feasible transport plan defined by (A2), i.e. γ is one feasible way to transport all the surplus mass $\{(a_\ell^s - b_\ell^s)_+\}_{\ell=1}^L$ respectively out of $\{\mathcal{A}_\ell\}_{\ell=1}^L$; while λ is one feasible choice according to (A4), i.e. a way for rearrangement. As Proposition A.3 has already established that η^s is a feasible transport plan, its cost is well-defined, which is further computed by the

sum of all d_ℓ^s , where, for each $\ell \in [L]$,

$$\begin{aligned}
 d_\ell^s &= \int_{\mathcal{A}_\ell \times \mathcal{A}} c(x, y) d\sigma^s(x, y) \\
 &= \int_{\mathcal{A}_\ell \times \mathcal{A}} c_l \left[\sum_{i \neq j} I_{\mathcal{A}_i \times \mathcal{A}_j}(x, y) \right] + c_t \left[\sum_{i=1}^L I_{\mathcal{A}_i \times \mathcal{A}_i}(x, y) \right] d\sigma^s(x, y) \\
 &= c_l \int_{\mathcal{A}_\ell \times \mathcal{A}} \left[\sum_{i \neq j} I_{\mathcal{A}_i \times \mathcal{A}_j}(x, y) \right] d\sigma^s(x, y) + c_t \int_{\mathcal{A}_\ell \times \mathcal{A}} \left[\sum_{i=1}^L I_{\mathcal{A}_i \times \mathcal{A}_i}(x, y) \right] d\sigma^s(x, y) \\
 &= c_l \left[\sum_{j \neq \ell} \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}_j}(x, y) d\sigma^s(x, y) \right] + c_t \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y) \\
 &= c_l \left[\sum_{j=1}^L \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}_j}(x, y) d\sigma^s(x, y) \right] - c_l \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y) + c_t \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y) \\
 &= c_l \int_{\mathcal{A}_\ell \times \mathcal{A}} \sum_{j=1}^L I_{\mathcal{A}_\ell \times \mathcal{A}_j}(x, y) d\sigma^s(x, y) + (c_t - c_\ell) \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y) \\
 &= c_l \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}}(x, y) d\sigma^s(x, y) + (c_t - c_\ell) \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y).
 \end{aligned}$$

Now for the first term, by the property of any feasible transport plan described in Definition A.1, we know

$$\begin{aligned}
 c_l \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}}(x, y) d\sigma^s(x, y) &= c_l \sigma^s(\mathcal{A}_\ell \times \mathcal{A}) \\
 &= c_l (\gamma^s + \lambda^s)(\mathcal{A}_\ell \times \mathcal{A}) \\
 &= c_l [\gamma^s(\mathcal{A}_\ell \times \mathcal{A}) + \lambda^s(\mathcal{A}_\ell \times \mathcal{A})] \\
 &= c_l [\rho^s(\mathcal{A}_\ell) + \lambda^s(\mathcal{A}_\ell \times \mathcal{A})] \\
 &= c_l \left[\sum_{i=1}^L (a_i^s - b_i^s)_+ \pi_{o,i}(\mathcal{A}_\ell | s) \right] + c_l \left[\sum_{i=1}^L \tau_i^s \lambda_i^s \right] (\mathcal{A}_\ell \times \mathcal{A}) \\
 &= c_l \left[\sum_{i=1}^L (a_i^s - b_i^s)_+ \pi_{o,i}(\mathcal{A}_\ell | s) \right] + c_l \left[\sum_{i=1}^L (a_i^s \wedge b_i^s) \pi_{o,i}(\mathcal{A}_\ell | s) \right] \\
 &= c_l [(a_\ell^s - b_\ell^s)_+ + a_\ell^s \wedge b_\ell^s].
 \end{aligned}$$

And then for the second term,

$$\begin{aligned}
 (c_t - c_\ell) \int_{\mathcal{A}_\ell \times \mathcal{A}_\ell} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y) &= (c_t - c_\ell) \sigma^s(\mathcal{A}_\ell \times \mathcal{A}_\ell) \\
 &= (c_t - c_\ell) [\gamma^s(\mathcal{A}_\ell \times \mathcal{A}_\ell) + \lambda^s(\mathcal{A}_\ell \times \mathcal{A}_\ell)].
 \end{aligned}$$

