

Appendix

A MORE DISCUSSIONS ON BAYESDICE

In this section, we provide more details about BayesDICE.

Remark (parametrization of $q(\zeta)$ and $q(\beta|\zeta)$): We parametrize both $q(\zeta)$ (and the resulting $q(\beta|\zeta)$) as Gaussians with the mean and variance approximated by a multi-layer perceptron (MLP), *i.e.*: $\zeta = \text{MLP}_w(s, a) + \sigma_{w'}\xi$, $\xi \sim \mathcal{N}(0, 1)$. w and w' denote the parameters of the MLP.

Remark (connection to Bayesian inference for stochastic processes): Recall the posterior can be viewed as the solution to an optimization (Zellner, 1988; Zhu et al., 2014; Dai et al., 2016),

$$q(\zeta|\mathcal{D}) = \underset{q \in \mathcal{P}}{\text{argmin}} \langle q(\zeta), \log p(\zeta, \mathcal{D}) \rangle + KL(q(\zeta) || p(\zeta))$$

The (I3) is equivalent to define the log-likelihood proportion to $\ell(\zeta, \mathcal{D})$, which is a stochastic process, including Gaussian process (\mathcal{GP}) by setting $f^*(\beta) = \frac{1}{2}\beta^\top \beta$. Specifically, plug $f(\beta) = \frac{1}{2}\beta^\top \beta$ back into (I3), we have $\beta^* = \hat{\mathbb{E}}_{\mathcal{D}}[\zeta(s, a) \cdot (\gamma\phi(s', a') - \phi(s, a))] + (1 - \gamma) \mathbb{E}_{\mu_0\pi}[\phi]$, resulting the optimization

$$\min_q KL(q||p) + \frac{\lambda}{\epsilon} \mathbb{E}_q \mathbb{E}_{\mu_0\pi} \hat{\mathbb{E}}_{\mathcal{D}} \left[\zeta(s_1, a_1)^\top k((s_1, a_1, s'_1, a'_1), (s_2, a_2, s'_2, a'_2)) \zeta(s_2, a_2) \right], \quad (15)$$

with the kernel $k(x_1, x_2) := (\gamma\phi(s'_1, a'_1) - \phi(s_1, a_1))^\top (\gamma\phi(s'_2, a'_2) - \phi(s_2, a_2)) + (1 - \gamma)^2 \phi(s_1^0, a_1^0)^\top \phi(s_2^0, a_2^0) + 2(1 - \gamma) \phi(s_1^0, a_1^0)^\top (\gamma\phi(s'_2, a'_2) - \phi(s_2, a_2))$, which is a \mathcal{GP} . Obviously, with different choices of $f^*(\cdot)$, the BayesDICE framework is far beyond \mathcal{GP} .

Although the \mathcal{GP} has been applied for RL (Engel et al., 2003; Ghavamzadeh et al., 2016; Aziz-zadenesheli et al., 2018), they all focus on prior on value function; while BayesDICE considers general stochastic processes likelihood, including \mathcal{GP} , for the stationary ratio modeling, which as we justified is more flexible for different selection criteria in downstream tasks.

Remark (auxiliary constraints and undiscounted MDP): As Yang et al. (2020) suggested, the non-negative and normalization constraints are important for optimization. We exploit positive neuron to ensure the non-negativity of the mean of the $q(\zeta)$. For the normalization, we consider the chance constraints $\mathbb{P}\left(\left(\hat{\mathbb{E}}_{\mathcal{D}}(\zeta) - 1\right)^2 \leq \epsilon_1\right) \geq \xi_1$. By applying the same technique, it leads to extra term $\frac{\lambda_1}{\epsilon_1} \mathbb{E}_q \left[\max_{\alpha \in \mathbb{R}} \alpha \cdot \hat{\mathbb{E}}_{\mathcal{D}}[\zeta - 1] \right]$ in (I3).

With the normalization condition introduced, the proposed BayesDICE is ready for undiscounted MDP by simply setting $\gamma = 1$ in (I3) together with the above extra term for normalization.

Remark (variants of log-likelihood): We apply the Markov's inequality to (I2) for the upper bound (I3). In fact, the optimization with chance constraint has rich literature (Ben-Tal et al., 2009), where plenty of surrogates can be derived with different safe approximation. For example, if the q is simple, one can directly calculate the CDF for the probability $\mathbb{P}_q(\ell(\zeta) \leq \epsilon)$; or one can also exploit different probability inequalities to derive other surrogates, *e.g.*, condition value-at-risk, *i.e.*,

$$\min_q KL(q||p) + \lambda \inf_t \left[t + \frac{1}{\epsilon} \mathbb{E}_q[\ell(\zeta) - t] \right]_+, \quad (16)$$

and Bernstein approximation (Nemirovski & Shapiro, 2007). These surrogates lead to better approximation to the chance probability $\mathbb{P}_q(\ell(\zeta) \leq \epsilon)$ with the extra cost in optimization.

