BATCH REINFORCEMENT LEARNING THROUGH CONTINUATION METHOD

Yijie Guo[1] **Shengyu Feng**[1] **Nicolas Le Roux**[2] **Ed Chi**[2] **Honglak Lee**[1,2] **Minmin Chen**[2]
[1]University of Michigan [2]Google AI
{guoyijie,shengyuf}@umich.edu {nlr,edchi,honglak,minminc}@google.com

ABSTRACT

Many real-world applications of reinforcement learning (RL) require the agent to learn from a fixed set of trajectories, without collecting new interactions. Policy optimization under this setting is extremely challenging as: 1) the geometry of the objective function is hard to optimize efficiently; 2) the shift of data distributions causes high noise in the value estimation. In this work, we propose a simple yet effective policy iteration approach to batch RL using global optimization techniques known as continuation. By constraining the difference between the learned policy and the behavior policy that generates the fixed trajectories, and continuously relaxing the constraint, our method 1) helps the agent escape local optima; 2) reduces the error in policy evaluation in the optimization procedure. We present results on a variety of control tasks, game environments and a recommendation task to empirically demonstrate the efficacy of our proposed method.

1 INTRODUCTION

While RL is fundamentally an *online* learning paradigm, many practical applications of RL algorithms, e.g., recommender systems [5, 7] or autonomous driving [36], fall under the *batch* RL setup. Under this setting, the agent is asked to learn its policy from a fixed set of interactions collected by a different (and possibly unknown) policy commonly referred to as the behavior policy, without the flexibility to gather new interactions. Realizing the interactive nature of online RL has been hindering its wider adoptions, researchers strive to bring these techniques offline [24, 11, 20, 23, 31, 12, 21, 2, 32, 8]. We focus on policy optimization under batch RL setup. As pointed out in [3, 26], even with access to the exact gradient, the loss surface of the objective function maximizing the expected return is difficult to optimize, leading to slow convergence. Chen et al. [8] show that the objective function of expected return exhibits sub-optimal plateaus and exponentially many local optima in the worst case. Batch setup makes the learning even harder as it adds large variance to the gradient estimate, especially when the learned policy differs from the behavior policy used to generate the fixed trajectories. Recent works propose to constrain the size of the policy update [27, 28] or the distance between the learned policy and the behavior policy [14, 21]. The strength of that constraint is a critical hyperparameter that can be hard to tune [28], as a loose constraint does not alleviate the distribution shift while a strict one results in conservative updates.

Here we propose to address the challenges using continuation methods [35, 6, 17]. Continuation methods attempt to solve the global optimization problem by progressively solving a sequence of new objectives that can be optimized more efficiently and then trace back the solutions to the original one. We change the objective function of policy optimization by including an additional term penalizing the KL divergence between the parameterized policy π_θ and the behavior policy. We then gradually decrease the weight of that penalty, eventually converging to optimizing the expected return. With this additional constraint, we benefit from more accurate policy evaluation in the early stage of training as the target policy is constrained to be close to the behavior policy. As training continues, we relax the constraint and allow for more aggressive improvement over the behavior policy as long as the policy evaluation is still stable and relatively reliable, i.e. with a small enough variance. By doing so, the proposed method exhaustively exploits the information in the collected trajectories while avoiding the overestimation of state-action pairs that lack support.

The contributions of this paper are as follows: (1) We propose a soft policy iteration approach to batch RL through the continuation method. (2) We theoretically verify that in the tabular setting with exact gradients, maximizing KL regularized expected return leads to faster convergence than optimizing the expected return alone. Also, our method converges to the globally optimal policy if there are sufficient data samples for accurate value estimation. (3) We demonstrate the effectiveness of our method in reducing errors in value estimation using visualization; (4) We empirically verify the advantages of our method over existing batch RL methods on various complex tasks.

2 Related Work

Batch Reinforcement Learning. Off-policy reinforcement learning has been extensively studied [11, 20, 30, 23, 31], with many works [12, 21, 2] focusing on variants of Q-learning. Fujimoto et al. [12], Kumar et al. [21] investigated the extrapolation error in batch RL resulting from the mismatch of state-action visitation distribution between the fixed dataset and the current policy, and proposed to address it by constraining the action distribution of the current policy from deviating much from the training dataset distribution. Recent works [29, 33] studied policy iteration under batch RL. The Q function is estimated in the policy evaluation step without special treatment while the policy updates are regularized to remain close to the prior policy with a *fixed* constraint. To further reduce uncertainty in Q learning, an ensemble of Q networks [21, 29] and distributional Q-function [2, 33] are introduced for the value estimation. [34, 18] use the KL divergence between the the target policy and the behavior policy as a regularization term in the policy update and/or value estimation. The constraint is controlled by a fixed weight of the KL regularization or a fixed threshold for the KL divergence. While all of these works apply a *fixed* constraint determined by a sensitive hyper-parameter to control the distance between the behavior/prior policy and the target policy, we focus on *gradually relaxed* constraints.

Constrained Policy Updates. Several works [27, 1, 15] studied constrained policy updates in online settings. Kakade & Langford [19] show that large policy updates can be destructive, and propose a conservative policy iteration algorithm to find an approximately optimal policy. Schulman et al. [27] constrain the KL divergence between the old policy and new policy to guarantee policy improvement in each update. Grau-Moya et al. [15] force the policy to stay close to a learned prior distribution over actions, deriving a mutual-information regularization between state and action. Cheng et al. [9] propose to regularize in the function space. Again these methods focused on a *fixed* constraint while we are interested in continuing relaxing the constraint to maximize the expected return eventually. Also none of these methods have been extensively tested for batch RL with fixed training data.

Continuation Method. Continuation method [35] is a global optimization technique. The main idea is to transform a nonlinear and highly non-convex objective function to a series of smoother and easier to optimize objective functions. The optimization procedure is successively applied to the new functions that are progressively more complex and closer to the original non-context problem, to trace their solutions back to the original objective function. Chapelle et al. [6] use the continuation method to optimize the objective function of semi-supervised SVMs and reach lower test error compared with algorithms directly minimizing the original objective. Hale et al. [17] apply the continuation method to l1-regularized problems and demonstrate better performance for compressed sensing problems. Inspired by prior works, we employ the continuation method to transform the objective of batch RL problems by adding regularization. We gradually decrease the regularization weight to trace the solution back to the original problem.

3 Method

In classical RL, an agent interacts with the environment while updating its policy. At each step t, the agent observes a state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$ according to its policy to receive a reward $r_t = r(s_t, a_t) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and transitions to the next state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. The state value of a policy π is $V^\pi(s) = \mathbb{E}_{s_0=s, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$. $\gamma \in [0, 1]$ is the discounting factor. At each step, the agent update its policy so that the expected return $V^\pi(\rho) = \mathbb{E}_{s\sim\rho}[V^\pi(s)]$ (where ρ is the initial state distribution) is maximized when acting accordingly.

In batch RL, the agent is not allowed to interact with the environment during learning. Instead it has access to a fixed set of trajectories sampled from the environment according to some behavior policy[1]. A trajectory $\{(s_0, a_0, r_0), (s_1, a_1, r_1), \cdots, (s_T, a_T, r_T)\}$ is generated by sampling s_0 from the initial state distribution ρ and the action $a_t \sim \beta(\cdot|s_t)$ at the state s_t and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$ for each step $t \in [0, 1, \cdots, T]$. The length T can vary between trajectories. We then convert the trajectories to a dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$, where s'_i is the next state after s_i in a trajectory.

The goal is to learn a parameterized policy π_θ with the provided dataset to maximize the expected return $V^\pi(\rho)$. In Sec. 3.1, we will first introduce a new objective function $\tilde{V}^{\pi,\tau}(\rho)$, i.e. the expected return of policy π with KL regularization weight τ. With exact gradients, $\tilde{V}^{\pi,\tau}(\rho)$ can be optimized more efficiently than the original objective $V^\pi(\rho)$. With the continuation method, solving a sequence of optimization problems for $\tilde{V}^{\pi,\tau}(\rho)$ with decaying value of τ converges toward optimizing $V^\pi(\rho)$

[1] If the behavior policy is not known in advance, it can be fitted from the data [30, 7].

and makes the optimization easier. In Sec. 3.2, we derive soft policy iteration with KL regularization to optimize $\tilde{V}^{\pi,\tau}(\rho)$, without the assumption of exact gradients. Finally, in Sec. 3.3, we propose a practical batch RL algorithm with value estimation for target policy based on this theory.

3.1 Optimizing Expected Return with KL regularization

In batch RL, the distribution of the trajectories generated by the behavior policy can be very different from that of the learned policy. We thus restrict the learned policy to stay close to the behavior policy via the regularization of KL divergence. Define the soft state value of policy π as

$$\tilde{V}^{\pi,\tau}(s) = \mathbb{E}_{s_0=s,a_t\sim\pi(\cdot|s_t),s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t) - \tau\log\frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}\right)\right], \quad (1)$$

where the temperature parameter τ controls the deviation from β, the new objective function becomes $\tilde{V}^{\pi,\tau}(\rho) = \mathbb{E}_{s\sim\rho}[\tilde{V}^{\pi,\tau}(s)]$. This KL regularized objective differs from the original objective $V^{\pi}(\rho)$, which however can be recovered as $\tau \to 0$.

As pointed out in [3], even with exact gradients, the objective function $V^{\pi}(\rho)$ is still difficult to optimize due to its highly non-smooth landscape. Mei et al. [26] further prove that, in a tabular setting with softmax parameterized policy and exact gradients, the vanilla policy gradient method (i.e. directly maximizing $V^{\pi}(\rho)$ with gradient descent) converges to the global optimal policy at a convergence rate $\mathcal{O}(1/t)$, while the entropy-regularized policy gradient enjoys a significantly faster linear convergence rate $O(e^{-t})$. Motivated by this line of work, we investigate the convergence of optimizing $\tilde{V}^{\pi,\tau}(\rho)$ with the exact gradient descent and compare it with the vanilla policy gradient method. We study the smoothness and Łojasiewicz inequality for the function $\tilde{V}^{\pi,\tau}(\rho)$ to prove the convergence rate, similar to [26]. The detailed proofs of the following theorems are in the appendix.

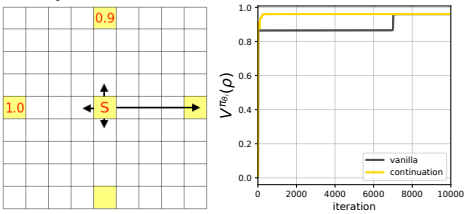
Theorem 1. *In the tabular setting with softmax parameterized policy π_θ, maximizing $\tilde{V}^{\pi,\tau}(\rho)$ using policy gradient with the learning rate $\eta = \frac{(1-\gamma)^3}{(8M+\tau(4+8\log A))}$, for all $t > 1$, we have*

$$\tilde{V}^{\pi^*_\tau,\tau}(\rho) - \tilde{V}^{\pi_{\theta_t},\tau}(\rho) \le C\cdot e^{-C_\tau(t-1)}\cdot\frac{M+\tau\log A}{(1-\gamma)^2}$$

where π^*_τ is the optimal policy maximizing $\tilde{V}^{\pi,\tau}(\rho)$, M is the bound of the absolute value of $r(s,a)+\tau\log\beta(a|s)$, A is the size of action space, S is the size of state space, $C_\tau \propto \frac{(1-\gamma)^4}{(8M/\tau+4+8\log A)\cdot S}$, and C is a constant independent with t and τ.

Theorem 1 states that KL regularized expected return can be optimized with a convergence rate $\mathcal{O}(e^{-t})$ rather than the $\mathcal{O}(1/t)$, the convergence rate of vanilla policy gradient for expected return alone. The faster convergence inspires us to optimize $\tilde{V}^{\pi,\tau}(\rho)$ to reach policy π^*_τ, then use π^*_τ as initialization, gradually decrease the temperature τ towards 0, and eventually move from π^*_τ to $\pi^* = \arg\max_\pi V^{\pi}(\rho)$. With a reasonable value of τ, we enjoy a linear convergence rate toward π^*_τ from the randomly initialized policy π_θ. As τ decreases, π^*_τ gets closer to π^*. The final optimization of $V^{\pi_\theta}(\rho)$ from π^*_τ can be much faster than from a randomly initialized π_θ.

We construct a toy example to illustrate this motivation. In the grid world (Fig. 1a), the start state, annotated with 'S', is in the center and the terminal states are marked in yellow. There are only two states with positive rewards (0.9 and 1). There are four actions {up, down, left, right}. A badly initialized policy π_{θ_0} is shown as arrows in Fig. 1a). The initialization results in a poor policy, having high tendency to go right toward a terminal state with zero reward. The vanilla policy gradient method (i.e. maximizing $V^{\pi}(\rho)$ with true gradient) starting from this initial point takes more than 7000 iterations to escape a sub-optimal solution (Fig. 1b). In contrast, we escape the sub-optimal solution much faster when applying the continuation method to update the policy with the gradients of $\tilde{V}^{\pi,\tau}(\rho)$, where the behavior policy $\beta(\cdot|s) = [u_1, u_2, u_3, u_4]$ with $u_i, i = 1,2,3,4$ randomly sampled from $\mathcal{U}[0,1]$ and normalized for each state s. In Fig. 1b, as we decrease τ, the value of learned policy π_{θ_i} for each iteration i quickly converges to the optimal value. In other words, optimizing a sequence of objective functions $\tilde{V}^{\pi,\tau}(\rho)$ can reach the optimal solution for $V^{\pi}(\rho)$ significantly faster.



(a) (b)

Figure 1: (a) A grid world with sparse rewards. (b) Learning curve of the value of parameterized policy π_{θ_i}. We conduct a hyper-parameter search for the learning rate 5,1,0.5,0.1,0.05,0.01,0.005,0.001 and report the best performance for each method.

Algorithm 1 Soft Policy Iteration through Continuation Method

Initialize: actor network π_θ, ensemble critic network $\{Q_{\phi^{(1)}}, Q_{\phi^{(2)}}, \cdots, Q_{\phi^{(K)}}\}$, behavior policy network β_ψ, penalty coefficient τ, decay rate λ, number of iterations I for each τ
Input: training dataset $D = \{(s_i, a_i, r_i, s'_i)\}_{i=0}^N$
for update $j = 0, 1, \cdots$ **do**
  Sample batch of data $\{(s_i, a_i, r_i)\}_{i=1}^B$ from D
  # Learn the behavior policy with behavior cloning objective
  Update ψ to maximize $\frac{1}{B}\sum_{i=1}^B \log \beta_\psi(a_i|s_i)$
  # Train the critic network
  Update $\phi^{(k)}$ to minimize the temporal difference $\frac{1}{B}\sum_{i=1}^B \left(r_i + \gamma V(s'_i) - Q_{\phi^{(k)}}(s_i, a_i)\right)^2$
  where $V(s) = \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}(Q_{\phi^{(k)}}(s,a)) - \tau KL(\pi_\theta(\cdot|s)|\beta_\psi(\cdot|s))$
  # Train the actor network
  Update θ to maximize $\frac{1}{B}\sum_{i=1}^B \left[\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{a\sim\pi_\theta(\cdot|s_i)}(Q_{\phi^{(k)}}(s_i,a)) - \tau KL(\pi_\theta(\cdot|s_i)|\beta_\psi(\cdot|s_i))\right]$
  # Decay the weight of KL regularization τ for every I updates
  if $j \mod I = 0$ **then**
    $\tau \leftarrow \tau * \lambda$
  end if
end for

3.2 Soft Policy Iteration with KL regularization

As explained in the previous section, we focus on the new objective function $\tilde{V}^{\pi,\tau}(\rho)$, which can be optimized more efficiently, and use continuation method to relax toward optimizing $V^\pi(\rho)$. Batch RL adds the complexity of estimating the gradient of $\tilde{V}^{\pi,\tau}(\rho)$ with respect to π from a fixed set of trajectories. We propose to adapt soft actor-critic[16], a general algorithm to learn optimal maximum entropy policies in batch RL for our use case. We change the entropy regularization to KL regularization and derive the soft policy iteration to learn KL regularized optimal policy. For a policy π and temperature τ, the soft state value is defined in Eq. 1 and soft Q function is defined as:

$$\tilde{Q}^{\pi,\tau}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\tilde{V}^{\pi,\tau}(s') \tag{2}$$

In the step of soft policy evaluation, we aim to compute the value of policy π according to the minimum KL divergence objective $\tilde{V}^{\pi,\tau}(\rho) = \mathbb{E}_{s\sim\rho}[\tilde{V}^{\pi,\tau}(s)]$. According to Lemma 1 in Appendix, the soft Q value can be computed by repeatedly applying the soft bellman backup operator.

