A Appendix : Additional Discussions

A.1 Extended Related Work

Other related works : Several other prior works have previously considered the batch RL setting (Lange et al., 2012) for off-policy evaluation, counterfactual risk minimization (Swaminathan & Joachims, 2015a;b), learning value based methods such as DQN (Agarwal et al., 2019), and others (Kumar et al., 2019; Wu et al., 2019b). Recently, batch off-policy optimization has also been introduced to reduce the exploitation error (Fujimoto et al., 2019) and for regularizing with arbitrary behaviour policies (Wu et al., 2019b). However, due to the per-step importance sampling corrections on episodic returns (Precup et al., 2000), off-policy batch RL methods is challenging. In this work, we instead consider marginalized importance sampling corrections and correct for the stationary state-action distributions (Nachum et al., 2019a; Uehara & Jiang, 2019; Zhang et al., 2020). Additionally, under the framework of Constrained MDPs (Altman & Asingleutility, 1999), risk-sensitive and constrained actor-critic algorithms have been proposed previously (Chow et al., 2017; Chow & Ghavamzadeh, 2014; Achiam et al., 2017). However, these works come with their own demerits, as they mostly require minimizing the risk (ie, variance) term, where finding the gradient of the variance term often leads a double sampling issue (Baird, 1995). We avoid this by instead using Fenchel duality (Boyd & Vandenberghe, 2004), inspired from recent works (Nachum & Dai, 2020; Dai et al., 2018) and cast risk constrained actor-critic as a max-min optimization problem. Our work is closely related to (Bisi et al., 2019), which also consider per-step variance of returns, w.r.t state occupancy measures in the on-policy setting, while we instead consider the batch off-policy optimization setting with per-step rewards w.r.t stationary distribution corrections.

Constrained optimization has previously been studied in in reinforcement learning for batch policy learning (Le et al., 2019), and optimization (Achiam et al., 2017), mostly under the framework of constrained MDPs (Altman & Asingleutility, 1999). In such frameworks, the cumulative return objective is augmented with a set of constraints, for safe exploration (García et al., 2015; Perkins & Barto, 2003; Ding et al., 2020), or to reduce risk measures (Chow et al., 2017; A. & Fu, 2018; Castro et al., 2012). Batch learning algorithms (Lange et al., 2012) have been considered previously for counterfactual risk minimization and generalization (Swaminathan & Joachims, 2015a;b) and policy evaluation (Thomas et al., 2015a; Li et al., 2015), although little has been done for constrained offline policy based optimization. This raises the question of how can we learn policies in RL from fixed offline data, similar to supervised or unsupervised learning.

A.2 What makes Offline Off-Policy Optimization Difficult?

Offline RL optimization algorithms often suffer from distribution mismatch issues, since the underlying data distribution in the batch data may be quite different from the induced distribution under target policies. Recent works (Fujimoto et al., 2019; Kumar et al., 2019; Agarwal et al., 2019; Kumar et al., 2020) have tried to address this, by avoiding overestimation of Q-values, which leads to the extraplation error when bootstrapping value function estimates. This leads to offline RL agents generalizing poorly for unseen regions of the dataset. Additionally, due to the distribution mismatch, value function estimates can also have large variance, due to which existing online off-policy algorithms (Haarnoja et al., 2018; Lillicrap et al., 2016; Fujimoto et al., 2018) may fail without online interactions with the environment. In this work, we address the later problem to minimize variance of value function estimates through variance related risk constraints.

B Appendix : Per-Step versus Episodic Variance of Returns

Following from (Castro et al., 2012; A. & Ghavamzadeh, 2016), let us denote the returns with importance sampling corrections in the off-policy learning setting as :

$$D^{\pi}(s,a) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t) \Big(\prod_{t=1}^{T} \frac{\pi(a_t \mid s_t)}{\mu(a_t \mid s_t)} \Big) \mid s_0 = s, a_0 = a, \tau \sim \mu \tag{17}$$

From this definition in equation 17, the action-value function, with off-policy trajectory-wise importance correction is $Q^{\pi}(s,a) = \mathbb{E}_{(s,a)\sim d_{\mu}(s,a)}[D^{\pi}(s,a)]$, and similarly the value function can be defined as : $V^{\pi}(s) = \mathbb{E}_{s\sim d_{\mu}(s)}[D^{\pi}(s)]$. For the trajectory-wise importance corrections, we can define the variance of the returns, similar to (A. & Fu, 2018) as :

$$\mathcal{V}_{\mathcal{P}}(\pi) = \mathbb{E}_{(s,a)\sim d_{\mu}(s,a)}[D^{\pi}(s,a)^2] - \mathbb{E}_{(s,a)\sim d_{\mu}(s,a)}[D^{\pi}(s,a)]^2 \tag{18}$$

where note that as in (Sobel, 1982), equation 18 also follows a Bellman like equation, although due to lack of monotonocitiy as required for dynamic programming (DP), such measures cannot be directly optimized by standard DP algorithms (A. & Fu, 2018).

In contrast, if we consider the variance of returns with stationary distribution corrections (Nachum et al., 2019a; Liu et al., 2018), rather than the product of importance sampling ratios, the variance term involves weighting the rewards with the distribution ratio $\omega_{\pi/\mu}$. Typically, the distribution ratio is approximated using a separate function class (Uehara & Jiang, 2019), such that the variance can be written as :

$$W^{\pi}(s,a) = \omega_{\pi/\mathcal{D}}(s,a) \cdot r(s,a) \mid s = s, a \sim \pi(\cdot \mid s), (s,a) \sim d_{\mathcal{D}}(s,a) \tag{19}$$

where we denote $\mathcal{D}$ as the data distribution in the fixed dataset, collected by either a known or unknown behaviour policy. The variance of returns under occupancy measures is therefore given by :

$$\mathcal{V}_{\mathcal{D}}(\pi) = \mathbb{E}_{(s,a)\sim d_{\mathcal{D}}(s,a)}\Big[W^{\pi}(s,a)^2\Big] - \mathbb{E}_{(s,a)\sim d_{\mathcal{D}}(s,a)}\Big[W^{\pi}(s,a)\Big]^2 \tag{20}$$

where note that the variance expression in equation 20 depends on the square of the per-step rewards with distribution correction ratios. We denote this as the dual form of the variance of returns, in contrast to the primal form of the variance of expected returns (Sobel, 1982).

Note that even though the variance term under episodic per-step importance sampling corrections in equation 18 is equivalent to the variance with stationary distribution corrections in equation 20, following from (Bisi et al., 2019), considering per-step corrections, we will show that the variance with distribution corrections indeed upper bounds the variance of importance sampling corrections. This is an important relationship, since constraining the policy improvement step under variance constraints with occupancy measures therefore allows us to obtain a lower bound to the offline optimization objective, similar to (Kumar et al., 2020).