B BAYESDICE FOR EXPLORATION VS. EXPLOITATION TRADEOFF

In main text, we mainly consider exploitin BayesDICE for estimating various ranking scores for both discounted MDP and undiscounted MDP. In fact, with the posterior of the stationary ratio computed, we can also apply it for better balance between exploration vs. exploitation for policy optimization.

Instead of selecting from a set of policy candidates, the policy optimization is considering all feasible policies and selecting optimistically. Specifically, the feasibility of the stationary state-action

distribution can be characterized as

$$\sum_a d(s, a) = (1 - \gamma) \mu_0 + \mathcal{P}_* d(s), \quad \forall s \in S, \quad (17)$$

where $\mathcal{P}_* d(s) := \sum_{\bar{s}, \bar{a}} T(s|\bar{s}, \bar{a}) d(\bar{s}, \bar{a})$. Apply the feature mapping for distribution matching, we obtain the constraint for $\zeta \cdot \pi$ with $\zeta(s, a) := \frac{d(s)}{d^{\mathcal{P}}(s, a)}$ as

$$\max_{\beta \in \mathcal{H}_\phi} \beta^\top \mathbb{E}_{d^{\mathcal{P}}} \left[\sum_a (\zeta(s, a) \pi(a|s)) \phi(s) - \gamma (\zeta(s, a) \pi(a|s)) \phi(s') \right] + (1 - \gamma) \mathbb{E}_{\mu_0} [\beta^\top \phi] - f^*(\beta) = 0. \quad (18)$$

Then, we have the posteriors for all valid policies should satisfies

$$\lambda \mathbb{P}_q (\ell(\zeta \cdot \pi, \mathcal{D}) \leq \epsilon) \geq \xi, \quad (19)$$

with $\ell(\zeta \cdot \pi, \mathcal{D}) := \max_{\beta \in \mathcal{H}_\phi} \beta^\top \hat{\mathbb{E}}_{\mathcal{D}} [\sum_a (\zeta(s, a) \pi(a|s)) \phi(s) - \gamma (\zeta(s, a) \pi(a|s)) \phi(s')] + (1 - \gamma) \mathbb{E}_{\mu_0} [\beta^\top \phi] - f^*(\beta)$. Meanwhile, we will select one posterior from among these posteriors of all valid policies optimistically, *i.e.*,

$$\max_{q(\zeta)q(\pi)} \mathbb{E}_q [U(\tau, r, \mathcal{D})] + \lambda_1 \xi - \lambda_2 KL(q(\zeta)q(\pi) || p(\zeta, \pi)) \quad (20)$$

$$\text{s.t.} \quad \mathbb{P}_q (\ell(\zeta \cdot \pi, \mathcal{D}) \leq \epsilon) \geq \xi \quad (21)$$

where $\mathbb{E}_q [U(\tau, r, \mathcal{D})]$ denotes the optimistic policy score to capture the upper bound of the policy value estimation. For example, the most widely used one is

$$\mathbb{E}_q [U(\tau, r, \mathcal{D})] = \mathbb{E}_q \hat{\mathbb{E}}_{\mathcal{D}} [\tau \cdot r] + \lambda_u \mathbb{E}_q \left[\left(\hat{\mathbb{E}}_{\mathcal{D}} [\tau \cdot r] - \mathbb{E}_q \hat{\mathbb{E}}_{\mathcal{D}} [\tau \cdot r] \right)^2 \right],$$

where the second term is the empirical variance and usually known as one kind of ‘‘exploration bonus’’.

Then the whole algorithm is iterating between solving (20) and use the obtain policy collecting data into \mathcal{D} in (20).

This Exploration-BayesDICE follows the same philosophy of Osband et al. (2019); ODonoghue et al. (2018) where the variance of posterior of the policy value is taken into account for exploration. However, there are several significant differences: **i)**, the first and most different is the modeling object, Osband et al. (2019); ODonoghue et al. (2018) is updating with Q -function, while we are handling the dual representation; **ii)**, BayesDICE is compatible with arbitrary nonlinear function approximator, while Osband et al. (2019); ODonoghue et al. (2018) considers tabular or linear functions; **iii)**, BayesDICE is considering infinite-horizon MDP, while Osband et al. (2019); ODonoghue et al. (2018) considers fixed finite-horizon case. Therefore, the exploration with BayesDICE pave the path for principle and practical exploration-vs-exploitation algorithm. The regret bound is out of the scope of this paper, and we leave for future work.