$\mathcal{T}^{\pi,\tau}Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}(V(s'))$, where $V(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q(s,a) - \tau \log \frac{\pi(a|s)}{\beta(a|s)}\right]$.

In the step of policy improvement, we maximize the expected return based on Q-value evaluation with the KL divergence regularization. The following policy update can be guaranteed to result in an improved policy in terms of its soft value (Lemma 2 in Appendix).

$$\pi_{new}(\cdot|s) = \arg\max_{\pi\in\Pi}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\left(\tilde{Q}^{\pi_{old},\tau}(s,a)\right) - \tau KL(\pi(\cdot|s)|\beta(\cdot|s))\right] \tag{3}$$

$$\text{where} \quad KL(\pi(\cdot|s)|\beta(\cdot|s)) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\log\frac{\pi(a|s)}{\beta(a|s)}\right]$$

The soft policy iteration algorithm alternates between the soft policy evaluation and soft policy improvement, and it will provably converge to the optimal policy maximizing the objective $\tilde{V}^{\pi,\tau}(\rho)$.

Theorem 2. *Repeated application of soft policy evaluation and soft policy improvement converges to a policy π^*_τ such that $\tilde{Q}^{\pi^*_\tau,\tau}(s,a) \geq \tilde{Q}^{\pi,\tau}(s,a)$ for any $\pi \in \Pi$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$.*

The soft policy iteration finds a policy π^*_τ with optimal soft Q value for each state-action pair and hence gets the optimal value of $\tilde{V}^{\pi,\tau}(\rho)$. Here we propose to use the soft policy iteration to solve objectives $\tilde{V}^{\pi,\tau}(\rho)$ with decreasing value of τ and move back to the objective $V^\pi(\rho)$ as $\tau = 0$. The method is guaranteed to asymptotically converge to the optimal policy π^* for the objective $V^\pi(\rho)$.

Theorem 3. *Let $\pi^*_\tau(a|s)$ be the optimal policy from soft policy iteration with fixed temperature τ. We have $\pi^*_\tau(a|s) \propto \exp\left(\frac{\tilde{Q}^{\pi^*_\tau,\tau}(s,a)}{\tau}\right)\beta(a|s)$. As $\tau \to 0$, $\pi^*_\tau(a|s)$ will take the optimal action a^* with optimal Q value for state s.*

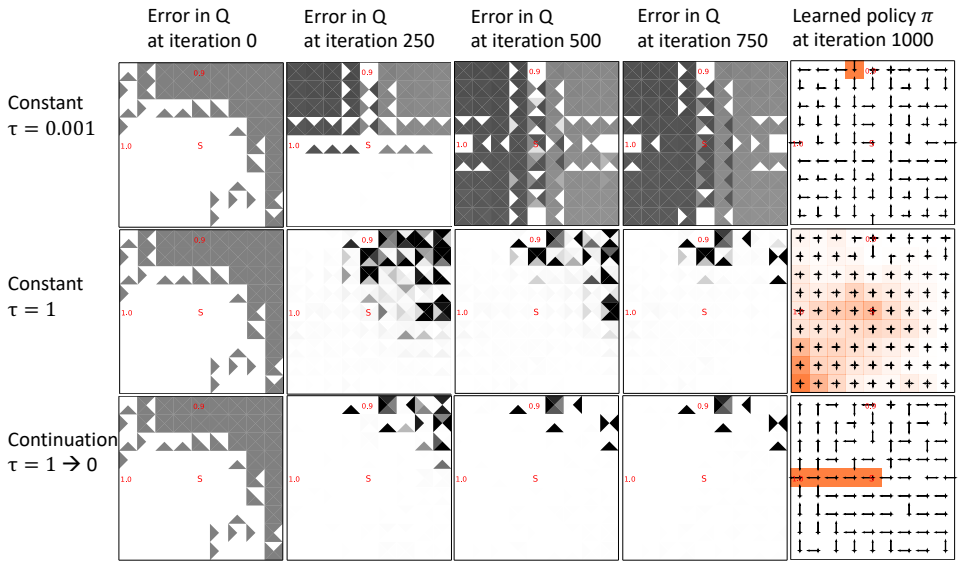


Figure 2: Visualization of the error in soft Q value estimation and quality of the learned policy. In the first four columns, triangles represent the error for actions that move in different directions. Darker color indicates higher error. To investigate the performance of the learned policy $\pi_{\theta_{1000}}$, the length of arrows represents the probability of taking each actions in each states. We run $\pi_{\theta_{1000}}$ in the grid world and visualize the visitation count in the last column (heatmap). Darker color means more visitation.

3.3 Error in Value Estimate

In the previous section, we show that the soft policy iteration with the continuation method provably converges to the global optimal policy maximizing expected return. However, in batch RL with a fixed dataset and limited samples, we cannot perform the soft policy iteration with KL regularization in its exact form. Specifically, in the policy evaluation step, when the learned policy π deviates for the behavior policy β, and chooses the state-action pair (s, a) rarely visited by β, the estimation of target $r(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}(V(s'))$ can be very noisy. The error in the value estimate $Q(s, a)$ will be further propagated to other state-action pairs through the bellman update. Finally, inaccurate value estimation will cause errors in the policy improvement step, resulting in a worse policy. On the other hand, if we constrain the learned policy π to be very close to the behavior policy β, we can expect the policy evaluation to be reliable and safely update the learned policy. The tight constraint however prevents π to be much better than β due to the conservative update.

On the grid world, we study this problem of value estimation with different values of τ. Figure 2 visualizes the propagation of Q value estimation errors and the learned policies. We assume a mediocre behavior policy tending to move left and down. For the rarely visited states in the upper right part of the grid, there are errors in the value estimation of $\tilde{Q}^{\pi,\tau}(s, a)$, i.e. $|Q(s, a) - \tilde{Q}^{\pi,\tau}(s, a)| > 0$ where $Q(s, a)$ is the Q value we learn during training and $\tilde{Q}^{\pi,\tau}(s, a)$ is the ground truth soft Q value. Because the bad initial policy (Fig. 1a) tends to move towards the right part, without a strong KL regularization, the policy evaluation can be problematic due to the errors of value estimation in the right part of the grid world. In Fig. 2, the first row shows that errors even propagate to the frequently visited states by the behavior policy. On the other hand, when we set a large value of $\tau = 1$ (second row), the error $|Q(s, a) - \tilde{Q}^{\pi,\tau}(s, a)|$ is smaller. Yet the performance of the learned policy is not much better than the behavior policy. Our continuation method gradually moves the policy update between these two spectra. The value estimation benefits from the gradually relaxed KL regularization and the errors remain small. The last column of Fig. 2 visualizes the learned policy in these methods. With constant $\tau = 0.001$, the wrong value estimates in some states mislead the agent. It fails to visit any terminal state and gets stuck at the state in dark orange. With constant $\tau = 1$, the tight constraint of KL divergence makes the learned policy close to the behavior policy, mostly visiting the left bottom part of the environment. With continuation method, the agent learns to always take the optimal path moving left directly and obtains the highest expected return. More details of this example are provided in the appendix.

In the toy example, gradually relaxing KL regularization towards zero alleviates the propagation of errors in the soft Q estimate and helps the agent converge to the optimal policy. In more complicated domains, we find that as τ decays close to 0, the policy evalua-

tion is still erroneous. To mitigate this issue, we introduce an ensemble of critic networks $\{Q_{\phi^{(1)}}, Q_{\phi^{(2)}}, \cdots, Q_{\phi^{(K)}}\}$ to approximate the soft Q value, and monitor the variance of value estimation in different critic networks to measure the uncertainty. Given a batch of data samples $\{s_i\}_{i=1}^{B} \subset \mathcal{D}$, $var(Q^\pi) = \frac{1}{B}\sum_{i=1}^{B} \mathbb{E}_{a\sim\pi(\cdot|s_i)} var(Q_{\phi^{(1)}}(s_i,a), Q_{\phi^{(2)}}(s_i,a), \cdots, Q_{\phi^{(k)}}(s_i,a))$ indicates whether the current policy π tends to take actions with highly noisy value estimation.

Our method is summarized in Algorithm 1. Instead of running the soft policy evaluation and policy improvement until convergence, we alternate between optimizing the critic network and actor network with stochastic gradient descent. We set τ to large value initially and let the KL divergence term dominate the objective, thus performing behavior cloning. We record a moving average of the Q value estimation variance $var(Q^{\pi,\tau_0})$ over 1000 updates at the end of the phase. After that, we decay the temperature gradually with $= 0.9$ every I steps. When the moving average of the Q value estimation variance $var(Q^{\pi,\tau})$ is large compared with the initial value $var(Q^{\pi,\tau_0})$ (i.e. at the end of behavior cloning), we no longer trust the value estimate under the current temperature τ and take the policy checkpointed before the temperature decays to this τ as our solution.

4 Experiments

4.1 Mujoco

We evaluate our method with several baselines on continuous control tasks. We train a Proximal Policy Optimization agent [28] with entropy regularization for 1000 million steps in the environments. We parameterize the policy using Gaussian policies where the mean is a linear function of the agent's state $\theta^T s$ and variance is an identity matrix to keep the policy simple as introduced in [3]. To generate training datasets $\mathcal{D}$ with varying quality, we construct the behavior policy by mixing the well-trained policy $\mathcal{N}(\theta_{opt}^T s, 0.5\mathbb{I})$, i.e. checkpoint with the highest score during training, and a poor policy $\mathcal{N}(\theta_0^T s, 0.5\mathbb{I})$, i.e. checkpoint at the beginning of the training, with the weight α. Then the behavior policy $\beta(\cdot|s)$ is $\mathcal{N}(((1-\alpha)\theta_{opt} + \alpha\theta_0)^T s, 0.5\mathbb{I})$. We generate trajectories and store a total of one million data samples from the mixed behavior for different values of the coefficient α.

The architecture of the target policy is the same as the behavior policy. We consider four baseline approaches: BCQ [14], BEAR [21], ABM+SVG [29], CRR [33], CQL [22], BRAC [34]. For a fair comparison, the architectures of the critic network and the policy network are the same in the baselines and our method, except BCQ which has no policy network. To evaluate and compare the methods, we run the learned policy in the environments for 100 episodes and report the average episode reward in Fig 3. As for the continuation method, we report the score of policy from checkpointed last with reasonable value estimation variance, as explained in Section 3.3. For the baselines, we report the score of the final policy when we terminate the training at 1.5M updates.

	α	BCQ	BEAR	ABM	CRR	CQL	BRAC	Ours
Hopper	0.2	1908.0 ±327.0	1733.1 ±138.0	1814.2 ±176.9	1861.5 ±112.8	1962.6 ±194.1	523.0 ±543.1	**2097.0 ±112.4**
	0.4	876.0 ±462.2	1132.0 ±316.6	1113.0 ±201.3	1281.0 ±218.2	1262.0±116.7	2.4 ±2.6	**1569.3 ±180.0**
	0.6	429.9 ±198.1	902.1 ±28.0	1402.6 ±350.5	791.1 ±87.0	609.1 ±35.1	459.9 ±93.0	**1648.0 ±340.2**
	0.8	**450.3 ±174.2**	4.2 ±4.3	3.6 ±0.0	304.2 ±36.3	295.7 ±41.9	185.1 ±143.3	226.3 ±22.0
Half Cheetah	0.2	1765.8 ±41.2	2130.4 ±23.2	2168.3 ±26.1	2154.1 ±6.6	2105.7 ±37.2	**2248.5 ±22.8**	2144.8 ±34.5
	0.4	1336.4 ±35.6	1826.4 ±24.6	1914.7 ±13.2	1860.1 ±38.2	1811.9 ±60.7	**1921.6 ±44.3**	1823.9 ±21.8
	0.6	513.6 ±35.6	1177.9 ±22.5	1457.9 ±36.4	1161.7 ±11.7	1156.6 ±43.3	1410.3 ±65.1	**1487.4±37.1**
	0.8	179.0 ±14.1	**672.9 ±13.9**	600.4 ±8.6	572.8 ±14.3	605.8 ±44.5	581.5 ±31.9	546.6± 16.2
Walker	0.2	1400.2 ±30.9	1421.1 ±37.9	1405.2 ±57.3	**1442.9 ±19.5**	1385.2 ±73.4	572.5 ±721.8	1441.3 ±45.9
	0.4	965.9 ±69.0	1201.1 ±57.3	**1259.3 ±27.7**	1233.5 ±35.8	979.6 ±83.5	1078.9 ±547.4	1222.7 ±26.6
	0.6	266.3 ±56.9	472.3 ±38.4	664.9 ±37.5	529.9 ±46.8	218.8 ±16.9	69.7 ±88.4	**852.7 ±240.5**
	0.8	3.5 ±5.0	10.5 ±1.0	5.9 ±0.6	**10.7 ±0.3**	3.3 ±6.5	5.7 ±3.4	6.7± 6.1

Table 1: Results on Mujoco. We show the average and standard deviation of the scores in 5 independent runs.

Tab. 1 shows that our method outperforms all the baselines on 5 settings. On the dataset with relatively reasonable quality (i.e. $\alpha = 0.2, 0.4, 0.6$), ours performs comparable or better than the baselines. With $\alpha = 0.2$, i.e., close to optimal behavior policy, all the methods perform similarly and one can achieve a good return by simply cloning the behavior policy. With $\alpha = 0.8$, i.e., low-quality behavior policy, there are few good trajectories in the dataset for any methods to learn. The advantage of our method is most obvious when $\alpha = 0.6$ (Fig. 3), as the dataset contains trajectories of both

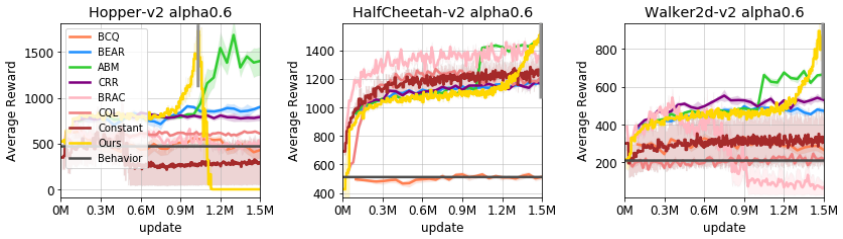


Figure 3: Learning curves of average reward over 5 runs on Mujoco tasks with $\alpha = 0.6$. The shaded area with light color indicates the standard deviation of the reward. The gray vertical lines on the yellow curves indicate where we take the checkpointed policy, according to the measure of Q value variance, as our final solution.

	Amidar	Asterix	Breakout	Enduro	MsPacman	Qbert	Seaquest	SpaceInvaders
BCQ	154.4 ±11.0	2466.7 ±273.4	203.8 ±19.6	604.7 ±39.7	2299.7 ±150.2	4088.3 ±332.8	4420.0 ±548.7	726.5 ±58.7
REM	32.0 ±3.5	741.4 ±236.7	3.1 ±0.0	244.2 ±19.8	1997.4 ±6.4	2062.9 ±511.5	474.0 ±61.9	678.4 ±41.3
CQL	145.0 ±1.6	2618.7 ±102.1	**253.7** **±14.4**	206.6 ±5.8	2234.6 ±203.2	4094.7 ±74.1	4652.6 ±2017.1	493.5 ±11.4
Ours	**174.5** **±7.1**	**3476.7** **±229.0**	199.0 ±32.0	**922.9** **±31.7**	**2494.0** **±301.3**	**4732.5** **±172.5**	**9935.0** **±1175.9**	**1070.3** **±137.1**

Table 2: Results on Atari, the mean and standard deviation of scores achieved in 3 indenpendent runs.

high and low cumulative rewards. Our method can learn from the relatively large number of good trajectories and at the same time deviate from the behavior policy to avoid those bad trajectories and achieve higher rewards. In Fig. 3, 'Constant' is the method of optimizing KL regularized expected reward with constant value of τ. We search several values of τ and report the best result. We can see that gradually relaxing constraint performs better than the fixed constraint. In Fig. 3(left), as τ decays to close to 0, the learned policy can degrade due to errors in the Q estimation, the stopping condition explained in Section 3.3 is however able to identify a good policy before the degenerating point. More experimental details are in the Appendix.

