Provable Benefits of Actor-Critic Methods for Offline Reinforcement Learning

Andrea Zanette [1] **Martin Wainwright** [2] **Emma Brunskill** [3]

Abstract

Actor-critic methods are widely used in offline reinforcement learning practice but are understudied theoretically. In this work we show that the pessimism principle can be naturally incorporated into actor-critic formulations. We create an offline actor-critic algorithm for a linear MDP model more general than the low-rank model. The procedure is both minimax optimal and computationally tractable.

1. Introduction

Learning a near-optimal policy is a core reinforcement learning (RL) task. Oftentimes, we need to find a good policy using the available data and without the possibility of further interaction with the environment; this is called the *policy learning* problem in *offline* RL. Offline RL has unique challenges due to the incomplete information about the Markov decision process (MDP) encoded in the available dataset. For example, due to the maximization bias, a naive offline algorithm can settle for a policy with a dangerously high estimated value even if such value is highly uncertain. To avoid this phenomenon, researchers have introduced the idea of *pessimism* in offline RL (Liu et al., 2020; Jin et al., 2020b; Buckman et al., 2020; Kumar et al., 2019; Kidambi et al., 2020; Yu et al., 2020). Additional literature is presented in Appendix A.

Pessimism prevents algorithms from settling down on uncertain policies whose value might be misleadingly high under the current dataset due to statistical errors. By using pessimism, uncertain policies are penalized and only those robust to statistical errors are returned. The principle can be implemented in two different ways: 1) by penalizing policies that are far from the one that generated the dataset or 2) by penalizing the value functions of policies not well covered by the dataset. We adopt the second view in this work.

Challenges Implementing pessimism with function approximation brings several challenges. First, *the uncertainty must be carefully estimated*: underestimating it may not lead to an effective algorithm and overestimating it leads to policies that are too conservative and thus underperform. Second, pessimism may *introduce complex, higher order perturbations* into the value function class handled by the algorithm (similar to adding optimistic bonuses in the exploration setting). Increasing the complexity of the function class often requires additional assumptions on the model, because the new class needs to interact 'nicely' with the Bellman operator. Prior art on pessimism with function approximation bypassed this problem by making strong model assumptions, like low rank transitions (Jin et al., 2020b) or algorithm-specific assumptions (Liu et al., 2020).

Actor-critic methods Moreover, most of these theoretically-justified algorithms are either model or value based (Liu et al., 2020; Jin et al., 2020b; Buckman et al., 2020; Kidambi et al., 2020; Yu et al., 2020), but in practice, actor-critic methods are widely used (Levine et al., 2020; Wu et al., 2019; 2021; Kumar et al., 2019; 2020). An actor-critic method generally consists of an *actor* that changes the policy in order to maximize its value as estimated by the *critic*. In relation to value and model-based methods we ask:

Do actor-critic methods provably offer any advantage in offline RL?

We give a positive answer: by separating the policy optimization from the policy evaluation, both tasks become simpler to design and the pessimism principle can be incorporated more naturally.

Contributions We consider the linear function approximation setting and assume that a batch dataset $\mathcal{D}$ of states, actions, rewards and successor states is available. Using $\mathcal{D}$ we can construct the set $\mathcal{M}$ of statistically plausible MDPs subject to the linearity assumption.

*Equal contribution [1]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, USA [2]Department of Electrical Engineering and Computer Sciences and Department of Statistics, University of California at Berkeley, Berkeley, USA [3]Department Computer Science, Stanford University, Stanford, USA. Correspondence to: Andrea Zanette <zanette@stanford.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Our objective is then to find the policy that performs the best in the face of uncertainty, namely the policy π with the highest minimum value function V_M^π across all plausible MDPs M in the set $\mathcal{M}$ of statistically plausible MDPs (e.g., (Mannor et al., 2012)):

$$\sup_\pi \inf_{M \in \mathcal{M}} V_M^\pi. \tag{1}$$

Actor-critic methods fit naturally in this framework: the *actor* solves the outer maximization problem over policies which are evaluated in the inner minimization problem by a pessimistic *critic*. This way, each algorithm solves a simple task: 1) the critic provides a pessimistic value function estimate for a fixed policy (the one currently examined by the actor) while 2) the actor ensures online learning-style guarantees with respect to a sequence of pessimistic MDPs implicitly identified by the critic. This is the first algorithmic idea and leads to a computationally tractable implementation.