C EXPERIMENT DETAILS AND ADDITIONAL RESULTS

C.1 ENVIRONMENTS AND POLICIES.

Bandit. We create a two-armed bandit where α controls the proportion of optimal arm ($\alpha = 0$ and $\alpha = 1$ means never and always choosing the optimal arm respectively). Our selection experiments are based on 5 target policies with $\alpha = [0.75, 0.8, 0.85, 0.9, 0.95]$.

Reacher. We modify the Reacher task to be infinite horizon, and sample trajectories of length 100 in the behavior data. To obtain different behavior and target policies, We first train a deterministic policy from OpenAI Gym (Brockman et al., 2016) until convergence, and define various policies by converting the optimal policy into a Gaussian policy with optimal mean with standard deviation $0.4 - 0.3\alpha$. Our selection experiments are based on 5 target policies with $\alpha = [0.75, 0.8, 0.85, 0.9, 0.95]$.

C.2 DETAILS OF NEURAL NETWORK IMPLEMENTATION

We parametrize the distribution correction ratio as a Gaussian using a deep neural network for the continuous control task. Specifically, we use feed-forward networks with two hidden-layers of 64

neurons each and ReLU as the activation function. The networks are trained using the Adam optimizer ($\beta_1 = 0.99$, $\beta_2 = 0.999$) with batch size 2048.

C.3 ADDITIONAL EXPERIMENTAL RESULTS

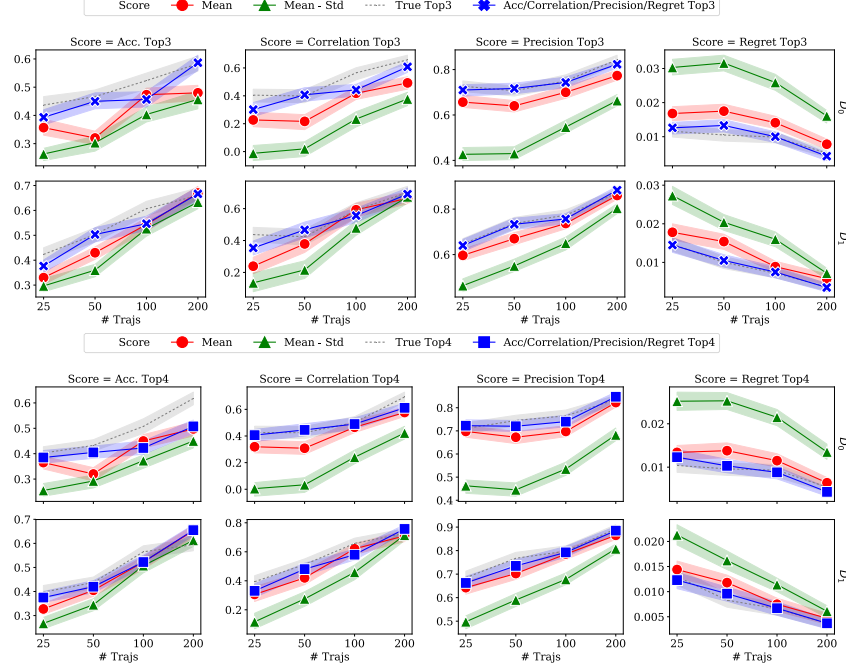


Figure 5: Additional k values for top- k ranking on bandit. Ranking results based on Algorithm 1 (blue lines) always perform better than using mean or high-confidence lower bound.