4.2 Atari

We further study our method on several Atari games from the Arcade Learning Environment (ALE) [4]. The rich observation space requires more complicated policies and makes policy optimization even more challenging. We focus on eight games and generate the datasets as discussed in Fujimoto et al. [13]. We use a mediocre DQN agent, trained online for 10 million timesteps (40 million frames). The performance of the DQN agent is shown as 'Online DQN' in Fig. 4. We add exploratory noise on the DQN agent (at 10 million timesteps) to gather a new set of 10 million transitions, similar to [13]. The line "Behavior" in Fig. 4 shows the average of trajectory reward in the dataset $\mathcal{D}$. The dataset $\mathcal{D}$ is used to train each off-policy agent. We compare with BCQ[13] and REM[2] because they are recently proposed offline RL algorithms and work well on Atari domain. For evaluation, we run 10 episodes on the Atari games with the learned policies and record the average episode reward (Fig. 4). Tab. 2 summarizes the performance of BCQ, REM and CQL after 6M updates. For our method, we report the score before the variance of Q estimate becomes too high.

Our approach achieves higher scores than the baselines on 7 out of 8 games, and perform comparably on the other one. Agarwal et al. [2] reports that REM performs well in the dataset consisting of the entire replay experiences collected in the online training of the DQN agent for 50M timesteps (200M frames). We hypothesize that learning on the entire replay experience makes the setup easier as the training dataset contains more exploratory and higher quality trajectories. With the dataset of much smaller size and worse quality, REM performs poorly in this single behavioral policy setting. We use the same architecture of the critic network for both our method and BCQ with ensemble of 4 Q networks. As mentioned in [13], BCQ only matches the performance of the online DQN on most games. In contrast, ours is able to outperforms online DQN significantly on several games.

4.3 Recommender

We also showcase our proposed method for building a softmax recommender agent. We use a publicly available dataset MovieLens-1M, a popular benchmark for recommender system. There are 1 million ratings of 3,900 movies (with the title and genre features) from 6,040 users (with demographic features). The problem of recommending movies for each user can be converted to a

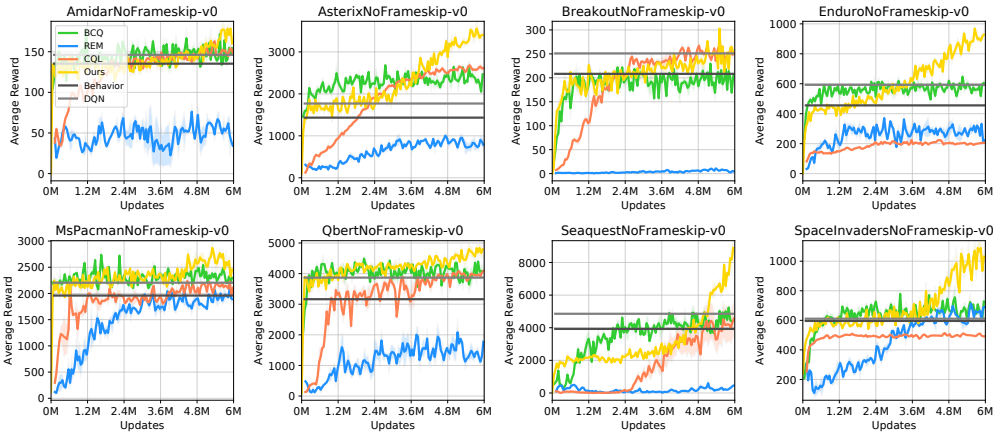


Figure 4: Learning curves of average reward over 3 runs on Atari games.

	data with avg. reward 0.53		data with avg. reward 0.65		data avg. reward 0.80	
sample size	30,000	60,000	30,000	60,000	30,000	60,000
Cross-Entropy	73.2±0.02	72.6±0.02	80.0±0.01	79.8±0.01	85.7±0.01	84.8±0.01
IPS	78.0±0.03	79.9±0.02	87.2±0.02	88.3±0.01	87.8 ± 0.01	89.4 ± 0.01
Ours	**82.6±0.01**	**83.9±0.01**	**89.8±0.01**	**90.5±0.00**	**90.1±0.01**	**91.9±0.00**

Table 3: Results on MovieLens dataset. The number is precision at 10, i.e., the percent of recommended movies getting positive feedback when we use the well-trained policy to recommend top-10 movies to the test users. We report the mean and standard deviation of the precision over 20 independent runs,

contextual bandit problem, where we aim to learn a target policy $\pi_\theta(a|s)$ selecting the proper action (movie) a for each state (user) s to get a high reward (rating) r in a single step. The ratings are converted to binary reward using a cutoff of 4. To evaluate whether a learned target policy works well, ideally we should run the learned policy in real recommendation environments. However, such environments for online test are rarely publicly available. Thus, we use the online simulation method. We train a simulator to predict the immediate binary feedback from user and movie features, and the well-trained simulator can serve as a proxy of the real online environments, because it outputs the feedback for any user-movie pair. Similar to [25], we train the simulator with all records of logged feedback in MovieLens-1M dataset. The behavior policy is trained with partial data in MovieLens-1M. We then construct the bandit datasets $\mathcal{D} = \{s_i, a_i, r_i, \beta(a_i|s_i)\}_{i=1}^N$ of different size and quality, by using different behavior policies β to select movies a_i for users s_i and getting the binary feedback r_i from the well-trained simulator. We train offline RL agents on the generated dataset $\mathcal{D}$ and use the simulator to evaluate the learned policies on a held-out test set of users.

We compare our method with two baselines, as they are commonly used in current industrial recommender systems [10, 7]. (1) Cross-Entropy: a supervised learning method for the softmax recommender where the learning objective is the cross-entropy loss $J_{CE}(\theta) = -\frac{1}{N}\sum_{i=1}^N r_i \log \pi_\theta(a_i|s_i)$ (2) IPS: the off-policy policy gradient method introduced in [7] with the learning objective $J_{IPS}(\theta) = -\frac{1}{N}\sum_{i=1}^N \frac{sg(\pi_\theta(a_i|s_i))}{\beta(s_i,a_i)} r_i \log \pi_\theta(a_i|s_i)$ where sg indicates a stop-gradient operation. $J_{IPS}(\theta)$ produces the same gradient as that of the function $\frac{1}{N}\sum_{i=1}^N r_i \frac{\pi_\theta(a_i|s_i)}{\beta(a_i|s_i)}$, the expected return with importance sampling. (3) Ours: in the bandit setting, we simply perform IPS with gradually decaying KL regularization since estimating the soft Q from bellman update is not needed. Tab. 3 clearly demonstrates the advantage of our proposed method over the baselines. IPS can be viewed as vanilla policy gradient with importance sampling to correct the distribution shift. Our method clearly outperforms it across datasets collected using different behavior policies.

5 Conclusion

We propose a simple yet effective approach, soft policy iteration algorithm through continuation method to alleviate the two challenges in policy optimization under batch reinforcement learning: (1) highly non-smooth objective function which is difficult to optimize (2) high variance in value estimates. We provide theoretical ground and visualization tools to help understand this technique. We demonstrate its efficacy on multiple complex tasks.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 22–31. JMLR. org, 2017.

[2] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. Striving for simplicity in off-policy deep reinforcement learning. *arXiv preprint arXiv:1907.04543*, 2019.

[3] Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pp. 151–160. PMLR, 2019.

[4] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[5] James Bennett, Stan Lanning, et al. The netflix prize. Citeseer, 2007.

[6] Olivier Chapelle, Mingmin Chi, and Alexander Zien. A continuation method for semi-supervised svms. In *Proceedings of the 23rd international conference on Machine learning*, pp. 185–192, 2006.

[7] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 456–464, 2019.

[8] Minmin Chen, Ramki Gummadi, Chris Harris, and Dale Schuurmans. Surrogate objectives for batch policy optimization in one-step decision making. In *Advances in Neural Information Processing Systems*, pp. 8825–8835, 2019.

[9] Richard Cheng, Abhinav Verma, Gabor Orosz, Swarat Chaudhuri, Yisong Yue, and Joel W Burdick. Control regularization for reduced variance reinforcement learning. *arXiv preprint arXiv:1905.05380*, 2019.

[10] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pp. 191–198, 2016.

[11] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

[12] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. *arXiv preprint arXiv:1812.02900*, 2018.

[13] Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. Benchmarking batch deep reinforcement learning algorithms. *arXiv preprint arXiv:1910.01708*, 2019.

[14] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062, 2019.

[15] Jordi Grau-Moya, Felix Leibfried, and Peter Vrancx. Soft q-learning with mutual-information regularization. 2018.

[16] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

[17] Elaine T Hale, Wotao Yin, and Yin Zhang. A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing. *CAAM TR07-07, Rice University*, 43:44, 2007.

[18] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

[19] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pp. 267–274, 2002.

[20] Shivaram Kalyanakrishnan and Peter Stone. Batch reinforcement learning in a complex domain. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pp. 94. ACM, 2007.

[21] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11784–11794, 2019.

[22] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

[23] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

[24] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[25] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H Chi. Off-policy learning in two-stage recommender systems. In *Proceedings of The Web Conference 2020*, pp. 463–473, 2020.

[26] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392*, 2020.

[27] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

[28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[29] Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

[30] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, pp. 2217–2225, 2010.

[31] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

[32] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

[33] Ziyu Wang, Alexander Novikov, Konrad Żołna, Jost Tobias Springenberg, Scott Reed, Bobak Shahriari, Noah Siegel, Josh Merel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *arXiv preprint arXiv:2006.15134*, 2020.

[34] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[35] Zhijun Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. *SIAM Journal on Optimization*, 6(3):748–768, 1996.

[36] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

APPENDIX:

A CONVERGENCE FOR SOFT POLICY ITERATION

Lemma 1. *Consider the soft Bellman backup operator* $\mathcal{T}^{\pi,\tau}Q(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}(V(s'))$ *where* $V(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q(s,a) - \tau\log\frac{\pi(a|s)}{\beta(a|s)}\right]$*, then the sequence* Q*,* $\mathcal{T}^{\pi,\tau}Q$*,* $(\mathcal{T}^{\pi,\tau})^2Q$*,* $\cdots$*,* $(\mathcal{T}^{\pi,\tau})^kQ$ *will converge to the soft value of* π *as* $k\to\infty$*.*

Proof. We prove the soft Bellman backup operator is a contraction. If we apply $\mathcal{T}^\pi$ to two different value functions Q and Q', the max norm distance $\|Q-Q'\| = \max_{s,a}|Q(s,a)-Q'(s,a)|$ shrinks.

$$\begin{aligned}
\|\mathcal{T}^\pi Q - \mathcal{T}^\pi Q'\| &= \|r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}\sim p}(V(s_{t+1})) - r(s_t,a_t) - \gamma\mathbb{E}_{s_{t+1}\sim p}(V'(s_{t+1}))\| \quad (4)\\
&= \gamma\|\mathbb{E}_{s_{t+1}\sim p}(V(s_{t+1})) - \mathbb{E}_{s_{t+1}\sim p}(V'(s_{t+1}))\|\\
&= \gamma\|\mathbb{E}_{s_{t+1}\sim p}\mathbb{E}_{a_{t+1}\sim\pi(\cdot|s_{t+1})}\left[Q(s_{t+1},a_{t+1}) - \tau\log\frac{\pi(a_{t+1}|s_{t+1})}{\beta(a_{t+1}|s_{t+1})}\right] -\\
&\quad - \mathbb{E}_{s_{t+1}\sim p}\mathbb{E}_{a_{t+1}\sim\pi(\cdot|s_{t+1})}\left[Q'(s_{t+1},a_{t+1}) - \tau\log\frac{\pi(a_{t+1}|s_{t+1})}{\beta(a_{t+1}|s_{t+1})}\right]\|\\
&= \gamma\|\mathbb{E}_{s_{t+1}\sim p}\mathbb{E}_{a_{t+1}\sim\pi(\cdot|s_{t+1})}\left[Q(s_{t+1},a_{t+1}) - Q'(s_{t+1},a_{t+1})\right]\|\\
&\le \gamma\max_{s_{t+1},a_{t+1}}|Q(s_{t+1},a_{t+1}) - Q'(s_{t+1},a_{t+1})|\\
&= \gamma\|Q-Q'\|
\end{aligned}$$

Therefore, the sequence $\mathcal{T}^{\pi,\tau}Q$, $(\mathcal{T}^{\pi,\tau})^2Q$, $\cdots$, $(\mathcal{T}^{\pi,\tau})^kQ$ only has one fixed point. Consider the soft Q value of policy π, $\tilde{Q}^{\pi,\tau}(s_t,a_t) = r_t + \mathbb{E}_{s_{t+1}\sim\mathcal{P}(\cdot|s_{t+1},a_{t+1}),a_{t+1}\sim\pi(\cdot|s_{t+1}),s_{t+l+1}\sim\mathcal{P}(\cdot|s_{t+l},a_{t+l})}\left[\sum_{l=1}^\infty\gamma^l(r(s_{t+l},a_{t+1}) - \tau\log\frac{\pi(a_{t+l}|s_{t+l})}{\beta(a_{t+l}|s_{t+l})})\right]$, we have $\mathcal{T}^{\pi,\tau}\tilde{Q}^{\pi,\tau}(s_t,a_t) = \tilde{Q}^{\pi,\tau}(s_t,a_t)$.

$$\|(\mathcal{T}^{\pi,\tau})^kQ - \tilde{Q}^{\pi,\tau}\| \le \gamma\|(\mathcal{T}^{\pi,\tau})^{k-1} - \tilde{Q}^{\pi,\tau}\| \le \gamma^2\|(\mathcal{T}^{\pi,\tau})^{k-2}Q - \tilde{Q}^{\pi,\tau}\| \le \cdots \le \gamma^k\|Q - \tilde{Q}^{\pi,\tau}\| \to 0$$

Thus, $(\mathcal{T}^{\pi,\tau})^kQ$ will converge to the fixed point $\tilde{Q}^{\pi,\tau}$.

□

Lemma 2. *let* $\pi_{old}\in\Pi$ *and* $\pi_{new}(\cdot|s) = \arg\max_{\pi\in\Pi}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\left(\tilde{Q}^{\pi_{old},\tau}(s,a) - \tau\log\frac{\pi(a|s)}{\beta(a|s)}\right)\right]$*. Then* $\tilde{Q}^{\pi_{new},\tau}(s,a)\ge\tilde{Q}^{\pi_{old},\tau}(s,a)$ *for all* (s,a)*.*

Proof. Due to the definition of π_{new}, for any state s_t, we have
$\mathbb{E}_{a\sim\pi_{new}(\cdot|s_t)}\tilde{Q}^{\pi_{old},\tau}(s_t,a) - \tau KL(\pi_{new}(\cdot|s_t)|\beta(\cdot|s_t)) \ge \mathbb{E}_{a\sim\pi_{old}(\cdot|s_t)}\tilde{Q}^{\pi_{old},\tau}(s_t,a) - \tau KL(\pi_{old}(\cdot|s_t)|\beta(\cdot|s_t))$

$$\begin{aligned}
\tilde{Q}^{\pi_{old},\tau}(s_t,a_t) &= r_t + \gamma\mathbb{E}_{s_{t+1}\sim p}\left[\tilde{V}^{\pi_{old},\tau}(s_{t+1})\right] \quad (5)\\
&= r_t + \mathbb{E}_{s_{t+1}\sim p}\left[\mathbb{E}_{a_{t+1}\sim\pi_{old}(\cdot|s_{t+1})}\tilde{Q}^{\pi_{old},\tau} - \tau KL(\pi_{old}(\cdot|s_{t+1})|\beta(\cdot|s_{t+1}))\right]\\
&\le r_t + \mathbb{E}_{s_{t+1}\sim p}\left[\mathbb{E}_{a_{t+1}\sim\pi_{new}(\cdot|s_{t+1})}\tilde{Q}^{\pi_{old},\tau}(s_{t+1},a_{t+1}) - \tau KL(\pi_{new}(\cdot|s_{t+1})|\beta(\cdot|s_{t+1}))\right]\\
&\cdots\\
&\cdots\\
&\le \tilde{Q}^{\pi_{new},\tau}(s_t,a_t)
\end{aligned}$$

□

Theorem 2. *Repeated application of soft policy evaluation and soft policy improvement converges to a policy* π^*_τ *such that* $\tilde{Q}^{\pi^*_\tau,\tau}(s,a)\ge\tilde{Q}^{\pi,\tau}(s,a)$ *for any* $\pi\in\Pi$ *and* $(s,a)\in\mathcal{S}\times\mathcal{A}$*.*