The second algorithmic idea is to introduce pessimism without altering the prescribed function class. This is achieved by perturbing the value function (in the critic) within its prescribed functional space without adding pessimistic bonuses or absorbing states. This has two core advantages:

- There are no additional model assumptions compared to the vanilla (i.e., without pessimism) version of our actor-critic method; this is because the original value function class is not modified by the injection of pessimism.
- The algorithm operates on value functions with the original statistical complexity, enabling the construction of tight confidence intervals and ultimately minimax statistical rates.

2. Preliminaries and Assumptions

We consider an undiscounted finite-horizon MDP (Puterman, 1994) $M = (\mathcal{S}, \mathcal{A}, p, r, H)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, and horizon length $H \in \mathbb{N}^+$. For every $h \in [H] = \{1, \dots, H\}$, every state-action pair is characterized by an expected reward $r_h(s,a)$ with an associated reward distribution $R_h(s,a)$ and a transition kernel $p_h(\cdot \mid s,a)$ over next state. For any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the state-action value function of a non-stationary policy $\pi = (\pi_1, \dots, \pi_H)$ is defined as $Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}_{s_l \sim \pi \mid (s,a)} \sum_{l=h+1}^H r_l(s_l, \pi_l(s_l))$, where the expectation is over the trajectories induced by π upon starting from (s,a). When we omit the starting state-action (s,a), the expectation is intended to start from a fixed state denoted by s_1. The value function associated to π is $V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$. Under some regularity conditions, e.g., (Shreve & Bertsekas, 1978), there always exists an optimal policy $\pi^\star$ whose value and action-value functions are defined as $V_h^\star(s) = V_h^{\pi^\star}(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^\star(s,a) = Q_h^{\pi^\star}(s,a) = \sup_\pi Q_h^\pi(s,a)$. We define the Bellman evaluation operator

$$\mathcal{T}_h^\pi(Q_{h+1})(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim p_h(s,a)} \mathbb{E}_{a' \sim \pi} Q_{h+1}(s', a').$$

We let $\mathcal{B}_d(r) = \{x \in \mathbb{R}^d \mid \|x\|_2 \le r\}$ denote the Euclidean ball of radius $r \in \mathbb{R}$ in dimension d; sometime we simply write $\mathcal{B}$ when there is no possibility of confusion. We use the $\widetilde{O}$ notation to suppress log factors in the input parameters $(\frac{1}{\delta}, d, H, \lambda)$, and the O and Ω notation to ignore constants in the upper and lower bound. The notation $\lesssim$ means $\le$ up to a constant while $\lessapprox, \approx, \gtrapprox, \asymp$ are used to highlight dominant terms in the proof sketch without rigorous mathematical definitions. For a vector $x \in \mathbb{R}^d$ we let $[x]_i$ denote its i component.

2.1. Assumptions on Data Generation

The algorithm operates on a dataset $\mathcal{D} = \{(s_{hk}, a_{hk}, r_{hk}, s_{hk}^+)\}_{h=1,2,\dots,H}^{k=1,2,\dots,K}$ of state-action-reward-next states generated by the underlying MDP, possibly in an adaptive fashion.

Assumption 1 (Data Generation)**.** *Assume that for every (s,a) the reward random variable with distribution $R_h(s,a)$ is 1-subgaussian. The dataset $\mathcal{D}$ is such that*

$$r_{hk} \sim R(s_{hk}, a_{hk}), \qquad s_{hk}^+ \sim p_h(s_{hk}, a_{hk}) \tag{2}$$

where each s_{hk}, a_{hk} is allowed to depend on the previously sampled $(s_{ij}, a_{ij}, r_{ij}, s_{ij}^+)$.

This allows considerable freedom: 1) the dataset may be generated from (mixture) policies or by another mechanism that collects information at different state-actions; 2) the dataset may be generated by an adversarial procedure that changes the data acquisition strategy as feedback is received.