B.1 Proof of Lemma 1 : Variance Inequality

Following from (Bisi et al., 2019), we show that the variance of per-step rewards under occupancy measures, denoted by $\mathcal{V}_{\mathcal{D}}(\pi)$ upper bounds the variance of episodic returns $\mathcal{V}_{\mathcal{P}}(\pi)$.

$$\mathcal{V}_{\mathcal{P}}(\pi) \leq \frac{\mathcal{V}_{\mathcal{D}}(\pi)}{(1-\gamma)^2} \tag{21}$$

Proof. Proof of Lemma 1 following from (Bisi et al., 2019) is as follows. Denoting the returns, as above, but for the on-policy case with trajectories under π, as $D^{\pi}(s,a) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$, and denoting the return objective as $J(\pi) = \mathbb{E}_{s_0\sim\rho, a_t\sim\pi(\cdot|s_t), s'\sim\mathcal{P}}\Big[D^{\pi}(s,a)\Big]$, the variance of episodic returns can be written as :

$$\mathcal{V}_{\mathcal{P}}(\pi) = \mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}\Big[\Big(D^{\pi}(s,a) - \frac{J(\pi)}{(1-\gamma)}\Big)^2\Big] \tag{22}$$

$$= \mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}\Big[(D^{\pi}(s,a))^2\Big] + \frac{J(\pi)}{(1-\gamma)^2} - \frac{2J(\pi)}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}\Big[D^{\pi}(s,a)\Big] \tag{23}$$

$$= \mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}\Big[D^{\pi}(s,a)^2\Big] - \frac{J(\pi)^2}{(1-\gamma)^2} \tag{24}$$

Similarly, denoting returns under occupancy measures as $W^{\pi}(s,a) = d_{\pi}(s,a)r(s,a)$, and the returns under occupancy measures, equivalently written as $J(\pi) = \mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}[r(s,a)]$ based on the primal and dual forms of the objective (Uehara & Jiang, 2019; Nachum & Dai, 2020), we can equivalently write the variance as :

$$\mathcal{V}_{\mathcal{D}}(\pi) = \mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}\Big[\Big(r(s,a) - J(\pi)\Big)^2\Big] \tag{25}$$

$$= \mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}\Big[r(s,a)^2\Big] + J(\pi)^2 - 2J(\pi)\mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}[r(s,a)] \tag{26}$$

$$= \mathbb{E}_{(s,a)\sim d_{\pi}(s,a)}\Big[r(s,a)^2\Big] - J(\pi)^2 \tag{27}$$

Following from equation 22 and 25, we therefore have the following inequality :

$$(1-\gamma)^2\mathbb{E}_{s_0\sim\rho,a\sim\pi}\Big[D^\pi(s,a)^2\Big] \leq (1-\gamma)^2\mathbb{E}_{s_0\sim\rho,a\sim\pi}\Big[\Big(\sum_{t=0}^{\infty}\gamma^t\Big)\Big(\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)^2\Big)\Big] \tag{28}$$

$$= (1-\gamma)\mathbb{E}_{s_0\sim\rho,a\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)^2\Big] \tag{29}$$

$$= \mathbb{E}_{(s,a)\sim d_\pi(s,a)}\Big[r(s,a)^2\Big] \tag{30}$$

where the first line follows from Cauchy-Schwarz inequality. This concludes the proof. □

We can further extend lemma 1, for off-policy returns under stationary distribution corrections (ie, marginalized importance sampling) compared importance sampling. Recall that we denote the variance under stationary distribution corrections as :

$$\mathcal{V}_\mathcal{D}(\pi) = \mathbb{E}_{(s,a)\sim d_\mathcal{D}(s,a)}\Big[\Big(\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a) - J(\pi)\Big)^2\Big] \tag{31}$$

$$= \mathbb{E}_{(s,a)\sim d_\mathcal{D}(s,a)}\Big[\omega_{\pi/\mathcal{D}}(s,a)^2\cdot r(s,a)^2\Big] - J(\pi)^2 \tag{32}$$

where $J(\pi) = \mathbb{E}_{(s,a)\sim d_\mathcal{D}(s,a)}\Big[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)\Big]$. We denote the episodic returns with importance sampling corrections as : $D^\pi = \sum_{t=0}^{T}\gamma^t r_t \rho_{0:t}$. The variance, as denoted earlier is given by :

$$\mathcal{V}_\mathcal{P}(\pi) = \mathbb{E}_{(s,a)\sim d_\pi(s,a)}\Big[D^\pi(s,a)^2\Big] - \frac{J(\pi)^2}{(1-\gamma)^2} \tag{33}$$

We therefore have the following inequality

$$(1-\gamma)^2\mathbb{E}_{s_0\sim\rho,a\sim\pi}\Big[D^\pi(s,a)^2\Big] \leq (1-\gamma)^2\mathbb{E}_{s_0\sim\rho,a\sim\pi}\Big[\Big(\sum_{t=0}^{T}\gamma^t\Big)\Big(\sum_{t=0}^{T}\gamma^t r(s_t,a_t)^2\Big)\Big(\prod_{t=0}^{T}\frac{\pi(a_t|s_t)}{\mu_\mathcal{D}(a_t|s_t)}\Big)^2\Big]$$

$$= (1-\gamma)\mathbb{E}_{s_0\sim\rho,a\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)^2\Big(\prod_{t=0}^{T}\frac{\pi(a_t|s_t)}{\mu_\mathcal{D}(a_t|s_t)}\Big)^2\Big] \tag{34}$$

$$= \mathbb{E}_{(s,a)\sim d_\mathcal{D}(s,a)}\Big[\omega_{\pi/\mathcal{D}}(s,a)^2\cdot r(s,a)^2\Big] \tag{35}$$

which shows that lemma 1 also holds for off-policy returns with stationary distribution corrections.

B.2 Double Sampling for Computing Gradients of Variance

The gradient of the variance term often leads to the double sampling issue, thereby making it impractical to use. This issue has also been pointed out by several other works (A. & Ghavamzadeh, 2016; Castro et al., 2012; Chow et al., 2017), since the variance involves the squared of the objective function itself. Recall that we have:

$$\mathcal{V}_\mathcal{D}(\theta) = \mathbb{E}_{(s,a)\sim d_\mathcal{D}}\Big[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)^2\Big] - \Big\{\mathbb{E}_{(s,a)\sim d_\mathcal{D}}\Big[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)\Big]\Big\}^2 \tag{36}$$

The gradient of the variance term is therefore :

$$\begin{aligned}\nabla_\theta\mathcal{V}_\mathcal{D}(\theta) &= \nabla_\theta\mathbb{E}_{(s,a)\sim d_\mathcal{D}}\Big[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)^2\Big] \\ &- 2\cdot\Big\{\mathbb{E}_{(s,a)\sim d_\mathcal{D}}\Big[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)\Big]\Big\}\cdot\nabla_\theta\Big\{\mathbb{E}_{(s,a)\sim d_\mathcal{D}}\Big[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)\Big]\Big\}\end{aligned} \tag{37}$$

where equation 37 requires multiple samples to compute the expectations in the second term. To see why this is true, let us denote

$$J(\theta) = \mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\underbrace{\omega_{\pi/\mathcal{D}}(s,a)}\cdot r(s,a)_{\text{IS}(\omega,\pi_\theta)}\Big]$$

where we have $\text{IS}(\omega,\pi_\theta)$ as the returns in short form. The variance of the returns with the stationary state-action distribution corrections can therefore be written as :

$$\mathcal{V}_\mathcal{D}(\theta) = \underbrace{\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)^2\Big]}_{\text{(a)}} - \underbrace{\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)\Big]^2}_{(b)} \tag{38}$$