Proof. Let π_i be the policy at iteration i. The sequence Q^{π_i} is monotonically increasing. Since Q^{π} is bounded above (because both reward and KL divergence are bounded), the sequence converges to some π^*. We will still need to show that π^* is indeed optimal. At convergence, it must be case that:

$$\pi^*(\cdot|s) = \arg\max_{\pi\in\Pi} \left[\mathbb{E}_{a\sim\pi(\cdot|s)} \left(\tilde{Q}^{\pi^*,\tau}(s,a) - \tau \log \frac{\pi(a|s)}{\beta(a|s)}\right)\right]$$

So for any $\pi \in \Pi$, we have

$$\mathbb{E}_{a\sim\pi^*(\cdot|s_t)}\tilde{Q}^{\pi^*,\tau}(s_t,a) - \tau KL(\pi^*(\cdot|s_t)|\beta(\cdot|s_t)) \geq \mathbb{E}_{a\sim\pi(\cdot|s_t)}\tilde{Q}^{\pi^*,\tau}(s_t,a) - \tau KL(\pi(\cdot|s_t)|\beta(\cdot|s_t))$$

$$\begin{aligned}
\tilde{Q}^{\pi^*,\tau}(s_t,a_t) &= r_t + \gamma\mathbb{E}_{s_{t+1}\sim p}\left[\tilde{V}^{\pi^*_\tau,\tau}(s_{t+1})\right] \\
&= r_t + \mathbb{E}_{s_{t+1}\sim p}\left[\mathbb{E}_{a_{t+1}\sim\pi^*(\cdot|s_{t+1})}\tilde{Q}^{\pi^*_\tau,\tau} - \tau KL(\pi^*(\cdot|s_{t+1})|\beta(\cdot|s_{t+1}))\right] \\
&\geq r_t + \mathbb{E}_{s_{t+1}\sim p}\left[\mathbb{E}_{a_{t+1}\sim\pi(\cdot|s_{t+1})}\tilde{Q}^{\pi^*,\tau}(s_{t+1},a_{t+1}) - \tau KL(\pi(\cdot|s_{t+1})|\beta(\cdot|s_{t+1}))\right] \\
&\cdots \\
&\cdots \\
&\geq \tilde{Q}^{\pi,\tau}(s_t,a_t)
\end{aligned} \tag{6}$$

We denote this optimal policy π^* as π^*_τ in the following. Consider the optimization problem with hard constraint:

$$\begin{aligned}
\max_{\pi(a_1|s),\pi(a_2|s),\cdots,\pi(a_{|A|}|s)} \quad & \sum_{i=1}^{|A|}\pi(a_i|s)\tilde{Q}^{\pi^*_\tau,\tau}(s,a) - \tau\sum_{i=1}^{|A|}\pi(a_i|s)\log\frac{\pi(a_i|s)}{\beta(a_i|s)} \\
\text{s.t.} \quad & \sum_{i=1}^{|A|}\pi(a_i|s) = 1
\end{aligned}$$

Due to KKT condition, let $\tilde{Q}^{\pi^*_\tau,\tau}(s,a) - \tau\log\frac{\pi(a_i|s)}{\beta(a_i|s)} - \tau - \lambda = 0$. We have $\pi^*_\tau(a|s) = \frac{\exp\left(\frac{\tilde{Q}^{\pi^*_\tau,\tau}(s,a)}{\tau}\right)\beta(a|s)}{\sum_{a_i}\exp\left(\frac{\tilde{Q}^{\pi^*_\tau,\tau}(s,a_i)}{\tau}\right)\beta(a_i|s)}$ □

Theorem 3. *Let $\pi^*_\tau(a|s)$ be the optimal policy from soft policy iteration with fixed temperature τ. We have $\pi^*_\tau(a|s) \propto \exp\left(\frac{\tilde{Q}^{\pi^*_\tau,\tau}(s,a)}{\tau}\right)\beta(a|s)$. As $\tau \to 0$, $\pi^*_\tau(a|s)$ will take the optimal action a^* with optimal Q value for state s.*

Assume there exists a set of optimal actions $X(s)$ for state s. For each action in $X(s)$, its soft Q value is the optimal for state s. To simplify the notation, in the following proof, we use Q^*_τ to replace $\tilde{Q}^{\pi^*_\tau,\tau}$.

Assume $a_j \in X(s)$, then $\pi^*_\tau(a_j|s) = \frac{1}{\sum_{a_i\in X(s)}\frac{\beta(a_i|s)}{\beta(a_j|s)} + \sum_{a_i\notin X(s)}\exp\left(\frac{Q^*_\tau(s,a_i)-Q^*_\tau(s,a_j)}{\tau}\right)\frac{\beta(a_i|s)}{\beta(a_j|s)}}$.

For the second term in the denominator, we show it converges to 0 as $\tau \to 0$.

$$0 \leq \sum_{a_i\notin X(s)}\exp\left(\frac{Q^*_\tau(s,a_i)-Q^*_\tau(s,a_j)}{\tau}\right)\frac{\beta(a_i|s)}{\beta(a_j|s)} \leq \sum_{a_i\notin X(s)}\exp\left(\frac{Q^*_\tau(s,a_{sub})-Q^*_\tau(s,a_{opt})}{\tau}\right)\frac{\beta(a_i|s)}{\beta(a_j|s)}$$

$$0 \leq \sum_{a_i\notin X(s)}\exp\left(\frac{Q^*_\tau(s,a_i)-Q^*_\tau(s,a_j)}{\tau}\right)\frac{\beta(a_i|s)}{\beta(a_j|s)} \leq \exp\left(\frac{Q^*_\tau(s,a_{sub})-Q^*_\tau(s,a_{opt})}{\tau}\right)\sum_{a_i\notin X(s)}\frac{\beta(a_i|s)}{\beta(a_j|s)}$$

As $\tau \to 0$, $\exp\left(\frac{Q^*_\tau(s,a_{sub})-Q^*_\tau(s,a_{opt})}{\tau}\right) \to 0$ and $\sum_{a_i\notin X(s)}\frac{\beta(a_i|s)}{\beta(a_j|s)}$ is a constant.

Thus, as $\tau \to 0$, $\pi^*_\tau(a_j|s) \to \frac{1}{\sum_{a_i\in X(s)}\frac{\beta(a_i|s)}{\beta(a_j|s)}} = \frac{\beta(a_j|s)}{\sum_{a_i\in X(s)}\beta(a_i|s)}$ if $a_j \in X(s)$.

Assume $a_j \notin X(s)$, then $\pi^*_\tau(a_j|s) = \frac{1}{\sum_{a_i\in X(s)}\exp\left(\frac{Q^*_\tau(s,a_{opt})-Q^*_\tau(s,a_j)}{\tau}\right)\frac{\beta(a_i|s)}{\beta(a_j|s)} + \sum_{a_i\notin X(s)}\exp\left(\frac{Q^*_\tau(s,a_i)-Q^*_\tau(s,a_j)}{\tau}\right)\frac{\beta(a_i|s)}{\beta(a_j|s)}}$.

Obviously, $\exp\left(\frac{Q^*_\tau(s,a_{opt})-Q^*_\tau(s,a_j)}{\tau}\right)\frac{\beta(a_i|s)}{\beta(a_j|s)} \rightarrow +\infty$ as $\tau \rightarrow 0$ and $\exp\left(\frac{Q^*_\tau(s,a_i)-Q^*_\tau(s,a_j)}{\tau}\right)\frac{\beta(a_i|s)}{\beta(a_j|s)} > 0$.

Thus, as $\tau \rightarrow 0$, $\pi^*_\tau(a_j|s) \rightarrow 0$ if $a_j \notin X(s)$.

B Convergence Rate of KL regularized Policy Gradient

Recall our definitions:

$$V^{\pi}(\rho) = \mathbb{E}_{s_0\sim\rho, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]. \tag{7}$$

Mei et al. [26] proved that policy gradient with a softmax parameterization and true gradient optimizes $V^{\pi}(\rho)$ at a $\mathcal{O}(1/t)$ convergence rate, while the entropy regularized objective $V^{\pi}(\rho) + \tau\mathbb{H}(\rho,\pi)$ with $\mathbb{H}(\rho,\pi) = \mathbb{E}_{s_0\sim\rho, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}(-\gamma^t\log\pi(a_t|s_t))\right]$ enjoys a significantly faster linear convergence rate $\mathcal{O}(e^{-t})$.

With a behavior policy β and the "temperature" $\tau > 0$ that determines the strength of the regularization, we define the value of the policy π with the KL divergence regularization as

$$\tilde{V}^{\pi,\tau}(\rho) = V^{\pi}(\rho) - \tau D_{KL}(\pi||\beta) = \mathbb{E}_{s_0\sim\rho, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t) - \tau\log\frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}\right)\right] \tag{8}$$

Similarly we prove policy gradient optimizes $\tilde{V}^{\pi,\tau}(\rho)$ at a $\mathcal{O}(e^{-t})$ convergence rate. Here we explain the sketch of the proof. $\tilde{V}^{\pi,\tau}(\rho)$ can be re-written as $\hat{V}^{\pi,\tau}(\rho) + \tau\mathbb{H}(\rho,\pi)$ where

$$\hat{V}^{\pi,\tau}(\rho) = \mathbb{E}_{s_0\sim\rho, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t) + \tau\log\beta(a_t|s_t))\right]. \tag{9}$$

$$\mathbb{H}(\rho,\pi) = \mathbb{E}_{s_0\sim\rho, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}-\gamma^t\log\pi(a_t|s_t)\right] \tag{10}$$

$\hat{V}^{\pi,\tau}(\rho)$ in eq.(8) is the same as $V^{\pi}(\rho)$ in eq.(7) if we replace $r(s_t,a_t)$ with $r(s_t,a_t)+\tau\log\beta(a_t|s_t)$. Mei et al. [26] assume that $r(s_t,a_t) \in [0,1]$. Here we assume $r(s_t,a_t)+\tau\log\beta(a_t|s_t) \in [-M,M]$ with a constant M, effectively requiring the behavior policy to span the entire state-action space. We can then adapt the proof in [26] for the new objective $\tilde{V}^{\pi,\tau}(\rho)$. We conclude that in t iterations, the learned policy π_{θ_t} is approaching the optimal policy π^*_τ for the new objective $\tilde{V}^{\pi,\tau}(\rho)$, satisfying $\tilde{V}^{\pi^*_\tau,\tau}(\rho) - \tilde{V}^{\pi_{\theta_t},\tau}(\rho) \leq \frac{1}{\exp\{C_\tau\Omega(1)t\}}\frac{1+\tau\log A}{(1-\gamma)^2}\|\frac{1}{\mu}\|_\infty$, where $C_\tau = \frac{(1-\gamma)^4}{(8M/\tau+4+8\log A)\cdot S}\cdot\min_s\mu(s)\cdot\|\frac{d_\mu^{\pi^*_\tau}}{\mu}\|_\infty^{-1}$ and μ is the initial state distribution used in the policy optimization algorithm.

To simplify the notation, in the following proofs, we use $\tilde{V}^{\pi}(\rho) = \tilde{V}^{\pi,\tau}(\rho)$ and $\hat{V}^{\pi}(\rho) = \hat{V}^{\pi,\tau}(\rho)$.

Assumption 1 (Bounded Reward and Behavior Probability). *$r(s,a) + \tau\log\beta(a|s) \in [-M,M]$, $\forall(s,a)$ where M is a positive constant.*

Definition 1 (Smoothness). *A function $f:\Theta\to\mathbb{R}$ is λ-smooth if for all $\theta,\theta'\in\Theta$,*

$$\left|f(\theta') - f(\theta) - \left\langle\frac{\partial f(\theta)}{\partial\theta}, \theta'-\theta\right\rangle\right| \leq \frac{\lambda}{2}\|\theta'-\theta\|_2^2$$

Lemma 3 (Monotone Increasing). *Assume a function $f:\Theta\to\mathbb{R}$ is λ-smooth, and it is updated with $\theta_{t+1} = \theta_t + \eta\frac{\partial f(\theta)}{\partial\theta}$ and $\eta = \frac{2}{\lambda}$, then the function $f(\theta)$ is monotone increasing.*

$$f(\theta_{t+1}) - f(\theta_t) - \left\langle\frac{\partial f(\theta)}{\partial\theta}, \theta_{t+1}-\theta_t\right\rangle \geq -\frac{\lambda}{2}\|\theta_{t+1}-\theta_t\|_2^2$$

$$f(\theta_{t+1}) - f(\theta_t) \geq \left\langle\frac{\partial f(\theta)}{\partial\theta}, \eta\frac{\partial f(\theta)}{\partial\theta}\right\rangle - \frac{\lambda\eta^2}{2}\|\frac{\partial f(\theta)}{\partial\theta}\|_2^2$$

$$= (1-\frac{\lambda\eta}{2})\eta\|\frac{\partial f(\theta)}{\partial\theta}\|_2^2 = 0$$

Definition 2. *Given a vector $\theta\in\mathbb{R}^{[K]}$ and the probability distribution $\pi_\theta = softmax(\theta) = \frac{exp\theta}{\sum_k exp\theta(k)}$, then $H(\pi_\theta)$ is the Jacobian of the $\theta\to\pi_\theta$ map: $\left(\frac{d\pi_\theta}{d\theta}\right)^T = H(\pi_\theta) = diag(\pi_\theta) - \pi_\theta\pi_\theta^T \in \mathbb{R}^{K\times K}$.*

Lemma 4 (Smoothness). *$\hat{V}^{\pi_\theta}(\rho)$ is $\frac{8M}{(1-\gamma)^3}$-smooth.*

Denote $\theta_\alpha = \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. For any $s \in S$,

$$\begin{aligned}
\sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}|_{\alpha=0} \right| &= \sum_a \left| \left\langle \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha}|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \\
&= \sum_a \left| \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta}, u \right\rangle \right| \\
&= \sum_a \left| \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \right\rangle \right| \left(\text{Because } \frac{\partial \pi_\theta(a|s)}{\partial \theta(s',\cdot)} = 0, \forall s' \neq s \right) \\
&= \sum_a \pi_\theta(a|s) \cdot \left| u(s,a) - \pi_\theta(\cdot|s)^T u(s,\cdot) \right| \text{ (Because softmax parameterization)} \\
&\leq \max_a |u(s,a)| + \left| \pi_\theta(\cdot|s)^T u(s,\cdot) \right| \\
&\leq 2\|u\|_2 \qquad (11)
\end{aligned}$$

$$\begin{aligned}
\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}|_{\alpha=0} \right| &= \sum_a \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \{ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \}|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \\
&= \sum_a \left| \left\langle \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha^2}|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \\
&= \sum_a \left| \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \right| \\
&= \sum_a \left| \sum_{i=1}^A \sum_{j=1}^A S_{i,j} u(s,i) u(s,j) \right| (S_{i,j} \text{ is the i,j-th element of } \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)}) \\
&= \sum_a \pi_\theta(a|s) | u(s,a)^2 - 2u(s,a) \pi_\theta(\cdot|s)^T u(s,\cdot) \\
&- \pi_\theta(\cdot|s)^T (u(s,\cdot) \odot u(s,\cdot)) + 2(\pi_\theta(\cdot|s)^T u(s,\cdot))^2 | \\
&\leq \max_a \{ u(s,a)^2 + 2 \left| u(s,a) \pi_\theta(\cdot|s)^T u(s,\cdot) \right| \} \\
&+ \pi_\theta(\cdot|s)^T (u(s,\cdot) \odot u(s,\cdot)) + 2(\pi_\theta(\cdot|s)^T u(s,\cdot))^2 \\
&\leq \|u(s,\cdot)\|_2^2 + 2\|u(s,\cdot)\|_2^2 + \|u(s,\cdot)\|_2^2 + 2\|u(s,\cdot)\|_2^2 \\
&\leq 6\|u\|_2^2 \qquad (12)
\end{aligned}$$

Define $P(\alpha) \in \mathcal{R}^{S \times S}$, $\forall (s,s')$, $[P(\alpha)]_{(s,s')} = \sum_a \pi_{\theta_\alpha}(a|s) \mathcal{P}(s'|s,a)$ For any vector $x \in \mathbb{R}^S$, the l_∞ norm is

$$\begin{aligned}
\left\| \frac{\partial P(\alpha)}{\partial \alpha}|_{\alpha=0} x \right\|_\infty &= \max_s \left| \left[\frac{\partial P(\alpha)}{\partial \alpha}|_{\alpha=0} x \right]_{(s)} \right| \\
&= \max_s \left| \sum_{s'} \left[\frac{\partial P(\alpha)}{\partial \alpha}|_{\alpha=0} \right]_{(s,s')} x(s') \right| \\
&= \max_s \left| \sum_{s'} \sum_a \left[\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}|_{\alpha=0} \right] P(s'|s,a) x(s') \right| \\
&\leq \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}|_{\alpha=0} \right| \cdot \|x\|_\infty \\
&= \max_s \sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}|_{\alpha=0} \right| \cdot \|x\|_\infty \\
&\leq 2 \cdot \|u\|_2 \cdot \|x\|_\infty \qquad (13)
\end{aligned}$$

$$\begin{aligned}
\left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2}|_{\alpha=0} x \right\|_\infty &= \max_s \left| \left[\frac{\partial^2 P(\alpha)}{\partial \alpha^2}|_{\alpha=0} x \right]_{(s)} \right| \\
&= \max_s \left| \sum_{s'} \left[\frac{\partial^2 P(\alpha)}{\partial \alpha^2}|_{\alpha=0} \right]_{(s,s')} x(s') \right| \\
&= \max_s \left| \sum_{s'} \sum_a \left[\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}|_{\alpha=0} \right] P(s'|s,a) x(s') \right| \\
&\leq \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}|_{\alpha=0} \right| \cdot \|x\|_\infty \\
&= \max_s \sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}|_{\alpha=0} \right| \cdot \|x\|_\infty \leq 6 \cdot \|u\|_2^2 \cdot \|x\|_\infty \qquad (14)
\end{aligned}$$

Consider the KL regularized value function of π_{θ_α}.
$\hat{V}^{\pi_{\theta_\alpha}}(s) = \sum_a \pi_{\theta_\alpha}(a|s)(r(s,a) + \tau \log \beta(a|s)) + \gamma \sum_a \pi_{\theta_\alpha}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \hat{V}^{\theta_\alpha}(s') = e_s^T M(\alpha) r_{\theta_\alpha}$
where e_s is an indicator vector for the starting state s, $M(\alpha) = (\mathbf{Id} - P(\alpha))^{-1} = \sum_{t=0}^{\infty} \gamma^t P(\alpha)^t$, and $r_{\theta_\alpha}(s) = \sum_a \pi_{\theta_\alpha}(a|s)(r(s,a) + \tau \log \beta(a|s))$.

