Performance of NPG in Countable State-Space Average-Cost RL

Yashaswini Murthy
Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Urbana, IL 61801
ymurthy2@illinois.edu

Isaac Grosof
Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Urbana, IL 61801
igrosof@illinois.edu

Siva Theja Maguluri
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
siva.theja@gatech.edu

R. Srikant
Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Urbana, IL 61801
rsrikant@illinois.edu

Abstract

We consider policy optimization methods in reinforcement learning settings where the state space is arbitrarily large, or even countably infinite. The motivation arises from control problems in communication networks, matching markets, and other queueing systems. We consider Natural Policy Gradient (NPG), which is a popular algorithm for finite state spaces. Under reasonable assumptions, we derive a performance bound for NPG that is independent of the size of the state space, provided the error in policy evaluation is within a factor of the true value function. We obtain this result by establishing new policy-independent bounds on the solution to Poisson's equation, i.e., the relative value function, and by combining these bounds with previously known connections between MDPs and learning from experts.

1 Introduction

We are motivated by control problems in queueing models of resource allocation, such as those arising in communication networks, cloud computing systems, and riding hailing services. Examples of such systems include the following:

(a) The switch fabric in Internet routers and data centers where packets have to be transported (or switched) from one of many input ports to one of many output ports [SY13]: the system is modeled as a bipartite graph with input ports on one side and output ports on the other side. Technological constraints dictate that at each time slot, a matching must be selected in the bipartite graph, and packets are transferred along the edges of the matching from each input to the corresponding output. The goal is to find a sequence of matchings to minimize either the average delay experienced by the packets in the switch or the probability that the delay exceeds some threshold.

(b) Scheduling problems at base stations in 5G networks [SY13]: at a central controller (typically the base station associated with a cell in a cellular network), packets arrive and are queued in a separate queue for each receiver. The goal is to schedule these packets over different frequencies and time slots to minimize the average delay of the packets in the system, while taking into account the time-varying channel conditions in a wireless network due to fading and other wireless medium effects.

Preprint. Under review.

(c) Scheduling workloads in cloud computing systems [MSV$^+$19]: a workload in such systems takes the form of a collection directed acyclic graphs, where each DAG represents a job, the nodes in the graphs represent tasks in the job and the directed edges represent precedence relationships among the tasks in the graph. The goal is to allocate resources to tasks from a sequence of arriving jobs, while respecting the precedence relationships of the tasks within each job and minimizing the average delay experienced by the jobs.

(d) Customer-driver matching in ride hailing platforms such as Uber and Lyft [ÖW20, VBMW23]: the role of such platforms can be modeled as controlling the number of nodes in a bipartite graph, where one side is the set of waiting customers and the other side in the set of available drivers. The goal of a ride hailing platform is to choose a set of prices and match customers to drivers so that a weighted combination of the average delay experienced by customers and the average profit is optimized.

The above problems exhibit several common features:

(i) The state space of these problems is discrete, typically consisting of the queue lengths of the various entities waiting in the system such as packets, customers, drivers, jobs and tasks, depending on the context. Discrete state spaces are commonly studied in the reinforcement learning (RL) literature; however, in our applications, the state space is also countably infinite for all practical purposes, since queue lengths can become unbounded. In some applications, such as communication networks, the packet buffers may be finite but it is well known that modeling them as infinite buffers leads to good scheduling algorithms [SY13]. It should be noted that even if one were to model the finiteness of the buffers explicitly in our model, our results will still hold, and our performance guarantees would not depend on the size of the buffers.

(ii) Because we are dealing with a vector of queue lengths as the state of the system, the problems have some limited amount of structure that can and should be exploited to design good algorithms. In particular, it is relatively straightforward to design algorithms that ensures that the system is stable, i.e., the queue length is finite with probability one [SY13]. On the other hand, algorithms to optimize performance objectives such as average delay are unknown except in limited regimes [ES12, MS16]. Therefore, data-driven approaches such as reinforcement learning (RL) are natural candidates to solve such problems.

(iii) Due to a well-known result called Little's law, minimizing average delay is equivalent to minimizing average queue lengths [LG08]. Thus, the natural instantaneous cost in such problems is the current total queue length. Note that unlike many RL models, this cost is unbounded and results which assume that the costs (or rewards) at each (state, action) pair are uniformly bounded do not hold for our problems.

Given the above background, our goal in this paper is to study policy optimization algorithms for such countable state space models with discrete, finite action spaces where the cost is proportional to the total queue length in the system, and can thus grow in an unbounded fashion. For this purpose, we study the natural policy gradient (NPG) algorithm. Our main contributions are the following:

(1) One of the standard regret results for NPG uses a connection to a fundamental learning-theoretic problem called the best-experts problem. We show that this standard analysis does not apply to our problem due to the unbounded nature of the instantaneous cost in our problem.

(2) We show that one can obtain nontrivial regret bounds by making a small, but critical, change to the step-size used in the best-experts algorithm, and by obtaining bounds on the relative value function (called the solution of the Poisson's equation in the applied probability literature).

(3) An important component of our work is to obtain bounds on the solution of Poisson's equation that are uniform across all policies. To the best of our knowledge, prior works on obtaining bounds on the solution of Poisson's equation are limited to specific policies. A key contribution of our paper is to show that uniform bounds can be obtained by exploiting certain structural properties of the mathematical models for the motivating applications mentioned earlier.

(4) Policy evaluation using temporal difference learning and Monte Carlo methods have been well studied in the literature, so we do not consider them explicitly in this paper. However, we do consider the error due to function approximation. Traditionally, for analytical purposes, it is assumed that there is a uniform bound on the function approximation error of the value function.

We argue that such an assumption does not make sense for queueing models and propose a more general model for the function approximation. Existing mathematical tools for the study of convergence of RL algorithms cannot handle our proposed model for the function approximation error. However, we show that, by exploiting the special structure of our queueing models and the associated bounds on the solution to Poisson's equation, we can obtain non-trivial regret bounds for policy optimization.

1.1 Related Work

The Natural Policy Gradient algorithm is a well-known and extensively studied algorithm for MDP optimization, in both the average-reward and discounted-reward settings [Kak01, AKLM21, GSP19, MS23, MMS23]. An important line of research on the NPG algorithm treats the MDP-optimization problem as many parallel instances of the expert advice problem, and treats the NPG algorithm as many parallel instances of the weighted majority algorithm. Even-dar et al. [EDKM09] use this approach to prove the first convergence result for NPG in the finite-state average-reward tabular setting, and [AYBB+19] expand upon that result to incorporate function approximation. Our result uses the same "parallel weighted majority" framing, but generalizes the result to the infinite-state-space setting by incorporating state-dependent learning rates.

Policy gradient algorithms have been studied in certain specialized settings with average-reward uncountably-infinite state spaces [FGKM18, KT08]: the Linear Quadratic Regulator and the base-stock inventory control problem, demonstrating rapid convergence to the optimal policy. However, follow-up study of these settings has demonstrated that they exhibit additional structure which is critical to these results, causing these policy-gradient algorithms to act like policy improvement algorithms [BR24]. Our result is the first to handle an infinite state-space setting without the specialized structure of these prior results.

Key to our result are novel bounds on the relative value function, building off of our drift assumption for the policy space. This drift assumption is reasonable in a queueing setting, as we discuss in section 3.2 [SY13]. Prior drift-based bounds on the relative-value function exist [GM96], but are policy-dependent. In contrast, we prove policy-independent bounds on the relative-value function using reasonable assumptions on the MDP structure, which are motivated by the structure of MDPs in queueing networks. Our policy-independent bounds are critical to implement our state-dependent learning rates, allowing us to generalize the NPG algorithm to the infinite-state setting.

Several works have studied applications of reinforcement learning to queueing problems, including policy-gradient-based algorithms. In [SXX20], discounted reward reinforcement learning is applied to queue length control. Their analysis assumes bounded single-step rewards and the presence of an oracle capable of generating samples from the MDP to learn the value functions. This combination of discounted rewards and bounded single-step rewards provides a uniform, state-independent upper bound on the value function. However, this does not apply to our analysis since the single-step cost, represented by the current queue length, is unbounded and our objective is the infinite horizon average cost.

In [LXM19], a model-based approach addresses the challenge of a countable state space by using thresholding. That is, the optimal policy is learned for a finite subset of low-cost states using model-based methods, while a fixed stabilizing policy is applied outside this finite set. The optimality gap is then determined based on the size of this finite set. Contrary to their approach, our assumptions accommodate a soft thresholding, allowing for the learning of better policies beyond a finite set. Additionally, our analysis is model-free.

Several results focus on the problem of learning the relative value function from samples, including variance reduction techniques [DG22] and sample augmentation techniques [WLWY23]. Our results complement these results, as we focus on the function approximation step, and prove results on overall algorithmic performance, while these results focus on the policy evaluation step, and empirically demonstrate performance improvements. Dai and Gluzman [DG22] in particular empirically demonstrate that with variance reduction techniques in use, policy gradient algorithms with function approximation rapidly converge to the optimal policy in an infinite-state-space queueing setting. Our results theoretically justify this empirical observation.

2 Model and Preliminaries

We consider the class of Markov Decision Processes (MDP) with countably infinite states $\mathcal{S}$ (possibly $\mathbb{Z}_+^K$), finite actions $\mathcal{A}$ and the infinite horizon average cost objective. Particularly, we consider the context of queuing systems where each state of the MDP is denoted by a vector $\mathbf{q} \in \mathcal{S}$. Each element of the vector $\mathbf{q}$ indicates the queue length for a specific type of job, with a total of K different job types. We assume the buffer associated with the queue is of infinite capacity and therefore we obtain a state space whose cardinality is countably infinite. We consider a randomized class of policies Π, where a policy $\pi \in \Pi$ maps each state to a probability vector over actions $\mathcal{A}$, that is, $\pi : \mathcal{S} \to \Delta\mathcal{A}$. The actions in the context of queues denote which particular job type to service, if any.

The underlying probability transition kernel is denoted by $\mathbb{P} : \mathcal{S} \to \mathcal{S}$ and the transition kernel corresponding to any policy π is denoted by $\mathbb{P}_\pi$, where $\mathbb{P}_\pi(\mathbf{q}'|\mathbf{q}) = \sum_{a\in\mathcal{A}} \pi(a|\mathbf{q})\mathbb{P}(\mathbf{q}'|\mathbf{q}, a)$ is the probability of transitioning from $\mathbf{q}$ to $\mathbf{q}'$ under policy π in a single step. Associated with each state $\mathbf{q}$ is a single step cost $c(\mathbf{q})$, which in the context of queues is the total queue length, that is, $c(\mathbf{q}) = \|\mathbf{q}\|_1$. Since we consider unbounded queue lengths in this formulation, the single step costs also are unbounded. In this scenario, it is important to note that the costs incurred in a single step do not depend on the action taken or the policy being used at that moment. This is because the policy only influences the queue length in the next time step, not the current one. The infinite horizon average reward associated with a policy π is denoted by J_π, and is defined as follows:

$$J_\pi = \lim_{T\to\infty} \frac{\mathbb{E}_\pi\left[\sum_{t=0}^{T-1} \|\mathbf{q}_t\|_1\right]}{T} \tag{1}$$

where the expectation is taken with respect to the trajectory generated by $\mathbb{P}_\pi$. If the transition kernel $\mathbb{P}_\pi$ admits a unique stationary distribution d_π over the state space, then the infinite horizon average reward can be reformulated as $J_\pi = \sum_{\mathbf{q}\in\mathcal{S}} d_\pi(\mathbf{q})\|\mathbf{q}\|_1$. If a function $V_\pi : \mathcal{S} \to \mathbb{R}$ associated with a policy π is absolutely integrable, that is it satisfies:

$$\sum_{\mathbf{q}'\in\mathcal{S}} \mathbb{P}_\pi(\mathbf{q}'|\mathbf{q})|V_\pi(\mathbf{q}')| < \infty \tag{2}$$

and is a solution to the Poisson's equation:

$$J_\pi + V_\pi(\mathbf{q}) = \|\mathbf{q}\|_1 + \sum_{\mathbf{q}'\in\mathcal{S}} \mathbb{P}_\pi(\mathbf{q}'|\mathbf{q})V_\pi(\mathbf{q}'), \tag{3}$$

then $V_\pi(\mathbf{q})$ is defined as the relative value function associated with the policy π [GI24]. Since $V_\pi(\mathbf{q})$ is unique upto an additive constant, any function of the form $V_\pi(\mathbf{q}) + C$, where C is a constant is also a solution to the Poisson's equation. However, the most frequently used representation of the value function, which is also unique, is given by:

$$V_\pi(\mathbf{q}) = \mathbb{E}_\pi\left[\sum_{i=0}^{\tau_0^\pi - 1} \left(\|\mathbf{q}_i\|_1 - J_\pi\right) \middle| \mathbf{q}_0 = \mathbf{q}\right] \tag{4}$$

where τ_0^π represents the first time to hit state $\mathbf{0}$ starting from any state $\mathbf{q}$ under policy π. Hence, from definition it follows that $V_\pi(\mathbf{0}) = 0$. The value function associated with a state $\mathbf{q}$ represents the expected difference between the total cost and the expected total cost obtained under policy π when starting from state $\mathbf{q}$ until state $\mathbf{0}$ is reached for the first time. The relative state action value function $Q_\pi(\mathbf{q})$ is analogously defined as the solution to the following equation:

$$J_\pi + Q_\pi(\mathbf{q}, a) = \|\mathbf{q}\|_1 + \sum_{\mathbf{q}'\in\mathcal{S}} \mathbb{P}(\mathbf{q}'|\mathbf{q}, a)V_\pi(\mathbf{q}') \tag{5}$$