We derive the gradient of each of the terms in (a) and (b) in equation 38 below. First, we find the gradient of the variance term w.r.t θ :

$$\begin{aligned}\nabla_\theta \mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)^2\Big] &= \nabla_\theta \sum_{s,a} d_\mathcal{D}(s,a)\text{IS}(\omega,\pi_\theta)^2 = \sum_{s,a} d_\mathcal{D}(s,a)\nabla_\theta \text{IS}(\omega,\pi_\theta)^2 \\ &= \sum_{s,a} d_\mathcal{D}(s,a)\cdot 2\cdot \text{IS}(\omega,\pi_\theta)\cdot\text{IS}(\omega,\pi_\theta)\cdot\nabla_\theta\log\pi_\theta(a\mid s) \\ &= 2\cdot\sum_{s,a} d_\mathcal{D}(s,a)\text{IS}(\omega,\pi_\theta)^2\nabla_\theta\log\pi_\theta(a\mid s) \\ &= 2\cdot\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)^2\cdot\nabla_\theta\log\pi_\theta(a\mid s)\Big]\end{aligned} \tag{39}$$

Equation 39 interestingly shows that the variance of the returns w.r.t π_θ has a form similar to the policy gradient term, except the critic estimate in this case is given by the importance corrected returns, since $\text{IS}(\omega,\pi_\theta) = [\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)]$. We further find the gradient of term (b) from equation 38. Finding the gradient of this second term w.r.t θ is therefore :

$$\nabla_\theta\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)\Big]^2 = \nabla_\theta J(\theta)^2 = 2\cdot J(\theta)\cdot\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\omega_{\pi/\mathcal{D}}\cdot\{\nabla_\theta\log\pi_\theta(a\mid s)\cdot Q^\pi(s,a)\}\Big] \tag{40}$$

Overall, the expression for the gradient of the variance term is therefore :

$$\begin{aligned}\nabla_\theta\mathcal{V}_\mathcal{D}(\theta) = 2\cdot\mathbb{E}_{d_\mathcal{D}(s,a)}&\Big[\text{IS}(\omega,\pi_\theta)^2\cdot\nabla_\theta\log\pi_\theta(a\mid s)\Big] \\ &- 2\cdot J(\theta)\cdot\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\omega_{\pi/\mathcal{D}}\cdot\{\nabla_\theta\log\pi_\theta(a\mid s)\cdot Q^\pi(s,a)\}\Big]\end{aligned} \tag{41}$$

The variance gradient in equation 41 is difficult to estimate in practice, since it involves both the gradient of the objective and the objective $J(\theta)$ itself. This is known to have the double sampling issue (Baird, 1995) which requires separate independent rollouts. Previously, (Castro et al., 2012) tackled the variance of the gradient term using simultaneous perturbation stochastic approximation (SPSA) (Spall, 1992), where we can keep running estimates of both the return and the variance term, and use a two time scale algorithm for computing the gradient of the variance regularizer with per-step importance sampling corrections.

B.3 Alternative Derivation : Variance Regularization via Fenchel Duality

In the derivation of our algorithm, we applied the Fenchel duality trick to the second term of the variance expression 25. An alternative way to derive the proposed algorithm would be to see what happens if we apply the Fenchel duality trick to both terms of the variance expression. This might be useful since equation 41 requires evaluating both the gradient terms and the actual objective $J(\theta)$, due to the analytical expression of the form $\nabla_\theta J(\theta)\cdot J(\theta)$, hence suffering from a double sampling issue. In general, the Fenchel duality is given by :

$$x^2 = \max_y(2xy - y^2) \tag{42}$$

and applying Fenchel duality to both the terms, since they both involve squared terms, we get :

$$\begin{aligned}\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)^2\Big] &\equiv \mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\max_y\Big\{2\cdot\text{IS}(\omega,\pi_\theta)\cdot y(s,a) - y(s,a)^2\Big\}\Big] \\ &= 2\cdot\max_y\Big\{\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)\cdot y(s,a)\Big] - \mathbb{E}_{d_\mathcal{D}(s,a)}\Big[y(s,a)^2\Big]\Big\}\end{aligned} \tag{43}$$

Similarly, applying Fenchel duality to the second (b) term we have :

$$\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)\Big]^2 = \max_\nu\Big\{2\cdot\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)\cdot\nu(s,a)\Big] - \nu^2\Big\} \tag{44}$$

Overall, we therefore have the variance term, after applying Fenchel duality as follows, leading to an overall objective in the form $\max_y \max_\nu \mathcal{V}_\mathcal{D}(\theta)$, which we can use as our variance regularizer

$$\begin{aligned}\mathcal{V}_\mathcal{D}(\theta) = 2\cdot\max_y&\Big\{\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)\cdot y(s,a)\Big] - \mathbb{E}_{d_\mathcal{D}(s,a)}\Big[y(s,a)^2\Big]\Big\} \\ - \max_\nu&\Big\{2\cdot\mathbb{E}_{d_\mathcal{D}(s,a)}\Big[\text{IS}(\omega,\pi_\theta)\cdot\nu(s,a)\Big] - \nu^2\Big\}\end{aligned} \tag{45}$$

Using the variance of stationary distribution correction returns as a regularizer, we can find the gradient of the variance term w.r.t θ as follows, where the gradient terms dependent on the dual variables y and ν are 0.

$$\nabla_\theta \mathcal{V}_\mathcal{D}(\theta) = 2 \cdot \nabla_\theta \mathbb{E}_{d_\mathcal{D}(s,a)} \Big[\text{IS}(\omega, \pi_\theta) \cdot y(s,a) \Big] - 0 - 2 \cdot \nabla_\theta \mathbb{E}_{d_\mathcal{D}(s,a)} \Big[\text{IS}(\omega, \pi_\theta) \cdot \nu(s,a) \Big] + 0$$

$$= 2 \cdot \mathbb{E}_{d_\mathcal{D}(s,a)} \Big[\text{IS}(\omega, \pi_\theta) \cdot y(s,a) \cdot \nabla_\theta \log \pi_\theta(a \mid s) \Big] - 2 \cdot \mathbb{E}_{d_\mathcal{D}(s,a)} \Big[\text{IS}(\omega, \pi_\theta) \cdot \nu(s,a) \cdot \nabla_\theta \log \pi_\theta(a \mid s) \Big]$$

$$= 2 \cdot \mathbb{E}_{d_\mathcal{D}(s,a)} \Bigg[\text{IS}(\omega, \pi_\theta) \cdot \nabla_\theta \log \pi_\theta(a \mid s) \cdot \Big\{ y(s,a) - \nu(s,a) \Big\} \Bigg] \tag{46}$$

Note that from equation 46, the two terms in the gradient is almost equivalent, and the difference comes only from the difference between the two dual variables $y(s,a)$ and $\nu(s,a)$. Note that our variance term also requires separately maximizing the dual variables, both of which has the following closed form updates :

$$\nabla_\nu \mathcal{V}_\mathcal{D}(\theta) = -2 \cdot \nabla_\nu \mathbb{E}_{d_\mathcal{D}(s,a)} \Big[\text{IS}(\omega, \pi_\theta) \cdot \nu(s,a) \Big] + \nabla_\nu \nu^2 = 0 \tag{47}$$