Taking derivative with respect to α:

$$\frac{\partial \hat{V}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} = \gamma e_s^T M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + e_s^T M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \qquad (15)$$

Taking second derivative with respect to α:

$$\begin{aligned}
\frac{\partial^2 \hat{V}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} &= 2\gamma^2 e_s^T M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \gamma^2 e_s^T M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \\
&\quad + 2\gamma e_s^T M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} + e_s^T M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \qquad (16) \\
\|M(\alpha)x\|_\infty &= \max_i \left| [M(\alpha)]_{i,:}^T x \right| \\
&\leq \max_i \|[M(\alpha)]_{i,:}\|_1 \cdot \|x\|_\infty \\
&= \frac{1}{1-\gamma} \cdot \|x\|_\infty (\text{Because } \mathbf{1} = \frac{1}{1-\gamma} \cdot (\mathbf{Id} - \gamma P(\alpha))\mathbf{1}, M(\alpha)\mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}) \qquad (17) \\
\|r_{\theta_\alpha}\|_\infty &= \max_s |r_{\theta_\alpha}(s)| \\
&= \max_s \left| \sum_a \pi_{\theta_\alpha}(a|s)(r(s,a) + \tau \log \beta(a|s)) \right| \\
&\leq M (\text{Because of the range of } r(s,a) + \tau \log \beta(a|s)) \qquad (18)
\end{aligned}$$

$$
\begin{aligned}
\left\| \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \right\|_\infty &= \max_s \left| \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right| \\
&= \max_s \left| \left(\frac{\partial r_{\theta_\alpha}(s)}{\partial \theta_\alpha} \right)^T \frac{\partial \theta_\alpha}{\partial \alpha} \right| \\
&= \max_s \left| \left(\frac{\partial \{\pi_{\theta_\alpha}(\cdot|s)^T (r(s,\cdot) + \tau \log \beta(\cdot|s))\}}{\partial \theta_\alpha} \right)^T u(s,\cdot) \right| \\
&= \max_s \left| (H(\pi_{\theta_\alpha}(\cdot|s))(r(s,\cdot) + \tau \log \beta(\cdot|s)))^T u(s,\cdot) \right| \\
&\leq \max_s \|H(\pi_{\theta_\alpha}(\cdot|s))(r(s,\cdot) + \tau \log \beta(\cdot|s))\|_1 \cdot \|u(s,\cdot)\|_\infty \\
&= \max_s (\sum_a \pi_{\theta_\alpha}(a|s) \cdot \left| r(s,a) + \tau \log \beta(a|s) - \pi_{\theta_\alpha}(\cdot|s)^T (r(s,\cdot) + \tau \log \beta(\cdot|s)) \right|) \|u(s,\cdot)\|_\infty \\
&\leq \max_s \max_a \left| r(s,a) + \tau \log \beta(a|s) - \pi_{\theta_\alpha}(\cdot|s)^T (r(s,\cdot) + \tau \log \beta(\cdot|s)) \right| \cdot \|u(s,\cdot)\|_\infty \\
&\leq 2M \|u\|_2 \qquad (19)
\end{aligned}
$$

$$
\begin{aligned}
\left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \right\|_\infty &= \max_s \left| \frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \alpha^2} \right| \\
&= \max_s \left| \left(\frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right\} \right)^T \frac{\partial \theta_\alpha}{\partial \alpha} \right| \\
&= \max_s \left| \left(\frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \theta_\alpha^2} \frac{\partial \theta_\alpha}{\partial \alpha} \right)^T \frac{\partial \theta_\alpha}{\partial \alpha} \right| \\
&= \max_s \left| u(s,\cdot)^T \frac{\partial^2 \{\pi_{\theta_\alpha}(\cdot|s)^T (r(s,\cdot) + \tau \log \beta(\cdot|s))\}}{\partial \theta_\alpha(s,\cdot)^2} u(s,\cdot) \right| \\
&= \max_s \left| \sum_{i=1}^{A} \sum_{j=1}^{A} S_{i,j} u(s,i) u(s,j) \right| (S_{i,j} \text{ is the i,j-th element of the second derivative}) \\
&= \max_s | \sum_i \pi_{\theta_\alpha}(s,\cdot,i)(r(s,i) + \tau \log \beta(i|s) - \pi_{\theta_\alpha}(\cdot|s)^T (r(s,\cdot) + \tau \log \beta(\cdot|s))) u(s,i)^2 \\
&- 2 \sum_i \pi_{\theta_\alpha}(i|s)(r(s,i) + \tau \log \beta(i|s) - \pi_{\theta_\alpha}(\cdot|s)^T (r(s,\cdot) + \tau \log \beta(\cdot|s))) u(s,i) \sum_j \pi_{\theta_\alpha}(j|s) u(s,j)| \\
&= \max_s |(H(\pi_{\theta_\alpha})(r(s,\cdot) + \tau \log \beta(\cdot|s)))^T (u(s,\cdot) \odot u(s,\cdot)) \\
&- 2(H(\pi_{\theta_\alpha})(r(s,\cdot) + \tau \log \beta(\cdot|s)))^T u(s,\cdot)(\pi_{\theta_\alpha}^T u(s,\cdot))| \\
&\leq \max_s (\|H(\pi_{\theta_\alpha})(r(s,\cdot) + \tau \log \beta(\cdot|s)))\|_\infty \cdot \|u(s,\cdot) \odot u(s,\cdot)\|_1 \\
&+ 2\|H(\pi_{\theta_\alpha})(r(s,\cdot) + \tau \log \beta(\cdot|s)))\|_1 \cdot \|u(s,\cdot)\|_\infty \cdot \|\pi_{\theta_\alpha}\|_1 \cdot \|u(s,\cdot)\|_\infty) \\
&\leq \max_s (\|H(\pi_{\theta_\alpha})(r(s,\cdot) + \tau \log \beta(\cdot|s)))\|_\infty \cdot \|u(s,\cdot)\|_2^2 (\text{Because } \|u(s,\cdot) \odot u(s,\cdot)\|_1 = \|u(s,\cdot)\|_2^2) \\
&+ 2\|H(\pi_{\theta_\alpha})(r(s,\cdot) + \tau \log \beta(\cdot|s)))\|_1 \cdot \|u(s,\cdot)\|_2^2) (\text{Because } \|\pi_{\theta_\alpha}\|_1 = 1, \|u(s,\cdot)\|_\infty \leq \|u(s,\cdot)\|_2) \\
&\leq \max_s (\max_i \left| H_{i,:}(\pi_{\theta_\alpha})^T (r(s,\cdot) + \tau \log \beta(\cdot|s))) \right| \|u(s,\cdot)\|_2^2 + 2 \cdot 2M \cdot \|u(s,\cdot)\|_2^2) \\
&\leq \max_s \max_i \|H_{i,:}(\pi_{\theta_\alpha})\|_1 \cdot \|r(s,\cdot) + \tau \log \beta(\cdot|s))\|_\infty \cdot \|u(s,\cdot)\|_2^2 + 4M \|u(s,\cdot)\|_2^2 \\
&\leq \max_s \max_i (\pi_{\theta_\alpha}(i|s) - \pi_{\theta_\alpha}(i|s)^2 + \pi_{\theta_\alpha}(i|s)(1 - \pi_{\theta_\alpha}(i|s))) \cdot M \cdot \|u(s,\cdot)\|_2^2 + 4M \|u(s,\cdot)\|_2^2 \\
&\leq \frac{1}{2} M \|u(s,\cdot)\|_2^2 + 4M \|u(s,\cdot)\|_2^2 \leq 6M \|u(s,\cdot)\|_2^2 \qquad (20)
\end{aligned}
$$

For the first term of Eq. 16, according to Eq. 17,13,18

$$\begin{aligned}\left|e_s^T M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)r_{\theta_\alpha}|_{\alpha=0}\right| &\leq \left|M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)r_{\theta_\alpha}|_{\alpha=0}\right|_\infty \\ &\leq \frac{1}{1-\gamma}\cdot 2\cdot\|u\|_2\cdot\frac{1}{1-\gamma}\cdot 2\cdot\|u\|_2\cdot\frac{1}{1-\gamma}\cdot M \\ &= \frac{4M}{(1-\gamma)^3}\|u\|_2^2 \end{aligned} \tag{21}$$

For the second term of Eq. 16, according to Eq. 17,14,18,

$$\begin{aligned}\left|e_s^T M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}|_{\alpha=0}\right| &\leq \left|M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}|_{\alpha=0}\right|_\infty \\ &\leq \frac{1}{1-\gamma}\left|\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}|_{\alpha=0}\right|_\infty \\ &\leq \frac{6\|u\|_2^2}{1-\gamma}\left|M(\alpha)r_{\theta_\alpha}|_{\alpha=0}\right|_\infty \\ &\leq \frac{6\|u\|_2^2}{(1-\gamma)^2}\left|r_{\theta_\alpha}|_{\alpha=0}\right|_\infty \leq \frac{6M}{(1-\gamma)^2}\|u\|_2^2 \end{aligned} \tag{22}$$

For the third term of Eq. 16,

$$\begin{aligned}\left|e_s^T M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}|_{\alpha=0}\right| &\leq \left|M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}|_{\alpha=0}\right|_\infty \\ &\leq \frac{1}{1-\gamma}\left|\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}|_{\alpha=0}\right|_\infty \\ &\leq \frac{2\|u\|_2}{1-\gamma}\left|M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}|_{\alpha=0}\right|_\infty \\ &\leq \frac{2\|u\|_2}{(1-\gamma)^2}\left|\frac{\partial r_{\theta_\alpha}}{\partial \alpha}|_{\alpha=0}\right|_\infty \\ &\leq \frac{4M}{(1-\gamma)^2}\|u\|_2^2 \end{aligned} \tag{23}$$

For the last term of Eq. 16,

$$\begin{aligned}\left|e_s^T M(\alpha)\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\right| &\leq \left|M(\alpha)\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\right|_\infty \\ &\leq \frac{1}{1-\gamma}\left|\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\right|_\infty \leq \frac{6M}{1-\gamma}\|u\|_2^2 \end{aligned} \tag{24}$$

Combining the four terms:

$$\begin{aligned}\left|\frac{\partial^2 \hat{V}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}|_{\alpha=0}\right| &\leq \left(2\gamma^2\cdot\frac{4M}{(1-\gamma)^3}+\gamma\frac{6M}{(1-\gamma)^2}+2\gamma\frac{4M}{(1-\gamma)^2}+\frac{6M}{1-\gamma}\right)\|u\|_2^2 \\ &\leq \frac{8M}{(1-\gamma)^3}\|u\|_2^2 \end{aligned} \tag{25}$$

which implies that for all $y \in \mathbb{R}^{SA}$ and θ:

$$\begin{aligned}\left|y^T \frac{\partial^2 \hat{V}^{\pi_\theta}(s)}{\partial \theta^2} y\right| &= \left|\left(\frac{y}{\|y\|_2}\right)^T \frac{\partial^2 \hat{V}^{\pi_\theta}(s)}{\partial \theta^2} \left(\frac{y}{\|y\|_2}\right)\right| \|y\|_2^2 \\ &\leq \max_{\|u\|_2=1} \left|\left\langle \frac{\partial^2 \hat{V}^{\pi_\theta}(s)}{\partial \theta^2} u, u \right\rangle\right| \|y\|_2^2 \\ &= \max_{\|u\|_2=1} \left|\left\langle \frac{\partial^2 \hat{V}^{\pi_{\theta_\alpha}}(s)}{\partial \theta_\alpha^2}|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle\right| \|y\|_2^2 \\ &= \max_{\|u\|_2=1} \left|\frac{\partial^2 \hat{V}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}|_{\alpha=0}\right| \|y\|_2^2 \leq \frac{8M}{(1-\gamma)^3}\|y\|_2^2 \end{aligned} \tag{26}$$

Denote $\theta_\xi = \theta + \xi(\theta' - \theta)$, where $\xi \in [0, 1]$, according to Taylor's theorem,

$$\begin{aligned}\left|\hat{V}^{\pi_{\theta'}}(s) - \hat{V}^{\pi_\theta}(s) - \left\langle \frac{\partial \hat{V}^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta \right\rangle\right| &= \frac{1}{2}\left|(\theta' - \theta)^T \frac{\partial^2 \hat{V}^{\pi_{\theta_\xi}}(s)}{\partial \theta_\xi^2}(\theta' - \theta)\right| \\ &\leq \frac{4M}{(1-\gamma)}\|\theta' - \theta\|_2^2 \end{aligned} \tag{27}$$

Lemma 5 (Smoothness)**.** $\mathbb{H}(\rho, \pi_\theta)$ *is* $\frac{4+8\log A}{(1-\gamma)^3}$*-smooth, where* $A = |\mathcal{A}|$ *is the total number of actions.*

Lemma 6 (Soft Policy Gradient)**.** *It holds that*

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \tilde{A}^{\pi_\theta}(s,a) \tag{28}$$

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(a|s)) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(\cdot|s)\right] = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(a|s)) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau\theta(s,\cdot)\right] \tag{29}$$

where $\tilde{A}^{\pi_\theta}(s,a)$ *is the 'soft' advantage function defined as:* $\tilde{A}^{\pi_\theta}(s,a) = \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s)$, $\tilde{Q}^{\pi_\theta}(s,a) = r(s,a) + \tau \log \beta(a|s) + \gamma \sum_s \mathcal{P}(s'|s,a)\tilde{V}^{\pi_\theta}(s')$

Lemma 7. $d_\mu^{\pi_\theta}(s) > (1-\gamma)\mu(s)$

$$\begin{aligned} d_\mu^{\pi_\theta}(s) &= \mathbb{E}_{s_0 \sim \mu}\left[(1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(s_t = s|s_0, \pi_\theta, \mathcal{P})\right] \\ &\geq \mathbb{E}_{s_0 \sim \mu}\left[(1-\gamma)P(s_t = s|s_0)\right] \\ &= (1-\gamma)\mu(s) \end{aligned}$$

Lemma 8 (Non-uniform Lojasiewicz)**.** *Suppose initial state distribution in the policy gradient algorithm* $\mu(s) > 0$ *for all state* $s \in \mathcal{S}$

$$\left\|\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \left\|\frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-\frac{1}{2}} \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}}$$

Lemma 9. *Using the algorithm with soft policy gradient, we have* $\inf_{t \geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$

The augmented value function $\tilde{V}^{\pi_\theta}(\rho)$ is monotonically increasing following the gradient update with proper η due to smoothness.