The state action value function $Q_\pi(\mathbf{q}, a)$ has a similar interpretation as state value function $V_\pi(\mathbf{q})$ except the action enacted at time 0 is a and not dictated by the policy π.

The goal of reinforcement learning is to determine the policy $\pi^* \in \Pi$, such that the infinite horizon average cost is minimized. That is, to solve for

$$J^* = \min_{\pi\in\Pi} J_\pi \tag{6}$$

where $\pi^* = \arg\min_{\pi\in\Pi} J_\pi$. The focus of this paper is to analyze the performance of Natural Policy Gradient in determining the optimal policy that minimizes the infinite horizon average reward.

2.1 Natural Policy Gradient Algorithm

Natural Policy Gradient algorithm is related to the mirror descent algorithm in the context of tabular policies. The objective of mirror descent involves minimizing the first order approximation of the average cost with KL regularizer. In the context of tabular policies, the NPG policy update is of the form below:

$$\pi_{i+1}(a|\mathbf{q}) \propto \pi_i(a|\mathbf{q}) \exp\left(-\eta_{\mathbf{q}} \widehat{Q}_{\pi_i}(\mathbf{q}, a)\right) \tag{7}$$

where $\eta_{\mathbf{q}} > 0$ is the state dependent step size. Since in the limit as $\eta_{\mathbf{q}} \to \infty$, the above update picks the action with the lowest state action value function, NPG is also considered to be a form of soft policy iteration. The magnitude of $\eta_{\mathbf{q}}$ determines the greediness of the policy.

With the state space being infinitely large, a common approach to evaluate value functions is through linear function approximations. This simplifies the complexity from infinity to the dimension of the parameter vector, although with some loss in accuracy. A popular method involves using neural networks, where the weights act as the parameter vector and the network itself serves as the feature space. For each policy π, the estimate $\widehat{Q}_\pi$ of the state-action value function Q_π is then computed using overparametrized neural networks and samples gathered from trajectories under policy π. For further details, please see Subsection 4.1.

Algorithm 1: Natural Policy Gradient Algorithm

Require : T, $\pi_0 \in \Delta\mathcal{A}$

for $i = 0, 1, 2, 3, \cdots, T-1$ **do**

  Generate trajectory $\{\mathbf{q}_0, a_0, \mathbf{q}_1, a_1, \ldots, \mathbf{q}_n, a_n\}$ using policy π_i. Evaluate $\widehat{Q}_{\pi_i}$ using neural network linear function approximation.

  Update policy as:

$$\pi_{i+1}(a|\mathbf{q}) = \frac{\pi_i(a|\mathbf{q}) \exp\left(-\eta_{\mathbf{q}} \widehat{Q}_{\pi_i}(\mathbf{q}, a)\right)}{\sum_{a' \in \mathcal{A}} \pi_i(a'|\mathbf{q}) \exp\left(-\eta_{\mathbf{q}} \widehat{Q}_{\pi_i}(\mathbf{q}, a')\right)} \tag{8}$$

end

return π_T

To aide our analysis, we make the following assumptions, which are typically met by queuing systems. The irreducibility of the Markov chain under any policy is a standard assumption in reinforcement learning. This ensures adequate exploration and visitation of all state-action pairs, which is crucial for learning policies with reasonable confidence.

Assumption 2.1. *For all policies $\pi \in \Pi$ the induced Markov Chain $\mathbb{P}_\pi$ is irreducible.*

In countable state Markov chains, irreducibility together with positive recurrence ensures the existence of the stationary distribution which aides in the proof of convergence of NPG. The next assumption ensures that the underlying Markov chain is positive recurrent (see Lemma A.2.

Assumption 2.2. *There exist constants ϵ, c independent of policy π such that for every policy $\pi \in \Pi$ the following drift equation is satisfied:*

$$\mathbb{E}_\pi\left[\|\mathbf{q}_{k+1}\|^2 - \|\mathbf{q}_k\|^2 | \mathbf{q}_k = \mathbf{q}\right] \leq -\epsilon\|\mathbf{q}\|_1 + c \tag{9}$$

In addition to ensuring positive recurrence, the drift equation (9) gives a uniform bound on the average cost of any policy. It turns out that we also need policy independent bounds on the value function, which is ensured by our next assumption on a finite set of low cost states. More precisely, consider the finite set B defined as:

$$B := \left\{\mathbf{q} \in \mathcal{S} : \|\mathbf{q}\|_1 \leq \frac{2c}{\epsilon}\right\}. \tag{10}$$

Clearly, B is a set of low cost states and is finite since c, ϵ are finite quantities.

Assumption 2.3. *Structural Assumptions on the underlying Markov Decision Process:*

- *We assume that there exist constants x_B and p_B, independent of policy π, such that*

$$\mathbb{P}_{\pi}^{x_B}\left(\mathbf{q}'|\mathbf{q}\right) \geq p_B \qquad \forall \mathbf{q} \in B, \forall \mathbf{q}' \in B, \forall \pi \in \Pi \tag{11}$$

  where $\mathbb{P}_{\pi}^{x_B}$ is the x_B-step probability transition matrix.

- *We assume that the number of jobs that can arrive and the number of jobs that can depart in a single time step are both bounded.*

Equation (11) indicates that any state $\mathbf{q} \in B$ can be reached from any state $\mathbf{q}' \in B$ in utmost x_B transitions with atleast p_B probability under any policy $\pi \in \Pi$.

Bounded arrivals and departures ensures that the length of the queues in a single transition under any policy π cannot grow unbounded. That is, there exists constants c_1, c_2 independent of policy π such that,

$$\mathbb{P}_{\pi}\left(\mathbf{q}'|\mathbf{q}\right) > 0 \implies \|\mathbf{q}'\|^2 \leq c_1\|\mathbf{q}\|^2 + c_2 \qquad \forall \pi \in \Pi \tag{12}$$

3 Main Result and Discussion

We now present the main result, which is the performance of NPG in the context of infinite state MDPs within the learning framework. We then contextualize Assumptions 2.1, 2.2 and 2.3 and elaborate on how they can be satisfied in the context of queuing systems.

3.1 Main Result

Theorem 3.1. *Consider the sequence of policies $\pi_1, \pi_2, \ldots, \pi_T$ obtained from Algorithm 1 with a state-dependent step size $\eta_{\mathbf{q}} = \sqrt{\frac{8 \log |\mathcal{A}|}{T}} \frac{1}{M_{\mathbf{q}}}$, where $M_{\mathbf{q}}$ is quadratic in $\mathbf{q}$. Let J_{π_k} be the average cost associated with policy π_k and let J_* be the minimum average cost across policy class Π. Let $\widehat{Q}_{\pi_k}$ be an estimate of state action value function Q_{π_k} associated with policy such that with probability at least $1 - \frac{\delta}{2T}$ it is true that,*

$$\left\| Q_{\pi}(\mathbf{q}, a) - \widehat{Q}_{\pi}(\mathbf{q}, a) \right\| \leq \kappa \|\mathbf{q}\|^2 \qquad \forall \mathbf{q} \in \mathcal{S}, \pi \in \Pi \tag{13}$$

Then, under Assumptions 2.1, 2.2 and 2.3, there exist constants c', c'' not depending on T or $\pi_1, \pi_2, \ldots, \pi_T$ such that with probability at least $1 - \delta$:

$$\sum_{k=1}^{T} \left(J_{\pi_k} - J_*\right) \leq c'\sqrt{T} + c''T \tag{14}$$

where $c' = \left(2\kappa + \frac{2c_1}{\epsilon}\right) \mathbb{E}_{\mathbf{q}\sim d^{\pi^*}} \|\mathbf{q}\|^2 \sqrt{\frac{\log |\mathcal{A}|}{2}} + \sqrt{\frac{\log |\mathcal{A}|}{2}} \left(\frac{2c_2}{\epsilon} + \frac{x_B c}{\epsilon}\left(1 + \frac{2}{p_B} + \frac{1}{p_B^2} + \frac{1}{p_B^3}\right)\right)$ and $c'' = \kappa \mathbb{E}_{\mathbf{q}\sim d^{\pi^*}} \|\mathbf{q}\|^2$.

Proof. The proof is in Appendix A.3. An outline is provided in Section 4.

3.2 Discussion on Assumptions: Applications in Stochastic Networks

Assumptions 2.1 and 2.3 are easily satisfied in queueing systems under mild conditions on the arrival and service process that are fairly standard in queuing literature [SY13]. For instance, suppose that the policy class allows for a non-zero probability of both zero job arrivals and zero job departures. Under such a policy class, it is possible to transition from any state $\mathbf{q}$ to the state $\mathbf{0}$ and from the state $\mathbf{0}$ to any state $\mathbf{q}$, ensuring irreducibility. Similarly, within a finite set B, it is possible to establish uniform lower bounds for transitioning from any state to any other state within the set under any policy, thereby satisfying both assumptions.

In a large class of queueing systems, the MaxWeight policy is known to ensure stability, i.e., positive recurrence (see Chapter 4, [SY13]). Assumption 2.2 is inspired by the so-called MaxWeight policy, which is known to satisfy the drift equation below:

$$\mathbb{E}_{\pi_{\mathsf{MW}}}\left[\|\mathbf{q}_{k+1}\|^2 - \|\mathbf{q}_k\|^2 | \mathbf{q}_k = \mathbf{q}\right] \leq -\epsilon\|\mathbf{q}\|_1 + d_1 \tag{15}$$

where the expectation is taken with respect to π_{MW} and ϵ, d_1 are some positive constants independent of policy. Assumption 2.2 is designed so that we explore a family of randomized policies that inherit stability from MaxWeight, while also enabling us to learn policies that outperform MaxWeight.

In particular, we consider policies obtained by using a combination of MaxWeight and arbitrary randomized acations by transforming the underlying MDP as follows. Let the policies obtained from update Equation 8 be referred to as π_{NPG}. Modify the underlying MDP such that the probability transition kernel corresponds to a policy π defined below:

$$\pi(a|\mathbf{q}) = \begin{cases} \pi_{\mathsf{NPG}}(a|\mathbf{q}), & \text{w.p.} \quad \min\left(1, \frac{1}{\lambda\|\mathbf{q}\|}\right) \\ \pi_{\mathsf{MW}}(a|\mathbf{q}), & \text{w.p.} \quad 1 - \min\left(1, \frac{1}{\lambda\|\mathbf{q}\|}\right) \end{cases} \tag{16}$$

where $\lambda > 0$ is a fixed parameter with a very small positive value.

