SCORE MODELS FOR OFFLINE GOAL-CONDITIONED REINFORCEMENT LEARNING

Harshit Sikchi[*θ] , **Rohan Chitnis**[† φ], **Ahmed Touati**[† φ], **Alborz Geramifard**[φ], **Amy Zhang**[θ,φ], **Scott Niekum**[ψ]
[θ] University of Texas at Austin, [φ] Meta AI, [ψ] UMass Amherst

ABSTRACT

Offline Goal-Conditioned Reinforcement Learning (GCRL) is tasked with learning to achieve multiple goals in an environment purely from offline datasets using sparse reward functions. Offline GCRL is pivotal for developing generalist agents capable of leveraging pre-existing datasets to learn diverse and reusable skills without hand-engineering reward functions. However, contemporary approaches to GCRL based on supervised learning and contrastive learning are often suboptimal in the offline setting. An alternative perspective on GCRL optimizes for occupancy matching, but necessitates learning a discriminator, which subsequently serves as a pseudo-reward for downstream RL. Inaccuracies in the learned discriminator can cascade, negatively influencing the resulting policy. We present a novel approach to GCRL under a new lens of mixture-distribution matching, leading to our discriminator-free method: SMORe. The key insight is combining the occupancy matching perspective of GCRL with a convex dual formulation to derive a learning objective that can better leverage suboptimal offline data. SMORe learns *scores* or unnormalized densities representing the importance of taking an action at a state for reaching a particular goal. SMORe is principled and our extensive experiments on the fully offline GCRL benchmark composed of robot manipulation and locomotion tasks, including high-dimensional observations, show that SMORe can outperform state-of-the-art baselines by a significant margin.

1 INTRODUCTION

A generalist agent will require a vast repertoire of skills, and large amounts of offline pre-collected data offer a way to learn useful skills without any environmental interaction. Many subfields of machine learning like vision and NLP have enjoyed great success by designing objectives to learn a general model from large and diverse datasets. In robot learning, offline interaction data has become more prominent in the recent past (Ebert et al., 2021), with the scale of the datasets growing consistently (Walke et al., 2023; Padalkar et al., 2023). Goal-conditioned reinforcement learning (GCRL) offers a principled way to acquire a variety of useful skills without the prohibitively difficult process of hand-engineering reward functions. In GCRL, the agent learns a policy to accomplish a variety of goals in the environment. The rewards are sparse and goal-conditioned: 1 when the agent's state is proximal to the goal and 0 otherwise. However, the benefit of not requiring the designer to hand-engineer dense reward functions can also be a curse, because learning from sparse rewards is difficult. Driving progress in fundamental offline GCRL algorithms thus becomes an important aspect of moving towards performant generalist agents whose skills scale with data.

Despite recent progress in developing methods for goal-reaching in the online setting (where environment interactions are allowed), a number of these methods are either suboptimal in the offline setting or suffer from learning difficulties. Prior GCRL algorithms can largely be classified into one of three categories: iterated behavior cloning, RL with sparse rewards, and contrastive learning. Iterated behavior cloning or goal-conditioned supervised learning approaches (Ghosh et al., 2019; Yang et al., 2019) have been shown to be provably suboptimal (Eysenbach et al., 2022a) for GCRL. Modifying single-task RL methods (Silver et al., 2014; Kostrikov et al., 2021) for GCRL with 0-1 reward implies learning a Q-function that predicts the discounted probability of goal reaching, which

[*] Work done partially during an internship at Meta AI. Correspondence to hsikchi@utexas.edu
[†] Equal Contribution. **Project page (Code and Videos):** hari-sikchi.github.io/smore/

makes it essentially a density model. Modeling density directly is a hard problem, an insight which has prompted the development of methods (Eysenbach et al., 2020) that learn density-ratio instead of densities, as classification is an easier problem than density estimation. Contrastive RL approaches to GCRL (Eysenbach et al., 2020; 2022b; Zheng et al., 2023) aim to do precisely this and are the main methods to enjoy success for applying GCRL in high-dimensional observation spaces. However, when dealing with offline datasets, contrastive RL approaches (Eysenbach et al., 2022b; Zheng et al., 2023) are suboptimal, as they only learn a policy that is a greedy improvement over the Q-function of the data generation policy. This begs the question: *How can we derive a performant GCRL method that learns near-optimal policies from offline datasets of suboptimal quality?*

In this work, we leverage the underexplored insight of formulating GCRL as an occupancy matching problem. Occupancy matching between the joint state-action-goal visitation distribution induced by the current policy and the distribution over state-actions that transition to goals can be shown to be equivalent to optimizing a max-entropy GCRL objective. Occupancy matching has been studied extensively in imitation learning (Ghasemipour et al., 2020) and often requires learning a discriminator and using the learned discriminator for downstream policy learning through RL. Indeed, a prior GCRL work (Ma et al., 2022) explores a similar insight. Unfortunately, errors in learned discriminators can compound and adversely affect the learned policy's performance, especially in the offline setting where these errors cannot be corrected with further interaction with the environment.

Going beyond the shortcomings of the previous methods, our proposed method combines the insight of formulating GCRL as an occupancy matching problem along with an efficient, discriminator-free dual formulation that learns from offline data. The resulting algorithm `SMORe` forgoes learning density functions or classifiers, but instead learns unnormalized densities or *scores* that allow it to produce near-optimal goal-reaching policies. The scores are learned via a Bellman-regularized contrastive procedure that makes our method a desirable candidate for GCRL with high-dimensional observations, avoiding the need for density modeling. Our experiments represent a wide variety of goal-reaching environments – consisting of robotic arms, anthropomorphic hands, and locomotion environments. We lay out the following contributions: 1) on the extended offline GCRL benchmark, our results demonstrate that `SMORe` significantly outperforms prior methods in the offline GCRL setting. 2) In line with our hypothesis, discriminator-free training makes `SMORe` particularly robust to decreasing goal-coverage in the offline dataset, a property we demonstrate in the experiments. 3) We test `SMORe` for zero-shot GCRL on a prior benchmark (Zheng et al., 2023) for high dimensional vision-based GCRL where contrastive RL approaches are the only class of GCRL methods that have been successful, and show improved performance over other state-of-the-art baselines.