Due to the definition of γ^s in (A2) and the property of $(a_\ell^s - b_\ell^s)_+$, we know there is at most one of $(a_\ell^s - b_\ell^s)_+$ and $(b_\ell^s - a_\ell^s)_+$ that is non-zero, and more importantly, for an arbitrarily given $\ell \in [L]$, without loss of generality, if we assume $a_\ell^s - b_\ell^s \geq 0$, then we immediately know $(a_\ell^s - b_\ell^s)_+ \geq 0$, while $(b_\ell^s - a_\ell^s)_+ = 0$. In addition,

$$\begin{aligned}
 \gamma^s(\mathcal{A}_i, \mathcal{A}_i) &\leq \gamma^s(\mathcal{A}, \mathcal{A}_\ell) \\
 &= \eta^s(\mathcal{A}_\ell) \\
 &= \sum_{i=1}^L (b_i^s - a_i^s)_+ \pi_{n,i}(\mathcal{A}_\ell | s) \\
 &= (b_\ell^s - a_\ell^s)_+ \pi_{n,\ell}(\mathcal{A}_\ell | s) = 0 \\
 &\Rightarrow \gamma^s(\mathcal{A}_i, \mathcal{A}_i) = 0, \quad \text{for any } i \in [L].
 \end{aligned}$$

On the other hand, by similarly noticing that, for any $\ell \in [L]$, λ_ℓ^s vanishes on any $\mathcal{A}_{\ell'} \times \mathcal{A}_{\ell'}$ with $\ell' \neq \ell$, we further know

$$\begin{aligned} \lambda^s(\mathcal{A}_\ell \times \mathcal{A}_\ell) &= \sum_{i=1}^L \tau_i^s \lambda_i^s(\mathcal{A}_\ell \times \mathcal{A}_\ell) = \tau_\ell^s \lambda_\ell^s(\mathcal{A}_\ell \times \mathcal{A}_\ell) = \tau_\ell^s = a_\ell^s \wedge b_\ell^s \\ &\Rightarrow (c_t - c_\ell) \int_{\mathcal{A}_\ell \times \mathcal{A}_\ell} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y) = (c_t - c_\ell) a_\ell^s \wedge b_\ell^s. \end{aligned}$$

Combining the above two terms,

$$d_\ell^s = c_l[(a_\ell^s - b_\ell^s)_+ + a_\ell^s \wedge b_\ell^s] + (c_t - c_\ell) a_\ell^s \wedge b_\ell^s = c_l(a_\ell^s - b_\ell^s)_+ + c_t(a_\ell^s \wedge b_\ell^s).$$

Therefore,

$$\int_{\mathcal{A} \times \mathcal{A}} c(x, y) d\sigma^s(x, y) = \sum_{l=1}^L d_l^s = c_l \sum_{\ell=1}^L (a_\ell^s - b_\ell^s)_+ + c_t \sum_{\ell=1}^L a_\ell^s \wedge b_\ell^s = F(\pi_o(\cdot|s), \pi_n(\cdot|s)).$$

Step 3 (Costs of other transport plans). As we have already derived in the above step, for any feasible transport plan σ (no longer to be restricted to the proposed ones due to (A2) and (A4)), we always have

$$\begin{aligned} \int_{\mathcal{A} \times \mathcal{A}} c(x, y) d\sigma^s(x, y) &= \sum_{l=1}^L d_l^s \quad \text{with} \\ d_\ell^s &= c_l \left[\sum_{j \neq \ell} \int_{\mathcal{A}_\ell \times \mathcal{A}_j} I_{\mathcal{A}_\ell \times \mathcal{A}_j}(x, y) d\sigma^s(x, y) \right] + c_t \int_{\mathcal{A}_\ell \times \mathcal{A}_\ell} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y). \end{aligned}$$

As in each problem setting, π_o, π_n and $\{\mathcal{A}_\ell\}$ are all given, we know $\{\pi_o(\mathcal{A}_\ell)\}_{\ell=1}^L$ are all nonnegative constants with a sum of 1. Further notice that

$$\begin{aligned} \sum_{j \neq \ell} \int_{\mathcal{A}_\ell \times \mathcal{A}_j} I_{\mathcal{A}_\ell \times \mathcal{A}_j}(x, y) d\sigma^s(x, y) + \int_{\mathcal{A}_\ell \times \mathcal{A}_\ell} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y) &= \int_{\mathcal{A}_\ell \times \mathcal{A}} \sum_{j=1}^L I_{\mathcal{A}_\ell \times \mathcal{A}_j}(x, y) d\sigma^s(x, y) \\ &= \int_{\mathcal{A}_\ell \times \mathcal{A}} I_{\mathcal{A}_\ell \times \mathcal{A}}(x, y) d\sigma^s(x, y) \\ &= \sigma(\mathcal{A}_\ell \times \mathcal{A}) \\ &= \pi_o(\mathcal{A}_\ell|s). \end{aligned}$$