As the queue length grows larger, the above transformed MDP enacts the Max-Weight policy with greater probability at higher queue lengths. The value of λ decides the threshold at which Max-Weight policy starts influencing the transition dynamics. Once queue lengths exceed $\frac{1}{\lambda}$, this soft thresholding compromises some optimality to prioritize stability. This differs from the hard thresholding approach taken in [LXM19].

We will now illustrate that this family of soft-thresholded policies satisfy Assumption 2.2. First note that from Assumption 2.3 and Equation 12 it is easy to show that π_{NPG} satisfies

$$\mathbb{E}_{\pi_{\mathsf{NPG}}}\left[\|\mathbf{q}_{k+1}\|^2 - \|\mathbf{q}_k\|^2 | \mathbf{q}_k = \mathbf{q}\right] \leq d_2\|\mathbf{q}\|_1 + d_3 \tag{17}$$

where the expectation is taken with respect to π_{NPG} and d_2, d_3 are some positive constants independent of policy. Thus the drift equation corresponding to policy π in Equation 16 is as follows:

$$\begin{aligned} &\mathbb{E}_{\pi}\left[\|\mathbf{q}_{k+1}\|^2 - \|\mathbf{q}_k\|^2 | \mathbf{q}_k = \mathbf{q}\right] \leq \\ &\begin{cases} \frac{d_2}{\lambda} + d_3, & \|\mathbf{q}\|_1 \leq \frac{1}{\lambda} \\ \frac{1}{\lambda\|\mathbf{q}\|_1}\left(d_2\|\mathbf{q}\|_1 + d_3\right) + \left(1 - \frac{1}{\lambda\|\mathbf{q}\|_1}\right)\left(-\epsilon\|\mathbf{q}\|_1 + d_1\right), & \|\mathbf{q}\|_1 > \frac{1}{\lambda} \end{cases} \end{aligned} \tag{18}$$

Combining the cases in Equations 18, we obtain the following drift relation for policy π for all $\mathbf{q} \in \mathcal{S}$:

$$\mathbb{E}_{\pi}\left[\|\mathbf{q}_{k+1}\|^2 - \|\mathbf{q}_k\|^2 | \mathbf{q}_k = \mathbf{q}\right] \leq -\epsilon\|\mathbf{q}\|_1 + D \tag{19}$$

where D is a constant independent of policy π but is a function of constants d_1, d_2, d_3, ϵ and λ. Note that the constant ϵ remains the same in both (15) and (19). This constitutes one such class of policies that satisfies the required the drift equation (9) for our analysis.

4 Proof outline and Key Insights

The difference in average cost associated with a policy π and the optimal average cost is linked to the Q_π function through the performance difference lemma ([Cao99]) as below:

$$J_\pi - J^* = \mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[Q_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right) - Q_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right)\right]. \tag{20}$$

Hence, the regret in LHS of Equation (14) can be captured in terms of difference in the state action action Q_π. However, in practise it is not possible to determine Q_π exactly since the model might be unknown or the state space is infinite. Hence, we incorporate the estimates $\widehat{Q}_\pi$ of the value function Q_π. If the estimates satisfy Equation (13), then from Equation (20) we obtain the following regret formulation:

$$\sum_{k=1}^{T} J_{\pi_k} - J^* \leq 2\kappa T \mathbb{E}_{\mathbf{q}\sim d^{\pi^*}}\|\mathbf{q}\|^2 + \underbrace{\mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\sum_{k=1}^{T} \widehat{Q}_{\pi_k}\left(\mathbf{q}, \pi^*(\mathbf{q})\right) - \widehat{Q}_{\pi_k}\left(\mathbf{q}, \pi_k(\mathbf{q})\right)\right]}_{(a)} \tag{21}$$

The term linear in T, i.e., $2\kappa T \mathbb{E}_{\mathbf{q}\sim d^{\pi^*}}\|\mathbf{q}\|^2$ is a consequence of function approximation and is generally unavoidable [AYBB+19].The primary task is to bound (a) in Equation (21). We approach this in four steps: (i) examining the link between NPG and prediction through expert advice as highlighted in prior literature, and identifying challenges specific to our countable state-space model and cost structure, (ii) deriving policy-independent bounds on the value functions, i.e., the solution to Poisson's Equation (3), (iii) accounting for policy evaluation errors and establishing policy-independent bounds on the estimates of the value function, and (iv) integrating all these steps to achieve the final result. We now proceed with the proof outline.

Step 1 (Connection to Weighted Averaging): This step involves connecting learning within Markov Decision Processes (MDPs) to prediction through expert advice. This connection was initially identified in [EDKM09] for MDPs and later extended to the learning setting in [AYBB$^+$19]. We now discuss this connection in some detail and why need our proof techniques to adapt this connection to the countable state-space setting. In the framework of prediction through expert advice, the agent selects an action a_t at time t, and the environment responds with a corresponding loss $l_t(a_t)$. Concurrently, an expert follows a predetermined strategy, which in our context can be simplified to a single action a^* taken at each time step, also experiencing a loss of $l_t(a^*)$. The agent's objective is to minimize the overall loss by considering all it's past observations when choosing an action. If the expert opts for a fixed strategy π^* over the available actions, the following holds true.

Theorem 4.1. *(Section 4.2, Corollary 4.2, [CBL06].) Consider the exponentially weighted average forecaster problem. Let the set of actions possible at each time step and each instance be denoted by $\mathcal{A} := \{1, \ldots, n\}$. For a fixed instance s, let $l_t(s, i)$ be the loss associated with action $i \in \mathcal{A}$ at time t such that for any pair of actions $i, i' \in \mathcal{A}$,*

$$|l_t(s, i) - l_t(s, i')| \leq M(s) \tag{22}$$

Consider the action strategy below:

$$\pi_t(i|s) = \frac{\pi_{t-1}(i|s) \exp\left(-\eta_s l_{t-1}(s, i)\right)}{\sum_{k=1}^{n} \pi_{t-1}(k|s) \exp\left(-\eta_s l_{t-1}(s, k)\right)} \tag{23}$$

Then, for any fixed policy π^*, setting $\eta_s = \sqrt{\frac{8 \log n}{T}} \frac{1}{M(s)}$ yields the following overall regret corresponding to instance s.

$$\sum_{k=1}^{T} \left(l_k\left(s, \pi_k(s)\right) - l_k\left(s, \pi^*(s)\right)\right) \leq M(s)\sqrt{\frac{T \log n}{2}} \tag{24}$$

The NPG algorithm can be interpreted as applying the weighted averaging algorithm to each state $\mathbf{q}$ in the state space, with the goal of learning the optimal policy for each state. In this context, the loss function associated with an action a in state $\mathbf{q}$ at time k is the estimate $\widehat{Q}_{\pi_k}(\mathbf{q}, a)$ of the state-action value function, where the policy in use at time k is π_k. However, as indicated by Equation (22), the loss function—$\widehat{Q}_{\pi_k}(\mathbf{q}, a)$—must be bounded for any given state $\mathbf{q}$. In finite-dimensional MDPs, a state-independent uniform bound on the state-action value function is typically assumed [AYBB$^+$19]. This is due to the fact that the step-size η is assumed to be independent of s. Note that, compared to [EDKM09, AYBB$^+$19], we have made a small, but critical, change to the best-experts algorithm by allowing the step-size η to be a function of s. When the state-space is countable, the state-action value function Q_π cannot be uniformly bounded and hence, a constant step-size cannot be assumed. With the introduction of a state-dependent step-size, we can choose a different step-size for each state using bounds on the solution to Poisson's equation, i.e., $Q_\pi(s, a)$, which depends on the state, but is uniform over all policies. Obtaining such bounds is one of the key contributions of the paper.

Step 2 (Value Function Bounds): To establish bounds on Poisson's Equation 5, we initially rely on Assumptions 2.1 and 2.2. In dealing with countable state space MDPs, along with irreducibility, we require the Markov chain to be positive recurrent for a unique stationary distribution to exist. The drift equation 9 within Assumption 2.2 ensures the positive recurrence of the underlying Markov chain. Since Q_π is related to the state value function V_π(see Equation 5), we initially constrain V_π using Assumptions 2.1 and 2.2. This leads to an upper bound on $V_\pi(\mathbf{q})$ for all $\mathbf{q} \in B^{\mathsf{c}}$, where B is defined in Equation 10.

$$V_\pi(\mathbf{q}) \leq \frac{2}{\epsilon}\|\mathbf{q}\|^2 + \underbrace{\max_{\substack{\mathbf{q}' \in B \\ \pi \in \Pi}} V_\pi(\mathbf{q}')}_{(b)} \tag{25}$$

Recall Equation (22) in the context of weighted expert averaging. The constraint on the loss function's bound $(M(s))$ must be independent of time. When applied to the NPG framework, this implies the necessity of a policy-independent upper bound on the state-action function Q_π, which, in turn, necessitates a policy-independent bound on the state value function V_π. For (b) to be well-defined, the drift alone is insufficient, as indicated in previous studies [GI24, GM96]. Addressing this is the second challenge in our analysis, which we navigate by introducing a mild structural Assumption 2.3 commonly satisfied in queuing systems.

These structural assumptions yield a uniform upper bound on the hitting time of state $\mathbf{0}$ when starting from any point within B. This uniform upper bound on hitting time aids in bounding the state value function V_π from below since the drift inequality (9) assists in bounding the value function V_π from above alone. As a consequence, we obtain the following,

$$|Q_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a')| \leq O(\|\mathbf{q}\|^2) \qquad \forall \pi \in \Pi, \forall a, a' \in \mathcal{A} \text{ and } \forall \mathbf{q} \in \mathcal{S} \tag{26}$$

As a result, we establish policy-independent bounds on the value function Q_π. While the drift assumption 2.2 played a crucial role in deriving policy-dependent bounds on the value function V_π, for the purpose of NPG, we need these bounds to be independent of the policy. The structural assumption 2.3 eliminates this policy dependence. Consequently, from Equation 5, this translates into policy-independent bounds on Q_π.

Step 3 (Handling Estimation Errors): Since our loss function in the context of Theorem 4.1 is $\widehat{Q}_\pi$, we need uniform bounds on $\widehat{Q}_\pi$. We leverage the bounds on Q_π obtained in Equation 26 and in conjunction with the evaluation error characterized in Subsection 4.1 Equation 34, we obtain the following:

$$\left|\widehat{Q}_\pi(\mathbf{q}, a) - \widehat{Q}_\pi(\mathbf{q}, a')\right| \leq O(\|\mathbf{q}\|^2) \qquad \forall \pi \in \Pi, \forall a, a' \in \mathcal{A} \text{ and } \forall \mathbf{q} \in \mathcal{S} \tag{27}$$

Adapting Equation 22 to the context of context of infinite state NPG, implies that $M_\mathbf{q} = O(\|\mathbf{q}\|^2)$.

Step 4 (Piecing it all together): The upper bound $M_\mathbf{q}$ on $\widehat{Q}_\pi$ in Step 3 is utilized to determine the state dependent step size as $\eta_\mathbf{q} = \sqrt{\frac{8 \log |\mathcal{A}|}{T}} \frac{1}{M_\mathbf{q}}$. With bounds over $\widehat{Q}$ quantified in Equation 27, (a) of Equation (21) is upper bounded by leveraging the connection to the prediction through expert advice Theorem 4.1. This yields the final result.

The detailed proof of all steps and the main theorem can be found in Appendix.