2 Problem Formulation

We consider an infinite horizon discounted Markov Decision Process denoted by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma, d_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, p is the transition probability function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in (0, 1)$ is the discount factor, and d_0 is the initial state distribution. We constrain ourselves to the goal-conditioned RL setting, where we additionally assume a goal space $\mathcal{G}$ where states in $\mathcal{S}$ are mapped to the goal space using a known mapping: $\phi : \mathcal{S} \to \mathcal{G}$. The reward function $r(s, a, g)$ in GCRL is sparse and also depends on the goal. A goal conditioned policy $\pi : \mathcal{S} \times \mathcal{G} \to \Delta(\mathcal{A})$ outputs a distribution over actions in a given state conditioned on a goal. Given a distribution over desired evaluation goals $q^{\texttt{test}}(g)$, the objective of goal-conditioned RL is to find a policy π_g[1] that maximizes the expected discounted return:

$$J(\pi_g) := \mathbb{E}_{g\sim q^{\texttt{test}}((g)), s_0\sim d_0, a_t\sim\pi_g}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, g)\right]. \tag{1}$$

We denote by P^{π_g} the transition operator induced by the policy π_g defined as $P^{\pi_g}S(s, a, g) := \mathbb{E}_{s'\sim p(\cdot|s,a), a'\sim\pi_g(\cdot|s',g)}[S(s', a', g)]$, for any *score* function $S : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}$. We use $d^\pi(s, a \mid g)$ to denote the discounted goal-conditioned state-action occupancy distribution of π_g, i.e $d^{\pi_g}(s, a \mid g) = (1 - \gamma)\pi(a|s, g) \sum_{t=1}^{\infty}[\gamma^t \mathrm{Pr}(s_t = s|\pi_g, d_0)]$. which represents the expected discounted time spent in each state-action pair by the policy π_g conditioned on the goal g. For complete generality, in GCRL, the distribution of goals the policy is trained on often differs from the test goal distribution. To make this distinction clear we define the training distribution $q^{\texttt{train}}(g)$, a uniform measure over goals

[1] We use the subscript g to make the policy's conditioning on g explicit.

we desire to learn to optimally reach during training. We write $d^{\pi_g}(s,a,g) = q^{\texttt{train}}(g)d^{\pi_g}(s,a \mid g)$ as the joint state-action-goal visitation distribution of the policy π_g under the training goal distribution. A state-action-goal occupancy distribution must satisfy the *Bellman flow constraint* in order for it to be a valid occupancy[2] distribution for some stationary policy $\pi_g, \forall s \in \mathcal{S}, a \in \mathcal{A}, g \in \mathcal{G}$:

$$d(s,a,g) = (1-\gamma)d_0(s,g)\pi_g(a \mid s,g) + \gamma \sum_{s',a'} p(s \mid s',a')d(s',a',g)\pi_g(a \mid s,g), \tag{2}$$

where $d_0(s,g) = d_0(s)q^{\texttt{train}}(g)$. Finally, given d^{π_g}, we can express the learning objective for the GCRL agent under the training goal distribution as $J^{\texttt{train}}(\pi_g) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a,g)\sim d^{\pi_g}}[r(s,a,g)]$.

In this work, we focus on the offline setup where the agent cannot interact with the environment $\mathcal{M}$ and instead has access to a offline dataset of $\mathcal{D} := \{\tau_i\}_{i=1}^N$, where each trajectory $\tau^{(i)} = (s_0^{(i)}, a_0^{(i)}, r_0^{(i)}, s_1^{(i)}, ...; g^{(i)})$ with $s_0^{(i)} \sim d_0$. The trajectories are usually relabelled with the $q^{\texttt{train}}(g)$ during learning. We denote the joint state-action-goal distribution of the offline dataset $\mathcal{D}$ as $\rho(s,a,g)$.

3 Score-models for Offline Goal Conditioned Reinforcement Learning

In this section, we introduce our method in two parts: First, we build up the equivalence of the GCRL objective to the occupancy matching problem in Section 3.1, and then we derive a discriminator-free dual objective for solving the occupancy matching problem using off-policy data in Section 3.2. Finally, we present the algorithm for `SMORe` under practical considerations in Section 3.3.

3.1 GCRL as an occupancy matching problem

Define a *goal-transition distribution* $q(s,a,g)$ in a stochastic MDP as $q(s,a,g) \propto q^{\texttt{train}}(g)\mathbb{E}_{s'\sim p(\cdot|s,a)}\big[\mathbb{I}_{\phi(s')=g}\big]$. Intuitively, the distribution has probability mass on each transition that leads to a goal. We formulate the GCRL problem as an occupancy matching problem by searching for the policy π_g that minimizes the discrepancy between its state-action-goal occupancy distribution and the goal-transition distribution $q(s,a,g)$:

$$\texttt{Occupancy matching problem:}\ \mathcal{D}_f(d^{\pi_g}(s,a,g)\|q(s,a,g)), \tag{3}$$

where D_f denotes an f-divergence with generator function f. Note that the q distribution is potentially unachievable by any goal-conditioned policy π_g. Firstly, it does not account for the initial transient phase that the policy must navigate to reach the desired goal. Secondly, even if we consider only the stationary regime (when $\gamma \to 1$), it may not be dynamically possible for the policy to continuously remain at the goal and rather necessitate cycling around the goal. However, in Proposition 1, we show that the occupancy matching in Eq. 3 offers a principled objective since it forms a lower bound to the max-entropy GCRL problem.

Proposition 1. *Consider a stochastic MDP, a stochastic policy π, and a sparse reward function $r(s,a,g) = \mathbb{E}_{s'\sim p(\cdot|s,a)}\big[\mathbb{I}(\phi(s') = g, q^{train}(g) > 0)\big]$ where $\mathbb{I}$ is an indicator function. Define a soft goal transition distribution to be $q(s,a,g) \propto exp(\alpha\, r(s,a,g))$. The following bounds hold for any f-divergence that upper bounds KL-divergence (eg. χ^2, Jensen-Shannon):*

$$J^{train}(\pi_g) + \frac{1}{\alpha}\mathcal{H}(d^{\pi_g}) \geqslant -\frac{1}{\alpha}\mathcal{D}_f(d^{\pi_g}(s,a,g)\|q(s,a,g)) + C, \tag{4}$$

where $\mathcal{H}$ denotes the entropy, α is a temperature parameter and C is the partition function for $e^{R(s,a,g)}$. Furthermore, the bound is tight when f is the KL-divergence.

Ma et al. (2022) (in Proposition 4.1) presented a similar result connecting state-goal distribution matching (i.e $D_{KL}(d^\pi(s,g)||q(s,g))$) to GCRL objective and Proposition 1 extends their results to goal-transition distribution matching. Matching action-free distributions necessitates constructing a loose lower bound that is tractable to optimize. By considering goal-transition distributions we sidestep constructing a loose lower bound and instead directly obtain a tractable distribution matching objective (Ghasemipour et al., 2020; Kostrikov et al., 2019) that is tight under KL-divergence.

How does converting a GCRL objective to an imitation learning objective make learning easier? Estimating the f-divergence still requires estimating the joint policy visitation probabilities $d^{\pi_g}(s,a,g)$, which itself presents a challenging problem. We show in the following section that we can leverage

[2] We will use "occupancy" and "visitation" interchangeably.

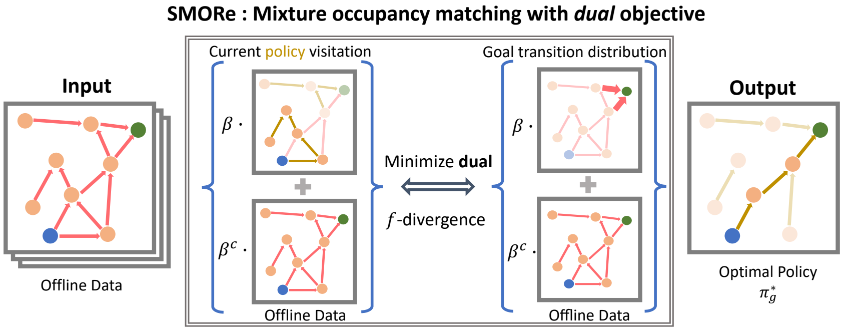

Figure 1: Illustration of the `SMORe` objective where $\beta^c = 1 - \beta$: `SMORe` matches a mixture distribution of current policy and offline data to a mixture of the goal-transition distribution and offline data in order to find the optimal goal reaching policy.

convex duality to transform the imitation learning problem into an off-policy optimization problem, removing the need to sample from $d^{\pi_g}(s, a, g)$ whilst being able to leverage offline data collected from arbitrary sources.

