OPT2023: 15th Annual Workshop on Optimization for Machine Learning

On Optimization Formulations of Finite Horizon MDPs

Rajat Vadiraj Dwaraknath RAJATVD@STANFORD.EDU
Lexing Ying LEXING@STANFORD.EDU
Institute for Computational and Mathematical Engineering (ICME), Stanford University

Abstract

In this paper, we extend the connection between linear programming formulations of MDPs and policy gradient methods for infinite horizon MDPs presented in [20] to finite horizon MDPs. The main tool we use for this extension is a reduction from optimization formulations of finite horizon MDPs to infinite horizon MDPs. Additionally, we show using a reparameterization argument that the KKT conditions for the non-convex policy optimization problem for the finite horizon setting are sufficient for global optimality. Further, we use the reduction to extend the Quasi-Newton policy gradient algorithm of [13] to the finite horizon case and achieve performance competitive with value iteration by exploiting backward induction for policy evaluation. To our knowledge, this serves as the first policy gradient-based method for finite horizon MDPs that is competitive with value iteration-based approaches.

1. Introduction

There are two main approaches to finding the optimal policy of a Markov Decision Process (MDP). The first approach essentially involves fixed point iterations of the Bellman equation [2] like value iteration, policy iteration, and Q learning [3, 4, 14]. The second approach leverages optimization formulations of the MDP problem to develop algorithms to find the optimal policy or value function by solving an optimization problem, which is usually a linear program or a convex program with entropic regularization [15, 16, 18]. A complete summary of the optimization formulations for infinite horizon MDPs and the connection to policy gradient approaches can be found in [20].

The focus of most of the prior work has been on the infinite horizon case. Optimization formulations for *finite* horizon MDPs have received lesser attention primarily due to the effectiveness of applying backward induction to directly obtain the optimal value function in a finite number of iterations. Prior work [5, 6] develops linear programming formulations for the finite horizon case, but the discussion does not include the regularized setting, nor does it link to policy optimization. However, recent advances in Quasi-Newton policy gradient-based approaches [11, 12] motivate studying this class of techniques in the finite horizon setting. In this article, we present LP formulations for the finite horizon MDP in a style similar to [20], and link these LPs to the policy optimization problem. We further show global optimality for the regularized problem, and apply the Quasi-Newton policy gradient method of [13] to obtain a policy gradient-based method for finite horizon MDPs. By using backward induction for policy evaluation, we significantly improve the performance of the policy gradient algorithm. We summarize our contributions below.

Contributions

- In Section 3, we present equivalences between the primal, dual, and policy optimization formulations of the finite horizon MDP in both the unregularized and regularized settings. The main tool to show these equivalences is a reduction from the finite horizon MDP to the infinite horizon MDP, which allows us to extend the corresponding equivalences in [20] to the finite horizon case.
- In Section 4, we show that the KKT conditions for the non-convex policy optimization problem for the finite horizon MDP are sufficient for optimality. As a consequence, we conclude that locally optimal policies are also globally optimal.
- In Section 5, we apply the Quasi-Newton policy gradient method from [13] for the finite horizon MDP via the reduction to the infinite horizon case. We improve the algorithm by using backward induction for the policy evaluation step, which is much faster than a matrix inversion. Since this method requires roughly a constant number of interations to converge (5-6), this is the first policy gradient-based approach for finite horizon MDPs that is competitive with direct backward induction.

2. Notation and Preliminaries

A finite horizon Markov decision process $\mathcal{M}$ with discrete state and action space is characterized by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, P, r, g, \gamma)$, where $\mathcal{S}$ is the discrete state space, with each state denoted by s; $\mathcal{A}$ is the discrete action space, with each action denoted by a; T is the finite time horizon; $P^{a,t}$ is a probability transition matrix for each action $a \in A$ and each timestemp $0 \leq t < T$. We use $P^{a,t}_{s \to s'}$ to denote the probability of transitioning from state s to s' if action a is taken at timestep t. $r \in \mathbb{R}^{|\mathcal{S}|(T-1) \times |\mathcal{A}|}$ is the matrix of rewards where $r^a_{(s,t)}$ denotes the reward for taking action a in state s at timestep t (we use the indexing shorthand of $r_{(s,t)}$ to denote $r_{(s+t\cdot|\mathcal{S}|)}$). $g \in \mathbb{R}^{|\mathcal{S}|}$ is the vector of terminal rewards where g_s denotes the reward for reaching state s at the final time T. $\gamma \in [0, 1]$ is the discount factor. We denote by Δ the probability simplex over the space of actions,

$$\Delta := \left\{ \eta \in \mathbb{R}^{|\mathcal{A}|} : \sum_{a \in \mathcal{A}} \eta_a = 1,\ \eta_a \geq 0\ \forall a \in \mathcal{A} \right\}.$$

A policy π is a set of probability distributions over the action space indexed by the state s and also the **timestep** $0 \leq t < T$. We denote by $\pi^a_{(s,t)}$ the probability of taking action a under policy π when in state s at timestep t. For notational convenience, we still treat π as a second-order tensor, with dimension $|\mathcal{A}| \times |\mathcal{S}|\,(T-1)$. We again use the notation $\pi_{(s,t)}$ as shorthand for $\pi_{(s+|\mathcal{S}|*t)}$. We use this indexing shorthand throughout the paper. The set of all valid policies is therefore $\Delta^{|\mathcal{S}|(T-1)}$. Note that both the policy π and transition probabilities P are non-stationary and can depend on the timestep t.

We now introduce policy-averaged quantities – the transition matrix $P^{\pi,t}$ under policy π and the reward r^π under policy π:

$$P^{\pi,t}_{s \to s'} := \sum_{a \in \mathcal{A}} P^{a,t}_{s \to s'} \pi^a_{(s,t)}, \quad \text{and} \quad r^\pi_{(s,t)} := \sum_{a \in \mathcal{A}} r^a_{(s,t)} \pi^a_{(s,t)}.$$

$P^{\pi,t}_{s\to s'}$ denotes the probability of transitioning from state s to state s' at timestep t under policy π, and $r^{\pi}_{(s,t)}$ denotes the expected reward at state s and timestep t under policy π.