2.2. Policy and Value Function Class

Next, we define the policy space Π and the action value function space $\mathcal{Q}$ where we seek solutions. For a fixed timestep h (which we omit here for brevity), consider a fixed feature extractor $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, $\|\phi(\cdot,\cdot)\|_2 \le 1$ and two radii, $r_w \in (0,1]$, $r_\theta > 0$ for the value function parameter w and for the policy parameter θ.

Definition 1 (Functional Spaces)**.**

$$\mathcal{Q}(r_w) = \{(s,a) \mapsto \phi(s,a)^\top w \mid \|w\|_2 \le r_w\},$$

$$\Pi(r_\theta) \stackrel{def}{=} \left\{ \frac{\exp[\phi(s,a)^\top \theta]}{\sum_{a'} \exp[\phi(s,a')^\top \theta]} \mid \|\theta\|_2 \le r_\theta \right\}.$$

The policy radius can be large $r_\theta \gg 1$ but we constrain $r_w \le 1$ so that $\sup_{(s,a,w)} |Q_w(s,a)| \le 1$. In finite horizon problems one can select different feature extractors

ϕ_h in every step h; this generates H functional spaces $\mathcal{Q}_1, \ldots, \mathcal{Q}_H$ and $\Pi_1, \ldots, \Pi_H$. We drop the dependence on the radii when referring to the functional spaces and implicitly assume that the terminal value function is zero.

2.3. Assumptions on Function Class

If we seek to find the policy $\pi \in \Pi$ with the highest value function, it seems reasonable to require that the following representation condition (approximately) holds. We assume a common feature extractor $\phi : \mathcal{S} \times \mathcal{A}$, $\|\phi(\cdot,\cdot)\|_2 \leq 1$ throughout this section.

Assumption 2 (Linear Q^π)**.** *We say the MDP admits a linear action-value function representation for all policies in* Π *if* Q^π *is linear, i.e.,*

$$\forall \pi \in \Pi, h \in [H], \exists w_h^\pi \text{ such that } Q_h^\pi(s,a) = \phi_h(s,a)^\top w_h^\pi. \tag{3}$$

Unfortunately, Corollary 1 in (Zanette, 2020) or Theorem 4.1 in (Wang et al., 2020a) establish that even under such assumption, we might need exponentially many samples to do better than a random policy. This suggests we need even stronger conditions. One such condition is the assumption we make in this work, which allows a classical temporal-difference critic to evaluate the policies in Π.

Assumption 3 (Restricted Closedness)**.** *The policy and value function spaces* $(\Pi, \mathcal{Q})$ *are closed up to* $\epsilon^{miss} \in \mathbb{R}$ *error in* ∞ *norm with respect to a finite horizon MDP if* $\forall h \in [H]$*:*

$$\sup_{\substack{Q_{h+1} \in \mathcal{Q}_{h+1} \\ \pi_{h+1} \in \Pi_{h+1}}} \inf_{Q_h \in \mathcal{Q}_h} \|Q_h - \mathcal{T}_h^{\pi_{h+1}} Q_{h+1}\|_\infty \leq \epsilon_h^{miss} \in \mathbb{R}. \tag{4}$$

The restricted closedness assumption measures how well we can fit the action-value function resulting from the application of the Bellman evaluation operator to an action value function in $\mathcal{Q}$ and for a policy in Π. It enables the analysis of the classical *Least Square Policy Evaluation* (LSPE) (Nedić & Bertsekas, 2003), which will be our starting point when constructing the critic.

A related model assumption is the *low-rank* or *linear* MDP model (Jin et al., 2020a; Yang & Wang, 2020) used by the state of the art for offline RL with pessimismistic guarantees (Jin et al., 2020b) and much of the online RL literature (Agarwal et al., 2020a; Modi et al., 2021; Zanette et al., 2020a). It is possible to show that the restricted closendess assumption is more general than low rank; details in appendix.

3. Main Result

Due to space reason, the algorithm is reported here but is described in Appendix B in appendix.