3.2 SMORe: A Dual Formulation for Occupancy Matching

The previous section establishes GCRL as an occupancy matching problem (Eq. 3) but provides no way to use offline data whose joint visitation distribution is given by $\rho(s, a, g)$. To leverage offline data to learn performant goal-reaching policies, we consider a surrogate objective to the occupancy matching learning problem by matching *mixture* distributions:

$$\min_{\pi_g} \mathcal{D}_f(\texttt{Mix}_\beta(d^{\pi_g}, \rho) \| \texttt{Mix}_\beta(q, \rho)), \tag{5}$$

where for any two distributions μ_1 and μ_2, $\texttt{Mix}_\beta(\mu_1, \mu_2)$ denotes the mixture distribution with coefficient $\beta \in (0, 1]$ defined as $\texttt{Mix}_\beta(\mu_1, \mu_2) = \beta\mu_1 + (1 - \beta)\mu_2$. Proposition 2 (in appendix) shows the matching mixture distribution[3] provably maximizes a lower bound to the Lagrangian relaxation of the max-entropy GCRL objective subject to a dataset regularization constraint. We can rewrite the mixture occupancy matching objective as a convex program with linear constraints (Manne, 1960; Nachum and Dai, 2020):

$$\max_{\pi_g, d} -\mathcal{D}_f(\texttt{Mix}_\beta(d, \rho) \| \texttt{Mix}_\beta(q, \rho))$$

$$\text{s.t } d(s, a, g) = (1 - \gamma)d_0(s, g)\pi(a|s) + \gamma \textstyle\sum_{s' \in \mathcal{S}} d(s', a', g)p(s|s', a')\pi(a'|s', g), \ \forall s \in \mathcal{S}. \tag{6}$$

An illustration of this objective can be found in Figure 1. Effectively, we have simply rewritten Eq. 5 into an equivalent problem by considering an arbitrary probability distribution $d(s, a, g)$ in the optimization objective, only to later constrain it to be a valid probability distribution induced by some policy π_g using the *Bellman-flow constraints*. The motivation behind this construction of the primal form is that we have made computing the Lagrangian-dual easier as this objective is convex with linear constraints. Theorem 1 shows that we can leverage tools from convex duality to obtain an unconstrained dual problem that does not require computing $d^{\pi_g}(s, a, g)$ or sampling from it, while effectively leveraging offline data.

Theorem 1. *The dual problem to the primal occupancy matching objective (Equation 6) is given by:*

$$\max_{\pi_g} \min_S \beta(1 - \gamma)\mathbb{E}_{d_0, \pi_g}[S(s, a, g)] + \mathbb{E}_{Mix_\beta(q,\rho)}[f^*(\gamma P^{\pi_g} S(s, a, g) - S(s, a, g))] \tag{7}$$

$$- (1 - \beta)\mathbb{E}_\rho[\gamma P^{\pi_g} S(s, a, g) - S(s, a, g)],$$

where f^* is conjugate function of f and S is the Lagrange dual variable defined as $S : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}$. Moreover, as strong duality holds from Slater's conditions the primal and dual share the same optimal solution π_g^* for any offline transition distribution ρ.

To our knowledge, the closest prior works to our proposed method are GoFAR (Ma et al., 2022) and Dual-RL (Sikchi et al., 2023). GoFAR considers the special case of KL-divergence for the imitation formulation and derives a dual objective that requires learning the density ratio $\frac{\rho(s,g)}{q(s,g)}$ in the

[3] Note that Eq. 5 shares the same global optima as the previous occupancy matching objective at $d_g^\pi(s, a, g) = q(s, a, g)$ when q is an achievable visitation under some policy and recovers the original objective in Eq. 3 when $\beta = 1$.

form of a discriminator and using this as a pseudo-reward. This leads to compounding errors in the downstream RL optimization when learning the density ratio is challenging, e.g. in the case of low coverage between $\rho(s, a, g)$ and $q(s, a, g)$. We show this phenomenon experimentally in Section 4.3. Dual-RL (Sikchi et al., 2023) uses convex duality for matching visitation distribution of realizable expert demonstrations and does not deal with the GCRL setting. *Our contribution is a novel method for GCRL that is discriminator-free, applicable for a number of f-divergences, and robust to low coverage of goals in the offline dataset.*

Sampling from the goal-transition distribution: Goal relabelling is an effective technique to address reward sparsity by widening the training goal distribution $q^{\texttt{train}}(g)$. It utilizes knowledge about reaching other goals, possibly unrelated to test goals, to help in reaching the test distribution of goals $q^{\texttt{test}}(g)$. In the most general case, $q^{\texttt{train}}(g)$ can be set to a uniform distribution over goals corresponding to all the states in the offline data. A common method, Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) chooses a training goal distribution that depends on the current sampled state from the offline dataset as well as the data-collecting policies. In this setting, the sampling distribution used for training Eq 7, $\rho(s, a, g)$, can no longer be factorized into $\rho(s, a)$ and $q^{\texttt{train}}(g)$, as goals are conditionally dependent on state-actions. However, our formulation can naturally account for learning from such relabelled data as the `SMORe` objective in Eq 7 is derived considering the joint distribution $\rho(s, a, g)$. In this setting, we construct our goal transition distribution $q(s, a, g)$ as the uniform distribution over all transitions that lead to the goals selected by the HER procedure — in practice, this amounts to first selecting g through HER and then selecting $\{s, a\}$ that transitions to the selected goal from the offline dataset to get a sample $\{s, a, g\}$ from goal transition distribution. We emphasize that relabelling does not change the test distribution of goals, which is an immutable property of the environment.

3.3 Practical Algorithm

To devise a stable learning algorithm we consider the Pearson χ^2 divergence. Pearson χ^2 divergence has been found to lead to distribution matching objectives that are stable to train as a result of a smooth quadratic generator function f (Garg et al., 2021; Al-Hafez et al., 2023; Sikchi et al., 2023). Our dual formulation `SMORe` simplifies to the following objective:

$$\max_{\pi_g} \min_S \overbrace{\beta(1-\gamma)\mathbb{E}_{(s,g)\sim d_0, a\sim\pi_g(\cdot|s,g)}[S(s,a,g)] + \beta\gamma\mathbb{E}_{(s,a,g)\sim q, s'\sim p(\cdot|s,a), a'\sim\pi_g(\cdot|s',g)}\left[S(s',a',g)\right]}^{\text{Decrease score at transitions under current policy } \pi_g}$$
$$- \underbrace{\beta\mathbb{E}_{(s,a,g)\sim q}[S(s,a,g)]}_{\text{Increase score at the proposed goal transition distribution}} + 0.25\,\underbrace{\mathbb{E}_{(s,a,g)\sim \texttt{Mix}_\beta(q,\rho)}\left[(\gamma S(s',\pi_g(s'),g) - S(s,a,g))^2\right]}_{\text{Smoothness/Bellman regularization}}. \quad (8)$$

Equation 8 suggests a contrastive procedure, maximizing the score at the goal-transition distribution and minimizing the score at the offline data distribution under the current policy with Bellman regularization. The Bellman regularization has the interpretation of discouraging neighboring S values from deviating far and smoothing the score landscape. Instantiating with KL divergence results in an objective with similar intuition while resembling an InfoNCE (Oord et al., 2018) objective. Although Propositions 1 and 2 suggest that KL divergence gives an objective that is a tighter bound to the GCRL objective, prior work has found KL divergence to be unstable in practice (Sikchi et al., 2023; Garg et al., 2023) for dual optimization. It is important to note that S-function is not grounded to any rewards and does not serve as a probability density of reaching goals, but is rather a score function learned via *a Bellman-regularized contrastive learning procedure*.