The value function v^{π} for an MDP under policy π is defined as the expected future cumulative reward starting from a state s at timestep t and taking actions under policy π,

$$v^{\pi}_{(s,t)} := \mathbb{E}\left[\gamma^{T-t} g_{s_T} + \sum_{k=t}^{T-1} \gamma^{k-t} r^{a_k}_{(s_k,k)} \mid s_t = s\right], \tag{1}$$

where the randomness in the expectation is from $a_k \sim \pi_{(s_k,k)}$ and $s_{k+1} \sim P^{a_k,k}_{s_k\to *}$ for $t \le k < T$. Note that $v^{\pi}_{(s,T)} = g_s$ for all policies π. The goal of the finite horizon MDP problem is to find a policy that maximizes the value function of all states at all timesteps $0 < t \le T$. For convenience, we define the following block quantities – $\widetilde{P}_a \in \mathbb{R}^{|\mathcal{S}|(T-1)\times|\mathcal{S}|(T-1)}$ and $g^a \in \mathbb{R}^{|\mathcal{S}|\cdot(T-1)}$ as

$$\widetilde{P}_a := \begin{bmatrix} 0 & P^{a,0} & 0 & \dots & 0 \\ 0 & 0 & P^{a,1} & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & P^{a,T-2} \\ 0 & \dots & \dots & 0 & 0 \end{bmatrix} \quad \text{and} \quad g^a := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ P^{a,T-1} g \end{bmatrix}.$$

The quantity g^a can be thought of as a reward correction term, which adds the expected terminal reward in the final step T to the reward at timestep $T-1$ if action a is always taken at timestep $T-1$. Additionally, we define $\mathcal{K}_a := I - \gamma \widetilde{P}_a$. Note that $\mathcal{K}_a$ is also a square matrix of size $|\mathcal{S}|\,(T-1)$. Finally, we define the *negative conditional entropy* for a non-negative vector $\rho \in \mathbb{R}^{|\mathcal{A}|}_{+}$ as $h(\rho) := \sum_{a\in\mathcal{A}} \rho^a \log \frac{\rho^a}{\sum_{b\in A}\rho^b}$. Note that this function is strictly convex on the simplex.

3. Main Results

We first present the main equivalences between optimization problems for the unregularized finite horizon MDP in Theorem 1, which are extensions of the equivalences for the infinite horizon case (Theorems 2.1 and 2.2 of [20]).

Theorem 1 *For a finite horizon MDP $\mathcal{M}$, the following optimization formulations are equivalent.*

$$\textbf{\textit{Primal}} \qquad \min_{v\in\mathbb{R}^{|\mathcal{S}|\cdot(T-1)}} e^T v,\ \textit{s.t.}\ \mathcal{K}_a v \ge r^a + \gamma g^a \quad \forall a \in \mathcal{A}, \tag{2}$$

$$\textbf{\textit{Dual}} \qquad \max_{\mu\in\mathbb{R}^{|A|\times|\mathcal{S}|\cdot(T-1)}_{+}} \sum_{a\in A} (\mu^a)^T (r^a + \gamma g^a),\ \textit{s.t.}\ e = \sum_{a\in\mathcal{A}} \mathcal{K}_a^T \mu^a, \tag{3}$$

$$\textbf{\textit{Policy Opt}} \qquad \max_{\pi\in\Delta^{|\mathcal{S}|\cdot(T-1)}} e^T v^{\pi},\ \textit{s.t.}\ \mathcal{K}_{\pi} v^{\pi} = r^{\pi} + \gamma g^{\pi}, \tag{4}$$

where $e \in \mathbb{R}^{|\mathcal{S}|(T-1)}_{++}$ is a fixed vector with positive entries.

Analogously, we present equivalences for the regularized case as extensions to Theorems 3.1 and 3.2 of [20].

Theorem 2 *For an entropy regularized finite horizon MDP $\mathcal{M}$, the following optimization formulations are equivalent.*

$$\textbf{\textit{Primal}} \qquad \min_{v\in\mathbb{R}^{|\mathcal{S}|\cdot(T-1)}} e^T v,\ \textit{s.t.}\ v \geq \beta^{-1}\log\left(\sum_{a\in\mathcal{A}} \exp\beta\left[r^a+\gamma g^a+\gamma\widetilde{P}_a v\right]\right), \tag{5}$$

$$\textbf{\textit{Dual}} \qquad \max_{\mu\in\mathbb{R}_+^{|A|\times|\mathcal{S}|\cdot(T-1)}} \sum_{a\in A}\left(\mu^a\right)^T\left(r^a+\gamma g^a\right)-\beta^{-1}\sum_{s\in\mathcal{S},0\leq t<T} h\left(\mu_{(s,t)}\right),\ \textit{s.t.}\ e=\sum_{a\in\mathcal{A}}\mathcal{K}_a^T\mu^a, \tag{6}$$

$$\textbf{\textit{Policy Opt}} \qquad \max_{\pi\in\Delta^{|\mathcal{S}|\cdot(T-1)}} e^T v^\pi,\ \textit{s.t.}\ \mathcal{K}_\pi v^\pi = r^\pi+\gamma g^\pi-\beta^{-1}h^\pi. \tag{7}$$

where β is the regularization coefficient and $e\in\mathbb{R}_{++}^{|\mathcal{S}|(T-1)}$ is a fixed vector with positive entries.

An important observation to highlight is that the policy optimization problems (4) and (7) are non-convex, while the primal and dual problems are convex. The main tool we use to prove Theorems 1 and 2 is a reduction from the finite horizon MDP $\mathcal{M}$ to an infinite horizon MDP $\widetilde{\mathcal{M}}$ that we present below as Theorem 3. Note that [6] presents a similar reduction, but in the setting without a terminal reward g.

Theorem 3 *Given a finite horizon MDP $\mathcal{M}=(\mathcal{S},\mathcal{A},T,P,r,g,\gamma)$, we can construct an infinite horizon MDP $\widetilde{\mathcal{M}}=\left(\widetilde{\mathcal{S}},\mathcal{A},\widetilde{P},\widetilde{r},\gamma\right)$ such that $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ have identical primal LP formulations where:*

- *$\widetilde{\mathcal{S}}$ has size $|\mathcal{S}|(T-1)+1$ with $T-1$ groups of $|\mathcal{S}|$ states, each corresponding to the state of the finite horizon MDP at a particular timestep, along with one additional terimnal state.*
- *The transition matrices and rewards are defined as*

$$\widetilde{P}^a := \begin{bmatrix} 0 & P^{a,0} & 0 & \dots & 0 & 0\\ 0 & 0 & P^{a,1} & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots\\ 0 & \dots & \dots & 0 & P^{a,T-2} & 0\\ 0 & \dots & \dots & 0 & 0 & \mathbf{1}_{|\mathcal{S}|}\\ 0 & \dots & \dots & 0 & 0 & 1 \end{bmatrix} \quad \textit{and} \quad \widetilde{r} := \begin{bmatrix} r^a_{(*,0)}\\ r^a_{(*,1)}\\ \vdots\\ r^a_{(*,T-1)}+\gamma P^{a,T-1}g\\ 0 \end{bmatrix}.$$

We present here a rough proof sketch for Theorem 3. A detailed proof can be found in the appendix. The transition probabilities are the natural probabilities from the finite horizon MDP, except for the final step which always goes into the terminal absorbing state. The final step transition probabilities $P^{a,T-1}$ instead show up in the reward definition. The reward for taking action a in state s in group $\mathcal{S}_t$ is naturally defined to be $r^a_{(s,t)}$ for $0\leq t<T-1$. The major difference is in the rewards for taking an action in the final group $\mathcal{S}_{T-1}$ which is defined as $r^a_{(s,T-1)}+\gamma\left(P^{a,T-1}g\right)_s$. The additional correction term to the reward can be interpreted as adding in the average contribution of the terminal reward g of taking action a at timestep $T-1$ in state s. Note that if $g=0$, this correction would also be 0. Finally, we also define the reward of taking any action at the terminal absorbing state to be 0.