$\tilde{V}^{\pi_\theta}(\rho)$ is upper bounded as:

$$\begin{aligned} \tilde{V}^{\pi_{\theta_t}}(\rho) &= \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_{\theta_t}}(s)\left[\sum_a \pi_{\theta_t}(a|s) \cdot (r(s,a) + \tau \log \beta(a|s) - \tau \log \pi_{\theta_t}(a|s))\right] \\ &\leq \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_{\theta_t}}(s)\,(M + \tau \log A) \text{ (Because } -\sum_a \pi_{\theta_t}(a|s)\log \pi_{\theta_t}(a|s) \leq \log A) \\ &\leq \frac{M + \tau \log A}{1-\gamma} \end{aligned} \tag{30}$$

$\tilde{V}^{\pi_\theta}(\rho)$ is lower bounded as:

$$\begin{aligned}\tilde{V}^{\pi_{\theta_t}}(\rho) &= \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_{\theta_t}}(s)\left[\sum_a \pi_{\theta_t}(a|s)\cdot(r(s,a)+\tau\log\beta(a|s)-\tau\log\pi_{\theta_t}(a|s))\right] \\ &\geq \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_{\theta_t}}(s)(-M)(\text{Because entropy function is positive}) \\ &\geq \frac{-M}{1-\gamma} \end{aligned} \tag{31}$$

$\tilde{Q}^{\pi_\theta}(s,a)$ is lower bounded as:

$$\begin{aligned}\tilde{Q}^{\pi_{\theta_t}}(s,a) &= r(s,a)+\tau\log\beta(a|s)+\gamma\sum_{s'}\mathcal{P}(s'|s,a)\tilde{V}^{\pi_{\theta_t}}(s') \\ &\geq -M-\frac{\gamma}{1-\gamma}M \geq \frac{-M}{1-\gamma}\end{aligned} \tag{32}$$

According to monotone convergence theorem, $\tilde{V}^{\pi_\theta}(\rho)$ converges to a finite value, $\pi_{\theta_t}(a|s)\to\pi_{\theta_\infty}(a|s)$. For any state $s\in\mathcal{S}$, define the following sets: $\mathcal{A}_0(s)=\{a:\pi_{\theta_\infty}(a|s)=0\}$, $\mathcal{A}_+(s)=\{a:\pi_{\theta_\infty}(a|s)>0\}$. We prove that $\mathcal{A}_0(s)=\emptyset$ by contradiction.

Suppose that $\exists s\in\mathcal{S}$,such that $\mathcal{A}_0(s)$ is non-empty. For any $a_0\in\mathcal{A}_0(s)$, we have $\pi^{\theta_t}(a_0|s)\to\pi^{\theta_\infty}(a_0|s)=0$, which implies $-\log\pi^{\theta_t}(a_0|s)\to\infty$. There exists $t_0>0$, such that $\forall t\geq t_0$, $-\log\pi^{\theta_t}(a_0|s)\geq\frac{2M+\tau\log A}{\tau(1-\gamma)}$.

According to Lemma 6, $\forall t\geq t_0$:

$$\begin{aligned}\frac{\partial\tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial\theta(s,a_0)} &= \frac{1}{1-\gamma}\cdot d_\mu^{\pi_{\theta_t}}(s)\cdot\pi_{\theta_t}(a_0|s)\cdot\tilde{A}^{\pi_{\theta_t}}(s,a_0) \\ &= \frac{1}{1-\gamma}\cdot d_\mu^{\pi_{\theta_t}}(s)\cdot\pi_\theta(a_0|s)\cdot\left[\tilde{Q}^{\pi_{\theta_t}}(s,a_0)-\tau\log\pi_{\theta_t}(a_0|s)-\tilde{V}^{\pi_{\theta_t}}(s)\right] \\ &\geq \frac{1}{1-\gamma}\cdot d_\mu^{\pi_{\theta_t}}(s)\cdot\pi_\theta(a_0|s)\cdot\left[\frac{-M}{1-\gamma}-\tau\log\pi_{\theta_t}(a_0|s)-\tilde{V}^{\pi_{\theta_t}}(s)\right] \\ &\geq \frac{1}{1-\gamma}\cdot d_\mu^{\pi_{\theta_t}}(s)\cdot\pi_\theta(a_0|s)\cdot\left[\frac{-M}{1-\gamma}+\tau\frac{2M+\tau\log A}{\tau(1-\gamma)}-\frac{M+\tau\log A}{1-\gamma}\right] \\ &\geq 0\end{aligned} \tag{33}$$

This means $\theta_t(s,a_0)$ is always increasing $\forall t\geq t_0$, which implies that $\theta_\infty(s,a_0)$ is lower bounds by a constant c, and thus $\exp\{\theta_\infty(s,a_0)\}\geq e^c>0$. According to $\pi_{\theta_\infty}(a_0|s)=\frac{\exp\{\theta_\infty(s,a_0)\}}{\sum_a\exp\{\theta_\infty(s,a)\}}=0$, we have $\sum_a\exp\{\theta_\infty(s,a)\}=\infty$. On the other hand, for any $a_+\in\mathcal{A}_+(s)$, according to $\pi_{\theta_\infty}(a_+|s)=\frac{\exp\{\theta_\infty(s,a_+)\}}{\sum_a\exp\{\theta_\infty(s,a)\}}>0$, we have $\exp\{\theta_\infty(s,a_+)\}=\infty,\forall a_+\in\mathcal{A}_+(s)$, which implies $\sum_{a_+\in\mathcal{A}_+(s)}\theta_\infty(s,a_+)=\infty$. Note that $\forall t$,the summation of logit incremental over all actions is zero.

$$\begin{aligned}\sum_a\frac{\partial\tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial\theta_t(s,a)} &= \sum_{a_0\in\mathcal{A}_0(s)}\frac{\partial\tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial\theta_t(s,a_0)}+\sum_{a_+\in\mathcal{A}_+(s)}\frac{\partial\tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial\theta_t(s,a_+)} \\ &= \frac{1}{1-\gamma}\cdot d_\mu^{\pi_{\theta_t}}(s)\cdot\sum_a\pi_{\theta_t}(a|s)\tilde{A}^{\pi_{\theta_t}}(s,a)=0\end{aligned} \tag{34}$$

$\forall t\geq t_0$, $\sum_{a_0\in\mathcal{A}_0(s)}\frac{\partial\tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial\theta_t(s,a_0)}\geq 0$, then $\sum_{a_+\in\mathcal{A}_+(s)}\frac{\partial\tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial\theta_t(s,a_+)}\leq 0$. So $\sum_{a_+\in\mathcal{A}_+(s)}\theta_\infty(s,a_+)$ will always decrease for all large enough $t>t_0$. This is contradiction with $\sum_{a_+\in\mathcal{A}_+(s)}\theta_\infty(s,a_+)=\infty$. To this point, we have shown $\mathcal{A}_0(s)=\emptyset$ for any state s.

At the convergence point $\pi_{\theta_\infty}(\cdot|s)$, the gradient is zero. $\frac{\partial\tilde{V}^{\pi_{\theta_\infty}}(\mu)}{\partial\theta_\infty(s,\cdot)}=\frac{1}{1-\gamma}\cdot d_\mu^{\pi_{\theta_\infty}}(s)\cdot H(\pi_{\theta_\infty}(\cdot|s))\left[\tilde{Q}^{\pi_{\theta_\infty}}(s,\cdot)-\tau\log\pi_{\theta_\infty}(\cdot|s)\right]=\mathbf{0}$. Because of Lemma 7, $d_\mu^{\pi_{\theta_\infty}}(s)>0$, for all state s. Therefore $H(\pi_{\theta_\infty}(\cdot|s))\left[\tilde{Q}^{\pi_{\theta_\infty}}(s,\cdot)-\tau\log\pi_{\theta_\infty}(\cdot|s)\right]=\mathbf{0}$. $H(\pi_{\theta_\infty}(\cdot|s))$ has eigenvalue 0 with multiplicity 1, and the corresponding eigenvector is $c\cdot\mathbf{1}$ for some constant $c\in\mathbb{R}$. Therefore,

$\tilde{Q}^{\pi_{\theta_\infty}}(s,\cdot) - \tau \log \pi_{\theta_\infty}(\cdot|s) = c \cdot \mathbf{1}$, which is equivalent to $\pi_{\theta_\infty}(\cdot|s) = \text{softmax}(\tilde{Q}^{\pi_{\theta_\infty}}(s,\cdot)/\tau)$. Because $\tau \in \Omega(1) > 0$, $\frac{-M}{1-\gamma} \leq \tilde{Q}^{\pi_{\theta_\infty}}(s,a) \leq \frac{M+\tau \log A}{1-\gamma}$, we have $\pi_{\theta_\infty}(a|s) \in \Omega(1)$.

Since $\pi^{\theta_t}(a|s) \to \pi^{\theta_\infty}(a|s) > 0$, there exists $t_0 > 0$ such that $\forall t \geq t_0$, $0.9\pi^{\theta_\infty}(a|s) \leq \pi^{\theta_t}(a|s) \leq 1.1\pi^{\theta_\infty}(a|s)$, which means $\inf_{t \geq t_0} \min_{s,a} \pi_{\theta_t}(a|s) \in \Omega(1)$.
$\inf_{t\geq 1} \min_{s,a} \pi_{\theta_t}(a|s) = \min\{\inf_{1\leq t\leq t_0} \min_{s,a} \pi_{\theta_t}(a|s), \inf_{t\geq t_0} \min_{s,a} \pi_{\theta_t}(a|s)\} = \min\{\Omega(1), \Omega(1)\} \in \Omega(1)$

Theorem 4. *Suppose $\mu_(s) > 0$ for all state s. Using entropy regularized softmax policy gradient algorithm with $\eta = \frac{(1-\gamma)^3}{(8M+\tau(4+8\log A))}$ and $\pi_{\theta_1}(a|s) \in \Omega(1)$, $\forall(s,a)$,*

$$\tilde{V}^{\pi^*_\tau}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \leq \frac{\|1/\mu\|_\infty}{\exp\{C_\tau \cdot \Omega(1) \cdot t\}} \cdot \frac{M + \tau \log A}{(1-\gamma)^2}$$

for all $t > 0$, where $C_\tau, \Omega(1) > 0$ are independent with t.

According to the soft sub-optimality lemma,

$$\begin{aligned}
\tilde{V}^{\pi^*_\tau}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) &= \frac{1}{1-\gamma} \sum_s \left[d_\rho^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{KL}(\pi_{\theta_t}(\cdot|s) \| \pi^*_\tau(\cdot|s)) \right] \\
&= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_{\theta_t}}(s)}{d_\mu^{\pi_{\theta_t}}(s)} \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{KL}(\pi_{\theta_t}(\cdot|s) \| \pi^*_\tau(\cdot|s)) \right] \\
&\leq \frac{1}{1-\gamma} \sum_s \frac{1}{(1-\gamma)\mu(s)} \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{KL}(\pi_{\theta_t}(\cdot|s) \| \pi^*_\tau(\cdot|s)) \right] \\
&\leq \frac{1}{(1-\gamma)^2} \| \frac{1}{\mu} \|_\infty \sum_s \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{KL}(\pi_{\theta_t}(\cdot|s) \| \pi^*_\tau(\cdot|s)) \right] \\
&\leq \frac{1}{1-\gamma} \| \frac{1}{\mu} \|_\infty \left[\tilde{V}^{\pi^*_\tau}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right] \qquad (35)
\end{aligned}$$

Because of the property of smoothness, $\tilde{V}^{\pi_\theta}(\mu) = \hat{V}^{\pi_\theta}(\mu) + \tau \mathbb{H}(\mu, \pi_\theta)$ is λ-smooth with $\lambda = \frac{8M+\tau(4+8\log A)}{(1-\gamma)^3}$.

Denote $\tilde{\delta}_t = \tilde{V}^{\pi^*_\tau}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu)$,
$\tilde{\delta}_{t-1} - \tilde{\delta}_t = \tilde{V}^{\pi_{\theta_t}}(\mu) - \tilde{V}^{\pi_{\theta_{t+1}}}(\mu)$

$$\begin{aligned}
&\leq -\left\langle \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{4M + \tau(2 + 4\log A)}{(1-\gamma)^3} \|\theta_{t+1} - \theta_t\|_2^2 \\
&= \left(-\eta + \frac{4M + \tau(2+4\log A)}{(1-\gamma)^3} \cdot \eta \right) \left\| \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 \\
&= -\frac{(1-\gamma)^3}{16M + \tau(8 + 16\log A)} \left\| \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 (\text{Because } \eta = \frac{(1-\gamma)^3}{8M + \tau(4+8\log A)}) \\
&\leq -\frac{(1-\gamma)^3}{16M + \tau(8 + 16\log A)} \cdot \frac{2\tau}{S} \cdot \min_s \mu(s) \cdot \left[\min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi^*_\tau}}{d_\mu^{\pi_{\theta_t}}} \right\|_\infty^{-1} \cdot \left[\tilde{V}^{\pi^*_\tau}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right] \\
&\leq -\frac{(1-\gamma)^3}{(8M/\tau + 4 + 8\log A) \cdot S} \cdot \min_s \mu(s) \left[\min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi^*_\tau}}{(1-\gamma)\mu} \right\|_\infty^{-1} \cdot \tilde{\delta}_t (\text{Because } d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma)\mu(s)) \\
&\leq -\frac{(1-\gamma)^4}{(8M/\tau + 4 + 8\log A) \cdot S} \cdot \min_s \mu(s) \cdot \left[\inf_{t\geq 1} \min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi^*_\tau}}{\mu} \right\|_\infty^{-1} \cdot \tilde{\delta}_t \qquad (36)
\end{aligned}$$

$\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)\in\Omega(1)$ is independent with t.

$$\begin{aligned}
\tilde{\delta}_t &\leq \left[1-\frac{(1-\gamma)^4}{(8M/\tau+4+8\log A)\cdot S}\cdot\min_s\mu(s)\cdot\left[\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)\right]^2\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{\mu}\right\|_\infty^{-1}\right]\tilde{\delta}_{t-1}\\
&\leq \exp\left\{-\frac{(1-\gamma)^4}{(8M/\tau+4+8\log A)\cdot S}\cdot\min_s\mu(s)\cdot\left[\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)\right]^2\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{\mu}\right\|_\infty^{-1}\right\}\tilde{\delta}_{t-1}\\
&\leq \exp\left\{-\frac{(1-\gamma)^4}{(8M/\tau+4+8\log A)\cdot S}\cdot\min_s\mu(s)\cdot\left[\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)\right]^2\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{\mu}\right\|_\infty^{-1}(t-1)\right\}\tilde{\delta}_1\\
&\leq \exp\left\{-\frac{(1-\gamma)^4}{(8M/\tau+4+8\log A)\cdot S}\cdot\min_s\mu(s)\cdot\left[\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)\right]^2\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{\mu}\right\|_\infty^{-1}(t-1)\right\}\frac{M+\tau\log A}{1-\gamma}
\end{aligned}\tag{37}$$

Thus

$$\tilde{V}^{\pi_\tau^*}(\mu)-\tilde{V}^{\pi_{\theta_t}}(\mu)\leq \exp\left\{-\frac{(1-\gamma)^4}{(8M/\tau+4+8\log A)\cdot S}\cdot\min_s\mu(s)\cdot\left[\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)\right]^2\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{\mu}\right\|_\infty^{-1}(t-1)\right\}\frac{M+\tau\log A}{1-\gamma}$$

$$\begin{aligned}
\tilde{V}^{\pi_\tau^*}(\rho)-\tilde{V}^{\pi_{\theta_t}}(\rho) &\leq \frac{1}{1-\gamma}\|\frac{1}{\mu}\|_\infty\left[\tilde{V}^{\pi_\tau^*}(\mu)-\tilde{V}^{\pi_{\theta_t}}(\mu)\right]\\
&\leq \frac{\|1/\mu\|_\infty}{\exp\{C_\tau\cdot[\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)]^2\cdot t\}}\cdot\frac{M+\tau\log A}{(1-\gamma)^2}
\end{aligned}\tag{38}$$

where $C_\tau=\frac{(1-\gamma)^4}{(8M/\tau+4+8\log A)\cdot S}\cdot\min_s\mu(s)\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{\mu}\right\|_\infty^{-1}$

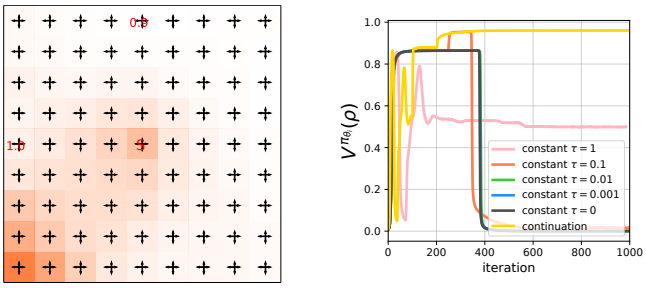


Figure 5: **Left:** visualization of the behavior policy and the dataset collected from the behavior policy. **Right:** learning curve of the value of the parameterized policy π_{θ_i} at each iteration i.