Algorithm 1: SMORe

```
1: Init $S_\phi$, $M_\psi$, and $\pi_\theta$
2: Params: expectile $\tau$, mixture ratio $\beta$,
   temperature $\alpha$
3: Let $\mathcal{D} = \widehat{\rho} = \{(s, a, s', g)\}$ be an offline
   dataset and $q$ be goal-transition distribution
4: for $t = 1..T$ iterations do
5:     Train $S_\phi$ via Eq. 10
6:     Train $M_\psi$ via Eq. 9
7:     Update $\pi_\theta$ via Eq. 11
8: end for
```

We now derive a practical approach for `SMORe` in the offline GCRL setting. We use parameterized functions: $S_\phi(s, a, g)$, $M_\psi(s, g)$, $\pi_\theta(a|s, g)$. The offline learning regime necessitates measures to constrain the learning policy to the offline data support in order to prevent overestimation due to maximizing π_g in Eq. 8 over potentially out-of-distribution actions. Inspired by prior work (Kostrikov et al., 2021), we use implicit maximization to constrain the learning algorithm to learn expectiles using the observed empirical samples. More concretely, we use expectile regression:

$$\min_{\psi} \mathcal{L}(\psi) := \mathbb{E}_{(s,a,g)\sim\rho}[L_2^{\tau}(M_{\psi}(s,g) - S_{\phi}(s,a,g))], \tag{9}$$

where $L_2^{\tau}(u) = |\tau - 1(u<0)|u^2$. Intuitively, this step implements the maximization w.r.t π by using expectile regression. With the above practical considerations, our objective for learning S_ϕ reduces to:

$$\min_{\phi} \mathcal{L}(\phi) := \beta(1-\gamma)\mathbb{E}_{(s,g)\sim\mathcal{D},a\sim\pi_g(\cdot|a,g)}[S_\phi(s,\pi_g(s),g)] + \beta\gamma\mathbb{E}_{(s,a,g)\sim q,s'\sim p(\cdot|s,a)}\big[S_\phi(s',\pi_g(s'),g)\big]$$
$$- \beta\mathbb{E}_{(s,a,g)\sim q}[S_\phi(s,a,g)] + \mathbb{E}_{(s,a,g)\sim\mathtt{Mix}_\beta(q,\rho)}\big[(\gamma M_\psi(s',g) - S_\phi(s,a,g))^2\big], \tag{10}$$

where we have set the offline data distribution as our initial state distribution. Finally, the policy is extracted via advantage-weighted regression that learns in-distribution actions maximizing the score $S(s,a,g)$:

$$\min_{\theta} \mathcal{L}(\theta) := \mathbb{E}_{(s,a,g)\sim\rho}[\exp(\alpha(S_\phi(s,a,g) - M_\psi(s,g)))\log(\pi_\theta(a|s,g))], \tag{11}$$

where α is the temperature parameter. Algorithm 1 details the practical implementation.

4 Experiments

Our experiments study the effectiveness of proposed GCRL algorithm `SMORe` on a set of simulated benchmarks against other GCRL methods that employ behavior cloning, RL with sparse reward, and contrastive learning. We also analyze if `SMORe` is robust to environment stochasticity — a number of prior methods are based on an assumption of deterministic dynamics. Then, we study if the discriminator-free nature of `SMORe` is indeed able to prevent performance degradation in the face of low expert coverage in offline data. Finally, we analyze if `SMORe`'s score-modeling approach helps `SMORe` scale to a vision-based manipulation offline GCRL benchmark, as density modeling and discriminator learning become increasingly difficult with high-dimensional observations. Hyperparameter ablations can be found in Appendix D.

4.1 Experimental Setup

Our experiments will use a suite of simulated goal-conditioned tasks extending the tasks from previous work (Ma et al., 2022; Plappert et al., 2018). In particular we consider the following environments: `Reacher`, Robotic arm environments - [`SawyerReach`, `SawyerDoor`, `FetchReach`, `FetchPick`, `FetchPush`, `FetchSlide`], Anthropomorphic hand environment - `HandReach` and Locomotion environments - [`CheetahTgtVel-me,CheetahTgtVel-re,AntTgtVel-me,AntTgtVel-re`]. Tasks in all environments are specified by a sparse reward function. Depending on whether the task involves object manipulation, the goal distribution is defined over valid configurations in robot or object space. The offline dataset for manipulation tasks consists of transitions collected by a random policy or mixture of 90% random policy and 10% expert policy. For locomotion tasks, we generate our dataset using the D4RL benchmark (Fu et al., 2020), combining a random or medium dataset with 30 episodes of expert data. Note that the policies used to collect the expert locomotion datasets have a different objective than the tasks here, which are to achieve and maintain a particular desired velocity.

4.2 Offline Goal-conditioned RL benchmark

Baselines. We compare to state-of-art offline GCRL algorithms, consisting of both regression-based and actor-critic methods. The occupancy-matching based methods are: (1) **GoFar** (Ma et al., 2022), which derives a dual objective for GCRL based on a coverage assumption. The behavior cloning based methods are: (1) **GCSL** (Ghosh et al., 2019), which incorporates hindsight relabeling in conjunction with behavior cloning to clone actions that lead to a specified goal, and (2) **WGCSL** (Yang et al., 2022), which improves upon GCSL by incorporating discount factor and advantage weighting into the supervised policy learning update. **Contrastive RL** (Eysenbach et al., 2022b) generalizes C-learning (Eysenbach et al., 2020) and represents contrastive GCRL approaches. The RL with sparse reward methods are (1) **IQL** (Kostrikov et al., 2021) where we use a state-of-the-art offline RL method repurposed for GCRL along with HER (Andrychowicz et al., 2017) goal sampling, and (2)