4.1 Policy Evaluation in Countable State MDPs

It is a well-known fact that neural networks with at least one hidden layer of sufficient width and a non-linear activation function can approximate any continuous function on a compact domain arbitrarily well [Cyb89, Fun89, HSW89]. A potential technique to evaluate value functions associated with infinite state spaces can be through neural network temporal difference learning. In order to do so, consider the following transformation to compactify the domain of the problem. Let the system comprise of K queues that is, $\mathbf{q} \in \mathbb{R}_+^K$, where $\mathbb{R}_+^K$ denotes the non-negative subspace of $\mathbb{R}^K$. Let q_i represent the number of jobs in the i^{th} queue. Then define a vector $\mathbf{x} \in [0,1]^K$ such that the i^{th} element is $x_i = \frac{1}{1+q_i}$. Given a policy π, consider a linear function approximation $\widehat{Q}_\pi(\mathbf{q}, a)$ of the state-action value function $Q_\pi(\mathbf{q}, a)$ as below:

$$\frac{\widehat{Q}_\pi(\mathbf{q}, a)}{\|\mathbf{q}\|^2} = {\theta_\pi}^\top \phi\left(\mathbf{x}(\mathbf{q}), a\right) \tag{28}$$

where the feature vector ϕ is defined as below,

$$\phi\left(\mathbf{x}(\mathbf{q}), a\right) = \begin{bmatrix} \mathbb{I}_{w_1^\top(\mathbf{x}(\mathbf{q}), a) \geq 0} \ (\mathbf{x}(\mathbf{q}), a) \\ \vdots \\ \mathbb{I}_{w_m^\top(\mathbf{x}(\mathbf{q}), a) \geq 0} \ (\mathbf{x}(\mathbf{q}), a) \end{bmatrix}. \tag{29}$$

Here, $w_i \sim \mathcal{N}(0, I)$ and $I \in \mathbb{R}^{(K+1)\times(K+1)}$ is the identity matrix. This linearized model is well-studied approximation to a neural network and is called the Neural Tangent kernel (NTK) approximation; see [JT19], for example. We will not discuss the merits of the NTK approximation here since that is irrelevant to our analysis, but we only introduce the NTK approximation to discuss we why we chose our model for function approximation. In the NTK approximation, $w_i \in \mathbb{R}^{K+1}$ is random initialization which chooses a random set of features. Each feature vector $\phi(\mathbf{x}(\mathbf{q}), a)$ is of length $m|\mathcal{A}|K$, where m represents the width of the hidden layer in the neural network. Finally, θ^*_π represents the optimal parameter vector, i.e., the parameter that best estimates $Q_\pi(\mathbf{q}, a)$.

The state action value function $Q_\pi(\mathbf{q}, a)$ can be approximated arbitrarily well if Q_π is a continuous function. This is indeed the case for some simple contexts such as the M/M/1 queue, where the value function is a quadratic function in queue length ([Mey08]). More generally, Equation (26) indicates that the $Q_\pi(\mathbf{q}, a)$ can be upper bounded by a quadratic function. Therefore, under the assumption that Q_π is continuous, the learning error due to policy evaluation using the neural network can be characterized as follows:

$$\begin{aligned}
\|Q_\pi(\mathbf{q}, a) - \widehat{Q}_\pi(\mathbf{q}, a)\| &= \|Q_\pi(\mathbf{q}, a) - {\theta_\pi}^\top \phi(\mathbf{x}(\mathbf{q}), a)\, \|\mathbf{q}\|^2\| && (30)\\
&\le \left\| Q_\pi(\mathbf{q}, a) - {\theta^*_\pi}^\top \phi(\mathbf{x}(\mathbf{q}), a)\, \|\mathbf{q}\|^2 \right\| && (31)\\
&+ \left\| {\theta^*_\pi}^\top \phi(\mathbf{x}(\mathbf{q}), a)\, \|\mathbf{q}\|^2 - {\theta_\pi}^\top \phi(\mathbf{x}(\mathbf{q}), a)\, \|\mathbf{q}\|^2 \right\| && (32)
\end{aligned}$$

The function approximation error is captured in Equation (31) as follows:

$$\left\| Q_\pi(\mathbf{q}, a) - {\theta^*_\pi}^\top \phi(\mathbf{x}(\mathbf{q}), a)\, \|\mathbf{q}\|^2 \right\| = \left\| \frac{Q_\pi(\mathbf{q}, a)}{\|\mathbf{q}\|^2} - {\theta^*_\pi}^\top \phi(\mathbf{x}(\mathbf{q}), a) \right\| \|\mathbf{q}\|^2 \le \kappa_1(m)\|\mathbf{q}\|^2 \quad (33)$$

where $\kappa(m)$ is a constant that is independent of the underlying policy but depends on the width of the hidden layer. In fact, it is shown in [SS21] that when approximating polynomials, as $m \to \infty$, $\kappa_1(m) \to 0$. The temporal difference (TD) learning error is captured in Equation (32) and is a function of the number of samples available and can be quantified as $\kappa_2\|\mathbf{q}\|^2$ with high probability. And thus, with high probability, the overall state dependent error can be quantified as follows:

$$\|Q_\pi(\mathbf{q}, a) - \widehat{Q}_\pi(\mathbf{q}, a)\| \le \kappa \|\mathbf{q}\|^2 \quad (34)$$

where $\kappa = \kappa_1(m) + \kappa_2$.

References

[AKLM21] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(1), jan 2021.

[AYBB+19] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

[BR24] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 2024.

[Bré13] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

[Cao99] Xi-Ren Cao. Single sample path-based optimization of markov chains. *Journal of optimization theory and applications*, 100:527–548, 1999.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[Cyb89] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[DG22] J. G. Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 12(1):30–67, 2022.

[EDKM09] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

[ES12] Atilla Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72:311–359, 2012.

[FGKM18] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

[Fun89] Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

[GI24] Peter W Glynn and Alex Infanger. Solution representations for poisson's equation, martingale structure, and the markov chain central limit theorem. *Stochastic Systems*, 14(1):47–68, 2024.

[GM96] Peter W Glynn and Sean P Meyn. A liapounov bound for solutions of the poisson equation. *The Annals of Probability*, pages 916–931, 1996.

[GSP19] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2160–2169. PMLR, 09–15 Jun 2019.

[Haj82] Bruce Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability*, 14(3):502–525, 1982.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[JT19] Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

[Kak01] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[KT08] Sumit Kunnumkal and Huseyin Topaloglu. Using stochastic approximation methods to compute optimal base-stock levels in inventory control problems. *Operations Research*, 56(3):646–664, 2008.

[LG08] John DC Little and Stephen C Graves. Little's law. *Building intuition: insights from basic operations management models and principles*, pages 81–100, 2008.

[LXM19] Bai Liu, Qiaomin Xie, and Eytan Modiano. Reinforcement learning for optimal control of queueing systems. In *2019 57th annual allerton conference on communication, control, and computing (allerton)*, pages 663–670. IEEE, 2019.

[Mey08] Sean Meyn. *Control techniques for complex networks*. Cambridge University Press, 2008.

[MMS23] Yashaswini Murthy, Mehrdad Moharrami, and R Srikant. Performance bounds for policy-based average reward reinforcement learning algorithms. *Advances in Neural Information Processing Systems*, 36:19386–19396, 2023.

[MS16] Siva Theja Maguluri and R Srikant. Heavy traffic queue length behavior in a switch under the maxweight algorithm. *Stochastic Systems*, 6(1):211–250, 2016.

[MS23] Yashaswini Murthy and R Srikant. On the convergence of natural policy gradient and mirror descent-like policy methods for average-reward mdps. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 1979–1984. IEEE, 2023.

[MSV+19] Hongzi Mao, Malte Schwarzkopf, Shaileshh Bojja Venkatakrishnan, Zili Meng, and Mohammad Alizadeh. Learning scheduling algorithms for data processing clusters. In *Proceedings of the ACM special interest group on data communication*, pages 270–288. Association for Computing Machinery, 2019.

[MT92] Sean P Meyn and Richard L Tweedie. Stability of markovian processes i: Criteria for discrete-time chains. *Advances in Applied Probability*, 24(3):542–574, 1992.

[ÖW20] Erhun Özkan and Amy R Ward. Dynamic matching for real-time ride sharing. *Stochastic Systems*, 10(1):29–70, 2020.

[SS21] Siddhartha Satpathi and R. Srikant. The dynamics of gradient descent for over-parametrized neural networks. In *Learning for Dynamics and Control*, pages 373–384. PMLR, 2021.

[SXX20] Devavrat Shah, Qiaomin Xie, and Zhi Xu. Stable reinforcement learning with unbounded state space. *arXiv preprint arXiv:2006.04353*, 2020.

[SY13] R. Srikant and Lei Ying. *Communication networks: an optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.

[VBMW23] Sushil Mahavir Varma, Pornpawee Bumpensanti, Siva Theja Maguluri, and He Wang. Dynamic pricing and matching for two-sided queues. *Operations Research*, 71(1):83–100, 2023.

[WLWY23] Honghao Wei, Xin Liu, Weina Wang, and Lei Ying. Sample efficient reinforcement learning in mixed systems through augmented samples and its applications to queueing networks. *NeurIPS*, 2023.

A Appendix / supplemental material

The proof of Step 1 can be found in Chapter 4 of [CBL06].

A.1 Proof of Step 2

The following lemmas are a consequence of Assumptions 2.1 and 2.2.

Lemma A.1. *Given Assumptions 2.1 and 2.2, there exists a positive constant α such that $\mathbb{E}_\pi[e^{\alpha\|q\|}] < \infty$. Consequently, for any $p \geq 1$, there exists a constant $\beta_p > 0$ such that for all policies π, it is true that $\mathbb{E}_\pi\left[\|q\|^p\right] < \beta_p$. [Haj82, ES12].*

This lemma ensures that for all policies $\pi \in \Pi$, all moments of $||\mathbf{q}||$ exist. The second moment is particularly important since final regret depends on $\mathbb{E}_{\mathbf{q}\sim d^{\pi^*}}\left[\|\mathbf{q}\|^2\right]$.

Lemma A.2. *Let $\mathbb{P}_\pi$ be an irreducible transition matrix on the countable state space $\mathcal{S}$. Suppose that* (9) *is satisfied. Then the corresponding homogenous Markov Chain is positive recurrent. Consequently, the stationary distribution d_π corresponding to $\mathbb{P}_\pi$ exists and is unique [Bré13].*

Lemmas A.1 and A.2 are utilized to establish a policy independent upper bound on the infinite-horizon average-cost. Such a bound is indicative of stable policies since the negative drift in Equation (9) prevents the average cost from being unbounded. Thus it is possible to show that all policies that satisfy the drift equation 9, have an uniform upper bound on the expected total queue length. This is proven in the Lemma below.

Lemma A.3. *Given Assumptions 2.1 and 2.2, for all policies $\pi \in \Pi$ it is true that,*

$$J_\pi \leq \frac{c}{\epsilon} \tag{35}$$

where $J_\pi = \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]$ is the average cost associated with policy π and constants c, ϵ are the drift parameters in Equation 9.

Proof. From Assumption 2.2, it follows that for any policy $\pi \in \Pi$, the following drift inequality is satisfied,

$$\mathbb{E}_\pi\left[\|\mathbf{q}_{k+1}\|^2 - \|\mathbf{q}_k\|^2 | \mathbf{q}_k = \mathbf{q}\right] \leq -\epsilon\|\mathbf{q}\|_1 + c. \tag{36}$$

Recall that d_π represents the stationary measure associated with policy π. Since we assume that all policies induce irreducible Markov chains and from Lemma A.2, we know that the drift equation (9) implies positive recurrence of the Markov chain, d_π exists and is unique. Since $d_\pi \geq 0$, consider the following weighted drift inequality:

$$\sum_{\mathbf{q}\in\mathcal{S}} d_\pi(\mathbf{q})\left[\mathbb{E}_\pi\left[\|\mathbf{q}_{k+1}\|^2 - \|\mathbf{q}_k\|^2 | \mathbf{q}_k = \mathbf{q}\right]\right] \leq -\epsilon \sum_{\mathbf{q}\in\mathcal{S}} d_\pi(\mathbf{q})\|\mathbf{q}\|_1 + c \tag{37}$$

From Lemma A.1, recall that the second moment of $\mathbf{q}$ is defined and exists for all policies $\pi \in \Pi$. Hence the left hand summation in Equation 37 is well defined. Since the expectation is taken with respect to $\mathbb{P}_\pi$ and since $d_\pi\mathbb{P}_\pi = d_\pi$, the left hand summation in Equation 37 is 0. Hence the expected drift in stationarity is zero.