Solving which exactly, leads to the closed form solution $\nu(s,a) = \mathbb{E}_{d_\mathcal{D}(s,a)} \Big[\text{IS}(\omega, \pi_\theta) \Big]$. Similarly, we can also solve exactly using a closed form solution for the dual variables y, such that :

$$\nabla_y \mathcal{V}_\mathcal{D}(\theta) = 2 \cdot \nabla_y \mathbb{E}_{d_\mathcal{D}(s,a)} \Big[\text{IS}(\omega, \pi_\theta) \cdot y(s,a) \Big] - 2 \cdot \nabla_y \mathbb{E}_{d_\mathcal{D}(s,a)} \Big[y(s,a)^2 \Big] = 0 \tag{48}$$

Solving which exactly also leads to the closed form solution, such that $y(s,a) = \frac{1}{2} \cdot \text{IS}(\omega, \pi_\theta) = \frac{1}{2} \cdot \frac{d_\pi(s,a)}{d_\mu(s,a)} \cdot r(s,a)$. Note that the exact solutions for the two dual variables are similar to each other, where $\nu(s,a)$ is the expectation of the returns with stationary distribution corrections, whereas $y(s,a)$ is only the return from a single rollout.

C Appendix : Monotonic Performance Improvement Guarantees under Variance Regularization

We provide theoretical analysis and performance improvements bounds for our proposed variance constrained policy optimization approach. Following from (Kakade & Langford, 2002; Schulman et al., 2015; Achiam et al., 2017), we extend existing performance improvement guarantees based on the stationary state-action distributions instead of only considering the divergence between the current policy and old policy. We show that existing conservative updates in algorithms (Schulman et al., 2015) can be considered for both state visitation distributions and the action distributions, as similarly pointed by (Achiam et al., 2017). We can then adapt this for the variance constraints instead of the divergence constraints. According to the performance difference lemma (Kakade & Langford, 2002), we have that, for all policies π and π' :

$$J(\pi') - J(\pi) = \mathbb{E}_{s \sim d_{\pi'}, a \sim \pi'} [A^\pi(s,a)] \tag{49}$$

which implies that when we maximize 49, it will lead to an improved policy π' with policy improvement guarantees over the previous policy π. We can write the advantage function with variance augmented value functions as :

$$A^\pi_\lambda = Q^\pi_\lambda(s,a) - V^\pi_\lambda(s) = \mathbb{E}_{s' \sim \mathcal{P}} \Big[r(s,a) - \lambda(r(s,a) - J(\pi))^2 + \gamma V^\pi_\lambda(s') - V^\pi_\lambda(s) \Big]$$

However, equation 49 is often difficult to maximize directly, since it additionally requires samples from π' and $d_{\pi'}$, and often a surrogate objective is instead proposed by (Kakade & Langford, 2002). Following (Schulman et al., 2015), we can therefore obtain a bound for the performance difference based on the variance regularized advantage function :

$$J(\pi') \geq J(\pi) + \mathbb{E}_{s \sim d_\pi(s), a \sim \pi'(a|s)} \Big[A^\pi_\lambda(s,a) \Big] \tag{50}$$

where we have the augmented rewards for the advantage function, and by following Fenchel duality for the variance, can avoid policy dependent reward functions. Otherwise, we have the augmented rewards for value functions as $\tilde{r}(s,a) = r(s,a) - \lambda(r(s,a) - J(\pi))^2$. This however suggests that the performance difference does not hold without proper assumptions (Bisi et al., 2019). We can therefore obtain a monotonic improvement guarantee by considering the KL divergence between policies :

$$\mathcal{L}_{\pi}(\pi') = J(\pi) + \mathbb{E}_{s\sim d_{\pi}, a\sim\pi'}[A^{\pi}(s,a)] \tag{51}$$

which ignores the changes in the state distribution $d_{\pi'}$ due to the improved policy π'. (Schulman et al., 2015) optimizes the surrogate objectives $\mathcal{L}_{\pi}(\pi')$ while ensuring that the new policy π' stays close to the current policy π, by imposing a KL constraint ($\mathbb{E}_{s\sim d_{\pi}}[\mathcal{D}_{\mathrm{KL}}(\pi'(\cdot \mid s)||\pi(\cdot \mid s)] \leq \delta$). The performance difference bound, based on the constraint between π and π' as in TRPO (Schulman et al., 2015) is given by :

Lemma 4. *The performance difference lemma in (Schulman et al., 2015), where* $\alpha = \mathcal{D}_{TV}^{max} = \max_s \mathcal{D}_{TV}(\pi, \pi')$

$$J(\pi') \geq \mathcal{L}_{\pi}(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2}(\mathcal{D}_{TV}^{max}(\pi'||\pi))^2 \tag{52}$$

where $\epsilon = \max_{s,a}|A^{\pi}(s,a)|$*, which is usually denoted with* α*, where*

The performance improvement bxound in (Schulman et al., 2015) can further be written in terms of the KL divergence by following the relationship between total divergence (TV) and KL, which follows from Pinsker's inequality, $\mathcal{D}_{\mathrm{TV}}(p||q)^2 \leq \mathcal{D}_{\mathrm{KL}}(p||q)$, to get the following improvement bound :

$$J(\pi') \geq \mathcal{L}_{\pi}(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2}\mathcal{D}_{\mathrm{KL}}(\pi'||\pi) \tag{53}$$

We have a performance difference bound in terms of the state distribution shift $d_{\pi'}$ and d_{π}. This justifies that $\mathcal{L}_{\pi}(\pi')$ is a sensible lower bound to $J(\pi')$ as long as there is a total variation distance between $d_{\pi'}$ and d_{π} which ensures that the policies π' and π stay close to each other. Finally, following from (Achiam et al., 2017), we obtain the following lower bound, which satisfies policy improvement guarantees :

$$J(\pi') \geq \mathcal{L}_{\pi}(\pi') - \frac{2\gamma\epsilon^{\pi}}{1-\gamma}\mathbb{E}_{s\sim d_{\pi}}[\mathcal{D}_{\mathrm{TV}}(\pi'(\cdot \mid s)||\pi(\cdot \mid s))] \tag{54}$$

Equation 53 and 54 assumes that there is no state distribution shift between π' and π. However, if we explicitly assume state distribution changes, $d_{\pi'}$ and d_{π} due to π' and π respectively, then we have the following performance improvement bound :