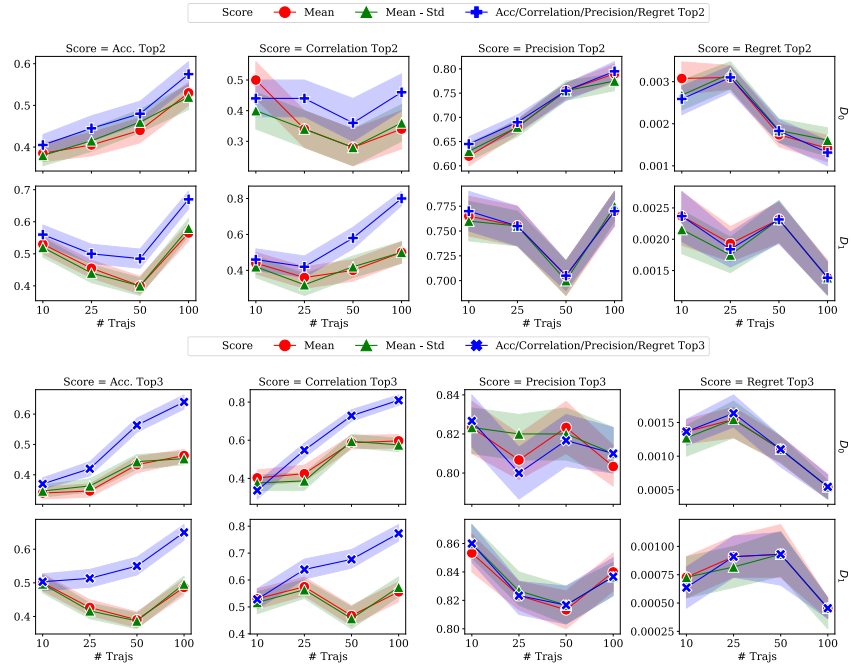


Figure 6: Additional k values for top- k ranking on reacher and additional scores (precision and regret). Ranking results based on Algorithm 1 (blue lines) generally perform much better than using mean or high-confidence lower bound for top- k accuracy and correlation. Precision and regret are similar between posterior samples and the mean/confidence bound based ranking.

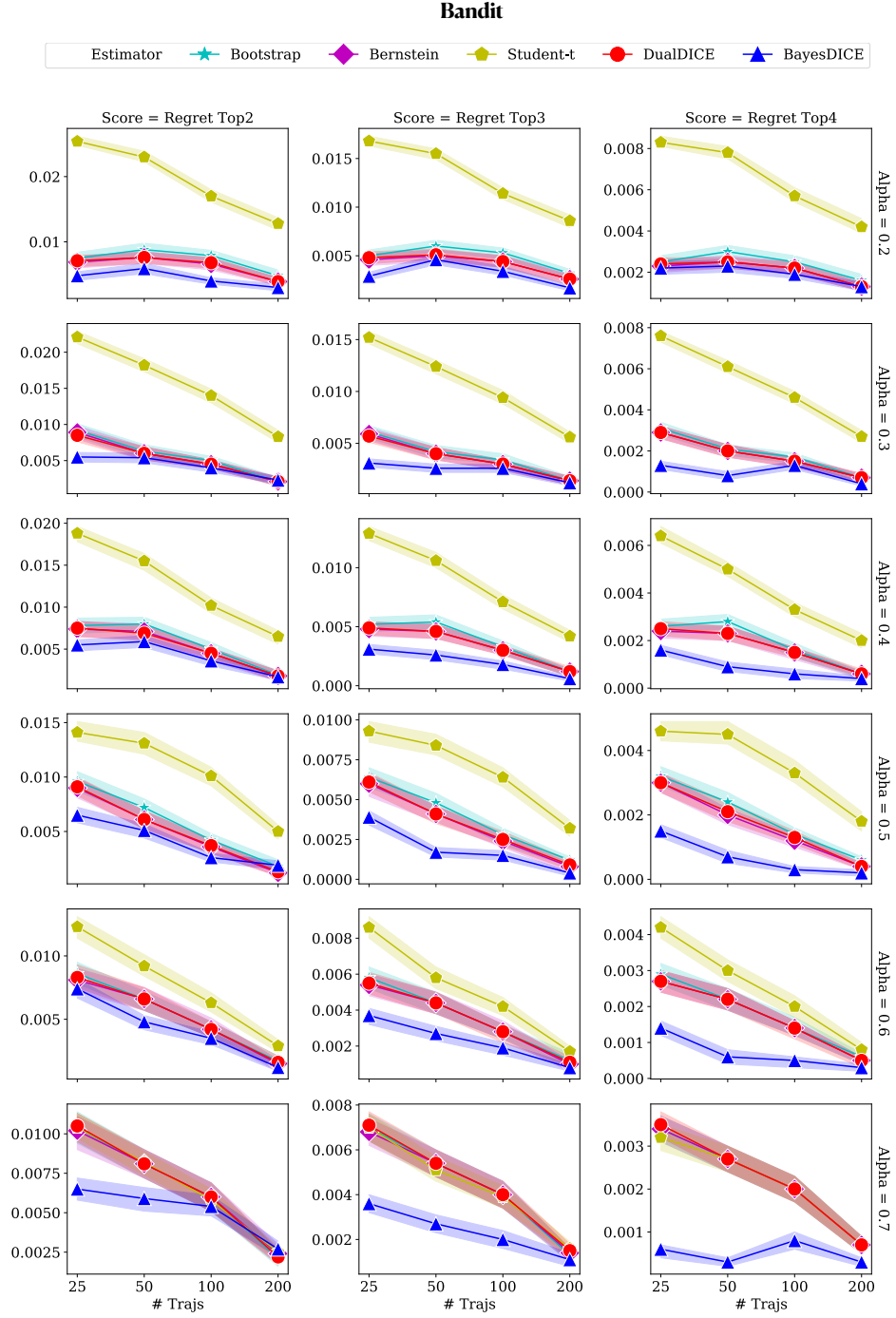


Figure 7: Improved regret using BayesDICE across all trajectory lengths, behavior data, and top- k values considered for the bandit task.