From Lemma A.1, we know that J_π exists and can be defined as $J_\pi = \sum_{\mathbf{q}\in\mathcal{S}} d_\pi(\mathbf{q})\left(\|\mathbf{q}\|_1\right)$. We thus obtain the following:

$$0 \leq -\epsilon J_\pi + c \tag{38}$$

Therefore,

$$J_\pi \leq \frac{c}{\epsilon} \tag{39}$$

Equation 39 is true for all policies π. Hence, the average cost is uniformly upper bounded by $\frac{c}{\epsilon}$.

Since Q_π is related to V_π through Equation (5), in order to bound Q_π, we first bound V_π. We now derive an upper bound on the value function V_π leveraging the drift equation 9 and the uniform upper bound on J_π in Equation 39. First we leverage Assumptions 2.1 and 2.2 to establish policy dependent upper bounds on the value function as elaborated in the following subsection.

A.1.1 Policy Dependent upper bound on State Value Function

Lemma A.4. *Consider a set B defined in Equation 10. Let $V_\pi(\mathbf{q})$ represent the state value function associated with state $\mathbf{q} \in \mathcal{S}$ and policy $\pi \in \Pi$. Under Assumptions 2.1 and 2.2, for all $\mathbf{q} \in B^{\mathsf{c}}$ and for all policies $\pi \in \Pi$, it is true that,*

$$V_\pi(\mathbf{q}) \leq \frac{2}{\epsilon}\|\mathbf{q}\|^2 + \max_{\substack{\mathbf{q}' \in B \\ \pi \in \Pi}} V_\pi(\mathbf{q}') \tag{40}$$

Proof. Define the following set:

$$A_\pi := \left\{ \mathbf{q} \in \mathcal{S} : \|\mathbf{q}\|_1 \leq \frac{2c}{\epsilon} - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right] \right\} \tag{41}$$

Since $J_\pi \leq \frac{c}{\epsilon}$, A_π is a finite non-empty set. Multiplying (36) throughout by $\frac{2}{\epsilon}$, we obtain the following:

$$\mathbb{E}_\pi\left[\frac{2}{\epsilon}\|\mathbf{q}_{k+1}\|^2 - \frac{2}{\epsilon}\|\mathbf{q}_k\|^2 \middle| \mathbf{q}_k = \mathbf{q} \right] \leq -2\|\mathbf{q}\|_1 + \frac{2c}{\epsilon} \tag{42}$$

Consider a $\mathbf{q} \in A_\pi^{\mathsf{c}}$. Then, from definition it is true that $-\|\mathbf{q}\|_1 \leq -\frac{2c}{\epsilon} + \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]$. Bounding $-\|\mathbf{q}\|_1$ from above, we obtain,

$$\mathbb{E}_\pi\left[\frac{2}{\epsilon}\|\mathbf{q}_{k+1}\|^2 - \frac{2}{\epsilon}\|\mathbf{q}_k\|^2 \middle| \mathbf{q}_k = \mathbf{q} \right] \leq -\|\mathbf{q}\|_1 + \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right] \tag{43}$$

Let τ_{A_π} be the first time to enter set A_π starting from state $\mathbf{q} \in A_\pi^{\mathsf{c}}$. Note that τ_{A_π} is almost surely finite, because stability follows from lemma A.1. Let $\mathbf{q}_{\tau_{A_\pi}} \in A^\pi$ represent the state first reached within A_π. Then, consider the drift equation summed up along the trajectory starting from $\mathbf{q}$ till the first time it enters the set A_π,

$$\underbrace{\mathbb{E}\left[\sum_{k=0}^{\tau_{A_\pi}-1} \|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right] \middle| \mathbf{q}_0 = \mathbf{q} \right]}_{(a)} \leq \underbrace{-\mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\sum_{k=0}^{\tau_{A_\pi}-1} \frac{2}{\epsilon}\|\mathbf{q}_{k+1}\|^2 - \frac{2}{\epsilon}\|\mathbf{q}_k\|^2 \middle| \mathcal{F}_k \right]\right]}_{(b)} \tag{44}$$

where $\mathcal{F}_k = \sigma\left(\{\mathbf{q}_0, a_0, \mathbf{q}_1, a_1, \ldots, \mathbf{q}_k, a_k\}\right)$ represents the filtration till time k. Consider the expression in (b). Since τ_{A_π} is a valid stopping time and since $\|\mathbf{q}\|^2$ is a non-negative random variable, from discrete Dynkin's formula [MT92] we have the following:

$$(b) = -\mathbb{E}_\pi\left[\sum_{k=0}^{\tau_{A_\pi}-1} \mathbb{E}_\pi\left[\frac{2}{\epsilon}\|\mathbf{q}_{k+1}\|^2 \middle| \mathcal{F}_k \right] - \frac{2}{\epsilon}\|\mathbf{q}_k\|^2 \right] \tag{45}$$

$$= \frac{2}{\epsilon}\|\mathbf{q}\|^2 - \mathbb{E}_\pi\left[\frac{2}{\epsilon}\|\mathbf{q}_{\tau_{A_\pi}}\|^2 \middle| \mathbf{q}_0 = \mathbf{q} \right] \tag{46}$$

where $\|\mathbf{q}_{\tau_{A_\pi}}\|^2 = \sum_{\mathbf{q}' \in A_\pi} \mathbb{P}_\pi\left(\mathbf{q}_{\tau_{A_\pi}} = \mathbf{q}' | \mathbf{q}_0 = \mathbf{q}\right) f(\mathbf{q}')$.

Since $\|\mathbf{q}_{\tau_{A_\pi}}\|^2 > 0$, we obtain the following:

$$(b) \leq \frac{2}{\epsilon}\|\mathbf{q}\|^2 \tag{47}$$

Consider the expression in (a). From definition of A_π in (41), we know that the state corresponding to zero queue length, that is $\mathbf{0} \in A_\pi, \forall \pi \in \Pi$. Let τ_0^π denote the time to hit state $\mathbf{0}$. Then,

$$(a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\tau_{A_\pi}-1} \left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right) + \sum_{k=\tau_{A_\pi}}^{\tau_0^\pi-1} \left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right) \middle| \mathbf{q}_0 = \mathbf{q} \right] \tag{48}$$

$$- \mathbb{E}_\Pi\left[\sum_{k=\tau_{A_\pi}}^{\tau_0^\pi-1} \left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right) \middle| \mathbf{q}_0 = \mathbf{q} \right] \tag{49}$$

Recall the definition of value function $V_\pi(\mathbf{q})$ from Equation (4). We thus obtain:

$$(a) = V_\pi(\mathbf{q}) - \mathbb{E}_\pi\left[\sum_{k=\tau_{A_\pi}}^{\tau_0^\pi - 1}\left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right)\Big|\mathbf{q}_0 = \mathbf{q}\right] \tag{50}$$

Let $\mathbb{I}$ denote the indicator function. Since τ_{A_π} is the first time to enter set A_π when starting from state $\mathbf{q}$,

$$(a) = V_\pi(\mathbf{q}) - \mathbb{E}_\pi\left[\left(\sum_{k=\tau_{A_\pi}}^{\tau_0^\pi - 1}\left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right)\right)\left(\sum_{\mathbf{q}'\in A_\pi}\mathbb{I}_{\mathbf{q}_{\tau_{A_\pi}}=\mathbf{q}'}\right)\Big|\mathbf{q}_0 = \mathbf{q}\right] \tag{51}$$

$$= V_\pi(\mathbf{q}) - \sum_{\mathbf{q}'\in A_\pi}\mathbb{E}_\pi\left[\sum_{k=\tau_{A_\pi}}^{\tau_0^\pi - 1}\left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right)\Big|\mathbf{q}_0 = \mathbf{q}, \mathbf{q}_{\tau_{A_\pi}=\mathbf{q}'}\right]\mathbb{P}_\pi\left(\mathbf{q}_{\tau_{A_\pi}=\mathbf{q}'}|\mathbf{q}_0 = \mathbf{q}\right) \tag{52}$$

Since τ_{A_π} is the first time to enter set A_π, it qualifies as a valid stopping time. Hence, from the strong Markov property we know that:

$$\mathbb{E}_\pi\left[\sum_{k=\tau_{A_\pi}}^{\tau_0^\pi - 1}\left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right)\Big|\mathbf{q}_0 = \mathbf{q}, \mathbf{q}_{\tau_{A_\pi}=\mathbf{q}'}\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\tau_0^\pi - 1}\left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right)\Big|\mathbf{q}_0 = \mathbf{q}'\right] \tag{53}$$

Since, $V_\pi(\mathbf{q}') := \mathbb{E}_\pi\left[\sum_{k=0}^{\tau_0^\pi - 1}\left(\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}_k\|_1\right]\right)\Big|\mathbf{q}_0 = \mathbf{q}'\right]$,

$$(a) = V_\pi(\mathbf{q}) - \sum_{\mathbf{q}'\in A_\pi}\mathbb{P}_\pi\left(\mathbf{q}_{\tau_{A_\pi}=\mathbf{q}'}|\mathbf{q}_0 = \mathbf{q}\right)V_\pi(\mathbf{q}') \tag{54}$$

Combining Equations 44,47 and 54, we obtain the following,

$$V_\pi(\mathbf{q}) \leq \frac{2}{\epsilon}\|\mathbf{q}\|^2 + \sum_{\mathbf{q}'\in A_\pi}\mathbb{P}_\pi\left(\mathbf{q}_{\tau_{A_\pi}=\mathbf{q}'}|\mathbf{q}_0 = \mathbf{q}\right)V_\pi(\mathbf{q}') \tag{55}$$

Recall the definition of set B in Equation 10

$$B := \left\{\mathbf{q}\in\mathcal{S} : \|\mathbf{q}\|_1 \leq \frac{2c}{\epsilon}\right\} \tag{56}$$

Since $\mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right] \geq 0$ for all policies $\pi\in\Pi$, it is evident from Equation 41 that $A_\pi \subset B$. Thus the following is true:

$$\max_{\substack{\mathbf{q}'\in A_\pi\\ \pi\in\Pi}} V_\pi(\mathbf{q}') \leq \max_{\substack{\mathbf{q}'\in B\\ \pi\in\Pi}} V_\pi(\mathbf{q}') \tag{57}$$

Hence Equation 55 reduces to,

$$\begin{aligned} V_\pi(\mathbf{q}) &\leq \frac{2}{\epsilon}\|\mathbf{q}\|^2 + \sum_{\mathbf{q}'\in A_\pi}\mathbb{P}_\pi\left(\mathbf{q}_{\tau_{A_\pi}=\mathbf{q}'}|\mathbf{q}_0 = \mathbf{q}\right)\max_{\substack{\mathbf{q}''\in A_\pi\\ \pi\in\Pi}} V_\pi(\mathbf{q}'') \\ &\leq \frac{2}{\epsilon}\|\mathbf{q}\|^2 + \sum_{\mathbf{q}'\in A_\pi}\mathbb{P}_\pi\left(\mathbf{q}_{\tau_{A_\pi}=\mathbf{q}'}|\mathbf{q}_0 = \mathbf{q}\right)\max_{\substack{\mathbf{q}''\in B\\ \pi\in\Pi}} V_\pi(\mathbf{q}'') \\ &\leq \frac{2}{\epsilon}\|\mathbf{q}\|^2 + \max_{\substack{\mathbf{q}'\in B\\ \pi\in\Pi}} V_\pi(\mathbf{q}') \end{aligned} \tag{58}$$

Recall that Equation 58 holds true for all $\mathbf{q}\in A_\pi^{\mathsf{c}}$. Since $B^{\mathsf{c}} \subset A_\pi^{\mathsf{c}}$ for all policies $\pi\in\Pi$, we thus obtain that for all $\mathbf{q}\in B^{\mathsf{c}}$, it is true that,

$$V_\pi(\mathbf{q}) \leq \frac{2}{\epsilon}\|\mathbf{q}\|^2 + \max_{\substack{\mathbf{q}'\in B\\ \pi\in\Pi}} V_\pi(\mathbf{q}') \tag{59}$$

In order to invoke the connection of NPG to prediction through expert advice, we need policy independent bounds on the estimate $\widehat{Q}_\pi$. As a step towards achieving that, we first need to establish policy independent bounds on the exact value function V_π and therefore on V_π. Since the drift provides us with a policy dependent upper bound alone, we exploit the structure of queuing systems in order to obtain a policy independent lower bound and upper bound.