C Experimental Details

C.1 Grid World

We design the grid world of size 9×9. The action space includes actions: up, down, left, right. The reward is 0 at most states, while there are positive rewards 0.9 and 1 at two terminal states. In each episode, the agent starts from the state in the center annotated with 'S', walks around the environment, and terminates the episode only when the agent visits any of the four terminal states (yellow squares at the edge of the grid work). As explained in Section 3.3, we study the performance of soft policy iteration algorithm with different value of τ. The behavior policy is mediocre and tends to move left and down with higher probability at each state. We collect 10000 transitions in the domain according to the behavior policy . We visualize the behavior policy in Fig. 5. In each state, the length of the arrow is proportional to the probability of taking the actions in four directions. We also visualize the visitation count of each state in the collected dataset. The darker color means more visitation. We can see the behavior policy mostly move around the left bottom corner.

We consider optimizing $\tilde{V}^{\pi,\tau}(\rho)$ with different value of τ. It could be constant value in $\{1, 0.1, 0.01, 0.001, 0\}$. For the continuation method, we initially set the value of τ as 1, and decay it with $\tau \leftarrow 0.1\tau$ at every 100 iteration. We search the hyper-parameter of learning rate in $\{5, 1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$. The learning curves in Fig. 5 verify that the continuation method has better performance than the baselines with a fixed constraint. Visualization of the learned policies for different methods is in Fig. 2. With a constant $\tau = 1$, the learned policy performs similarly to the behavior policy. With a constant $\tau = 0.001$, the learned policy crashes because the value estimate is noisy. With decaying value of τ, the error in Q value estimate is reduced and the relaxing constraint allows the agent to deviate much from the behavior policy.

C.2 Mujoco

As mentioned in the main paper, the architecture of the critic network and the policy network is kept the same in the baselines and our method. The critic network consists of 4 independent Q–networks, which share the same feedforward network architecture. The state and action are concatenated and then passed through 2 hidden layers of size 400 and 300. The policy network is a simple linear network, with the input and output corresponding to the state and the action respectively. ReLU is used as the activation function in all hidden layers. We use 2 separate Adam optimizers for policy and critic network learning and the learning rates are set to 0.001. For all the baselines, we do a grid search of their algorithm–specific hyper–parameters for each game and each dataset. This parameter is fixed in 5 independent runs.

C.2.1 Continuous BCQ

For continuous BCQ, we mostly follow their official implementation[2] and hyper–parameter settings. BCQ uses a perturbation model to adjust the action output with an added residual, which is in the range of $[-\Phi, \Phi]$. We do a grid search of the max perturbation Φ over $\{0.1, 0.2, 0.3\}$.

[2] https://github.com/sfujim/BCQ

C.2.2 BEAR

For BEAR, we also follow the official implementation[3]. In BEAR, a threshold ϵ is used to constrain the MMD distance between the unknown behaviour policy β and the target policy π. It is set to be 0.05 in the original experiments. We instead do a grid search of ϵ over $\{0.05, 0.1, 0.2\}$. BEAR also uses a VAE network to generate actor with distribution similar as the unknown behavior policy. We keep the same architecture as their implementation for VAE and use an Adam optimizer with learning rate 0.001

C.2.3 ABM+SVG

For ABM, we use the batch size of 100, target network update period 1. The KL divergence between the learned policy and the prior policy is constrained by ϵ in ABM, which is set to 0.2 for SVG in the original work. We do a grid search of ϵ over $\{0.05, 0.1, 0.2\}$. For the additional prior policy network, we also use an Adam optimizer with learning rate 0.001.

C.2.4 CRR

In CRR training, we use the batch size of 128, target network update period 1. We use 4 samples to compute the advantage as the original paper does. For fair comparison, we keep the critic architecture the same as others, instead of the distributional one. We focus on the 'mean exp' variant of CRR, because it performs well in CRR paper, while another variant 'max binary' is roughly equivalent to ABM+SVG, which we have run as a separate baseline. To decide on the appropriate temperature in the scalar function f, we swept β over $\{0.01, 0.1, 1, 10\}$, where β is fixed as 1 in the original paper.

C.2.5 CQL

For CQL, we mostly follow the official implementation[4] but make the network architecture same as the other methods. In CQL, α is the coefficient to control the conservative, lower bound Q–function and there are two variants, $CQL(\rho)$ and $CQL(\mathcal{H})$. We search the hyper-parameter for both variants on our datasets. For $CQL(\rho)$, we search the Lagrange threshold τ over $\{0.1, 0.5, 2, 5, 10\}$. For $CQL(\mathcal{H})$, we search the fixed tradeoff factor α over $\{0.1, 1, 10, 100, 1000, 10000\}$. Besides, since our buffer size is 1 million, a half of the buffer size in the original D4RL experiments, we also reduce the batch size from 256 to 100 to keep it consistent with the other methods. The remaining hyper parameters are kept the same as the default value in CQL.

C.2.6 CONSTANT

To compare with the cases that there's no decay of the KL weight in our method, we set KL weight to be 10, 1, 0.1 separately and fix the decay to be 1. We run this method on the dataset with $\alpha = 0.6$ on all three games.

C.2.7 BRAC

For BRAC, we adapt the code of the official implementation[5] to use the same network architecture and datasets as the other methods. BRAC has different design choices (value penalty and/or policy regularization). We use both value penalty in the Q function objective and policy update with KL regularization, which is reported to have better performance than policy regularization only. We use the KL divergence in the primal form to keep the setting the same as our method. We search the hyper-parameter for both fixed and adaptive version to set the regularization coefficient α. For the fixed version, we search α over the range $\{0.1, 1, 10\}$. For the adaptive version, we search the threshold ϵ, which is the same as the threshold used in BEAR, over the range $\{0.05, 0.5, 2\}$. We find the fixed version has a better performance, which is consistent with the conclusion from BRAC paper. Besides, we also change the learning rate of the policy network to 0.001, which is the same as ours. Other hyper-parameters are kept the same as the default value.

[3] https://github.com/aviralkumar2907/BEAR

[4] https://github.com/aviralkumar2907/CQL

[5] https://github.com/google-research/google-research/tree/master/behavior_regularized_offline_rl

The fixed version of BRAC is equivalent to our "Constant" baseline, though there are many differences in implementation details. For example, following the official implementation of BRAC, there is a regularization for the weight of network parameters in BRAC baseline but "Constant KL" baseline does not have this term. BRAC has two training phases, the training of the behavior policy for 30K updates, and the training of BRAC with the fixed behavior policy trained before. However, "Constant" trains the behavior policy simultaneously throughout the whole procedure. In the computation of KL divergence, BRAC samples 4 actions for one state to approximate the KL divergence but "Constant" uses the closed–form solution to calculate KL divergence between two normal distributions.

C.2.8 OURS

Initially in the first J iterations we learn behavior policy β_ψ from the data and train the target policy π_θ only minimizing KL divergence between π_θ and β_ψ. To find the well-trained policy with good performance before the value estimate becomes quite noisy, we record the variance in Q estimate $var(Q^{\pi_{\theta_i}})$ for each update i. In each run, we calculate the average of the variance in 1000 iterations at the end of behavior cloning, i.e. $x = \frac{1}{1000}\sum_{i=J-1000}^{J-1} var^{\pi_{\theta_i}}(Q)$. This is a reference point of the Q variance for this run. As training continues, at the iteration j, we monitor the average of Q variance in the most recent 1000 iterations $y = \frac{1}{1000}\sum_{i=j-1000}^{j-1} var(Q^{\pi_{\theta_i}})$. If the average of Q variance is larger than 1.5 times the reference point, i.e. $\frac{y}{x} > 1.5$, the current value of τ may be not reliable and we report the score of policy checkpointed with the previous value of τ. Until the end of training, if we always have $\frac{y}{x} \leq 1.5$, then we find the policy iteration is stable and we report the final performance of the trained agent.

For our continuation method, accross the different datasets, we set most hyper-parameters the same. For the first $J = 500K$ updates, we train the policy network with only KL divergence loss to conduct behavior cloning, which is equivalent to the case that $\tau \to \infty$. We set the decay rate $\lambda = 0.9$ and decay the weight KL term with $\tau \leftarrow \lambda * \tau$ for every $I = 10000$ updates. We conduct hyper-parameter search over the initial value of τ in the set $\{10000, 1000, 100\}$.

C.3 ATARI

To generate the dataset $\mathcal{D}$ for the Atari games, we train the DQN agents using the standard online procedure to 10 million timesteps on the environment with sticky action. To generate the trajectory, with probability of 0.8, we uses ϵ-greedy with noise $\epsilon = 0.2$ for the whole episode to take actions. With probability of 0.2, we use the noise $\epsilon = 0.001$ for the whole trajectory. As explained in [13], we can ensure the dataset includes trajectories reaching the good performance of the DQN agent as well as trajectories with exploratory behavior.

C.3.1 DISCRETE BCQ

We refer to the official implementation of discrete BCQ[13], training the the behavioral policies, generating datasets and training the BCQ model. The $84 \times 84 \times 3$ RGB image is fed into a convolutional neural network. The input image first goes through an 8×8 convolution with 32 filters and stride 4, then a 4×4 convolution with 64 filters and 2 stride, followed by a 3×3 convolution with 64 filters and 1 stride, with ReLU activations. There are 2 heads, each being a fully connected layer, for Q–network output and generative model output after the convolutional neural network. Both fully connected layers have hidden size 512 and use the ReLU activation. The convolutional neural network is shared between the Q–network and the generative model. A final softmax layer is used after the output of the generative model to recover the probability for each action. We search of BCQ threshold over $\{0.1, 0.3, 0.5\}$ to find the best parameter for each game. We use the same hyper–parameter as the original implementation for training and only change the evaluation ϵ in the testing time to keep consistent with other baselines and our method.

C.3.2 REM

For REM training, we completely follow the implementation from the official codebase[6] released by [2]. The architecture is exactly the same as [2] while the dataset is from a single behavior policy. shared with discrete BCQ and our method. We use the multi-head architecture with 200 heads and do not change any hyper parameter.

C.3.3 CQL

The official implementation of CQL[7] is based on REM, so we use the same way to adapt the code and run experiments on our dataset. We follow CQL experiments on Atari, using $CQL(H)$ with a fixed tradeoff factor α. We search α over $\{0.5, 1, 4\}$ for the best hyper-parameter for each game. Other hyper-parameters are kept the same as the default setting in the official code.

C.3.4 OURS

In our method, we also use an ensemble critic network as we do in Mujoco. The architecture of each Q-network in the critic network is the same as the Q-network architecture used by discrete BCQ, with a 3–layer convolution and a fully connect layer. The policy network also uses the same architecture.

For the first $J = 500K$ updates, we train the policy network with only KL divergence loss to conduct behavior cloning, which is equivalent to the case that $\tau \to \infty$. We set initial weight of the KL term as 1, the decay rate $\lambda = 0.9$ and decay the weight KL term with $\tau \leftarrow \lambda * \tau$ for every $I = 100000$ updates. We pick the learned policy with a reasonable value of τ using the same way as in Mujoco experiment.

C.4 RECOMMENDER

Here we explain more details about the experiments for a softmax recommender. In the MovieLen-10M, the rating is of 5-score. Similarly to [25], we view the scores 4 or 5 as positive feedback from users but the ratings less than 4 as negative feedback. In the bandit setting, the expected return of the policy π can be expressed as $V^{\pi}(\rho) = \mathbb{E}_{s\sim\rho, a\sim\beta(\cdot|s)}(\frac{\pi(a|s)}{\beta(a|s)} r(s,a))$. So we do not use the critic network to estimate the soft Q value. We convert the objective $V^{\pi,\tau}(\rho)$ to $\mathbb{E}_{s\sim\rho, a\sim\beta(\cdot|s)}(\frac{\pi(a|s)}{\beta(a|s)} r(s,a)) - \tau\mathbb{E}_{s\sim\rho} KL(\pi(\cdot|s)|\beta(\cdot|s))$ in the bandit setting. Here the initial state distribution should be sampling the users randomly from the dataset, and we set the objective function as $J_{Ours}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\pi_\theta(a_i|s_i)}{\beta(a_i|s_i)} r_i - \tau KL(\pi_\theta(\cdot|s_i)|\beta(\cdot|s_i))\right]$.

The architecture of the simulator, the behavior policies and target policy are the same as introduced in [25]. The simulator is trained with all 1 million records of (user, movie, feedback) in the MovieLen-1M dataset. We train the behavior policy with 500, 5000, 10000 samples respectively to get the behavior policy with different performance. Then 6040 users in Movie-Len dataset, the movie can be selected from the behavior policy, and the simulator outperforms the feedback. If the behavior policy recommends 5 movies out of the 3900 choices for each user, we have a dataset $\mathcal{D}$ with around 30,000 samples. If the behavior policy recommend 10 movies for each user, we have a dataset $\mathcal{D}$ with around 60,000 samples

With the generated datasets of different sizes and different qualities, we train the target policy π_θ using three methods: 'Cross-Entropy', 'IPS' and 'Ours'. The objective functions are optimized for 400 epochs, with Adagrad optimizer and the learning rate 0.05 and the batch size 256. For our continuation method, the value of τ is initialized as 1 and decay with the rate $\lambda = 0.9$ in every epoch. The users in MovieLens-1M dataset are split into validation set (2000 users) and test set (4040 users). In each epoch, we measure the precision at 10 for the learned policy on the validation set. We take the policy achieving best performance on the validataion set during training, and test it for the held out users to report the performances in Table 3.

[6] https://github.com/google-research/batch_rl
[7] https://github.com/aviralkumar2907/CQL

D VARIANCE OF ENSEMBLE CRITIC NETWORKS

We found the error in Q estimation is highly correlated with the variance of ensemble critic network. If the learned policy π tends to select the state-action pairs with high variance in ensemble networks, the error in the Q estimation on these state-action pairs will propagate to the other state-action pairs and mislead the policy update. So we measure $\mathbb{E}_{a\sim\pi}\left[var(Q_{\phi^{(1)}}(s,a), Q_{\phi^{(2)}}(s,a), \cdots, Q_{\phi^{(K)}}(s,a))\right]$ as the criteria to determine proper value of τ. When this statistics is quite large, the policy tends to select actions with highly erroneous value estimation and we no longer trust the policy with smaller value of τ.

D.1 TRAIN THE CRITIC NETWORKS WITH THE SAME DATA

When we train the critic networks with the same data, the discrepancy in value estimation comes from the initialization of different critic networks. For state-action pairs frequently visited by the behavior policy, these data are frequently sampled to update the Q network. Therefore, the error in Q estimation and variance in Q ensemble on these state-action pairs diminishes as training goes on. But for state-action pairs less frequently visited by the behavior data, they are rarely used to update the Q network. The variance in Q ensemble are larger compared to the other state-action pairs. In Figure 6, for each state-action pair in the grid world environment, we visualize the error $|\frac{1}{K}\sum_{k=1}^{K} Q_{\Phi^{(k)}}(s,a) - \tilde{Q}^{\pi,\tau}(s,a)|$ and the variance $var(Q_{\phi^{(1)}}(s,a), Q_{\phi^{(2)}}(s,a), \cdots, Q_{\phi^{(K)}}(s,a))$ in the grid world for each state-action pair. It shows that the error becomes highly correlated with the variance as training goes on. Obviously, the frequently visited state-action pairs in the dataset has lower error and variance during training.