Lemma 5. *For all policies* π' *and* π*, we have the performance improvement bound based on the total variation of the state-action distributions* $d_{\pi'}$ *and* d_{π}

$$J(\pi') \geq \mathcal{L}_{\pi}(\pi') - \epsilon^{\pi}\mathcal{D}_{TV}(d_{\pi'}||d_{\pi}) \tag{55}$$

where $\epsilon^{\pi} = \max_s |\mathbb{E}_{a\sim\pi'(\cdot|s)}[A^{\pi}(s,a)]|$

which can be further written in terms of the surrogate objective $\mathcal{L}_{\pi}(\pi')$ as :

$$\begin{aligned} J(\pi') &\geq J(\pi) + \mathbb{E}_{s\sim d_{\pi}, a\sim\pi'}[A^{\pi}(s,a)] - \epsilon^{\pi}\mathcal{D}_{\mathrm{TV}}(d_{\pi'}||d_{\pi}) \\ &= \mathcal{L}_{\pi}(\pi') - \epsilon^{\pi}\mathcal{D}_{\mathrm{TV}}(d_{\pi'}||d_{\pi}) \end{aligned} \tag{56}$$

C.1 Proof of Theorem 1 : Policy Improvement Bound with Variance Regularization

Proof. We provide derivation for theorem 1. Recall that for all policies π' and π, and corresponding state visitation distributions $d_{\pi'}$ and d_{π}, we can obtain the performance improvement bound in terms of the variance of state-action distribution corrections

$$J(\pi') - J(\pi) \geq \mathbb{E}_{s\sim d_{\pi}, a\sim\pi'}\Big[A^{\pi}(s,a)\Big] - \mathrm{Var}_{s\sim d_{\pi}, a\sim\pi}\Big[f(s,a)\Big] \tag{57}$$

where $f(s,a)$ is the dual function class, for the divergence between $d_{\pi'}(s,a)$ and $d_{\pi}(s,a)$ Following from Pinsker's inequality, the performance difference lemma written in terms of the state visitation distributions can be given by :

$$\begin{aligned} J(\pi') &\geq \mathcal{L}_{\pi}(\pi') - \epsilon^{\pi}\mathcal{D}_{\mathrm{TV}}(d_{\pi'}||d_{\pi}) \\ &\geq J(\pi) + \mathbb{E}_{s\sim d_{\pi}, a\sim\pi'}[A^{\pi}(s,a)] - \epsilon^{\pi}\mathcal{D}_{\mathrm{TV}}(d_{\pi'}||d_{\pi}) \\ &\geq J(\pi) + \mathbb{E}_{s\sim d_{\pi}, a\sim\pi'}[A^{\pi}(s,a)] - \epsilon^{\pi}\sqrt{\mathcal{D}_{\mathrm{KL}}(d_{\pi'}||d_{\pi})} \end{aligned} \tag{58}$$

Following from (Schulman et al., 2015), we can alternately write this follows, where we further apply the variational form of TV

$$
\begin{aligned}
&J(\pi') \geq J(\pi) + \mathbb{E}_{s\sim d_\pi, a\sim\pi'}\Big[A^\pi(s,a)\Big] - C \cdot \mathbb{E}_{s\sim d_\pi}\Big[\mathcal{D}_{\text{TV}}(d_{\pi'}||d_\pi)^2\Big] \\
&= J(\pi) + \mathbb{E}_{s\sim d_\pi, a\sim\pi'}\Big[A^\pi(s,a)\Big] - C \cdot \mathbb{E}_{s\sim d_\pi}\Big[\Big(\max_f\{\mathbb{E}_{s\sim d_{\pi'}, a\sim\pi}[f(s,a)] - \mathbb{E}_{s\sim d_\pi, a\sim\pi}[f(s,a)]\}\Big)^2\Big] \\
&\geq J(\pi) + \mathbb{E}_{s\sim d_\pi, a\sim\pi'}\Big[A^\pi(s,a)\Big] - C \cdot \max_f \mathbb{E}_{s\sim d_\pi}\Big[\Big(\mathbb{E}_{s\sim d_{\pi'}, a\sim\pi}[f(s,a)] - \mathbb{E}_{s\sim d_\pi, a\sim\pi}[f(s,a)]\Big)^2\Big] \\
&= J(\pi) + \mathbb{E}_{s\sim d_\pi, a\sim\pi'}\Big[A^\pi(s,a)\Big] - C \cdot \max_f \Big\{\Big(\mathbb{E}_{s\sim d_\pi, a\sim\pi}[f(s,a)] - \mathbb{E}_{s\sim d_\pi, a\sim\pi}[\mathbb{E}_{s\sim d_\pi, a\sim\pi}[f(s,a)]]\Big)^2\Big\} \\
&= J(\pi) + \mathbb{E}_{s\sim d_\pi, a\sim\pi'}\Big[A^\pi(s,a)\Big] - C \cdot \max_f \text{Var}_{s\sim d_\pi, a\sim\pi}\Big[f(s,a)\Big] \qquad (59)
\end{aligned}
$$

Therefore the policy improvement bound depends on maximizing the variational representation $f(s,a)$ of the f-divergence to guaranetee improvements from $J(\pi)$ to $J(\pi')$. This therefore leads to the stated result in theorem 1. □

D Appendix : Lower Bound Objective with Variance Regularization

D.1 Proof of Lemma 3

Recalling lemma 3 which states that, the proof of this follows from (Metelli et al., 2018). We extend this for marginalized importance weighting, and include here for completeness. Note that compared to importance weighting, which leads to an unbiased estimator as in (Metelli et al., 2018), correcting for the state-action occupancy measures leads to a biased estimator, due to the approximation $\hat{\omega}_{\pi/\mathcal{D}}$. However, for our analysis, we only require to show a lower bound objective, and therefore do not provide any bias variance analysis as in off-policy evaluation.

$$
\text{Var}_{(s,a)\sim d_{\mathcal{D}}(s,a)}\Big[\hat{\omega}_{\pi/\mathcal{D}}\Big] \leq \frac{1}{N}||r||_\infty^2 \mathcal{F}_2(d_\pi||d_{\mathcal{D}}) \qquad (60)
$$

Proof. Assuming that state action samples are drawn i.i.d from the dataset $\mathcal{D}$, we can write :

$$
\begin{aligned}
&\text{Var}_{(s,a)\sim d_{\mathcal{D}}(s,a)}\Big[\hat{\omega}_{\pi/\mathcal{D}}(s,a)\Big] \leq \frac{1}{N}\text{Var}_{(s_1,a_1)\sim d_{\mathcal{D}}(s,a)}\Big[\frac{d_\pi(s_1,a_1)}{d_{\mathcal{D}}(s_1,a_1)} \cdot r(s_1,a_1)\Big] \\
&\leq \frac{1}{N}\mathbb{E}_{(s_1,a_1)\sim d_{\mathcal{D}}(s,a)}\Big[\Big(\frac{d_\pi(s_1,a_1)}{d_{\mathcal{D}}(s_1,a_1)} \cdot r(s_1,a_1)\Big)^2\Big] \\
&\leq \frac{1}{N}||r||_\infty^2 \mathbb{E}_{(s_1,a_1)\sim d_{\mathcal{D}}(s,a)}\Big[\Big(\frac{d_\pi(s_1,a_1)}{d_{\mathcal{D}}(s_1,a_1)} \cdot r(s_1,a_1)\Big)^2\Big] = \frac{1}{N}||r||_\infty^2 \mathcal{F}_2(d_\pi||d_{\mathcal{D}}) \qquad (61)
\end{aligned}
$$

□

D.2 Proof of Theorem 2:

First let us recall the stated theorem 2. By constraining the off-policy optimization problem with variance constraints, we have the following lower bound to the optimization objective with stationary state-action distribution corrections

$$
J(\pi) \geq \mathbb{E}_{(s,a)\sim d_{\mathcal{D}}(s,a)}[\frac{d_\pi(s,a)}{d_{\mathcal{D}}(s,a)} r(s,a)] - \sqrt{\frac{1-\delta}{\delta}\text{Var}_{(s,a)\sim d_\mu(s,a)}[\frac{d_\pi(s,a)}{d_{\mathcal{D}}(s,a)} r(s,a)]} \qquad (62)
$$

Proof. The proof for the lower bound objective can be obtained as follows. We first define a relationship between the variance and the α-divergence with $\alpha = 2$, as also similarly noted in (Metelli et al., 2018). Given we have batch samples $\mathcal{D}$, and denoting the state-action distribution correction with $\omega_{\pi/\mathcal{D}}(s,a)$, we can write from lemma 3 :

$$
\text{Var}_{(s,a)\sim d_{\mathcal{D}}(s,a)}\Big[\hat{\omega}_{\pi/\mathcal{D}}\Big] \leq \frac{1}{N}||r||_\infty^2 \mathcal{F}_2(d_\pi||d_{\mathcal{D}}) \qquad (63)
$$

where the per-step estimator with state-action distribution corrections is given by $\omega_{\pi/\mathcal{D}}(s,a) \cdot r(s,a)$. Here, the reward function $r(s,a)$ is a bounded function, and for any $N > 0$ the variance of the per-step reward estimator with distribution corrections can be upper bounded by the Renyi-divergence ($\alpha = 2$). Finally, following from (Metelli et al., 2018) and using Cantelli's inequality, we have with probability at least $1 - \delta$ where $0 < \delta < 1$:

$$\Pr\Big(\omega_{\pi/\mathcal{D}} - J(\pi) \geq \lambda\Big) \leq \frac{1}{1 + \frac{\lambda^2}{\text{Var}_{(s,a)\sim d_{\mathcal{D}}(s,a)}[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)]}} \tag{64}$$

and by using $\delta = \frac{1}{1+\frac{\lambda^2}{\text{Var}_{(s,a)\sim d_{\mathcal{D}}(s,a)}[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)]}}$ we get that with probability at least $1-\delta$, we have:

$$J(\pi) = \mathbb{E}_{(s,a)\sim d_\pi(s,a)} \geq \mathbb{E}_{(s,a)\sim d_{\mathcal{D}}(s,a)}[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)] - \sqrt{\frac{1-\delta}{\delta}\text{Var}_{(s,a)\sim d_{\mathcal{D}}(s,a)}[\omega_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)]} \tag{65}$$

where we can further replace the variance term with $\alpha = 2$ for the Renyi divergence to conclude the proof for the above theorem. We can further write the lower bound for for α-Renyi divergence, following the relation between variance and Renyi-divergence for $\alpha = 2$ as :

$$J(\pi) = \mathbb{E}_{(s,a)\sim d_\pi(s,a)}[r(s,a)] \geq \mathbb{E}_{(s,a)\sim d_{\mathcal{D}}(s,a)}[\frac{d_\pi(s,a)}{d_{\mathcal{D}}(s,a)}\cdot r(s,a)] - ||r||_\infty\sqrt{\frac{(1-\delta)d_2(d_\pi||d_{\mathcal{D}})}{\delta N}}$$

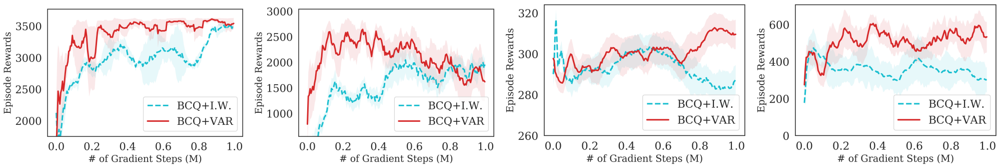
This hints at the similarity between our proposed variance regularized objective with that of other related works including AlgaeDICE (Nachum et al., 2019b) which uses a f-divergence $D_{f(d_\pi||d_{\mathcal{D}})}$ between stationary distributions. □

E Appendix : Additional Experimental Results

E.1 Experimental Ablation Studies

In this section, we present additional results using state-action experience replay weightings on existing offline algorithms, and analysing the significance of our variance regularizer on likelihood corrected offline algorithms. Denoting $\omega(s,a)$ for the importance weighting of state-action occupancy measures based on samples in the experience replay buffer, we can modify existing offline algorithms to account for state-action distribution ratios.

The ablation experimental results using the Hopper control benchmark are summarized in figure 2. The same base BCQ algorithm is used with a modified objective for BCQ (Fujimoto et al., 2019) where the results for applying off-policy importance weights are denoted as "BCQ+I.W.". We employ the same technique to obtain $\omega(s,a)$ for both the baseline and for adding variance regularization as described. The results suggest that adding the proposed per-step variance regularization scheme significantly outperforms just importance weighting the expected rewards for off-policy policy learning.



(a) Hopper Expert ablation (b) Hopper Medium ablation (c) Hopper Random ablation (d) Hopper Mixed ablation

Figure 2: Ablation performed on Hopper. The mean and standard deviation are reported over 5 random seeds. The offline datasets for these experiments are same as the corresponding ones in Fig 1 of the main paper.

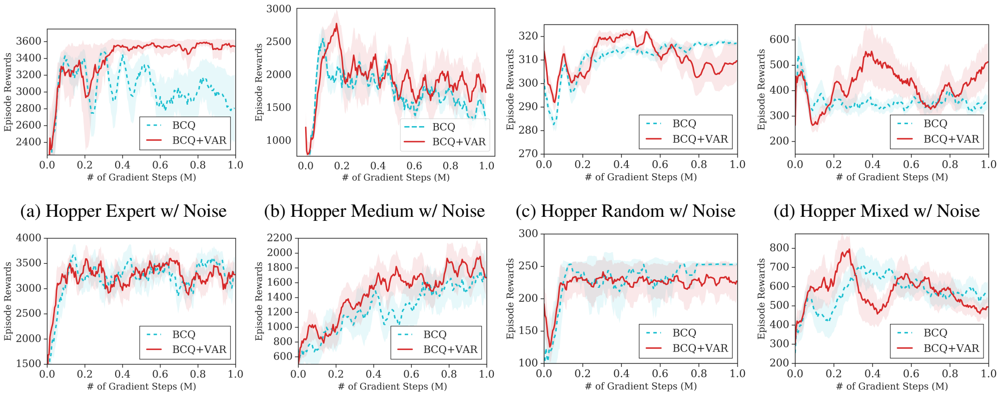
E.2 Experimental Results in Corrupted Noise Settings

We additionally consider a setting where the batch data is collected from a noisy environment, i.e, in settings with *corrupted rewards*, $r \rightarrow r + \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$. Experimental results are presented in figures 1, 3. From our results, we note that using OVR on top of BCQ (Fujimoto et al., 2019), we can achieve significantly better performance with variance minimization, especially when the agent is given sub-optimal demonstrations. We denote it as *medium* (when the dataset was collected by a half trained SAC policy) or a *mixed* behaviour logging setting (when the data logging policy is a mixture of random and SAC policy). This is also useful for practical scalability, since often data collection is