A.1.2 Policy Independent bounds on the State Value Function

The structural assumption 2.3 aids in obtaining policy independent bounds by providing an uniform upper bound on the hitting time of state $\mathbf{0}$ from any state within a finite set B across all policies π.

Lemma A.5. *Consider the set B in Equation (10). By definition, the state $\mathbf{0}$, which represents zero queue length, is included in the set B, i.e., $\mathbf{0} \in B$. Let τ_B^{bound} represent the maximum time taken to hit $\mathbf{0}$ starting from any state $\mathbf{q} \in B$ across all policies $\pi \in \Pi$. Then under Assumption 2.3, for any policy $\pi \in \Pi$, τ_B^{bound} is bounded from above as follows:*

$$\tau_B^{bound} \leq \frac{x_B}{p_B^2} + x_B. \tag{60}$$

Proof. For each state $\mathbf{q} \in B$, let $t_{\pi,B}(\mathbf{q})$ be the expected amount of time until policy π next reaches a state $\mathbf{q}' \in B$, starting in state $\mathbf{q}$. Let $t_{\pi,B}^{\max}$ be the maximum over $\mathbf{q} \in B$ of $t_{\pi,B}(\mathbf{q})$. Let $\mathbf{q}^*$ be the state in B which achieves this maximum. By Assumption 2.3, every state $\mathbf{q} \in B$ has at least p_B probability of moving to state $\mathbf{q}^*$ in the next x_B steps, and therefore spending at most x_B of the next $t_{\pi,B}^{\max} + x_B$ steps inside B.

Let us use a renewal-reward argument, where the renewal moments are steps on which the system visits state $\mathbf{q}^*$. Let us subdivide this renewal period into three segments:

1. Starting at $\mathbf{q}^*$, the time until a state $\mathbf{q} \in B$ is next visited.
2. Starting from a state $\mathbf{q} \in B$, the next x_B steps, during which there is at least a p_B probability of reaching the state $\mathbf{q}^*$.
3. Any time not in either of these two segments, which must be spent outside of B.

From these segments, we can upper-bound the fraction of time during the renewal period that the system spends in B: The system can only be in B during item 2, and the system can spend at most $\frac{1}{p_B}$ periods of length at most x_B in that segment. The fraction of time spent in B is maximized if no time is spent in item 3, resulting in the following bound:

$$\mathbb{P}\left(\mathbf{q}_t \in B\right) \leq \frac{x_B/p_B}{t_{\pi,B}^{\max} + x_B/p_B}$$

Since the system accrues a cost of atleast $\frac{2c}{\epsilon}$ outside B and a cost of atleast 0 inside B, the average cost J_π can be bounded as below:

$$J_\pi \geq \left(\frac{2c}{\epsilon}\right)\left(\frac{t_{\pi,B}^{\max}}{t_{\pi,B}^{\max} + x_B/p_B}\right) \tag{61}$$

However, from Lemma A.3, we know that $J_\pi \leq \frac{c}{\epsilon}$. We can use this fact to bound $t_{\pi,B}^{\max}$ as follows:

$$\left(\frac{2c}{\epsilon}\right)\left(\frac{t_{\pi,B}^{\max}}{t_{\pi,B}^{\max} + x_B/p_B}\right) \leq \frac{c}{\epsilon} \tag{62}$$

$$\implies t_{\pi,B}^{\max} \leq \frac{x_B}{p_B} \tag{63}$$

Now, we can bound the time until we reach B, starting from $\mathbf{q} \in B$. Whenever the system is in a state in B, there is a p_B chance of reaching B in the next x_B steps, and a maximum expected time of $t_{\pi,B}^{\max}$ until we next get the opportunity. Let $p_{\pi,B}(\mathbf{q})$ be the probability that the system reaches $\mathbf{0}$ from $\mathbf{q}$ in at most x_B steps. Let $t'_{\pi,B}(\mathbf{q})$ be the expected time until the system re-enters B, conditional on not reaching $\mathbf{0}$ in at most x_B steps. Note that $t'_{\pi,B}(\mathbf{q}) \leq \frac{t_{\pi,B}(\mathbf{q})}{1-p_{\pi,B}(\mathbf{q})}$, because the probability of the event of not reaching B in at most x_B steps is $1 - p_{\pi,B}(\mathbf{q})$.

We can now start to bound $\tau_\pi(\mathbf{q})$ as follows:

$$\tau_\pi(\mathbf{q}) \leq p_{\pi,B}(\mathbf{q})x_B + (1 - p_{\pi,B}(\mathbf{q}))(t'_{\pi,B}(\mathbf{q}) + \tau_\pi(\mathbf{q}'))$$

where $\mathbf{q}'$ is the state at which we re-enter B. Thus,

$$\begin{aligned}\tau_\pi(\mathbf{q}) &\le p_{\pi,B}(\mathbf{q})x_B + (1-p_{\pi,B}(\mathbf{q}))\left(\frac{t_{\pi,B}(\mathbf{q})}{1-p_{\pi,B}(\mathbf{q})} + \tau_\pi(\mathbf{q}')\right)\\ &= x_B + t_{\pi,B}(\mathbf{q}) + (1-p_{\pi,B}(\mathbf{q}))(\tau_\pi(\mathbf{q}') - x_B)\end{aligned}$$

In particular, letting τ_π^{bound} be the maximum over $\mathbf{q}\in B$ of $\tau_\pi(\mathbf{q})$, and letting $\mathbf{q}_\tau^*$ be the state in which that maximum occurs, we have

$$\begin{aligned}\tau_\pi^{\text{bound}} &\le x_B + t_{\pi,B}(\mathbf{q}_\tau^*) + (1-p_{\pi,B}(\mathbf{q}_\tau^*))(\tau_\pi^{\text{bound}} - x_B)\\ \tau_\pi^{\text{bound}} &\le x_B + \frac{t_{\pi,B}(\mathbf{q}_\tau^*)}{p_{\pi,B}(\mathbf{q}_\tau^*)} \le x_B + \frac{t_{\pi,B}^{\max}}{p_B}\\ \tau_\pi^{\text{bound}} &\le \frac{x_B}{p_B^2} + x_B\end{aligned}$$

□

The uniform hitting bound provides with a uniform lower bound on the value function V_π. This lower bound is further leveraged to get a policy independent upper bound on the quantity $\max_{\substack{\mathbf{q}\in B\\ \pi\in\Pi}} V_\pi(\mathbf{q})$. This leads to the upper bound on the value functions being policy independent.

Lemma A.6. *Let x_B, p_B be policy independent constants that satisfy Assumption 2.3 and c, ϵ be policy independent constants that satisfy Assumptions 2.1 and 2.2. Then, the following are true.*

1. *The value function $V_\pi(\mathbf{q})$ is lower bounded $\forall \mathbf{q}\in\mathcal{S}$ and for all policies $\pi\in\Pi$ as follows:*

$$V_\pi(\mathbf{q}) \ge -\frac{c}{\epsilon}\left(\frac{x_B}{p_B^2} + x_B\right) \tag{64}$$

2. *For all $\mathbf{q}\in B$, where B is a set defined in Equation (10), the value function $V_\pi(\mathbf{q})$ is uniformly bounded from above across all policies $\pi\in\Pi$ as follows:*

$$\max_{\substack{\mathbf{q}\in B\\ \pi\in\Pi}} V_\pi(\mathbf{q}) \le \frac{c}{p_B\epsilon}\left(\frac{x_B}{p_B^2} + 2x_B\right). \tag{65}$$

3. *The value function $V_\pi(\mathbf{q})$ is upper bounded $\forall \mathbf{q}\in\mathcal{S}$ and for all policies $\pi\in\Pi$ as follows:*

$$V_\pi(\mathbf{q}) \le \frac{2}{\epsilon}\|\mathbf{q}\|^2 + \frac{c}{p_B\epsilon}\left(\frac{x_B}{p_B^2} + 2x_B\right) \tag{66}$$

Proof of Part 1:

Proof. Recall the definition of the state value function $V_\pi(\mathbf{q})$ in Equation 4. Consider any state $\mathbf{q}\in\mathcal{S}$ and policy $\pi\in\Pi$, such that τ_0^π represents the time to hit state $\mathbf{0}$ when starting at $\mathbf{q}$. Then,

$$\begin{aligned}V_\pi(\mathbf{q}) &= \mathbb{E}_\pi\left[\sum_{k=0}^{\tau_0^\pi - 1}\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\Big|\mathbf{q}_0 = \mathbf{q}\right] &(67)\\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\tau_0^\pi - 1} c(\mathbf{q}_k) - J_\pi\Big|\mathbf{q}_0 = \mathbf{q}\right] &(68)\end{aligned}$$

where $c(\mathbf{q})$ is the single step cost associated with state $\mathbf{q}$, particularly the total queue length at that instant.

Consider the following definitions:

- $\tau_{\pi,B}^{\text{out}}$ is the first time to enter B when starting at $\mathbf{q}$.

- $\tau^{\text{in}}_{\pi,B}$ is the first time to hit state 0 from the first time to enter state B.
- $\overline{c}^{\text{out}}_{\pi}$ be the aggregate average cost obtained till time $\tau^{\text{out}}_{\pi,B}$.
- $\overline{c}^{\text{in}}_{\pi}$ be the aggregate average cost in the next time interval $\tau^{\text{in}}_{\pi,B}$.

Then the value function $V_\pi(\mathbf{q})$ can be decomposed as follows:

$$V_\pi(\mathbf{q}) = V^{\text{out}}_\pi(\mathbf{q}) + V^{\text{in}}_\pi(\mathbf{q}) \tag{69}$$

When further expanded, we obtain,

$$V_\pi(\mathbf{q}) = \left(\overline{c}^{\text{out}}_\pi - J_\pi\right)\tau^{\text{out}}_{\pi,B} + \left(\overline{c}^{\text{in}}_\pi - J_\pi\right)\tau^{\text{in}}_{\pi,B} \tag{70}$$

From definition $\forall \mathbf{q} \in B^{\mathsf{c}}$, $\|\mathbf{q}\|_1 > \frac{2c}{\epsilon}$. Hence the aggregate cost outside B is also greater than $\frac{2c}{\epsilon}$, that is $\overline{c}^{\text{out}}_\pi > \frac{2c}{\epsilon}$. Since from Lemma A.3 we know that $J_\pi \leq \frac{c}{\epsilon}$, it is true that $(\overline{c}^{\text{out}}_\pi - J_\pi) > 0$. Hence, we obtain,

$$V_\pi(\mathbf{q}) \geq \left(\overline{c}^{\text{in}}_\pi - J_\pi\right)\tau^{\text{in}}_{\pi,B} \tag{71}$$

Since all the costs are non-negative,

$$V_\pi(\mathbf{q}) \geq -J_\pi \tau^{\text{in}}_{\pi,B} \tag{72}$$

Since from Lemma A.3 we know that $J_\pi \leq \frac{c}{\epsilon}$,

$$V_\pi(\mathbf{q}) \geq -\frac{c}{\epsilon}\tau^{\text{in}}_{\pi,B} \tag{73}$$

Since $\tau^{\text{in}}_{\pi,B}$ represents the time to reach $\mathbf{0}$ starting from the first state in B to be reached from $\mathbf{q}$, this quantity can be bounded from above using Lemma A.5. That is $\tau^{\text{in}}_{\pi,B} \leq \tau^{\text{bound}}_B$. Hence for all $\mathbf{q} \in \mathcal{S}$, we obtain the following policy-independent uniform lower bound on the state value function,

$$V_\pi(\mathbf{q}) \geq -\frac{c}{\epsilon}\left(\frac{x_B}{p_B^2} + x_B.\right) \tag{74}$$

□

Proof of Part 2:

The proof of part 2 relies on the result in Part 1. Part 2 is essential in establishing policy independent upper bounds on the value function, and relies on assumptions 2.1,2.2 and 2.3. This analysis is one of the key contributions of this paper and is presented below.