Task	Occupancy Matching		Behavior cloning		Contrastive RL	RL+sparse reward	
	SMORe	GoFAR	WGCSL	GCSL	CRL	AM	IQL
Reacher (⋆)	**28.40**±0.88	19.74±1.35	17.57±0.53	15.87±1.31	16.44±0.60	23.26 ±0.14	11.70 ±1.97
SawyerReach (⋆)	**37.67**±0.12	15.34±0.64	15.15±0.44	14.25±0.7	22.32 ±0.34	23.34±0.17	35.18 ±0.29
SawyerDoor (⋆)	**31.48**±0.46	18.94±0.01	20.01±1.55	20.88±0.22	12.96±5.19	22.12 ±0.13	25.52 ±1.45
FetchReach (⋆)	**35.08**± 0.54	28.2 ± 0.61	21.9± 2.13	20.91 ± 2.78	30.07±0.07	30.1 ± 0.32	34.43 ± 1.00
FetchPick (⋆)	**26.47** ± 1.34	19.7 ± 2.57	9.84 ± 2.58	7.58±1.85	0.42±0.29	8.94 ± 3.09	16.8 ± 3.10
FetchPush (⋆)	**26.83**± 1.21	18.2 ± 3.00	14.7 ± 2.65	13.4 ± 3.02	2.40 ±1.28	14.0 ± 2.81	22.40 ± 0.74
FetchSlide	**4.99**± 0.40	2.47 ± 1.44	2.73 ± 1.64	1.75 ± 1.3	0.0±0.0	1.46 ± 1.38	**4.80** ± 1.59
HandReach (⋆)	**18.68** ± 3.35	11.5 ± 5.26	5.97 ± 4.81	1.37 ± 2.21	0.0±0.0	0.0 ± 0.0	1.44 ± 1.77
CheetahTgtVel-m-e (⋆)	**136.71** ± 10.59	0.0± 0.0	0.0± 0.0	95.98± 15.72	0.0±0.0	0.0± 0.0	100.38± 1.22
CheetahTgtVel-r-e (⋆)	**60.01** ± 39.40	0.0± 0.0	0.0± 0.0	11.56 ± 13.47	0.0±0.0	0.0± 0.0	0.0± 0.0
AntTgtVel-m-e	154.95± 19.44	**168.27**± 9.58	0.0± 0.0	164.54± 7.69	0.0±0.0	0.0± 0.0	148.17 ± 5.43
AntTgtVel-r-e (⋆)	**126.22**± 14.40	74.36± 15.97	0.0± 0.0	104.95± 6.00	0.0±0.0	0.0± 0.0	3.06 ± 2.64

Table 1: Discounted Return for the offline GCRL benchmark. Results are averaged over 10 seeds.'m-e' and 'r-e' stands for medium-expert mixture and random-expert mixture respectively. (⋆) denotes statistically significant improvements.

ActionableModel (AM) (Chebotar et al., 2021), which incorporates conservative Q-Learning (Kumar et al., 2020) as well as goal-chaining on top of an actor-critic method.

The results for all baselines are tuned individually, particularly the best HER ratio was searched among $\{0.2, 0.5, 0.8, 1.0\}$ for each task. SMORe shares the same network architecture for baselines and uses a mixture ratio of $\beta = 0.5$. Each method is trained for 10 seeds. Complete architecture and hyperparameter table as well as additional training details are provided in Appendix C.

Table 1 reports the **discounted return** obtained by the learned policy with a sparse binary task reward. (⋆) denotes statistically significant improvement over the second best method under a Mann-Whitney U test with a significance level of 0.05. This metric allows us to compare the algorithms on a finer scale to understand which methods reach the goal as fast as possible and stay in the goal region thereafter for the longest time. Additional results on metrics like success rate and final distance to goal can be found in the appendix. These additional metrics do not take into consideration how *precisely* and *consistently* a goal is being reached. In Table 1, we see that SMORe enjoys a high-performance gain consistently across all tasks in the extended offline GCRL benchmark.

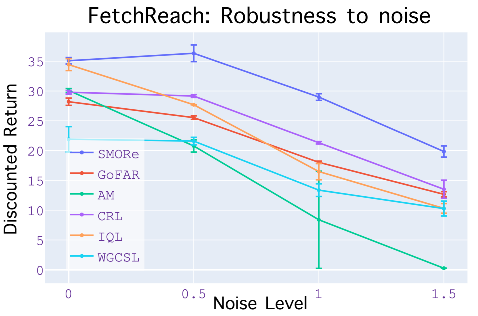

Figure 2: SMORe is robust in stochastic environments. With increasing noise, SMORe still outperforms prior methods.

Robustness to environment stochasticity: We consider a noisy version of the FetchReach environment in this experiment. Gaussian zero-mean noise is added before executing an action to generate different variants of the environment with standard deviations of $\{0.5, 1.0, 1.5\}$. Datasets for these environments are obtained from prior work (Ma et al., 2022). As we see in Figure 2, SMORe is robust to stochasticity in the environment, outperforming baselines in terms of discounted return. Behavior cloning based approaches assume deterministic dynamics and are therefore over-optimistic in stochastic environments.

4.3 Robustness of Occupancy-Matching Methods to Decreasing Expert Coverage

We posit that the discriminator-free nature of SMORe makes it more robust to decreasing goal coverage, as it does not suffer from cascading errors stemming from a learned discriminator. In this section, we set out to test this hypothesis by decreasing the amount of expert data in the offline goal-reaching dataset. We compare with GoFAR in Table 2 due to the similarity between methods and GoFAR's restrictive assumption on coverage of expert data in the suboptimal dataset. Comparison against all the baselines can be found in Appendix D.

Our hypothesis holds true as we see in Table 2, the performance of the discriminator-based method GoFar rapidly decays as expert data is decreased in the offline dataset – 28.4% with 2.5% and 36.15% with 1% expert data(i.e. optimal policy's coverage) respectively. SMORe shows a much slower decay

Task	5 % expert data		2.5 % expert data		1 % expert data	
	SMORe	GoFAR	SMORe	GoFAR	SMORe	GoFAR
Reacher	22.43±3.46	16.86 ±1.26	17.92 ± 0.93	12.20±0.81	19.61± 1.56	11.52 ± 0.52
SawyerReach	36.35±0.37	13.20 ±1.36	36.74±0.62	11.57 ±1.79	35.44 0.27	9.34± 0.17
SawyerDoor	32.82±0.88	20.07±0.01	25.69±0.21	19.54±1.32	23.78±2.88	18.04 ±1.80
FetchReach	36.00± 0.01	27.66 ± 0.55	35.58 ± 0.47	27.84 ± 0.82	35.97 ± 0.25	28.01 ± 0.20
FetchPick	26.43± 1.95	16.21 ± 1.46	26.17± 3.37	3.21 ± 2.22	15.38 ± 1.52	0.31 ± 0.31
FetchPush	23.81± 0.37	18.2 ± 3.00	22.75±1.08	5.17 ± 2.01	19.04± 2.79	4.23± 3.96
FetchSlide	4.05± 1.12	1.08 ± 0.06	3.11 ± 1.61	0.96 ± 0.73	3.50± 0.97	0.86 ± 1.22
Average Performance	25.98	16.18	23.99	11.49	21.81	10.33
Avg. Perf. Drop	0	0	-7.6%	-28.4%	-16%	-36.15%

Table 2: Discounted Return for the offline GCRL benchmark with 5%, 2.5% and 1% expert data in offline dataset. Results are averaged over 10 seeds.

in performance, 7.6% with 2.5% and 16% with 1% expert data, attesting to the method's robustness under decreasing expert coverage in the offline dataset.