Domain	Task Name	BCQ+OVR	BCQ	BEAR	BRAC-p	aDICE	SAC-off
Adroit	pen-human	**64.12**	56.58	-1	8.1	-3.3	6.3
	hammer-human	**1.05**	0.75	0.3	0.3	0.3	0.5
	door-human	0.00	0.00	-0.3	-0.3	0	**3.9**
	relocate-human	-0.13	**-0.08**	-0.3	-0.3	-0.1	0
	pen-cloned	40.84	**41.09**	26.5	1.6	-2.9	23.5
	hammer-cloned	**0.78**	0.35	0.3	0.3	0.3	0.2
	door-cloned	**0.03**	**0.03**	-0.1	-0.1	0	0
	relocate-cloned	-0.22	-0.26	-0.3	-0.3	-0.3	**-0.2**
	pen-expert	**99.32**	89.42	**105.9**	-3.5	-3.5	6.1
	hammer-expert	**119.32**	108.38	**127.3**	0.3	0.3	25.2
	door-expert	**100.39**	101.33	**103.4**	-0.3	0	7.5
	relocate-expert	31.31	23.55	**98.6**	-0.3	-0.1	-0.3

Table 2: The results on D4RL tasks compare BCQ (Fujimoto et al., 2019) with and without OVR, bootstrapping error reduction (BEAR) (Kumar et al., 2019), behavior-regularized actor critic with policy (BRAC-p) (**?**), AlgeaDICE (aDICE) (Nachum et al., 2019b) and offline SAC (SAC-off) (Haarnoja et al., 2018). The results presented are the normalized returns on the task as per Fu et al. (2020) (see Table 3 in Fu et al. (2020) for the unnormalized scores on each task).

expensive from an expert policy. We add noise to the dataset, to examine the significance of OVR under a noisy corrupted dataset setting.



(a) Hopper Expert w/ Noise (b) Hopper Medium w/ Noise (c) Hopper Random w/ Noise (d) Hopper Mixed w/ Noise

(e) Walker Expert w/ Noise (f) Walker Medium w/ Noise (g) Walker Random w/ Noise (h) Walker Mixed w/ Noise

Figure 3: Evaluation of the proposed approach and the baseline BCQ on a suite of three OpenAI Gym environments. We consider the setting of rewards that are corrupted by a Gaussian noise. Results for the uncorrupted version are in Fig. 1. Experiment results are averaged over 5 random seeds

E.3 Experimental Results on Safety Benchmark Tasks

Safety Benchmarks for Variance as Risk : We additionally consider safety benchmarks for control tasks, to analyse the significance of variance regularizer as a risk constraint in offline policy optimization algorithms. Our results are summarized in table 3.

E.4 Discussions on Offline Off-Policy Optimization with State-Action Distribution Ratios

In this section, we include several alternatives by which we can compute the stationary state-action distribution ratio, borrowing from recent works (Uehara & Jiang, 2019; Nachum et al., 2019a).

Off-Policy Optimization with Minimax Weight Learning (MWL) : We discuss other possible ways of optimizing the batch off-policy optimization objective while also estimating the state-action density ratio. Following from (Uehara & Jiang, 2019) we further modify the off-policy optimization part of the objective $J(\theta)$ in $\mathcal{L}(\theta, \lambda)$ as a min-max objective, consisting of weight learning $\omega_{\pi/\mathcal{D}}$

Table 3: Results on the Safety-Gym environments Ray et al.. We report the mean and S.D. of episodic returns and costs over five random seeds and 1 million timesteps. The goal of the agent is to maximize the episodic return, while minimizing the cost incurred.

	PointGoal1		PointGoal2	
	Reward	Cost	Reward	Cost
BCQ	43.1 ± 0.3	137.0 ± 3.6	32.7± 0.7	468.2 ± 9.1
BCQ+OVR	**44.2** ± 0.3	**127.1** ± 4.0	**33.2** ± 0.7	**453.9** ± 7.3

	PointButton1		PointButton2	
	Reward	Cost	Reward	Cost
BCQ	**30.9** ± 2.2	330.8 ± 8.3	18.1 ± 1.1	321.6 ± 4.1
BCQ+OVR	30.7 ± 2.3	**321.5** ± 6.8	**19.6** ± 1.0	**305.7** ± 6.1

and optimizing the resulting objective $J(\theta, \omega)$. We further propose an overall policy optimization objective, where a single objective can be used for estimating the distribution ratio, evaluating the critic and optimizing the resulting objective. We can write the off-policy optimization objective with its equivalent starting state formulation, such that we have :

$$\mathbb{E}_{d_{\mathcal{D}}(s,a)}\Big[\omega_{\pi_\theta/\mathcal{D}}(s,a)\cdot r(s,a)\Big] = (1-\gamma)\mathbb{E}_{s_0\sim\beta_0(s),a_0\sim\pi(\cdot|s_0)}\Big[Q^\pi(s_0,a_0)\Big] \tag{66}$$

Furthermore, following Bellman equation, we expect to have $\mathbb{E}[r(s,a)] = \mathbb{E}[Q^\pi(s,a) - \gamma Q^\pi(s',a')]$

$$\mathbb{E}_{d_{\mathcal{D}}(s,a)}\Big[\omega_{\pi_\theta/\mathcal{D}}(s,a)\cdot\{Q^\pi(s,a)-\gamma Q^\pi(s',a')\}\Big] = (1-\gamma)\mathbb{E}_{s_0\sim\beta_0(s),a_0\sim\pi(\cdot|s_0)}\Big[Q^\pi(s_0,a_0)\Big] \tag{67}$$

We can therefore write the overall objective as :

$$\begin{aligned} J(\omega,\pi_\theta,Q) = \mathbb{E}_{d_{\mathcal{D}}(s,a)}\Big[\omega_{\pi_\theta/\mathcal{D}}(s,a)\cdot\{Q^\pi(s,a) - \gamma Q^\pi(s',a')\}\Big] \\ - (1-\gamma)\mathbb{E}_{s_0\sim\beta_0(s),a_0\sim\pi(\cdot|s_0)}\Big[Q^\pi(s_0,a_0)\Big] \end{aligned} \tag{68}$$