Proof. Let $\mathbf{q}_0 = \mathbf{0}$, i.e., the trajectory is starting from state $\mathbf{0}$. Let τ^K_π represent the first time at which the trajectory hits the state $\mathbf{0}$ for the K^{th} time. That is,

$$\tau^K_\pi = \inf_k \left\{ \left(\sum_{i=1}^{k} \mathbb{I}(\mathbf{q}_i = \mathbf{0}) \right) = K \middle| \mathbf{q}_0 = \mathbf{0} \right\} \tag{75}$$

τ^K_π represents the minimum time to complete K recurrent cycles with the recurrent state being $\mathbf{0}$ when starting from $\mathbf{0}$. Recall x_B from Assumption 2.3. Consider the following quantity.

$$\widetilde{N} = \min\left\{N \in \mathbb{N}, \tau^N_\pi \geq x_B\right\} \tag{76}$$

Consider $\widetilde{V}_\pi(\mathbf{0})$ defined below:,

$$\widetilde{V}_\pi(\mathbf{0}) = \mathbb{E}_\pi \left[\sum_{i=0}^{\tau^{\widetilde{N}}_\pi - 1} \|\mathbf{q}_i\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right] \middle| \mathbf{q}_0 = \mathbf{0} \right] \tag{77}$$

Note that since $\widetilde{N}$ is a random quantity, Equation 77 does not correspond to the recurrent state characterization of the state value function in Equation 4. Hence, we need to consider the following. The above equation can be decomposed into sum of recurrent cycles as below:

$$\widetilde{V}_\pi(\mathbf{0}) = \mathbb{E}_\pi \left[\sum_{i=1}^{\widetilde{N}} \sum_{k=\tau^{i-1}_\pi}^{\tau^i_\pi - 1} \|\mathbf{q}_i\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right] \middle| \mathbf{q}_0 = \mathbf{0} \right] \tag{78}$$

Since from definition of $\tau_\pi^{\widetilde{N}}$ in Equation 76 there can be a maximum of x_B such recurrent cycles, $\widetilde{N}$ is bounded from above by x_B.

$$\widetilde{V}_\pi(\mathbf{0}) = \mathbb{E}_\pi\left[\sum_{i=1}^{x_B}\sum_{k=\tau_\pi^{i-1}}^{\tau_\pi^{i}-1}\left(\|\mathbf{q}_i\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\right)\mathbb{I}\left(\widetilde{N}\geq i\right)\Big|\mathbf{q}_0=\mathbf{0}\right] \tag{79}$$

$$= \sum_{i=1}^{x_B}\mathbb{E}_\pi\left[\sum_{k=\tau_\pi^{i-1}}^{\tau_\pi^{i}-1}\left(\|\mathbf{q}_i\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\right)\mathbb{I}\left(\widetilde{N}\geq i\right)\Big|\mathbf{q}_0=\mathbf{0}\right] \tag{80}$$

$$= \sum_{i=1}^{x_B}\mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\sum_{k=\tau_\pi^{i-1}}^{\tau_\pi^{i}-1}\left(\|\mathbf{q}_i\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\right)\mathbb{I}\left(\widetilde{N}\geq i\right)\Big|\mathcal{F}_{i-1}\right]\right] \tag{81}$$

where $\mathcal{F}_{i-1} = \sigma\left(\left\{\mathbf{q}_0, a_0, \mathbf{q}_1, a_1, \ldots, \mathbf{q}_{\tau_\pi^{i-1}}, a_{\tau_\pi^{i-1}}\right\}\right)$ is the filtration till the end of the $i-1^{\text{th}}$ cycle. Since $\mathbb{I}\left(\widetilde{N}\geq i\right)$ is a deterministic function when conditioned on $\mathcal{F}_{i-1}$, we obtain the following:

$$\widetilde{V}_\pi(\mathbf{0}) = \sum_{i=1}^{x_B}\mathbb{E}_\pi\left[\mathbb{I}\left(\widetilde{N}\geq i\right)\underbrace{\mathbb{E}_\pi\left[\sum_{k=\tau_\pi^{i-1}}^{\tau_\pi^{i}-1}\left(\|\mathbf{q}_i\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\right)\Big|\mathcal{F}_{i-1}\right]}_{(a)}\right] \tag{82}$$

(a) now corresponds to $V_\pi(\mathbf{0})$, which from definition in Equation 4 is 0. Hence,

$$\widetilde{V}_\pi(\mathbf{0}) = 0 \tag{83}$$

From definition in Equation (76), we can decompose $\widetilde{V}_\pi(\mathbf{0})$ as follows:

$$\widetilde{V}_\pi(\mathbf{0}) = \mathbb{E}_\pi\left[\sum_{k=0}^{x_B-1}\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\Big|\mathbf{q}_0=\mathbf{0}\right] + \sum_{\mathbf{q}'\in\mathcal{S}}\mathbb{P}_\pi^{x_B}\left(\mathbf{q}'|\mathbf{q}_0=\mathbf{0}\right)V_\pi(\mathbf{q}') \tag{84}$$

where $\mathbb{P}_\pi^{x_B}$ is the x_B step probability transition matrix under policy π. Let $\mathbf{q}\in B$. Since $\widetilde{V}_\pi(\mathbf{0})=0$ from Equation 83, we obtain the following,

$$-\mathbb{E}_\pi\left[\sum_{k=0}^{x_B-1}\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\Big|\mathbf{q}_0=\mathbf{0}\right] = \mathbb{P}_\pi^{x_B}\left(\mathbf{q}|\mathbf{q}_0=\mathbf{0}\right)V_\pi(\mathbf{q}) + \sum_{\mathbf{q}'\in\mathcal{S}/\{\mathbf{q}\}}\mathbb{P}_\pi^{x_B}\left(\mathbf{q}'|\mathbf{q}_0=\mathbf{0}\right)V_\pi(\mathbf{q}') \tag{85}$$

$$\mathbb{P}_\pi^{x_B}\left(\mathbf{q}|\mathbf{q}_0=\mathbf{0}\right)V_\pi(\mathbf{q}) = -\mathbb{E}_\pi\left[\sum_{k=0}^{x_B-1}\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\Big|\mathbf{q}_0=\mathbf{0}\right] - \sum_{\mathbf{q}'\in\mathcal{S}/\{\mathbf{q}\}}\mathbb{P}_\pi^{x_B}\left(\mathbf{q}'|\mathbf{q}_0=\mathbf{0}\right)V_\pi(\mathbf{q}') \tag{86}$$

From Part 1 of this lemma, we know that $-V_\pi(\mathbf{q}) \leq \frac{c}{\epsilon}\left(\frac{x_B}{p_B^2}+x_B\right)\forall\mathbf{q}\in\mathcal{S}$.

$$\mathbb{P}_\pi^{x_B}\left(\mathbf{q}|\mathbf{q}_0=\mathbf{0}\right)V_\pi(\mathbf{q}) \leq -\mathbb{E}_\pi\left[\sum_{k=0}^{x_B-1}\|\mathbf{q}_k\|_1 - \mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\Big|\mathbf{q}_0=\mathbf{0}\right] \tag{87}$$

$$+ \sum_{\mathbf{q}'\in\mathcal{S}/\{\mathbf{q}\}}\mathbb{P}_\pi^{x_B}\left(\mathbf{q}'|\mathbf{q}_0=\mathbf{0}\right)\frac{c}{\epsilon}\left(\frac{x_B}{p_B^2}+x_B\right) \tag{88}$$

$$\leq \mathbb{E}_\pi\left[\sum_{k=0}^{x_B-1}\mathbb{E}_\pi\left[\|\mathbf{q}\|_1\right]\Big|\mathbf{q}_0=\mathbf{0}\right] + \frac{c}{\epsilon}\left(\frac{x_B}{p_B^2}+x_B\right) \tag{89}$$

$$\leq x_B\left(\frac{c}{\epsilon}\right) + \frac{c}{\epsilon}\left(\frac{x_B}{p_B^2}+x_B\right) \tag{90}$$

where the last inequality follows from Lemma A.3.

If $V_\pi(\mathbf{q})$ is negative, then 0 is an upper bound on $V_\pi(\mathbf{q})$. However, if $V_\pi(\mathbf{q})$ is positive,

$$V_\pi(\mathbf{q}) \leq \frac{1}{\mathbb{P}_\pi^{x_B}(\mathbf{q}|\mathbf{q}_0 = \mathbf{0})} \frac{c}{\epsilon} \left(\frac{x_B}{p_B^2} + 2x_B \right) \tag{91}$$

Recall from Assumption 2.3, since $\mathbf{0}, \mathbf{q} \in B$, $\mathbb{P}_\pi^{x_B}(\mathbf{q}|\mathbf{q}_0 = \mathbf{0}) \geq p_B$. That is the probability of reaching any $\mathbf{q} \in B$ from $\mathbf{0}$ in x_B steps is at least p_B. Hence for all $\mathbf{q} \in B$, it is true that,

$$V_\pi(\mathbf{q}) \leq \frac{c}{p_B \epsilon} \left(\frac{x_B}{p_B^2} + 2x_B \right) \tag{92}$$

Since the above bound is uniform across all policies $\pi \in \Pi$ and $\mathbf{q} \in B$,

$$\max_{\substack{\mathbf{q} \in B \\ \pi \in \Pi}} V_\pi(\mathbf{q}) \leq \frac{c}{p_B \epsilon} \left(\frac{x_B}{p_B^2} + 2x_B \right) \tag{93}$$

□

Proof of Part 3:

Proof. Combining Equation (93) from Part 2 and Equation (59) from Lemma A.4 yields the results. That is $\forall \mathbf{q} \in \mathcal{S}$ and for all $\pi \in \Pi$,

$$V_\pi(\mathbf{q}) \leq \frac{2}{\epsilon} \|\mathbf{q}\|^2 + \frac{c}{p_B \epsilon} \left(\frac{x_B}{p_B^2} + 2x_B \right) \tag{94}$$

We thus obtain a policy independent upper bound on the value function for all states in the state space. □

A.1.3 Policy Independent bounds on the State-Action Value Function

In order to obtain policy independent bounds on the estimate $\widehat{Q}_\pi$ of the state action value function associated with some policy π, it is necessary to first obtain bounds on the exact state action value function Q_π. The following lemma provides with state-dependent, policy-independent bounds on the state action value function Q.

Lemma A.7. *There exists constant $c_4 > 0$, such that under Assumptions 2.1,2.2 and 2.3, the state action value function Q_π for all policies $\pi \in \Pi$ and forall $\mathbf{q} \in \mathcal{S}$ satisfies:*

$$|Q_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a')| \leq \frac{2}{\epsilon} c_1 \|\mathbf{q}\|^2 + c_4 \qquad a, a' \in \mathcal{A} \tag{95}$$

where ϵ is the drift parameter and c_1 is defined in Assumption 2.3.