Theorems 1 and 2 directly follow by applying the reduction Theorem 3 to Theorems 2.1 and 2.2 and Theorems 3.1 and 3.2 of [20] respectively. For completeness, we detail the full proofs in the appendix.

4. Global Optimality

In section 3, we showed that the non-convex policy optimization formulation for a finite horizon MDP is equivalent to a linear program. As a consequence, one would expect that the policy optimization problem possesses some structure that allows us to characterize optimal policies. This is formalized in the following result.

Theorem 4 *The KKT conditions for the non-convex policy optimization problem* (4) *are* ***sufficient*** *for optimality.*

A rough proof sketch is as follows. The equivalence between (3) and (4) is shown via a non-degenerate reparameterization of the dual LP from μ to $\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}$ and a weight variable $w \in \mathbb{R}_+^{|\mathcal{S}|\cdot(T-1)}$. Specifically, the policy π is equivalent to the normalized dual variables μ, and w are the normalizing constants. The constraints in the problem allow the weight variable to be eliminated, resulting in the final policy optimization problem (4). Since the reparameterization is non-degenerate and invertible, KKT points of the dual LP map to KKT points of the intermediate problem and consequently the policy optimization problem. Since KKT conditions are sufficient for LPs, the result follows. We present a similar result for the regularized case.

Theorem 5 *The KKT conditions for the non-convex regularized policy optimization problem* (7) *are* ***sufficient*** *for optimality.*

The proof is almost identical to the proof of Theorem 4, but we use the fact that (6) is a convex problem (since h is a strictly convex function) to conclude that the KKT conditions for the reparameterizated problem are sufficient.

Note that the KKT conditions are necessary for optimality for general constrained optimization problems, but few classes of problems also enjoy sufficiency of these conditions, like convex problems that satisfy certain regularity conditions or linear programs like the dual LP formulation (3). This result shows the sufficiency of the KKT conditions for a non-convex problem by effectively lifting into a higher space and then performing a non-degenerate reparameterization that results in a linear program in the unregularized case and a convex program in the regularized case.

5. Quasi-Newton Policy Gradient

Using the reduction presented in section 3 Theorem 3, we can apply the Quasi-Newton Policy Gradient Algorithm (Algorithm 2.1 in [13]) to the finite horizon MDP, by using $\widetilde{P}^a$ as the state transition matrix for the infinite horizon MDP. This algorithm uses the diagonal of the policy Hessian to precondition policy gradient steps, which leads to quadratic convergence to the optimal policy. Details of the algorithm can be found in [13]. A summary of the convergence for a synthetic finite horizon MDP can be found in Section 5.1.

Since the MDP is a finite horizon MDP, the policy evaluation step, which is line 6 in Algorithm 2.1 of [13], can be performed more efficiently via backward induction (Algorithm 1) instead of solving a linear system (see [2] for a detailed treatment of dynamic programming and backward induction). This takes $O(|S|^2\,(T-1))$ time compared to solving the system which would take $O(n^3)$ time where $n = |S|\,(T-1)$ is the size of the transition matrix. We additionally observe empirically that the algorithm converges in very few Newton steps, effectively requiring only a constant number of calls to the backward induction routine in Algorithm 1. Therefore, this policy gradient-based

algorithm is competitive with the direct value iteration approach for finding the optimal policy. We summarize the modified Quasi-Newton Policy Gradient algorithm for finite horizon MDPs in Algorithm 2.

Algorithm 1 Backward Induction for Policy Evaluation in a Regularized Finite Horizon MDP

Input: Finite Horizon MDP $\mathcal{M}$ with entropy regularization β, input policy $\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}$.
Output: Value function under policy π, that we denote $v^\pi \in \mathbb{R}^{|\mathcal{S}|\cdot(T-1)}$.

Set $v^\pi_{(s,T)} \leftarrow g_s \quad \forall s \in \mathcal{S}$.
Initialize $t \leftarrow T-1$.
while $t \geq 0$
    Set $v^\pi_{(s,t)} \leftarrow \sum_a \pi^a_{(s,t)} \left(r^a_{(s,t)} + \gamma \sum_{s'} P^{a,t}_{s \to s'} v^\pi_{(s',t+1)} \right) - \beta^{-1} h\left(\pi_{(s,t)}\right) \quad \forall s \in \mathcal{S}$.
    Update $t \leftarrow t-1$.
end
return v^*

Algorithm 2 Quasi-Newton Policy Gradient for Finite Horizon MDPs

Input: Finite Horizon MDP $\mathcal{M}$ with entropy regularization β, initial policy $\pi_{\text{init}} \in \Delta^{|\mathcal{S}|\cdot(T-1)}$, learning rate η, convergence threshold ϵ_{tol}.
Output: Estimate of the optimal policy $\hat{\pi} \in \Delta^{|\mathcal{S}|\cdot(T-1)}$.

Construct the infinite horizon MDP $\widetilde{\mathcal{M}}$ from $\mathcal{M}$ according to the reduction in Theorem 3.
Obtain $\hat{\pi}$ by runnning Algorithm 2.1 from [13] on input $\widetilde{\mathcal{M}}$, π_{init}, η, ϵ_{tol}, with the policy evaluation step (line 6) replaced by Algorithm 1.
return $\hat{\pi}$

5.1. Experimental results

We implement Algorithm 2 on synthetic MDPs similar to the experiments conducted by [13]. In Section 5.1, we use a finite horizon MDP with 100 states, 50 actions, and a time horizon of $T = 10$ steps. For each state, the transition probabilities for σ fraction of target states are chosen uniformly at random. The transition probabilities for the remaining $1 - \sigma$ fraction of transitions are set to 0. The rewards are chosen as $r^t_{(s,a)} = U_s U_{(s,a)}$ where U_s and $U_{(s,a)}$ are sampled uniformly at random from $[0, 1]$. We use a regularization coefficient of 0.001. This is similar to the setup used by [13]. We observe from the results in Section 5.1 that 6 iterations of the Quasi-Newton method are sufficient for convergence. Therefore, this empirically shows that Algorithm 2 is competitive with direct backward induction.

6. Conclusion

In this work, we derive equivalences between primal, dual, and policy optimization formulations for finite horizon MDPs along similar lines as [20]. The main technique development is a reduction from the finite horizon MDP to the infinite horizon MDP that allows us to immediately obtain the equivalences from the results of [20]. We exploit these equivalences to prove sufficiency of the KKT conditions for the non-convex policy optimization problem in both unregularized and regularized