4.4 Offline GCRL with image observations

SMORe provides an effective algorithm for offline GCRL in high-dimensional observation spaces by learning unnormalized scores using a contrastive procedure as opposed to prior works that learn normalized densities (Eysenbach et al., 2020) which are difficult to learn or density ratios (Eysenbach et al., 2022b; Zheng et al., 2023) which do not optimize for the optimal goal-conditioned policy in the offline GCRL setting. Similar to prior work (Eysenbach et al., 2022b), we consider the following structure in S-function parameterization to learn performant and generalizable policies: $S(s, a, g) = \phi(s, a)^T \psi(g)$. The S-function can be interpreted as the similarity between the two representations given by ϕ and ψ. Our network architecture for both representations is similar to Zheng et al. (2023) and is kept the same across all baselines to ensure a fair comparison of the underlying GCRL method.

We use the offline GCRL benchmark from (Zheng et al., 2021) which learns goal-reaching policies from an image-observation dataset of 250K transitions with the horizon ranging from 50-100. The benchmark adds another layer of complexity by testing on goals absent from the dataset — the dataset contains primitive behaviors like picking up objects and pushing drawers but no behavior that completes the compound task we consider from the initial state. The observations and goals are 48x48x3 RGB images.

Baselines We compare to the best performing GCRL algorithms from Section 1 as well as a recent state-of-the-art work, stable contrastive RL Zheng et al. (2023). Stable contrastive RL features a number of improvements over contrastive RL by changing design decisions in neural network architecture, layer normalization, and data augmentation. Since our objective is to compare the quality of the underlying GCRL algorithm, we keep these design decisions consistent across the board.

Results Figure 3 shows the success rate on a variety of unseen tasks for all the methods. SMORe achieves highest success rates across all the tasks, even for the most challenging task of pick, place

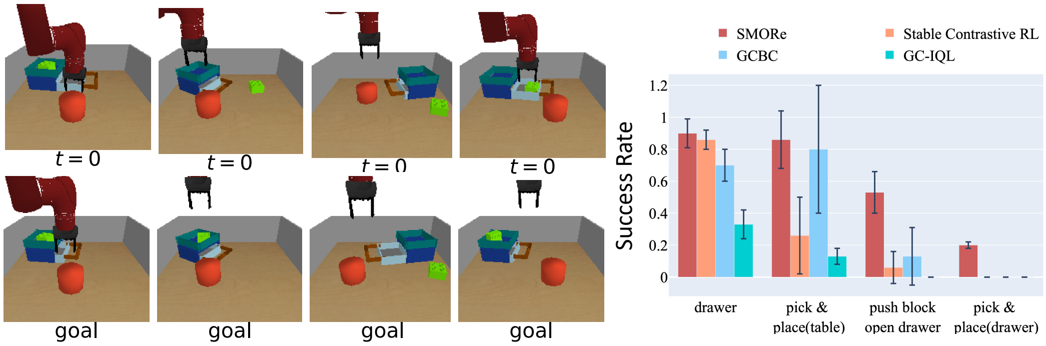

Figure 3: Evaluation on simulated manipulation tasks with image observations. The left image shows the starting state at the top and the goal at the bottom for evaluation tasks. The error bars show the standard deviation with 5 random seeds. SMORe is competitive or outperforms prior methods on all the tasks we considered.

and closing the drawer. We note that our results differ from Zheng et al. (2023) for the baselines as we apply the same design decisions for all methods whereas Zheng et al. (2023) focuses on ablating design decisions.

5 Related Works

Offline Goal Conditioned Reinforcement Learning. Learning to achieve goals in the environment optimally forms the basis of goal-condition RL problems. Studies in cognitive science (Molinaro and Collins, 2023) underscore the importance goal-achieving plays in human development. Offline GCRL approaches are typically catered to designing learning algorithms for addressing the sparsity of reward function in the offline setting. One of the most successful techniques in this setting has been hindsight relabelling. Hindsight-experience relabelling (HER) (Kaelbling, 1993; Andrychowicz et al., 2017) suggests relabelling any experience with some commanded goal to the goal that was actually achieved in order to leverage generalization. HER has been investigated in the setting of learning from demonstrations (Ding et al., 2019) and exploration (Fang et al., 2019) to validate its effectiveness. A number of prior works (Ghosh et al., 2019; Yang et al., 2019; Chen et al., 2020; Ding et al., 2019; Lynch et al., 2020; Paster et al., 2020; Srivastava et al., 2019; Hejna et al., 2023) have investigated using goal-conditioned behavior cloning, a strategy that uses relabelling to learn goal-conditioned policies, as a way to learn performant policies. Eysenbach et al. (2022a) shows that this line of work has a limitation of learning suboptimal policies that do not consistently improve over the policy that collected the dataset. The simplest strategy of applying single-task RL to the problem of multi-task goal reaching requires learning a Q-function which represents normalized densities over the state-action space. Contrastive RL (Eysenbach et al., 2022b; 2020; Zheng et al., 2023) emerged as another alternative for GCRL which relabels trajectories and, rather than use that relabelling to learn policies, learns a Q-function using a contrastive procedure. While these approaches learn optimal policies in the online setting, they fall behind in the offline setting where they only learn a policy that greedily improves over the Q-function of the data collecting policy. Our work learns optimal policies by presenting an off-policy objective that solves GCRL and furthermore learns scores (or unnormalized densities) that alleviate the learning challenges of normalized density estimation.

Distribution matching. Our approach is inspired by the distribution matching approach (Ghasemipour et al., 2020; Ni et al., 2021; Sikchi et al., 2022; Swamy et al., 2021; Sikchi et al., 2023) prominent in imitation learning. Ghasemipour et al. (2020); Ni et al. (2021) takes the problem of imitating an expert demonstrator in the environment and converts it into a problem of distribution matching between the current policy's state-action visitation distribution and the expert policy's visitation distribution. Indeed, a prior work f-PG (Agarwal et al., 2024) proposes a distribution matching approach to GCRL but is restricted to the on-policy setting. Another, prior work (Ma et al., 2022) creates one such distribution matching problem and presents a new optimization problem for GCRL in the form of an off-policy dual (Nachum and Dai, 2020; Sikchi et al., 2023). Such an off-policy dual is very appealing for the offline RL setup, as optimizing for this dual only requires sampling from the offline data distribution. A limitation of their dual construction is the fact that they require learning a discriminator and use that discriminator as the pseudo-reward for solving the GCRL objective. Our approach presents a new construction for GCRL as a distribution matching problem along with a dual construction that leads to a more performant discriminator-free off-policy approach for GCRL.

6 Conclusion

Prior work in performant online goal-conditioned RL often relies on iterated behavior cloning or contrastive RL. However, these approaches are suboptimal for the offline setting. Existing methods specifically derived for offline GCRL require learning a discriminator and using it as a pseudo-reward, enabling compounding errors that make the resulting policy ineffective. We present an occupancy-matching approach to offline GCRL that provably optimizes a lower bound to the regularized GCRL objective. Our method is discriminator-free, applicable to a number of f-divergences, and learns unnormalized scores over actions at a state to reach the goal. We show that these positive aspects of our algorithm allow us to empirically outperform prior methods, stay robust under decreasing goal coverage, and scale to high-dimensional observation space for GCRL.