This is similar to the MWL objective in (Uehara & Jiang, 2019) except we instead consider the bias reduced estimator, such that accurate estimates of Q or ω will lead to reduced bias of the value function estimation. Furthermore, note that in the first part of the objective $J(\pi_\theta,\omega,Q)^2$, we can further use entropy regularization for smoothing the objective, since instead of $Q^\pi(s',a')$ in the target, we can replace it with a log-sum-exp and considering the conjugate of the entropy regularization term, similar to SBEED (Dai et al., 2018). This would therefore give the first part of the objective as an overall min-max optimization problem :

$$\begin{aligned} J(\omega,\pi_\theta) = \mathbb{E}_{d_\mu(s,a)}\Big[\omega_{\pi_\theta/\mathcal{D}}(s,a)\cdot\{r(s,a) + \gamma Q^\pi(s',a') + \tau\log\pi(a\mid s) - Q^\pi(s,a)\}\Big] \\ + (1-\gamma)\mathbb{E}_{s_0\sim\beta_0(s),a_0\sim\pi(\cdot|s_0)}\Big[Q^\pi(s_0,a_0)\Big] \end{aligned} \tag{69}$$

such that from our overall constrained optimization objective for maximizing θ, we have turned it into a min-max objective, for estimating the density ratios, estimating the value function and maximizing the policies

$$\omega^*_{\pi/\mathcal{D}}, Q^*, \pi^* = \underset{\omega,Q}{\text{argmin}}\ \underset{\pi}{\text{argmax}}\ J(\pi_\theta,\omega,Q)^2 \tag{70}$$

where the fixed point solution for the density ratio can be solved by minimizing the objective :

$$\begin{aligned} \omega^*_{\pi/\mathcal{D}} = \underset{\omega}{\text{argmin}}\ \mathcal{L}(\omega_{\pi/\mathcal{D}},Q)^2 = \mathbb{E}_{d_\mu(s,a)}\Big[\{\gamma\omega(s,a)\cdot Q^\pi(s',a') - \omega(s,a)Q^\pi(s,a)\}+ \\ (1-\gamma)\mathbb{E}_{\beta(s,a)}Q^\pi(s_0,a_0)\Big] \end{aligned} \tag{71}$$

DualDICE : In contrast to MWL (Uehara & Jiang, 2019), DualDICE (Nachum et al., 2019a) introduces dual variables through the change of variables trick, and minimizes the Bellman residual of the dual variables $\nu(s,a)$ to estimate the ratio, such that :

$$\nu^*(s,a) - \mathcal{B}^\pi\nu^*(s,a) = \omega_{\pi/\mathcal{D}}(s,a) \tag{72}$$

the solution to which can be achieved by optimizing the following objective

$$\min_\nu \mathcal{L}(\nu) = \frac{1}{2}\mathbb{E}_{d_{\mathcal{D}}}\Big[(\nu - \mathcal{B}^\pi\nu)(s,a)^2\Big] - (1-\gamma)\mathbb{E}_{s_0,a_0\sim\beta(s,a)}\Big[\nu(s_0,a_0)\Big] \tag{73}$$

Minimizing Divergence for Density Ratio Estimation : The distribution ratio can be estimated using an objective similar to GANs (Goodfellow et al., 2014; Ho & Ermon, 2016), as also similarly proposed in (Kostrikov et al., 2019).

$$\max_h \mathcal{G}(h) = \mathbb{E}_{(s,a)\sim d_{\mathcal{D}}}\Big[\log h(s,a)\Big] + \mathbb{E}_{(s,a)\sim d_\pi}\Big[\log(1-h(s,a))\Big] \tag{74}$$

where h is the discriminator class, discriminating between samples from $d_{\mathcal{D}}$ and d_π. The optimal discriminator satisfies :

$$\log h^*(s,a) - \log(1-h^*(s,a)) = \log\frac{d_{\mathcal{D}}(s,a)}{d_\pi(s,a)} \tag{75}$$

The optimal solution of the discriminator is therefore equivalent to minimizing the divergence between d_π and $d_{\mathcal{D}}$, since the KL divergence is given by :

$$-D_{\text{KL}}(d_\pi||d_{\mathcal{D}}) = \mathbb{E}_{(s,a)\sim d_\pi}\Big[\log\frac{d_{\mathcal{D}}(s,a)}{d_\pi(s,a)}\Big] \tag{76}$$

Additionally, using the Donsker-Varadhan representation, we can further write the KL divergence term as :

$$-D_{\text{KL}}(d_\pi||d_{\mathcal{D}}) = \min_x \log\mathbb{E}_{(s,a)\sim d_{\mathcal{D}}}\Big[\exp x(s,a)\Big] - \mathbb{E}_{(s,a)\sim d_\pi}\Big[x(s,a)\Big] \tag{77}$$

such that now, instead of the discriminator class h, we learn the function class x, the optimal solution to which is equivalent to the distribution ratio plus a constant

$$x^*(s,a) = \log\frac{d_\pi(s,a)}{d_{\mathcal{D}}(s,a)} \tag{78}$$

However, note that both the GANs like objective in equation 74 or the DV representation of the KL divergence in equation 77 requires access to samples from both d_π and $d_{\mathcal{D}}$. In our problem setting however, we only have access to batch samples $d_{\mathcal{D}}$. To change the dependency on having access to both the samples, we can use the change of variables trick, such that : $x(s,a) = \nu(s,a) - \mathcal{B}^\pi\nu(s,a)$, to write the DV representation of the KL divergence as :

$$-D_{\text{KL}}(d_\pi||d_{\mathcal{D}}) = \min_\nu \log\mathbb{E}_{(s,a)\sim d_{\mathcal{D}}}\Big[\exp\nu(s,a) - \mathcal{B}^\pi\nu(s,a)\Big] - \mathbb{E}_{(s,a)\sim d_\pi}\Big[\nu(s,a) - \mathcal{B}^\pi\nu(s,a)\Big] \tag{79}$$

where the second expectation can be written as an expectation over initial states, following from DualDICE, such that we have

$$-D_{\text{KL}}(d_\pi||d_{\mathcal{D}}) = \min_\nu \quad \log\mathbb{E}_{(s,a)\sim d_{\mathcal{D}}}\Big[\exp\nu(s,a) - \mathcal{B}^\pi\nu(s,a)\Big] - (1-\gamma)\mathbb{E}_{(s,a)\sim\beta_0(s,a)}\Big[\nu(s_0,a_0)\Big] \tag{80}$$

By minimizing the above objective w.r.t ν, which requires only samples from the fixed batch data $d_{\mathcal{D}}$ and the starting state distribution. The solution to the optimal density ratio is therefore given by :

$$x^*(s,a) = \nu^*(s,a) - \mathcal{B}^\pi\nu^*(s,a) = \log\frac{d_\pi(s,a)}{d_{\mathcal{D}}(s,a)} = \log\omega^*(s,a) \tag{81}$$

Empirical Likelihood Ratio : We can follow Sinha et al. (2020) to compute the state-action likelihood ratio, where they use a binary a classifier to classify samples between an on-policy and off-policy distribution. The proposed classifier, ϕ, is trained on the following objective, and takes as input the state-action tuples (s,a) to return a probability score that the state-action distribution is from the target policy. The objective for ϕ can be formulated as

$$\mathcal{L}_{cls} = \max_\phi -\mathbb{E}_{s,a\sim\mathcal{D}}[\log(\phi(s,a))] + \mathbb{E}_{s\sim\mathcal{D}}[\log(\phi(s,\pi(s))] \tag{82}$$

where $s,a\sim\mathcal{D}$ are samples from the behaviour policy, and $s,\pi(s)$ are samples from the target policy. The density ratio estimates for a given $s,a\sim\mathcal{D}$ are simply $\omega(s,a) = \sigma(\phi(s,a))$ like in Sinha et al. (2020). We then use these $\omega(s,a)$ for density ratio corrections for the target policy in equantion **??**.