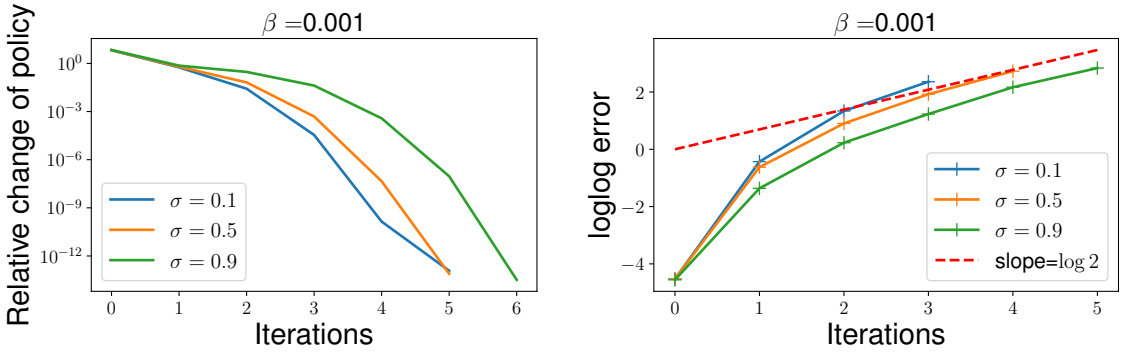

Figure 1: Convergence of Quasi-Newton policy gradient (Algorithm 2) for a synthetic **finite horizon MDP** with $|\mathcal{S}| = 100$, $T = 10$ and $|\mathcal{A}| = 50$, for different values of the sparsity parameter σ in the MDP transition matrix. The figure on the right plots the quantity $\log|\log|\text{policy error}||$ vs iteration count. For quadratic convergence, this plot should be parallel to a line with slope $\log 2$, which is what we roughly observe.

settings. Additionally as a consequence of the reduction, we develop the first policy gradient-based optimization algorithm that is competitive with direct backward induction for finite horizon MDPs by applying the Quasi-Newton approach introduced in [13], with faster policy evaluation via backward induction.

An interesting avenue of future work is to connect this analysis to the study of policy gradient flow for finite horizon stochastic optimal control [21] in the continuous setting, perhaps by taking a limit of the timesteps in the finite horizon MDP analysis. Additionally, further specializing policy gradient-based approaches to surpass backward induction for finite horizon MDPs is also a direction for future work.

Acknowledgements

The authors thank Haoya Li for sharing code to implement the Quasi-Newton policy gradient method.

References

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation , and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

[2] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[3] Dimitri Bertsekas. *Abstract dynamic programming*. Athena Scientific, 2022.

[4] Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[5] Arnab Bhattacharya and Jeffrey P Kharoufeh. Linear programming formulation for non-stationary, finite-horizon markov decision process models. *Operations Research Letters*, 45 (6):570–574, 2017.

[6] Cyrus Derman and Morton Klein. Some remarks on finite horizon markovian decision models. *Operations research*, 13(2):272–278, 1965.

[7] Archis Ghate and Robert L Smith. A linear programming approach to nonstationary infinite-horizon markov decision processes. *Operations Research*, 61(2):413–425, 2013.

[8] Soumyajit Guin and Shalabh Bhatnagar. A policy gradient approach for Finite Horizon Constrained Markov Decision Processes, October 2022. URL http://arxiv.org/abs/2210.04527. arXiv:2210.04527 [cs].

[9] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[10] Alec Koppel, Amrit Singh Bedi, Bhargav Ganguly, and Vaneet Aggarwal. Convergence rates of average-reward multi-agent reinforcement learning via randomized linear programming. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4545–4552. IEEE, 2022.

[11] Haoya Li, Samarth Gupta, Hsiangfu Yu, Lexing Ying, and Inderjit Dhillon. Quasi-newton policy gradient algorithms. *arXiv preprint arXiv:2110.02398*, 2021.

[12] Haoya Li, Hsiang-fu Yu, Lexing Ying, and Inderjit Dhillon. Accelerating Primal-dual Methods for Regularized Markov Decision Processes, February 2022. URL http://arxiv.org/abs/2202.10506. arXiv:2202.10506 [cs, math, stat].

[13] Haoya Li, Samarth Gupta, Hsiangfu Yu, Lexing Ying, and Inderjit Dhillon. Approximate Newton policy gradient algorithms, March 2023. URL http://arxiv.org/abs/2110.02398. arXiv:2110.02398 [cs, math].

[14] Hamid Maei, Csaba Szepesvari, Shalabh Bhatnagar, Doina Precup, David Silver, and Richard S Sutton. Convergent temporal-difference learning with arbitrary smooth function approximation. *Advances in neural information processing systems*, 22, 2009.

[15] Alan Malek, Yasin Abbasi-Yadkori, and Peter Bartlett. Linear programming for large-scale markov decision problems. In *International conference on machine learning*, pages 496–504. PMLR, 2014.

[16] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[17] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[18] Mengdi Wang. Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.

[19] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[20] Lexing Ying and Yuhua Zhu. A NOTE ON OPTIMIZATION FORMULATIONS OF MARKOV DECISION PROCESSES.

[21] Mo Zhou and Jianfeng Lu. A Policy Gradient Framework for Stochastic Optimal Control Problems with Global Convergence Guarantee, February 2023. URL http://arxiv.org/abs/2302.05816. arXiv:2302.05816 [cs, eess, math].

Appendix A. Related Work

LP Formulations of MDPs Linear programming formulations of MDPs have received much attention in the literature [10, 15, 16]. The primary benefit of these formulations is to use approximate LP solvers to obtain significant speedups over traditional value iteration-based methods. [7] extend this discussion to the non-stationary but infinite horizon setting. The finite horizon setting, however, has received much lesser attention. [5, 6] formulate LPs for the finite horizon setting, but their discussion is limited to LP solvers and does not connect to policy gradient. Additionally, they do not study the regularized setting.

Policy Gradient Methods Policy gradient methods [9, 17, 19] are a popular class of algorithms for reinforcement learning owing to their empirical efficacy, flexibility in dealing with large state and action spaces, and amenability with function approximation. However, much of the progress in this class of algorithms has been empirically motivated, and the theory behind these algorithms lags behind significantly. In the context of infinite horizon MDPs, [1] have developed convincing theories with regards to convergence rates, sample complexities, and approximation guarantees of policy gradient under tabular and parameteric settings. However, the theory behind policy gradient for finite horizon MDPs is much more limited. [8] present policy gradient methods for finite horizon-constrained MDPs, and much of their theory is focused on constraint analysis.

Appendix B. Proofs of Theoretical Results

B.1. Proof of Theorem 3

Proof Let $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ be as in the theorem. The primal LP formulation of the infinite horizon MDP $\widetilde{\mathcal{M}}$ (for example, see eq. (2.2) of [20]) is:

$$\min_{v \in \mathbb{R}^{|\mathcal{S}|\cdot(T-1)}, v_{\text{end}}} \sum_{s \in \mathcal{S}, 0 \le t < T} e_{(s,t)} v_{(s,t)} + e_{\text{end}} v_{\text{end}} \tag{8}$$

subject to:

$$\begin{bmatrix} r^a_{(*,0)} \\ r^a_{(*,1)} \\ \vdots \\ r^a_{(*,T-1)} + \gamma P^{a,T-1} g \\ 0 \end{bmatrix} \le \begin{bmatrix} I & -\gamma P^{a,0} & 0 & \dots & 0 & 0 \\ 0 & I & -\gamma P^{a,1} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & I & -\gamma P^{a,T-2} & 0 \\ 0 & \dots & \dots & 0 & I & -\gamma \mathbf{1}_{|\mathcal{S}|} \\ 0 & \dots & \dots & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_{(*,0)} \\ v_{(*,1)} \\ \vdots \\ v_{(*,T-1)} \\ v_{\text{end}} \end{bmatrix} \quad \forall a \in \mathcal{A}.$$

where v_{end} corresponds to the value function at the additional terminal state, and e and e_{end} are positive. Note that v is the value function for the infinite horizon MDP in this problem. The final row of the constraint is simply $v_{\text{end}} \ge 0$, while the remaining constraints can be succinctly written as

$$r^a + \gamma g^a \le \mathcal{K}_a v - \gamma v_{\text{end}} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{1}_{|\mathcal{S}|} \end{bmatrix} \quad \forall a \in \mathcal{A}. \tag{9}$$