Algorithm 1 ACTOR (MIRROR DESCENT)

Input: Dataset $\mathcal{D}$, starting state s_1
Set $\theta_1 = (\vec{0}, \ldots, \vec{0})$
for $k = 1, 2, \ldots, K$ **do**
  $\underline{w}_k \leftarrow \text{CRITIC}(\mathcal{D}, \pi_{\theta_k}, s_1)$
  $\theta_{k+1} = \theta_k + \eta \underline{w}_k$
end for
Mixture policy $\pi_{\theta_1}, \ldots, \pi_{\theta_K}$

Algorithm 2 CRITIC (PLSPE)

Input: Dataset $\mathcal{D}$, target policy π, starting state s_1
Solve the optimization program (7)
Return $\underline{w}$

Let us introduce the optimization error $\mathcal{R}(K)$, function of the number of actor's iterations K, and the uncertainty function $U(\pi)$ for a policy π where $\sqrt{\alpha_h} = \widetilde{O}(\sqrt{d_h + d_{h+1}}) + \epsilon_h^{miss}\sqrt{K} + \sqrt{\lambda}$ is fully defined in Lemma 5 and Definition 6 in appendix:

$$U(\pi) \overset{def}{=} 2\sum_{h=1}^{H} \left[\epsilon_h^{miss} + \sqrt{\alpha_h}\|\mathbb{E}_{(s_h,a_h)\sim\pi}\phi(s_h,a_h)\|_{\Sigma_h^{-1}}\right], \tag{5}$$

$$\mathcal{R}(K) \overset{def}{=} 4H\sqrt{\frac{\ln|\mathcal{A}|}{K}}. \tag{6}$$

The amount of information from the dataset $\mathcal{D}$ is fully encoded in the uncertainty function U through the cumulative covariance matrix Σ_h. The more data are available, the more positive definite Σ_h is and the smaller the uncertainty function $U(\pi)$ becomes for a fixed policy π. If the sampling distribution is fixed, then $U(\pi) \lessapprox C/\sqrt{n}$ where C can be interpreted as the condition number of Σ_h^{-1} and n is the number of samples.

Our main result holds under Assumption 1, when the learning rate is $\eta = \sqrt{\ln|\mathcal{A}|/K}$, the radii for the action value function[1] parameters are in $(0, 1]$, the regularization is $\lambda \geq 1$ and the number of iterations is $K \geq \ln|\mathcal{A}|$; Π_{all} is the class of all stochastic policies.

Theorem 1 (Main Result)**.** *Algorithms 1 and 2 return a policy* π_{ALG} *such that*

$$\mathbb{P}\Big(V_1^{\pi_{\text{ALG}}}(s_1) \geq \sup_{\pi \in \Pi_{all}} \Big[V_1^\pi(s_1) - U(\pi)\Big] - \mathcal{R}(K)\Big) \geq 1 - \delta.$$

[1] This represents a setting where both the reward and the value function can be as large as 1 in absolute value. One easily recovers the setting with value functions in $[0, H]$ using a rescaling argument.

The result provides a lower bound on the quality of the returned policy and highlights a tradeoff between the suboptimality of the comparator π and its uncertainty $U(\pi)$. Note that the optimization error $\mathcal{R}(K)$ goes to zero as $K \to \infty$; different choices of the learning rate are possible and they only affect the optimization error $\mathcal{R}(K)$ (i.e, the computational cost). Thus, ignoring the optimization error, regularization and misspecification and assuming $d_h = d, \forall h \in [H]$ we obtain with high probability $V_1^{\pi_{\text{ALG}}}(s_1) \gtrapprox \sup_{\pi \in \Pi_{all}} V_1^{\pi}(s_1) - \sqrt{d} \sum_{h=1}^{H} \|\mathbb{E}_{(s_h,a_h)\sim\pi_h} \phi(s_h, a_h)\|_{\Sigma_h^{-1}}$.

The result is complemented by a matching worst-case upper bound on the quality of the returned policy, excluding constants and log factors. The upper bound already arises in a setting that is easier for the learner, as it holds (1) when the MDP is *low-rank* (thus it applies when Assumption 3 holds), and (2) when the mechanism that generates the dataset is *non-adaptive* (thus it applies when Assumption 1 holds).