Acknowledgements

We thank the MIDI lab members and ICLR reviewers for valuable feedback on this work. This work has taken place in the Safe, Correct, and Aligned Learning and Robotics Lab (SCALAR) at The University of Massachusetts Amherst and Machine Intelligence through Decision-making and Interaction (MIDI) Lab at The University of Texas at Austin. SCALAR research is supported in part by the NSF (IIS-2323384), AFOSR (FA9550-20-1-0077), and ARO (78372-CS, W911NF-19-2-0333), and the Center for AI Safety (CAIS). This research was also sponsored by the Army Research Office under Cooperative Agreement Number W911NF-19-2-0333. HS and AZ are funded in part by a sponsored research agreement with Cisco Systems Inc. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019.

S. Agarwal, I. Durugkar, P. Stone, and A. Zhang. f-policy gradients: A general framework for goal-conditioned rl using f-divergences. *Advances in Neural Information Processing Systems*, 36, 2024.

F. Al-Hafez, D. Tateo, O. Arenz, G. Zhao, and J. Peters. Ls-iq: Implicit reward regularization for inverse reinforcement learning. *arXiv preprint arXiv:2303.00599*, 2023.

M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. Pieter Abbeel, and W. Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Y. Chebotar, K. Hausman, Y. Lu, T. Xiao, D. Kalashnikov, J. Varley, A. Irpan, B. Eysenbach, R. Julian, C. Finn, et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. *arXiv preprint arXiv:2104.07749*, 2021.

L. Chen, R. Paleja, and M. Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. *arXiv preprint arXiv:2010.11723*, 2020.

Y. Ding, C. Florensa, P. Abbeel, and M. Phielipp. Goal-conditioned imitation learning. *Advances in neural information processing systems*, 32, 2019.

F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

B. Eysenbach, R. Salakhutdinov, and S. Levine. C-learning: Learning to achieve goals via recursive classification. *arXiv preprint arXiv:2011.08909*, 2020.

B. Eysenbach, S. Udatha, R. R. Salakhutdinov, and S. Levine. Imitating past successes can be very suboptimal. *Advances in Neural Information Processing Systems*, 35:6047–6059, 2022a.

B. Eysenbach, T. Zhang, S. Levine, and R. R. Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022b.

M. Fang, T. Zhou, Y. Du, L. Han, and Z. Zhang. Curriculum-guided hindsight experience replay. *Advances in neural information processing systems*, 32, 2019.

J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

D. Garg, S. Chakraborty, C. Cundy, J. Song, and S. Ermon. Iq-learn: Inverse soft-q learning for imitation. *Advances in Neural Information Processing Systems*, 34:4028–4039, 2021.

D. Garg, J. Hejna, M. Geist, and S. Ermon. Extreme q-learning: Maxent rl without entropy. *arXiv preprint arXiv:2301.02328*, 2023.

S. K. S. Ghasemipour, R. Zemel, and S. Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pages 1259–1277. PMLR, 2020.

D. Ghosh, A. Gupta, A. Reddy, J. Fu, C. Devin, B. Eysenbach, and S. Levine. Learning to reach goals via iterated supervised learning. *arXiv preprint arXiv:1912.06088*, 2019.

J. Hejna, J. Gao, and D. Sadigh. Distance weighted supervised learning for offline interaction data. *arXiv preprint arXiv:2304.13774*, 2023.

L. P. Kaelbling. Learning to achieve goals. In *IJCAI*, volume 2, pages 1094–8. Citeseer, 1993.

I. Kostrikov, O. Nachum, and J. Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

I. Kostrikov, A. Nair, and S. Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

C. Lynch, M. Khansari, T. Xiao, V. Kumar, J. Tompson, S. Levine, and P. Sermanet. Learning latent plans from play. In *Conference on robot learning*, pages 1113–1132. PMLR, 2020.

Y. J. Ma, J. Yan, D. Jayaraman, and O. Bastani. How far i'll go: Offline goal-conditioned reinforcement learning via f-advantage regression. *arXiv preprint arXiv:2206.03023*, 2022.

A. S. Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

G. Molinaro and A. G. Collins. A goal-centric outlook on learning. *Trends in Cognitive Sciences*, 2023.

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

T. Ni, H. Sikchi, Y. Wang, T. Gupta, L. Lee, and B. Eysenbach. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pages 529–551. PMLR, 2021.

A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

K. Paster, S. A. McIlraith, and J. Ba. Planning from pixels using inverse dynamics models. *arXiv preprint arXiv:2012.02419*, 2020.

M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.

H. Sikchi, A. Saran, W. Goo, and S. Niekum. A ranking game for imitation learning. *arXiv preprint arXiv:2202.03481*, 2022.

H. Sikchi, Q. Zheng, A. Zhang, and S. Niekum. Dual rl: Unification and new methods for reinforcement and imitation learning. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr, 2014.

R. K. Srivastava, P. Shyam, F. Mutz, W. Jaśkowski, and J. Schmidhuber. Training agents using upside-down reinforcement learning. *arXiv preprint arXiv:1912.02877*, 2019.

G. Swamy, S. Choudhury, J. A. Bagnell, and S. Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pages 10022–10032. PMLR, 2021.

H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, et al. Bridgedata v2: A dataset for robot learning at scale. *arXiv preprint arXiv:2308.12952*, 2023.

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

H. Xu, L. Jiang, J. Li, Z. Yang, Z. Wang, V. W. K. Chan, and X. Zhan. Offline rl with no ood actions: In-sample learning via implicit value regularization. *arXiv preprint arXiv:2303.15810*, 2023.

C. Yang, X. Ma, W. Huang, F. Sun, H. Liu, J. Huang, and C. Gan. Imitation learning from observations by minimizing inverse dynamics disagreement. *arXiv preprint arXiv:1910.04417*, 2019.

R. Yang, Y. Lu, W. Li, H. Sun, M. Fang, Y. Du, X. Li, L. Han, and C. Zhang. Rethinking goal-conditioned supervised learning and its connection to offline rl. *arXiv preprint arXiv:2202.04478*, 2022.

C. Zheng, B. Eysenbach, H. Walke, P. Yin, K. Fang, R. Salakhutdinov, and S. Levine. Stabilizing contrastive rl: Techniques for offline goal reaching. *arXiv preprint arXiv:2306.03346*, 2023.

L. Zheng, T. Fiez, Z. Alumbaugh, B. Chasnov, and L. J. Ratliff. Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. *arXiv preprint arXiv:2109.12286*, 2021.

A Appendix

A.1 Theory

In this section, we first show the equivalence of the GCRL problem and the distribution-matching objective of imitation learning. Then, we show how the mixture distribution objective relates to offline GCRL objective. Finally, we derive the dual objective for mixture distribution matching that leads to our method `SMORe`.

A.1.1 Reduction of GCRL to distribution matching

Proposition 1. *Consider a stochastic MDP, a stochastic policy π, and a sparse reward function $r(s,a,g) = \mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\mathbb{I}(\phi(s') = g, q^{train}(g) > 0)\right]$ where $\mathbb{I}$ is an indicator function. Define a soft goal transition distribution to be $q(s,a,g) \propto exp(\alpha\, r(s,a,g))$. The following bounds hold for any f-divergence that upper bounds KL-divergence (eg. χ^2, Jensen-Shannon):*

$$J^{train}(\pi_g) + \frac{1}{\alpha}\mathcal{H}(d^{\pi_g}) \geqslant -\frac{1}{\alpha}\mathcal{D}_f(d^{\pi_g}(s,a,g)\|q(s,a,g)) + C, \tag{4}$$

where $\mathcal{H}$ denotes the entropy, α is a temperature parameter and C is the partition function for $e^{R(s,a,g)}$. Furthermore, the bound is tight when f is the KL-divergence.