These constraints directly imply that $v^*_{\text{end}} = 0$. To see why, assume for the sake of contradiction that $v^*_{\text{end}} > 0$. Since it is feasible, it must satisfy the remaining constraints (9). Therefore, $v_{\text{end}} = 0$ must also be feasible. But the objective value is necessarily smaller by taking $v_{\text{end}} = 0$, so we arrive at a contradiction. Consequently, v^*_{end} must be 0. As a result, we recover the primal LP for the finite horizon problem (2) by eliminating v_{end} from (8), and this completes the proof.

■

B.2. Proof of Theorem 1

Proof The statement directly follows by first applying the reduction in Theorem 3, and then using Theorems 2.1 and 2.2 from [20] on the resulting infinite horizon MDP. The resuling dual and policy optimization problems are exactly (3) and (4). For completeness, we present a direct proof of the equivalences below, without needing to appeal to the reduction.

We follow a similar presentation to [20] but derive the LP formulations for the finite horizon MDP instead. Note that much of the derivation is identical, except for the fact that the value function, policy, averaged reward, and transition probabilities are now also indexed by the timestep along with the state, effectively multiplying the state space by a factor of the time horizon T. We begin with the Bellman optimality equation for the **optimal** value function of an MDP with a finite state horizon.

$$v^*_{(s,t)} = \max_{a \in \mathcal{A}} \left[r^a_{(s,t)} + \gamma \sum_{s' \in \mathcal{S}} P^{a,t}_{s \to s'} v^*_{(s,t+1)} \right] \quad \forall s \in \mathcal{S}, 0 \le t < T, \tag{10}$$
$$v^*_{(s,T)} = g_s \quad \forall s \in \mathcal{S}.$$

Note that the primary differences between this and the Bellman optimality equation for the infinite horizon case is the presence of the boundary condition on $v^*_{(s,T)}$, and the $t+1$ index in the recursive term. We turn to a linear programming characterization of the optimal value function that is equivalent to the Bellman optimality equation for the finite horizon MDP. This equivalence is known and derived in [16] for example. The optimal value function v^* is the solution to the following LP

$$\min_{v \in \mathbb{R}^{|\mathcal{S}| \cdot T}} \sum_{s \in \mathcal{S}, 0 \le t \le T} e_{(s,t)} v_{(s,t)} \tag{11}$$
subject to:
$$v_{(s,t)} \ge r^a_{(s,t)} + \gamma \sum_{s' \in \mathcal{S}} P^{t,a}_{s \to s'} v_{(s',t+1)} \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}, 0 \le t < T$$
$$v_{(s,T)} = g_s \quad \forall s \in \mathcal{S},$$

where $e \in \mathbb{R}^{|\mathcal{S}|(T-1)}$ is an arbitrary positive fixed weight vector. In matrix notation, it takes the simpler form (2)

$$\min_{v \in \mathbb{R}^{|\mathcal{S}| \cdot (T-1)}} e^T v$$
subject to:
$$\mathcal{K}_a v \ge r^a + \gamma g^a \quad \forall a \in \mathcal{A}.$$

Next, we introduce a matrix of dual variables $\mu \in \mathbb{R}_+^{|A|\times|\mathcal{S}|\cdot(T-1)}$ corresponding to each of the inequality constraints in (2). The **Lagrangian** $\mathcal{L}$ takes the form

$$\mathcal{L}(v,\mu) = \sum_{s\in\mathcal{S},0\le t<T} e_{(s,t)}v_{(s,t)} + \sum_{s\in S,0\le t<T,a\in\mathcal{A}} \mu^a_{(s,t)}\left[r^a_{(s,t)} + \gamma\sum_{s'\in\mathcal{S}} P^{a,t}_{s\to s'}v_{(s',t+1)} - v_{(s,t)}\right] \tag{12}$$

$$= \sum_{s\in\mathcal{S},0\le t<T}\left(e_{(s,t)} - \sum_{a\in\mathcal{A}}\mu^a_{(s,t)} + \gamma\sum_{s',a}\mu^a_{(s',t-1)}P^{a,t-1}_{s'\to s}\right)v_{(s,t)} + \sum_{s\in S,0\le t<T,a\in\mathcal{A}}\mu^a_{(s,t)}r^a_{(s,t)}$$

$$+ \sum_{s\in S,a\in\mathcal{A}}\left(\mu^a_{(s,T-1)}\cdot\gamma\sum_{s'\in\mathcal{S}}P^{a,T-1}_{s\to s'}v_{(s',T)}\right).$$

In the second step above, we collected terms that multiply a factor of $v_{(s,t)}$ for a given state s and timestep t. This explicitly shows that the Lagrangian is an affine function of the primal variables v. We can write the Lagrangian succinctly in matrix notation as follows:

$$\mathcal{L}(v,\mu) = e^Tv + \sum_{a\in\mathcal{A}}(\mu^a)^T\left(r^a + \gamma g^a - \left(I - \gamma\widetilde{P}_a\right)v\right) \tag{13}$$

$$= \left(e - \sum_{a\in\mathcal{A}}\left(I - \gamma\widetilde{P}_a^T\right)\mu^a\right)^T v + \sum_{a\in\mathcal{A}}(\mu^a)^T(r^a + \gamma g^a).$$

Again, we explicitly write the Lagrangian in two forms where it is clear that it is an affine function of both μ and v. The primal-dual problem is then simply the following min-max optimization problem

$$\max_{\mu\in\mathbb{R}_+^{|A|\times|\mathcal{S}|\cdot(T-1)}} \quad \min_{v\in\mathbb{R}^{|\mathcal{S}|\cdot(T-1)}} \mathcal{L}(v,\mu). \tag{14}$$

To obtain the dual LP of (2), we explicitly solve the inner minimization of the primal-dual formulation (14). As we noted earlier, $\mathcal{L}$ is an affine function of v. Therefore, its minimum is $-\infty$ unless the linear coefficient in front of v is 0. We can find this coefficient by taking a gradient,

$$\left(\frac{\partial\mathcal{L}}{\partial v}\right)_{s,t} = e_{(s,t)} - \sum_{a\in\mathcal{A}}\mu^a_{(s,t)} + \gamma\sum_{s'\in\mathcal{S},a\in\mathcal{A}}\mu^a_{(s',t-1)}P^{a,t-1}_{s'\to s} \quad \forall s\in\mathcal{S}, 0\le t<T. \tag{15}$$