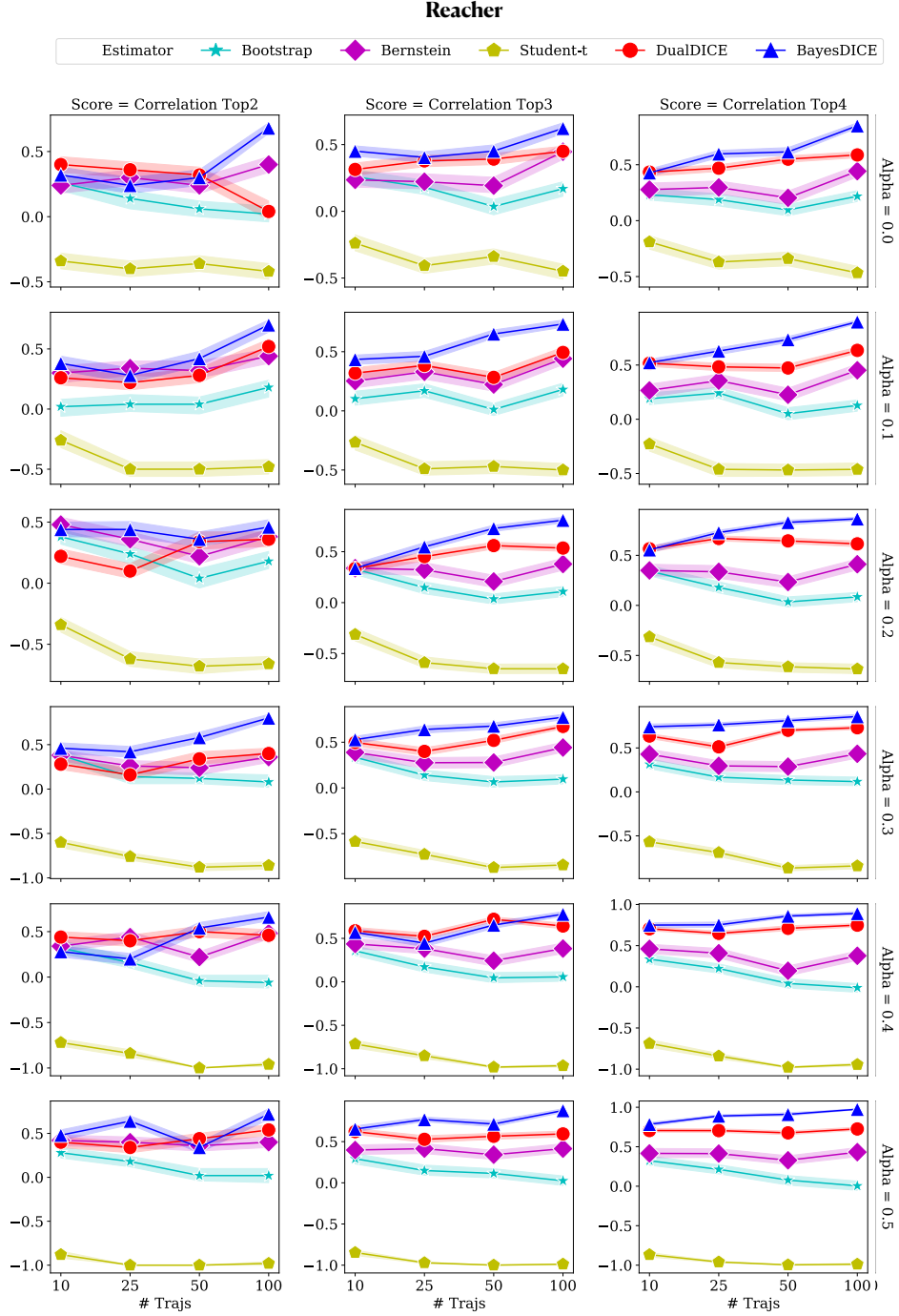


Figure 8: Improved correlation using BayesDICE across all trajectory lengths, behavior data, and top- k values considered for the reacher task.