We assume $d_h = d, \forall h \in [H]$ and $\epsilon_h^{miss} = 0$ for simplicity, as well as $\lambda = 1$ when referring to the uncertainty function U; $\mathbb{E}_M$ indicates that the expectation is with respect to MDP M.

Theorem 2 (Information-Theoretic Upper Bound)**.** *Fix any choice of horizon H, of dimension d and of number of samples n collected at each timestep. There exists an MDP class $\mathcal{M}$ such that*

$$\sup_{\text{ALG}} \inf_{M \in \mathcal{M}} \mathbb{E}_M V_{1M}^{\pi_{\text{ALG}}} \leq \sup_{\pi \in \Pi_{all}} \left[V_{1M}^{\pi}(s_1) - \frac{\Omega(1)}{\log\left(\frac{1}{\delta}, K\right)} \times U(\pi) \right].$$

Comparison with literature Theorem 1 automatically implies the typical bound $\mathbb{P}[V_1^{\pi_{\text{ALG}}}(s_1) \geq V_1^{\star}(s_1) - U(\pi^{\star})] \geq 1 - \delta$ when the comparator policy is the optimal policy $\pi^{\star}$, e.g., (Jin et al., 2020b; Rashidinejad et al., 2021; Kidambi et al., 2020; Kumar et al., 2019; Buckman et al., 2020). The guarantee can be written as $V_1^{\pi_{\text{ALG}}}(s_1) \gtrapprox V_1^{\star}(s_1) - C/\sqrt{n}$ where n is the number of samples and C is the (scaled) condition number of Σ_h^{-1}. One could interpret C as a concentrability coefficient that expresses the coverage of dataset — through Σ_h — with respect to the average direction in feature space $\mathbb{E}_{(s_h,a_h)\sim\pi_h^{\star}} \phi(s_h, a_h)$ of the optimal policy $\pi^{\star}$. As in (Jin et al., 2020b), such factor C can be small even when traditional concentrability coefficients are large because they depend on state-action visit ratios (see the literature in Appendix A, e.g., (Chen & Jiang, 2019)).

Even ignoring the concentrability coefficient, the form of our result is significantly stronger as our algorithm competes with all comparator policies simultaneously; these policies are not necessarily in the prescribed policy class Π. To highlight the strength of our formulation (see also (Yu et al., 2020; Liu et al., 2020) for results in a similar form), suppose that the optimal policy is not well covered, i.e., $U(\pi^{\star})$ infinite, but there exists a near-optimal policy π^{+} i.e., such that $V_1^{\pi^{+}}(s_1) \geq V_1^{\star}(s_1) - \epsilon$ for some small ϵ, which is well covered by the dataset, i.e., $U(\pi^{+}) \approx 0$. In this case, Theorem 1 ensures with high probability $V_1^{\text{ALG}}(s_1) \gtrapprox V_1^{\star}(s_1) - \epsilon$. In contrast, traditional analyses that use only $\pi^{\star}$ as comparator cannot return meaningful guarantees.

The work closest to ours is (Jin et al., 2020b); our work directly improves on theirs by closing the dH gap between their upper and lower bound while working under the more permissive Assumption 3 which includes low-rank MDPs. A $\sqrt{d}$-improvement is due to the algorithm we use and the remaining is due to a more refined analysis and construction to certify optimality in Theorem 2 (notice that our upper and lower bounds differ from theirs by a factor of H due to a different normalization in the value function). The result of (Liu et al., 2020) can be specialized to the low-rank MDP setting but would give a suboptimal bound while additionally requiring density estimates.

Deriving a computationally tractable model-free algorithm without low-rank dynamics but subject to value function perturbations (e.g., optimistic or pessimistic perturbations) is an open problem even in the more heavily studied exploration setting: there the current state of the art (Zanette et al., 2020b; Jin et al., 2021; Du et al., 2021; Jiang et al., 2017) only present computationally *intractable* algorithms with the exception of (Zanette et al., 2020c) for a PAC setting with low inherent Bellman error which however requires an additional 'explorability' condition.