Where we have implicitly defined $\mu^*_{(*,-1)} = 0$ for notational convenience. Observe that the states indexing the transition probability have switched (indicated in red). This subtle difference is important and will play a crucial role in the connection to the policy optimization formulation. The optimal value of the inner minimization is simply the affine constant in the expression of $\mathcal{L}$. This is equal to $\sum_{s\in\mathcal{S},0\le t<T,a\in\mathcal{A}}\mu^a_{(s,t)}\cdot r^a_{(s,t)} + \sum_{s\in S,a\in\mathcal{A}}\mu^a_{(s,T-1)}\cdot\gamma\sum_{s'\in\mathcal{S}}P^{a,T-1}_{s\to s'}g_{s'}$. Solving the inner

problem, we obtain the dual LP,

$$\max_{\mu} \sum_{s\in\mathcal{S}, 0\le t<T, a\in\mathcal{A}} \mu^a_{(s,t)} \cdot r^a_{(s,t)} + \sum_{s\in S, a\in\mathcal{A}} \left(\mu^a_{(s,T-1)} \cdot \gamma \sum_{s'\in\mathcal{S}} P^{a,T-1}_{s\to s'} g_{s'} \right) \tag{16}$$

subject to:

$$\mu \ge 0,$$

$$e_{(s,t)} = \sum_a \mu^a_{(s,t)} - \gamma \sum_{s',a} \mu^a_{(s',t-1)} P^{a,t-1}_{s'\to s} \quad \forall s \in \mathcal{S}, 0 \le t < T.$$

Again, we simplify the presentation by rewriting in matrix notation. First, we can write the equality constraints as a single vector constraint as follows

$$\begin{bmatrix} e_{(*,0)} \\ e_{(*,1)} \\ \vdots \\ e_{(*,T-1)} \end{bmatrix} = \sum_{a\in\mathcal{A}} \begin{bmatrix} I & 0 & 0 & \dots & 0 \\ -\gamma \left(P^{a,0}\right)^T & I & 0 & \dots & 0 \\ 0 & -\gamma \left(P^{a,1}\right)^T & I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -\gamma \left(P^{a,T-2}\right)^T & I \end{bmatrix} \begin{bmatrix} \mu^a_{(*,0)} \\ \mu^a_{(*,1)} \\ \vdots \\ \mu^a_{(*,T-1)} \end{bmatrix}.$$

The probability transition matrices appear transposed precisely because of the switch in the indices s' and s that we highlighted previously. Notice that the block matrix in the sum is exactly equal to $\mathcal{K}_a^T$. Therefore, the dual LP in matrix form can be succinctly written as

$$\max_{\mu\in\mathbb{R}_+^{|A|\times|\mathcal{S}|\cdot(T-1)}} \sum_{a\in A} (\mu^a)^T (r^a + \gamma g^a)$$

subject to:

$$e = \sum_{a\in\mathcal{A}} \mathcal{K}_a^T \mu^a,$$

which is exactly (3).

Finally, we prove the equivalence between the dual LP (3) and the policy optimization formulation. To do this, we first introduce an intermediate optimization problem by applying a non-degenerate reparameterization to the dual problem (3). Specifically, we apply the transformation $(w, \pi) = T(\mu)$, where $w \in \mathbb{R}_+^{|\mathcal{S}|\cdot(T-1)}$ and $\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}$. Note that the total number of effective variables is the same – w is $|\mathcal{S}|\,(T-1)$ variables and π can be described by $|\mathcal{S}|\,(T-1)\cdot(|\mathcal{A}|-1)$ due to the simplex constraint, resulting in a total of $|\mathcal{S}|\,(T-1)\,|\mathcal{A}|$ variables, which is the same as $\mu \in \mathbb{R}_+^{|A|\times|\mathcal{S}|\cdot(T-1)}$. The transformation T is given by:

$$w_{(s,t)} = \sum_{a\in\mathcal{A}} \mu^a_{(s,t)} \quad \text{and} \quad \pi^a_{(s,t)} = \frac{\mu^a_{(s,t)}}{w_{(s,t)}}. \tag{17}$$

It is easy to see that for $w > 0$, this transformation is invertible and differentiable, and is therefore non-degenerate. The optimization problem that results from applying this reparameterization to the

dual problem (3) is (after expanding out matrices in terms of indices),

$$\max_{\pi,w} \sum_{s\in\mathcal{S},0\le t<T,a\in\mathcal{A}} w_{(s,t)}\pi^a_{(s,t)}r^a_{(s,t)} + \sum_{s\in\mathcal{S},a\in\mathcal{A}} \left(w_{(s,T-1)}\pi^a_{(s,T-1)} \cdot \gamma \sum_{s'\in\mathcal{S}} P^{a,T-1}_{s\to s'} g_{s'} \right) \tag{18}$$

subject to:

$w \ge 0,$

$\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)} \quad \forall s \in \mathcal{S}, 0 \le t < T,$

$$e_{(s,t)} = \sum_a w_{(s,t)}\pi^a_{(s,t)} - \gamma \sum_{s',a} w_{(s',t-1)}\pi^a_{(s',t-1)} \cdot P^{a,t}_{s'\to s} \quad \forall s \in \mathcal{S}, 0 \le t < T.$$

Simplifying the objective, we get

$$\begin{aligned}
&\sum_{s\in\mathcal{S},0\le t<T,a\in\mathcal{A}} w_{(s,t)}\pi^a_{(s,t)}r^a_{(s,t)} + \sum_{s\in\mathcal{S},a\in\mathcal{A}} \left(w_{(s,T-1)}\pi^a_{(s,T-1)} \cdot \gamma \sum_{s'\in\mathcal{S}} P^{a,T-1}_{s\to s'} g_{s'} \right) \\
&= w^T r^\pi + \gamma \left(w_{(*,T-1)}\right)^T P^{\pi,T-1} g \\
&= w^T \left(r^\pi + \gamma g^\pi\right).
\end{aligned}$$

Similarly, the constraint simplifies to

$$\begin{aligned}
e_{(s,t)} &= w_{(s,t)} - \gamma \sum_{s',a} w_{(s',t-1)}\pi^a_{(s',t-1)} P^{a,t}_{s'\to s} \\
&= w_{(s,t)} - \gamma \sum_{s'} w_{(s',t-1)} P^{\pi,t-1}_{s'\to s} \\
\implies e &= \mathcal{K}^T_\pi w.
\end{aligned}$$

Now, we recall the Bellman equation for a finite horizon MDP to describe the value function under the policy given by π,

$$v^\pi_{(s,t)} = \sum_a \pi^a_{(s,t)} \left(r^a_{(s,t)} + \gamma \sum_{s'} P^{a,t}_{s\to s'} v^\pi_{(s',t+1)} \right) \quad \forall s \in \mathcal{S}, 0 \le t < T. \tag{19}$$

We can simplify this to,

$$\begin{aligned}
v^\pi_{(s,t)} &= r^\pi_{(s,t)} + \gamma \sum_{s'} P^{\pi,t}_{s\to s'} v^\pi_{(s',t+1)} \\
\implies r^\pi_{(s,t)} &= v^\pi_{(s,t)} - \gamma \sum_{s'} P^{\pi,t}_{s\to s'} v^\pi_{(s',t+1)}.
\end{aligned}$$