Proof. Recall the Poisson Equation (5) corresponding to the state action value function Q_π:

$$Q_\pi(\mathbf{q}, a) = \|\mathbf{q}\|_1 + \mathbb{E}_{\mathbf{q}' \sim \mathbb{P}(\cdot|\mathbf{q}, a)} V_\pi(\mathbf{q}') - J_\pi \tag{96}$$

For any pair of actions $a, a' \in \mathcal{A}$

$$\begin{aligned} Q_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a') &= \mathbb{E}_{\mathbf{q}' \sim \mathbb{P}(\cdot|\mathbf{q}, a)} V_\pi(\mathbf{q}') - \mathbb{E}_{\mathbf{q}' \sim \mathbb{P}(\cdot|\mathbf{q}, a')} V_\pi(\mathbf{q}') \\ &\leq \mathbb{E}_{\mathbf{q}' \sim \mathbb{P}(\cdot|\mathbf{q}, a)} \left(\frac{2}{\epsilon} f(\mathbf{q}') + \frac{c}{p_B \epsilon} \left(\frac{x_B}{p_B^2} + 2x_B \right) \right) \\ &+ \mathbb{E}_{\mathbf{q}' \sim \mathbb{P}(\cdot|\mathbf{q}, a')} \left(\frac{c}{\epsilon} \left(\frac{x_B}{p_B^2} + x_B . \right) \right) \end{aligned} \tag{97}$$

where the last inequality follows from Lemma A.6. Hence we obtain,

$$Q_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a') \leq \mathbb{E}_{\mathbf{q}' \sim \mathbb{P}(\cdot|\mathbf{q}, a)} \left(\frac{2}{\epsilon} f(\mathbf{q}') \right) + \frac{c}{\epsilon} \left(\frac{x_B}{p_B^2} + x_B \right) + \frac{c}{p_B \epsilon} \left(\frac{x_B}{p_B^2} + 2x_B \right) \tag{98}$$

Let $c_3 = \frac{c}{\epsilon}\left(\frac{x_B}{p_B^2} + x_B\right) + \frac{c}{p_B \epsilon}\left(\frac{x_B}{p_B^2} + 2x_B\right)$.

We know from Assumption 2.3, that $\forall \mathbf{q}' \in \mathcal{S}$ such that $\mathbb{P}_\pi(\mathbf{q}'|\mathbf{q}, a) > 0$ it is true that, $\|\mathbf{q}'\|^2 \leq c_1\|\mathbf{q}\|^2 + c_2$. Since $\|\mathbf{q}\|^2 = \|\mathbf{q}\|^2$,,

$$Q_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a') \leq \frac{2}{\epsilon}\left(c_1\|\mathbf{q}\|^2 + c_2\right) + c_3 \tag{99}$$

Since the above inequality is true for all $a, a' \in \mathcal{A}$,

$$|Q_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a')| \leq \frac{2}{\epsilon}\left(c_1\|\mathbf{q}\|^2 + c_2\right) + c_3 \tag{100}$$

Setting $c_4 = \frac{2}{\epsilon}c_2 + c_3$ yields the result. □

A.2 Proof of Step 3

The previous step provided us with bounds over the exact state action value function. Here we incorporate the policy evaluation error in Equation 34 to obtain bounds over the state action value function estimate.

Lemma A.8. *There exists a constant c_5 such that for all states $\mathbf{q} \in \mathcal{S}$, all pairs of actions $a, a' \in \mathcal{A}$, and all policies π,*

$$\left|\widehat{Q}_\pi(\mathbf{q}, a) - \widehat{Q}_\pi(\mathbf{q}, a')\right| \leq c_5\|\mathbf{q}\|^2 + c_4$$

where $\widehat{Q}_\pi(\mathbf{q}, a)$ is the estimate of $Q_\pi(\mathbf{q}, a)$ such that $\left|\widehat{Q}_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a)\right| \leq \kappa\|\mathbf{q}\|^2, \forall a \in \mathcal{A}$.

Proof.

$$\left|\widehat{Q}_\pi(\mathbf{q}, a) - \widehat{Q}_\pi(\mathbf{q}, a')\right| = \left|\widehat{Q}_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a) + Q_\pi(\mathbf{q}, a) - \widehat{Q}_\pi(\mathbf{q}, a') + Q_\pi(\mathbf{q}, a') - Q_\pi(\mathbf{q}, a')\right| \tag{101}$$

$$\leq \left|\widehat{Q}_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a)\right| + \left|\widehat{Q}_\pi(\mathbf{q}, a') - Q_\pi(\mathbf{q}, a')\right| \tag{102}$$

$$+ |Q_\pi(\mathbf{q}, a) - Q_\pi(\mathbf{q}, a')| \tag{103}$$

From Equation 34 and Lemma A.7, it follows that,

$$\left|\widehat{Q}_\pi(\mathbf{q}, a) - \widehat{Q}_\pi(\mathbf{q}, a')\right| \leq \left(2\kappa + \frac{2}{\epsilon}c_1\right)\|\mathbf{q}\|^2 + c_4 \tag{104}$$

Defining $c_5 := 2\kappa + \frac{2}{\epsilon}c_1$ yields the result. □

A.3 Proof of Main theorem (Step 4)

The proof requires utilizing the performance difference lemma to establish a connection between the difference in average cost associated with a policy π and the optimal average cost in terms of the state-action value function Q_π.

Lemma A.9. *Let J_π and $J_{\pi'}$ be the expected infinite horizon average cost associated with policies π and π' respectively. Let d_π be the stationary distribution over state space $\mathcal{S}$ associated with $\mathbb{P}_\pi$. Then it is true that,*

$$J_\pi - J_{\pi'} = \sum_{\mathbf{q} \in \mathcal{S}} d_\pi(\mathbf{q})\left[Q_{\pi'}(\mathbf{q}, \pi(\mathbf{q})) - Q_{\pi'}(\mathbf{q}, \pi'(\mathbf{q}))\right] \tag{105}$$

where $Q_{\pi'}(\mathbf{q}, \pi(\mathbf{q})) = \sum_{a \in \mathcal{A}} \pi(a|\mathbf{q})Q_{\pi'}(\mathbf{q}, a)$ and $Q_{\pi'}(\mathbf{q}, \pi'(\mathbf{q})) = V_{\pi'}(\mathbf{q})$.

Proof. The proof can be found in [Cao99]. □

We restate the theorem for convenience.

Theorem 3.1. *Consider the sequence of policies $\pi_1, \pi_2, \ldots, \pi_T$ obtained from Algorithm 1 with a state-dependent step size $\eta_{\mathbf{q}} = \sqrt{\frac{8\log|\mathcal{A}|}{T}}\frac{1}{M_{\mathbf{q}}}$, where $M_{\mathbf{q}}$ is quadratic in $\mathbf{q}$. Let J_{π_k} be the average cost associated with policy π_k and let J_* be the minimum average cost across policy class Π. Let $\widehat{Q}_{\pi_k}$ be an estimate of state action value function Q_{π_k} associated with policy such that with probability at least $1 - \frac{\delta}{2T}$ it is true that,*

$$\left\| Q_\pi(\mathbf{q}, a) - \widehat{Q}_\pi(\mathbf{q}, a) \right\| \leq \kappa \|\mathbf{q}\|^2 \qquad \forall \mathbf{q} \in \mathcal{S}, \pi \in \Pi \tag{13}$$

Then, under Assumptions 2.1, 2.2 and 2.3, there exist constants c', c'' not depending on T or $\pi_1, \pi_2, \ldots, \pi_T$ such that with probability at least $1 - \delta$:

$$\sum_{k=1}^{T} (J_{\pi_k} - J_*) \leq c'\sqrt{T} + c''T \tag{14}$$

where $c' = \left(2\kappa + \frac{2c_1}{\epsilon}\right)\mathbb{E}_{\mathbf{q}\sim d^{\pi^*}}\|\mathbf{q}\|^2\sqrt{\frac{\log|\mathcal{A}|}{2}} + \sqrt{\frac{\log|\mathcal{A}|}{2}}\left(\frac{2c_2}{\epsilon} + \frac{x_B c}{\epsilon}\left(1 + \frac{2}{p_B} + \frac{1}{p_B^2} + \frac{1}{p_B^3}\right)\right)$ and $c'' = \kappa\mathbb{E}_{\mathbf{q}\sim d^{\pi^*}}\|\mathbf{q}\|^2$.

Proof. Let J^* be the optimal average cost. Let $\pi^* \in \Pi$ be the optimal policy. For any policy $\pi \in \Pi$, performance difference lemma provides the following,

$$J_\pi - J^* = -\mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[Q_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right) - Q_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right)\right] \tag{106}$$

$$= -\mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\Big[Q_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right) - \widehat{Q}_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right) + \widehat{Q}_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right) - Q_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right) \tag{107}$$

$$+ \widehat{Q}_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right) - \widehat{Q}_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right)\Big] \tag{108}$$

$$\leq \mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\left|Q_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right) - \widehat{Q}_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right)\right|\right] + \mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\left|Q_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right) - \widehat{Q}_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right)\right|\right] \tag{109}$$

$$+ \mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\widehat{Q}_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right) - \widehat{Q}_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right)\right] \tag{110}$$

From Equation 34, we know that $\left|Q_\pi\left(\mathbf{q}, a\right) - \widehat{Q}_\pi\left(\mathbf{q}, a\right)\right| \leq \kappa\|\mathbf{q}\|^2$ with probability $1 - \frac{\delta}{2T}$. Hence we obtain the following:

$$J_\pi - J^* \leq 2\kappa\mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\|\mathbf{q}\|^2 + \mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\widehat{Q}_\pi\left(\mathbf{q}, \pi(\mathbf{q})\right) - \widehat{Q}_\pi\left(\mathbf{q}, \pi^*(\mathbf{q})\right)\right] \tag{111}$$

The total regret across time horizon T, with probability $1 - \delta$, can be expressed by summing the above inequality as follows,

$$\sum_{k=1}^{T} J_{\pi_k} - J^* \leq 2\kappa T\mathbb{E}_{\mathbf{q}\sim d^{\pi^*}}\|\mathbf{q}\|^2 + \mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\sum_{k=1}^{T}\left(\widehat{Q}_{\pi_k}\left(\mathbf{q}, \pi_k(\mathbf{q})\right) - \widehat{Q}_{\pi_k}\left(\mathbf{q}, \pi^*(\mathbf{q})\right)\right)\right] \tag{112}$$

where π_k are policy iterates obtained through the NPG policy update below:

$$\pi_k(a|\mathbf{q}) = \frac{\pi_{k-1}(a|\mathbf{q})\exp\left(-\eta_{\mathbf{q}}\widehat{Q}_{\pi_{k-1}}(\mathbf{q}, a)\right)}{\sum_{l\in\mathcal{A}}\pi_{k-1}(l|\mathbf{q})\exp\left(-\eta_{\mathbf{q}}\widehat{Q}_{\pi_{k-1}}(\mathbf{q}, l)\right)} \tag{113}$$

The above update is performed for all $\mathbf{q}$ and $a \in \mathcal{A}$. Let the update parameter $\eta_{\mathbf{q}} = \sqrt{\frac{8\log|\mathcal{A}|}{T}}\frac{1}{M_{\mathbf{q}}}$, where $M_{\mathbf{q}} = c_5\|\mathbf{q}\|^2 + c_4$. Then from Theorem 4.1, it follows that,

$$\sum_{k=1}^{T} J_{\pi_k} - J^* \leq 2\kappa T\mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\|\mathbf{q}\|^2 + \mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\sqrt{\frac{T\log|\mathcal{A}|}{2}}M_{\mathbf{q}}\right] \tag{114}$$

$$= \left(\kappa T + \sqrt{\frac{T\log|\mathcal{A}|}{2}}c_5\right)\mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\|\mathbf{q}\|^2\right] + c_4\sqrt{\frac{T\log|\mathcal{A}|}{2}} \tag{115}$$

From Lemma A.1, it follows that $\mathbb{E}_{\mathbf{q}\sim d_{\pi^*}}\left[\|\mathbf{q}\|^2\right] \leq \beta_2$. Hence,with probability $1-\delta$, we obtain,

$$\sum_{k=1}^{T} J_{\pi_k} - J^* \leq \kappa\beta_2 T + \sqrt{T}\left(\sqrt{\frac{\log|\mathcal{A}|}{2}}c_5\beta_2 + \sqrt{\frac{\log|\mathcal{A}|}{2}}c_4\right) \tag{116}$$

Setting $c' = \left(\sqrt{\frac{\log|\mathcal{A}|}{2}}c_5\beta_2 + \sqrt{\frac{\log|\mathcal{A}|}{2}}c_4\right)$ and $c'' = \kappa\beta_2$ yields the result in the theorem. □