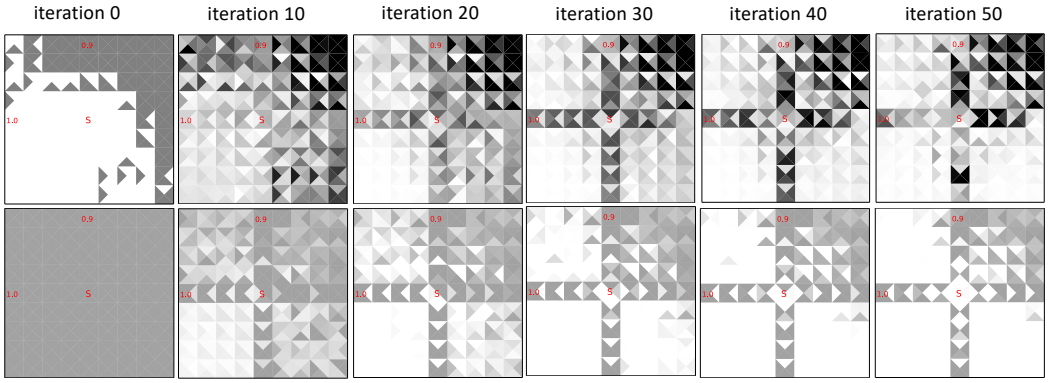


Figure 6: Visualization of the error in soft Q value estimation (the first row) and variance in the ensemble of critic networks (the second row). Triangles represent the error/variance for actions that move in different directions. Darker color indicates higher error/variance.

D.2 TRAIN THE CRITIC NETWORKS WITH THE DIFFERENT DATA

We have an alternative way to train the critic network with delete-d jackknife resampling. It is a more standard approach to estimate the variance through data perturbation. To train each of 4 critic networks, we leave out $\frac{1}{4}$ data samples in the dataset.

We assume $\hat{Q}(s,a)$ is an estimator of $\tilde{Q}^{\pi,\tau}(s,a)$. We consider the expected squared estimation error.

$$
\begin{aligned}
\mathbb{E}\left[(\hat{Q}(s,a) - \tilde{Q}^{\pi,\tau}(s,a))^2\right] &= \mathbb{E}\left[\hat{Q}(s,a)^2 + \tilde{Q}^{\pi,\tau}(s,a)^2 - 2\hat{Q}(s,a)\tilde{Q}^{\pi,\tau}(s,a)\right] \\
&= \mathbb{E}\left[\hat{Q}(s,a)^2\right] + \tilde{Q}^{\pi,\tau}(s,a)^2 - 2\mathbb{E}\left[\hat{Q}(s,a)\right]\tilde{Q}^{\pi,\tau}(s,a) \\
&= \mathbb{E}\left[\hat{Q}(s,a)\right]^2 + \tilde{Q}^{\pi,\tau}(s,a)^2 - 2\mathbb{E}\left[\hat{Q}(s,a)\right]\tilde{Q}^{\pi,\tau}(s,a) + \mathbb{E}\left[\hat{Q}(s,a)^2\right] - \mathbb{E}\left[\hat{Q}(s,a)\right]^2 \\
&= \left[\mathbb{E}\left[\hat{Q}(s,a)\right] - \tilde{Q}^{\pi,\tau}(s,a)\right]^2 + \left[\mathbb{E}\left[\hat{Q}(s,a)^2\right] - \mathbb{E}\left[\hat{Q}(s,a)\right]^2\right]
\end{aligned}
$$

The first term is the bias of the estimator and the second term is the variance of the estimator. In our case, we use the average of the ensemble Q networks $\hat{Q}(s,a) = \frac{1}{K}\sum_{k=1}^{K} Q_{\phi^{(k)}}(s,a)$ to estimate the soft Q value. With jackknife re-sampling, the variance of the estimator $\hat{Q}(s,a) = \frac{1}{K}\sum_{k=1}^{K} Q_{\phi^{(k)}}(s,a)$ can be approximated by $var(Q_{\phi^{(1)}}(s,a), Q_{\phi^{(2)}}(s,a), \cdots, Q_{\phi^{(K)}}(s,a)) = \frac{1}{K(K-1)}\sum_{k=1}^{K}\left[Q_{\phi^{(k)}}(s,a) - \frac{1}{K}\sum_{j=1}^{K} Q_{\phi^{(j)}}(s,a)\right]^2$. Thus, the squared estimation error is closely related with the variance in the ensemble $Q_{\phi^{(1)}}(s,a), Q_{\phi^{(2)}}(s,a), \cdots, Q_{\phi^{(K)}}(s,a)$.

With this method, the empirical results on Mujoco and Atari are similar to the scores we achieved when training critic networks with the same data (Table 4& 5). Therefore, we validate the use of variance of the ensemble network to detect the Q estimation error and then stop annealing τ on Mujoco and Atari as well.

	Hopper				HalfCheetah				Walker			
α	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
Same	2097	1569	1648	226	2145	1824	1487	547	1441	1223	853	7
Jackknife	2072	1718	1404	230	2206	1840	1529	589	1462	1078	920	8

Table 4: Performance of our method on Mujoco datasets. "same" is the variant to train each critic network with the same batch of data, which is reported in the main text. "jackknife" is the variant to train each critic network with data from jackknife resampling. The score is averaged over 5 independent runs for each method on each dataset, except that we only have one run on HalfCheetah and Walker with $\alpha = 0.2$ or $\alpha = 0.2$. We will complete the experiments with more runs.

	Amidar	Asterix	Breakout	Enduro	MsPacman	Qbert	Seaquest	SpaceInvaders
Same	175	3477	199	923	2494	4733	9935	1070
Jackknife	171	3890	217	908	2305	5389	9636	914

Table 5: Performance of our method on Atari datasets. "same" is the variant to train each critic network with the same batch of data, which is reported in the main text. "jackknife" is the variant to train each critic network with data from jackknife resampling. Due to time limit, we only have one run for the variant with jackknife resampling. We will complete the experiment with 3 independent runs on each dataset.

E SOFT POLICY ITERATION VIA LAGRANGIAN DUALITY

The discounted state-action visitations d^π of π can be defined as:

$$d^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t \mathcal{P}(s_t = s, a_t = a | s_0 \sim \rho, \forall t, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(s_t, a_t))$$

The visitation satisfies the single-step transpose Bellman recurrence:

$$d^\pi(s,a) = \rho(s)\pi(a|s) + \gamma \mathcal{P}_*^\pi d^\pi(s,a)$$

where $\mathcal{P}_*^\pi$ is the transpose policy transition operator:

$$\mathcal{P}_*^\pi d(s,a) = \pi(a|s) \sum_{\tilde{s},\tilde{a}} \mathcal{P}(s|\tilde{s},\tilde{a}) d(\tilde{s},\tilde{a})$$

Note that if we define the policy transition operator $\mathcal{P}^\pi y(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')} y(s',a')$. We have $\sum_{s,a} y(s,a) \mathcal{P}_*^\pi x(s,a) = \sum_{s,a} x(s,a) \mathcal{P}^\pi y(s,a)$.

The objective function $\tilde{V}^{\pi,\tau}(\rho)$ can be expressed as $\tilde{V}^{\pi,\tau}(\rho) = \sum_{s,a} d^\pi(s,a) r(s,a) - \tau \sum_{s,a} d^\pi(s,a) \log \frac{\pi(a|s)}{\beta(a|s)}$.

Consider the optimization problem:

$$\max_\pi \sum_{s,a} d^\pi(s,a) r(s,a) - \tau \sum_{s,a} d^\pi(s,a) \log \frac{\pi(a|s)}{\beta(a|s)}$$
$$s.t \quad d^\pi(s,a) = \rho(s)\pi(a|s) + \gamma \mathcal{P}_*^\pi d^\pi(s,a), \forall s, a$$

Application of Lagrange duality to the above problem yields $\max_\pi \min_Q L(Q,\pi)$ where

$$L(Q,\pi) = \sum_{s,a} d^\pi(s,a) r(s,a) - \tau \sum_{s,a} d^\pi(s,a) \log \frac{\pi(a|s)}{\beta(a|s)} + \sum_{s,a} Q(s,a) \left[\rho(s)\pi(a|s) + \gamma \mathcal{P}_*^\pi d^\pi(s,a) - d^\pi(s,a)\right]$$

and $Q(s,a)$ are dual variables.

We can rewrite $L(Q,\pi)$:

$$\begin{aligned}
L(Q,\pi) &= \sum_{s,a} d^\pi(s,a) r(s,a) - \tau \sum_{s,a} d^\pi(s,a) \log \frac{\pi(a|s)}{\beta(a|s)} \\
&\quad + \sum_{s,a} Q(s,a)\rho(s)\pi(a|s) + \gamma \sum_{s,a} Q(s,a) \mathcal{P}_*^\pi d^\pi(s,a) - \sum_{s,a} Q(s,a) d^\pi(s,a) \\
&= \sum_{s,a} d^\pi(s,a) r(s,a) - \tau \sum_{s,a} \rho(s)\pi(a|s) \log \frac{\pi(a|s)}{\beta(a|s)} - \tau \sum_{s,a} \gamma \mathcal{P}_*^\pi d^\pi(s,a) \log \frac{\pi(a|s)}{\beta(a|s)} \\
&\quad + \sum_{s,a} Q(s,a)\rho(s)\pi(a|s) + \gamma \sum_{s,a} Q(s,a) \mathcal{P}_*^\pi d^\pi(s,a) - \sum_{s,a} Q(s,a) d^\pi(s,a) \\
&= \sum_{s,a} Q(s,a)\rho(s)\pi(a|s) - \tau \sum_{s,a} \rho(s)\pi(a|s) \log \frac{\pi(a|s)}{\beta(a|s)} \\
&\quad + \sum_{s,a} d^\pi(s,a) r(s,a) + \gamma \sum_{s,a} (Q(s,a) - \tau \log \frac{\pi(a|s)}{\beta(a|s)}) \mathcal{P}_*^\pi d^\pi(s,a) - \sum_{s,a} d^\pi(s,a) Q(s,a) \\
&= \sum_{s,a} Q(s,a)\rho(s)\pi(a|s) - \tau \sum_{s,a} \rho(s)\pi(a|s) \log \frac{\pi(a|s)}{\beta(a|s)} \\
&\quad + \sum_{s,a} d^\pi(s,a) r(s,a) + \gamma \sum_{s,a} d^\pi(s,a) \mathcal{P}^\pi (Q(s,a) - \tau \log \frac{\pi(a|s)}{\beta(a|s)}) - \sum_{s,a} d^\pi(s,a) Q(s,a) \\
&= \mathbb{E}_{s \sim \rho} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} Q(s,a) - KL(\pi(\cdot|s) \| \beta(\cdot|s))\right] \\
&\quad + \sum_{s,a} d^\pi(s,a) \left[r(s,a) + \gamma \mathcal{P}^\pi (Q(s,a) - \tau \log \frac{\pi(a|s)}{\beta(a|s)}) - Q(s,a)\right]
\end{aligned}$$

In our soft policy evaluation step, we repeatedly apply the soft Bellman operator. At the convergence, the Q value satisfies $\mathcal{T}^{\pi,\tau}Q(s,a) = Q(s,a)$, i.e. $r(s,a)+\gamma\mathcal{P}^{\pi}(Q(s,a)-\tau\log\frac{\pi(a|s)}{\beta(a|s)})-Q(s,a) = 0$. Thus, the second part of $L(Q,\pi)$ is 0.

In our soft policy improvement step, for each state s, we optimize $\pi(\cdot|s)$ to maximize $\mathbb{E}_{a\sim\pi(\cdot|s)}Q(s,a) - KL(\pi(\cdot|s)\|\beta(\cdot|s))$. Thus, the first part of $L(Q,\pi)$ is maximized.

Therefore, the dual objective function bears some resemblance to the our soft policy iteration.

F PERFORMANCE ON DATASETS OF DIFFERENT QUALITY

We study the performance of our method on datasets of different quality. As shown in Figure 7, our method is able to learn a policy superior to the behavior policy and the gap is obvious especially when $\alpha \in [0.4, 0.6]$. In other words, when the behavior policy is mediocre, our method has a good chance to learn from the good trajectories in the dataset and achieve a better score than the average score of the dataset.

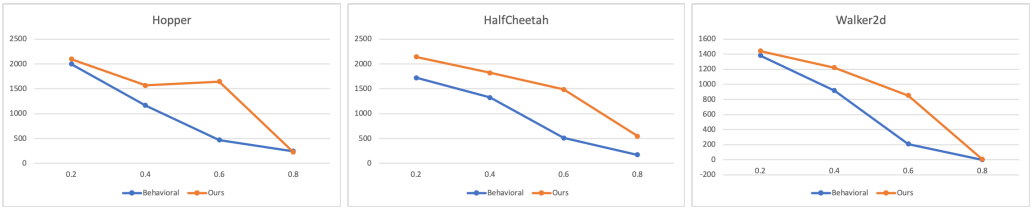


Figure 7: Performance of our learned policy and the behavior policy for different datasets. The x-axis, y-axis represents α and average reward separately. For $\alpha \in [0, 1]$, the behavior policy $\mathcal{N}(((1-\alpha)\theta_{opt} + \alpha\theta_0)^T s, 0.5\mathbb{I})$. If $\alpha = 0$, the behavior policy is the optimal. If $\alpha = 1$, the behavior policy is a random policy.

We further investigate the distribution of the trajectory rewards in the datasets. In Figure 7, we observe that our method does not perform so well in comparison with the behavior policy on the datasets (Hopper, $\alpha = 0.2$), (Hopper, $\alpha = 0.8$), (Walker, $\alpha = 0.2$), (Walker, $\alpha = 0.8$). In Figure 8, we find the problem of these four dataset is that there is only a small fraction of trajectories with cumulative rewards better than the average trajectory reward plus the standard deviation in the dataset. We study the percentage of trajectories with reward better than one standard deviation of the mean (i.e. $\%1\sigma$ trajectory). In Figure 9, we see the percentage of 1σ trajectory highly correlates with the improvement, where the improvement is the average trajectory reward of our learned policy minus the average trajectory reward of the dataset. If there is a reasonable proportion (i.e. $> 10\%$) of $\%1\sigma$ trajectory in the dataset, our method can learn from these good trajectories and perform significantly better than the average trajectory reward in the dataset.

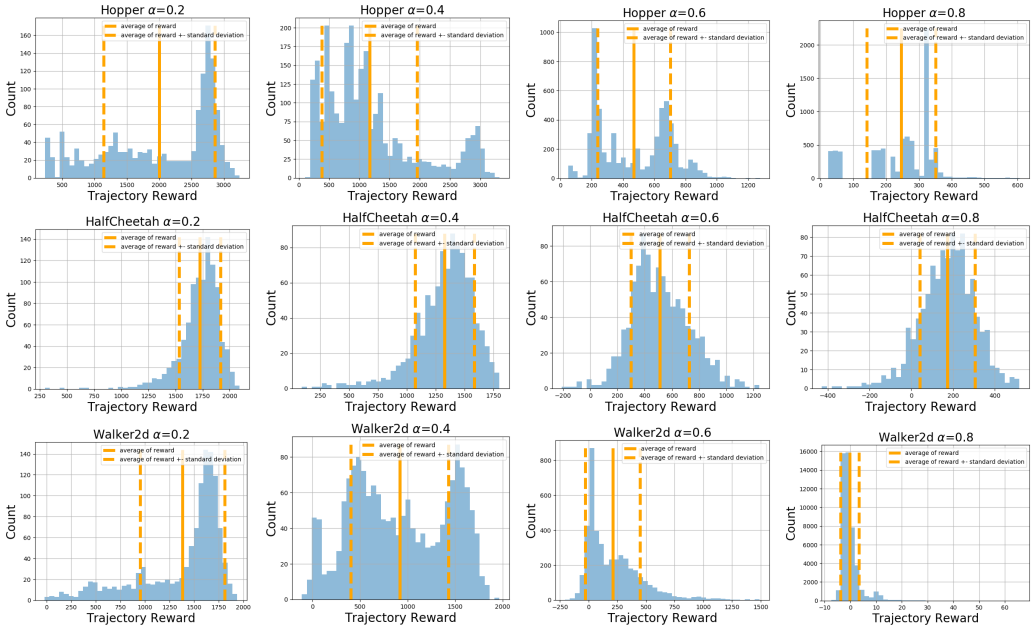


Figure 8: Histogram of the trajectory rewards in our Mujoco datasets.

Figure 9: On 12 datasets in Mujoco environment, we create the scatter plot of good trajectory percentage versus the improvement, where X-axis represents the percentage of 1σ trajectory and Y-axis represents the improvement of our method over the average trajectory reward in the dataset.