In matrix notation, this becomes

$$\begin{aligned}
r^\pi &= \mathcal{K}_\pi v^\pi - \gamma g^\pi \\
\implies r^\pi + \gamma g^\pi &= \mathcal{K}_\pi v^\pi.
\end{aligned}$$

Therefore, we can eliminate the weight variable from the reparameterization of the LP to get the following optimization problem:

$$\begin{aligned} &\max_{\pi} \quad e^T v^\pi \\ &\text{subject to:} \\ &\mathcal{K}_\pi v^\pi = r^\pi + \gamma g^\pi, \\ &\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}. \end{aligned}$$

This is exactly the **non-convex** policy optimization problem (4). Reformulating in terms of only the policy, we get

$$\max_{\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}} \quad e^T \left(\mathcal{K}_\pi\right)^{-1} \left(r^\pi + \gamma g^\pi\right). \tag{20}$$

This completes the proof. ■

B.3. Proof of Theorem 2

Proof Similar to the proof of Theorem 1, the statement directly follows by first applying the reduction in Theorem 3, and then using Theorems 3.1 and 3.2 from [20] on the resulting infinite horizon MDP. The resuling dual and policy optimization problems are exactly (6) and (7). Again for completeness, we present a direct proof of the equivalences below, without needing to appeal to the reduction.

We begin by introducing entropy regularization in the Lagrangian (13):

$$\begin{aligned} \mathcal{L}_\beta\left(v, \mu\right) &= e^T v + \sum_{a \in \mathcal{A}} \left(\mu^a\right)^T \left(r^a + \gamma g^a - \left(I - \gamma \widetilde{P}_a\right) v\right) - \beta^{-1} \sum_{s \in \mathcal{S}, 0 \le t < T} h\left(\mu_{(s,t)}\right) \\ &= \left(e - \sum_{a \in \mathcal{A}} \left(I - \gamma \widetilde{P}_a^T\right) \mu^a\right)^T v + \sum_{a \in \mathcal{A}} \left(\mu^a\right)^T \left(r^a + \gamma g^a\right) - \beta^{-1} \sum_{s \in \mathcal{S}, 0 \le t < T} h\left(\mu_{(s,t)}\right). \end{aligned} \tag{21}$$

Note that since h is 1-strongly convex (and consequently $-h$ is 1-strongly concave), the regularized Lagrangian is $\beta^{-1}-$strongly concave in μ. Additionally, it is also affine (and therefore convex) in v. The primal-dual problem is again simply:

$$\max_{\mu \in \mathbb{R}_+^{|A| \times |\mathcal{S}| \cdot (T-1)}} \quad \min_{v \in \mathbb{R}^{|\mathcal{S}| \cdot (T-1)}} \mathcal{L}_\beta\left(v, \mu\right). \tag{22}$$

Using this regularized primal-dual as a starting point, we can derive primal and dual formulations of the regularized MDP. Since $\mathcal{L}_\beta$ is convex in v and concave in μ, we can switch the order of the min and max in (22). To get the primal problem, we explicitly compute the maximum over μ. To do this, we apply the reparameterization (17), $\mu^a_{(s,t)} = w_{(s,t)} \pi^a_{(s,t)}$ with $\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}$. The primal-dual problem becomes

$$\min_{v \in \mathbb{R}^{|\mathcal{S}|\cdot(T-1)}} \max_{\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}, w \ge 0} e^T v + w^T \left(r^\pi + \gamma g^\pi - \mathcal{K}_\pi v - \beta^{-1} h^\pi\right).$$

The maximum over π only applies to the expression in parenthesis multiplying w. In fact, this maximum over π is in the form of the Gibbs variational principle, and the solution can be expressed in closed form as

$$\max_{\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}} r^\pi + \gamma g^\pi - \mathcal{K}_\pi v - \beta^{-1} h^\pi = \beta^{-1} \log \left(\sum_{a \in \mathcal{A}} \exp \left(\beta \left[r^a + \gamma g^a - \mathcal{K}_a v \right] \right) \right).$$

Now, one can interpret the variables w as Lagrange multipliers corresponding to inequality constraints given by $\beta^{-1} \log \left(\sum_{a \in \mathcal{A}} \exp \left(\beta \left[r^a + \gamma g^a - \mathcal{K}_a v \right] \right) \right) \leq 0$. We can expand $\mathcal{K}_a$, and bring out a term $\exp(\beta v)$ from the sum, to transform this constraint into

$$v \geq \beta^{-1} \log \left(\sum_{a \in \mathcal{A}} \exp \beta \left[r^a + \gamma g^a + \gamma \widetilde{P}_a v \right] \right).$$

The primal problem takes the form

$$\begin{aligned} &\min_{v \in \mathbb{R}^{|\mathcal{S}|\cdot(T-1)}} e^T v \\ &\text{subject to:} \\ &v \geq \beta^{-1} \log \left(\sum_{a \in \mathcal{A}} \exp \beta \left[r^a + \gamma g^a + \gamma \widetilde{P}_a v \right] \right), \end{aligned}$$

which is exactly (5).

An alternative starting point for the regularized MDP problem is the regularized Bellman optimality equation. This is essentially the standard Bellman optimality equation (10), but with the max over actions replaced by a softmax. Specifically, we use the logsumexp function with inverse temperature $\beta > 0$ to represent the softmax. The regularized Bellman optimality equation is therefore

$$\begin{aligned} v^*_{(s,t)} &= \beta^{-1} \log \left(\sum_{a \in \mathcal{A}} \exp \beta \left[r^a_{(s,t)} + \gamma \sum_{s' \in \mathcal{S}} P^{a,t}_{s \to s'} v^*_{(s',t+1)} \right] \right) \quad \forall s \in \mathcal{S}, 0 \leq t < T, \qquad (23) \\ v^*_{(s,T)} &= g_s \quad \forall s \in \mathcal{S}. \end{aligned}$$

In matrix notation,

$$v^* = \beta^{-1} \log \left(\sum_{a \in \mathcal{A}} \exp \beta \left[r^a + \gamma g^a + \gamma \widetilde{P}_a v^* \right] \right).$$

Notice that this form of the regularized Bellman optimality equation corresponds naturally with the constraints in the primal problem derived above.

To obtain the regularized dual problem, we use a similar argument as in the proof of Theorem 1. Specifically, we observe that $\mathcal{L}_\beta$ is an affine function of v. So, the inner minimization has optimal value $-\infty$ unless the linear multiplier in front of v is 0, in which case the optimal value is the affine constant. In (21), the affine constant is exactly that in (13) but with the additional entropy

regularization term that does not depend on v. So, we can write the dual problem as

$$\max_{\mu \in \mathbb{R}_+^{|A| \times |\mathcal{S}| \cdot (T-1)}} \sum_{a \in A} \left(\mu^a\right)^T \left(r^a + \gamma g^a\right) - \beta^{-1} \sum_{s \in \mathcal{S}, 0 \leq t < T} h\left(\mu_{(s,t)}\right)$$

subject to:

$$e = \sum_{a \in \mathcal{A}} \mathcal{K}_a^T \mu^a,$$

which is exactly (6).