4. Discussion

A key idea of this paper is to introduce pessimism while remaining in the prescribed function class. Doing so allows us to avoid making additional model assumptions, and achieves minimax optimality. Similar ideas have appeared before in the exploration setting (e.g., (Zanette et al., 2020b; Jin et al., 2021; Du et al., 2021)) with similar advantages (batch-style assumptions + minimax regret) *but at the expense of computational tractability*.

Fortunately, the offline RL setting differs from the online setting and we are able to maintain computational tractability by clearly separating the actor's update from the critic evaluation. In this way, *each algorithm solves a simpler task*, and computational tractability is retained.

The numerical evaluation of this procedure and the extension to more general function classes are important next steps, and it will be interesting to see if any of these ideas can be translated to the more challenging exploration setting.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*, 2020a.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66, 2020b.

Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020c.

Antos, A., Munos, R., and Szepesvári, C. Fitted q-iteration in continuous action-space mdps. 2007.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Bhandari, J. and Russo, D. A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*, 2020.

Bubeck, S. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051, 2019.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Duan, Y. and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2002.09516*, 2020.

Duan, Y., Jin, C., and Li, Z. Risk bounds and rademacher complexity in batch reinforcement learning. *arXiv preprint arXiv:2103.13883*, 2021.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Fu, Z., Yang, Z., and Wang, Z. Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*, 2020.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvári, C., and Wang, M. Bootstrapping statistical inference for off-policy evaluation. *arXiv preprint arXiv:2102.03607*, 2021.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*, 2020.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In Precup, D. and Teh, Y. W. (eds.), *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/jiang17c.html.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020a.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020b.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*, 2019.

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*, 2021.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

Kumar, A., Fu, J., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

Laroche, R., Trichelair, P., and Des Combes, R. T. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pp. 3652–3661. PMLR, 2019.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Liao, P., Qi, Z., and Murphy, S. Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Mannor, S., Mebel, O., and Xu, H. Lightning does not strike twice: Robust mdps with coupled uncertainty. *arXiv preprint arXiv:1206.4643*, 2012.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Munos, R. Error bounds for approximate policy iteration. In *ICML*, volume 3, pp. 560–567, 2003.

Munos, R. Error bounds for approximate value iteration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2005.

Nachum, O. and Dai, B. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*, 2019a.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

Nair, A., Dalal, M., Gupta, A., and Levine, S. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Nedić, A. and Bertsekas, D. P. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1):79–110, 2003.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Raskutti, G. and Mukherjee, S. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

Shreve, S. E. and Bertsekas, D. P. Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM Journal on control and optimization*, 16(6):953–978, 1978.

Siegel, N. Y., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R., Heess, N., and Riedmiller, M. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pp. 1057–1063. Citeseer, 1999.

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Voloshin, C., Jiang, N., and Yue, Y. Minimax model learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1612–1620. PMLR, 2021.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020a.

Wang, Z., Novikov, A., Żołna, K., Springenberg, J. T., Reed, S., Shahriari, B., Siegel, N., Merel, J., Gulcehre, C., Heess, N., et al. Critic regularized regression. *arXiv preprint arXiv:2006.15134*, 2020b.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Wu, Y., Zhai, S., Srivastava, N., Susskind, J., Zhang, J., Salakhutdinov, R., and Goh, H. Uncertainty weighted actor-critic for offline reinforcement learning. *arXiv preprint arXiv:2105.08140*, 2021.

Xie, T. and Jiang, N. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. volume 124 of *Proceedings of Machine Learning Research*, pp. 550–559, Virtual, 03–06 Aug 2020a. PMLR. URL http://proceedings.mlr.press/v124/xie20a.html.

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020b.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pp. 9668–9678, 2019.

Yang, L. F. and Wang, M. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning (ICML)*, 2020.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR, 2020.

Yin, M., Bai, Y., and Wang, Y.-X. Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*, 2020.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Zanette, A. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. *arXiv preprint arXiv:2012.08005*, 2020.

Zanette, A., Brandfonbrener, D., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *AISTATS*, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning (ICML)*, 2020b.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. Provably efficient reward-agnostic navigation with linear value iteration. In *Advances in Neural Information Processing Systems*, 2020c.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*, 2021.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*, 2020a.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020b.