So if we denote

$$J_\ell^s := \sum_{j \neq \ell} \int_{\mathcal{A}_\ell \times \mathcal{A}_j} I_{\mathcal{A}_\ell \times \mathcal{A}_j}(x, y) d\sigma^s(x, y),$$

we immediately have

$$d_\ell^s = c_l J_\ell^s + c_t [\pi_o(\mathcal{A}_\ell|s) - J_\ell^s].$$

Before proceeding, we first show an intermediate statement:

$$J_\ell^s \geq (a_\ell^s - b_\ell^s)_+.$$

This is because: first, if $a_\ell^s - b_\ell^s \leq 0$, then as J_ℓ^s is defined through a sum of nonnegative measures, thus the statement holds. Then if $a_\ell^s - b_\ell^s > 0$,

$$\begin{aligned} J_\ell^s &= \sum_{j \neq \ell} \int_{\mathcal{A}_\ell \times \mathcal{A}_j} I_{\mathcal{A}_\ell \times \mathcal{A}_j}(x, y) d\sigma^s(x, y) \\ &= \pi_o(\mathcal{A}_\ell|s) - \int_{\mathcal{A}_\ell \times \mathcal{A}_\ell} I_{\mathcal{A}_\ell \times \mathcal{A}_\ell}(x, y) d\sigma^s(x, y) \\ &= a_\ell^s - \sigma^s(\mathcal{A}_\ell \times \mathcal{A}_\ell). \end{aligned}$$

In addition, since

$$\sigma^s(\mathcal{A}_\ell \times \mathcal{A}_\ell) \leq \sigma^s(\mathcal{A} \times \mathcal{A}_\ell) = \pi_n(\mathcal{A}_\ell|s) = b_\ell^s,$$

we finally know $J_\ell^s \geq (a_\ell^s - b_\ell^s) = (a_\ell^s - b_\ell^s)_+$.

Now by knowing $d_\ell^s = c_l J_\ell^s + c_t[\pi_o(\mathcal{A}_\ell|s) - J_\ell^s]$, together with the fact that $J_\ell^s \geq (a_\ell^s - b_\ell^s)_+$ and $c_l \geq c_t$, we know d_ℓ^s attains its minimum when $J_\ell^s = (a_\ell^s - b_\ell^s)_+$, and such minimum is

$$\begin{aligned} c_l(a_\ell^s - b_\ell^s)_+ + c_t[\pi_o(\mathcal{A}_\ell|s) - (a_\ell^s - b_\ell^s)_+] &= c_l(a_\ell^s - b_\ell^s)_+ + c_t[a_\ell^s - (a_\ell^s - b_\ell^s)_+] \\ &= c_l(a_\ell^s - b_\ell^s)_+ + c_t(a_\ell^s \wedge b_\ell^s). \end{aligned}$$

As a result, for any feasible transport plan σ , its corresponding cost is lower bounded by

$$c_l \sum_{\ell=1}^L (a_\ell^s - b_\ell^s)_+ + c_t \sum_{\ell=1}^L a_\ell^s \wedge b_\ell^s = F(\pi_o(\cdot|s), \pi_n(\cdot|s)),$$

which shows that $F(\pi_o(\cdot|s), \pi_n(\cdot|s))$ is the solution to the OT problem. □

D.9. Global and Local Switching Costs are Special Cases

Lemma D.2. *Let π_1 and π_2 be two policies on a state space \mathcal{S} with finite cardinality. When considering policy switching from π_1 to π_2 , recall that the induced global and the local switching costs are defined as*

$$C^{\text{gl}}(\pi_1, \pi_2) = I_{\{\pi_1 \neq \pi_2\}}, \quad C^{\text{loc}}(\pi_1, \pi_2) = \sum_{s \in \mathcal{S}} I_{\{\pi_1(\cdot|s) \neq \pi_2(\cdot|s)\}}.$$

Then one can get back C^{gl} and C^{loc} as a specific case of the cost family (4) by the parameters given in Table 1.

Proof. One can recover the global switching cost as follows.

$$\begin{aligned} C(\pi_1, \pi_2) &= \sigma \left(\int_{\mathcal{S}} f(s) F(\pi_1(\cdot|s), \pi_2(\cdot|s)) d\mu(s) \right) = I_{\{\sum_{s \in \mathcal{S}} \frac{1}{|\mathcal{S}|} I_{\{\pi_1(\cdot|s) \neq \pi_2(\cdot|s)\}} > 0\}} \\ &= I_{\{\sum_{s \in \mathcal{S}} I_{\{\pi_1(\cdot|s) \neq \pi_2(\cdot|s)\}} > 0\}} = C^{\text{gl}}(\pi_1, \pi_2). \end{aligned}$$

One can get back the local switching cost as follows.

$$\begin{aligned} C(\pi_1, \pi_2) &= \sigma \left(\int_{\mathcal{S}} f(s) F(\pi_1(\cdot|s), \pi_2(\cdot|s)) d\mu(s) \right) = |\mathcal{S}| \sum_{s \in \mathcal{S}} \frac{1}{|\mathcal{S}|} I_{\{\pi_1(\cdot|s) \neq \pi_2(\cdot|s)\}} \\ &= \sum_{s \in \mathcal{S}} I_{\{\pi_1(\cdot|s) \neq \pi_2(\cdot|s)\}} = C^{\text{loc}}(\pi_1, \pi_2). \end{aligned}$$

□