Before moving on, we introduce the regularized Bellman equation for the value function under a given policy π. We obtain this equation by adding the entropy regularizer term to the recursive definition of the value function in the standard Bellman equation (19).

$$v^\pi_{(s,t)} = \sum_a \pi^a_{(s,t)} \left(r^a_{(s,t)} + \gamma \sum_{s'} P^{a,t}_{s \to s'} v^\pi_{(s',t+1)} \right) - \beta^{-1} h\left(\pi_{(s,t)}\right) \quad \forall s \in \mathcal{S}, 0 \leq t < T. \tag{24}$$

We introduce a vector $h^\pi \in \mathbb{R}^{|\mathcal{S}|(T-1)}$ with each entry equal to the entropy of policy distribution at each state and timestep. Specifically, $h^\pi_{(s,t)} = h\left(\pi_{(s,t)}\right)$. We rewrite (24) in matrix notation:

$$\mathcal{K}_\pi v^\pi = r^\pi + \gamma g^\pi - \beta^{-1} h^\pi. \tag{25}$$

We can now show that the regularized dual problem (6) and the regularized policy optimization problem (7) are equivalent. We again introduce the reparameterization (17) in the regularized dual to obtain

$$\max_{\substack{\pi \in \Delta^{|\mathcal{S}| \cdot (T-1)}, \\ w \in \mathbb{R}_+^{|\mathcal{S}| \cdot (T-1)}}} \sum_{\substack{a \in A, \\ s \in \mathcal{S}, \\ 0 \leq t < T}} w_{(s,t)} \pi^a_{(s,t)} r^a_{(s,t)} + \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \left(w_{(s,T-1)} \pi^a_{(s,T-1)} \cdot \gamma \sum_{s' \in \mathcal{S}} P^{a,T-1}_{s \to s'} g_{s'} \right) - \beta^{-1} \sum_{s \in \mathcal{S}, 0 \leq t < T} w_{(s,t)} h\left(\pi_{(s,t)}\right)$$

subject to:

$$e_{(s,t)} = w_{(s,t)} - \gamma \sum_{s',a} w_{(s',t-1)} \pi^a_{(s',t-1)} P^{a,t}_{s' \to s}.$$

Simplifying with matrix notation, we get

$$\max_{\pi \in \Delta^{|\mathcal{S}| \cdot (T-1)}, w \in \mathbb{R}_+^{|\mathcal{S}| \cdot (T-1)}} w^T \left(r^\pi + \gamma g^\pi - \beta^{-1} h^\pi\right)$$

subject to:

$$e = \mathcal{K}_\pi^T w.$$

Using a similar argument as in the proof of Theorem 1, we can eliminate the weight variable w by solving the constraint and plugging into the objective to get,

$$\max_{\pi \in \Delta^{|\mathcal{S}| \cdot (T-1)}} e^T \left(\mathcal{K}_\pi\right)^{-1} \left(r^\pi + \gamma g^\pi - \beta^{-1} h^\pi\right).$$

Now compare the objective to the regularized Bellman equation (24). Indeed, the objective simplifies exactly to $e^T v^\pi$. We get

$$\begin{aligned}&\max_{\pi \in \Delta^{|\mathcal{S}|\cdot(T-1)}, v^\pi \in \mathbb{R}^{|\mathcal{S}|\cdot(T-1)}} e^T v^\pi \\ &\text{subject to:} \\ &\mathcal{K}_\pi v^\pi = r^\pi + \gamma g^\pi - \beta^{-1} h^\pi,\end{aligned}$$

which is exactly the regularized policy gradient optimization problem (7), and this completes the proof. ■

B.4. Proof of Theorem 4

Proof Let π^* be a KKT point of (4). Based on the proof of Theorem 1, we know that (π^*, w^*) is a KKT point of (18) where $w^* = \left(\mathcal{K}_{\pi^*}^T\right)^{-1} e$. Let μ^* be uniquely defined by the reparameterization (17) applied to (π^*, w^*). Since this reparameterization is invertible and differentiable, μ^* is a KKT point of the dual LP (3). But (3) is an LP, so μ^* is optimal. Since the objective functions of (3), (18) and (4) are all equivalent, we conclude that π^* is optimal for the policy optimization problem (4).

We briefly justify why a differentiable and invertible reparameterization preserves the KKT conditions below. Consider the general constrained problem in n dimensions with m constraints given by functions $g_i : \mathbb{R}^n \to \mathbb{R}$,

$$\begin{aligned}&\min_{x \in \mathbb{R}^n} f(x) \\ &\text{subject to:} \\ &g_i(x) \leq 0, \forall i \in [m]\,.\end{aligned}$$

The pair (x^*, λ^*) for $\lambda^* \in \mathbb{R}^m$ is a KKT point if it satisfies the KKT conditions,

$$\begin{aligned}&\nabla f(x^*) + \sum_{i \in [m]} \lambda_i^* \nabla g_i(x^*) = 0, \\ &g_i(x^*) \leq 0, \forall i \in [m]\,, \\ &\lambda_i^* \geq 0, \forall i \in [m]\,, \\ &\lambda_i^* g_i(x^*) = 0, \forall i \in [m]\,.\end{aligned}$$

Now consider the equivalent problem obtained by applying the reparameterization or change of variables $x = T(y)$ where $T : \mathbb{R}^n \to \mathbb{R}^n$ is differentiable and invertible,

$$\begin{aligned}&\min_{y \in \mathbb{R}^n} f(T(y)) \\ &\text{subject to:} \\ &g_i(T(y)) \leq 0, \forall i \in [m]\,.\end{aligned}$$

The corresponding KKT conditions for a pair (y^*, μ^*) with $\mu^* \in \mathbb{R}^m$ are,

$$
\begin{aligned}
&J(y^*)^T \nabla f(T(y^*)) + \sum_{i \in [m]} \mu_i^* J(y^*)^T \nabla g_i(T(y^*)) = 0, \\
&g_i(T(y^*)) \leq 0, \forall i \in [m]\,, \\
&\mu_i^* \geq 0, \forall i \in [m]\,, \\
&\mu_i^* g_i(T(y^*)) = 0, \forall i \in [m]\,,
\end{aligned}
$$

where $J(y^*) = \frac{\partial T}{\partial y}(y^*)$ is the Jacobian of the transformation T evaluated at y^*. Since J is invertible by assumption, the first stantionarity condition reduces to $\nabla f(T(y^*)) + \sum_{i \in [m]} \mu_i^* \nabla g_i(T(y^*)) = 0$. Therefore, it is clear that $(T(y*^*), \mu^*)$ is a KKT point for the original problem. Additionally, since the objective is the same for both optimization problems, a KKT point (y^*, μ^*) for the transformed problem has the same objective value as the equivalent KKT point $(T(y^*), \mu^*)$ for the original problem. So finally, if the KKT conditions are sufficient for global optimality in the original problem, they will be sufficient for global optimality in the transformed problem. ■

B.5. Proof of Theorem 5

Proof The argument is identical to that in the proof of Theorem 4, but we work with the regularized dual problem (6) and policy optimization problem (7). We use the same reparameterization T as defined in (17). ■