Proof. This proof is adapted from Ma et al. (2022) for goal transition distributions and state-action distributions. Let $Z = \int e^{R(s,a,g)}\; ds\; da\; dg$ and $\alpha > 0$ be the temperatue parameter. Note that $q(s,a,g) = e^{r(s,a,g)}$ where r is defined in the proposition, strictly generalizes the original definition $q(s,a,g) = q^{\mathtt{train}}(g)\mathbb{E}_{s'\sim p(\cdot|s,a)}[\mathbb{I}(\phi(s') = g)]$ and recovers it when $\alpha \to \infty$. Starting with the true GCRL objective:

$$\begin{aligned}
\alpha J(\pi_g) &= \mathbb{E}_{d^{\pi_g}}[\alpha R(s,a,g)] && (12)\\
&= \mathbb{E}_{d^{\pi_g}}\left[\log e^{\alpha R(s,a,g)}\right] && (13)\\
&= \mathbb{E}_{d^{\pi_g}}\left[\log(\frac{e^{\alpha R(s,a,g)}}{Z}\frac{d^{\pi_g}(s,a,g)}{d^{\pi_g}(s,a,g)}Z)\right] && (14)\\
&= \mathbb{E}_{d^{\pi_g}}\left[\log(\frac{q(s,a,g)}{d^{\pi_g}(s,a,g)}Z)\right] + \mathbb{E}_{d^{\pi_g}}[\log d^{\pi_g}] && (15)\\
&= -D_{KL}(d^{\pi_g}(s,a,g)\|q(s,a,g)) - \mathcal{H}(d^{\pi_g}) + \log(Z) && (16)
\end{aligned}$$

Rearranging terms we get:

$$J(\pi_g) + \frac{1}{\alpha}\mathcal{H}(d^{\pi_g}) = -\frac{1}{\alpha}D_{KL}(d^{\pi_g}(s,a,g)\|q(s,a,g)) + C \tag{17}$$

For any f-divergence that upper bounds the KL divergence we have:

$$J(\pi_g) + \frac{1}{\alpha}\mathcal{H}(d^{\pi_g}) = -\frac{1}{\alpha}D_{KL}(d^{\pi_g}(s,a,g)\|q(s,a,g)) + C \geqslant -\frac{1}{\alpha}D_f(d^{\pi_g}(s,a,g)\|q(s,a,g)) + C \tag{18}$$

□

A (dataset) regularized GCRL objective: Define a regularized objective for GCRL as follows:

$$J_{offline}(\pi) = \alpha_1\mathbb{E}_{d^\pi}\left[e^{r(s,a,g)}\right] + \alpha_2\mathbb{E}_{d^\pi(s,a,g)}[\rho(s,a,g)]. \tag{19}$$

The second term in the objective $\mathbb{E}_{d^\pi(s,a,g)}[\rho(s,a,g)]$ above is maximized when the policy visitation places more probability mass on the most visited transitions in the dataset. To see why this is, consider two probability distributions represented as vectors d^π and ρ with individuals elements of the vector indexed by i:

$$\langle d^\pi, \rho\rangle \leqslant \max_i \rho_i \tag{20}$$

The equality holds only when d^π places probability mass on all state-action-goal tuples which are most visited in the offline dataset ρ. The first term maximizes the true GCRL objective while the second term prefers staying close to transitions that are most frequently observed in the offline dataset. A constraint of $\langle d^\pi, \rho\rangle \geqslant 1 - \delta$ implies that the agent visitation places atleast half of the probability mass on state-action-goal tuples whose average visitation in offline dataset is greater than or equal to $1-\delta$. With weights α_1 and α_2, the objective above reflects an lagrangian relaxation to this constraint. Thus the above offline objective presents an alternative offline objective when compared to the classical offline RL objectives Wu et al. (2019); Nachum and Dai (2020).

Proposition 2 derives the connection between the dataset regularized GCRL objective and `SMORe`:

Proposition 2. *Consider a stochastic MDP, a stochastic policy π, and a sparse reward function $r(s,a,g) = \mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\mathbb{I}(\phi(s')=g, q^{train}(g)>0)\right]$ where $\mathbb{I}$ is an indicator function, define a soft goal transition distribution to be $q(s,a,g) \propto exp(\alpha\ r(s,a,g))$ the following bounds hold for any f-divergence that upper bounds KL-divergence (eg. χ^2, Jensen-Shannon):*

$$\log J_{offline}(\pi_g) + \mathcal{H}(\mathtt{Mix}_\beta(d,\rho)(s,a,g)) + C \geqslant -\mathcal{D}_f(\mathtt{Mix}_\beta(d,\rho)(s,a,g)\|\mathtt{Mix}_\beta(q,\rho)(s,a,g)), \tag{21}$$

where $\mathcal{H}$ denotes the entropy, α is a temperature parameter, $\alpha_1=\beta^2$, $\alpha_2=\beta(1-\beta)Z$ and C is a positive constant. Furthermore, the bound is tight when f is the KL-divergence.

Proof. We first consider the following two objectives for GCRL and show that they are equivalent. This reduction will later help in proving a connection to mixture occupancy matching. We consider $\alpha=1$ w.l.o.g. Here are two objectives we consider:

$$J(\pi) = \mathbb{E}_{d^\pi}[r(s,a,g)] \tag{22}$$

$$J'(\pi) = \mathbb{E}_{d^\pi}\left[e^{r(s,a,g)}\right] \tag{23}$$

In GCRL reward functions are sparse and binary. We show the equivalence of first two objectives in find the optimal goal conditioned policy via two arguments. First, notice that the rewards for goal transition states for objective $J'(\pi)$ is e and 1 for all other transitions. This is in contrast to $J(\pi)$ which considers a reward function 1 at goal transitions states and 0 otherwise. Under our assumption of infinite horizon discounted MDP, we can translate the rewards while keeping the optimal policy same in MDP considered by $J'(\pi)$ to $e-1$ at goal transitions states and 0 otherwise. Further we can scale the rewards by $1/(e-1)$ and recover and MDP with same optimal policy that has reward of 1 at goal-transition states and 0 otherwise. This concludes the equivalence of maximizing $J'(\pi)$ as an alternative to $J(\pi)$ while recovering the same optimal policy.

We now consider a regularized (pessimistic/offline) GCRL problem with the shifted reward functions $e^{r(s,a,g)}$ that maximizes the reward while ensuring the policy visitation stays close to offline data visitation in cosine similarity.

$$J_{offline}(\pi) = \alpha_1\mathbb{E}_{d^\pi}\left[e^{r(s,a,g)}\right] + \alpha_2\mathbb{E}_{d^\pi(s,a,g)}[\rho(s,a,g)]. \tag{24}$$

With a particular instantiation of hyperparameters we show that the $J_{offline}(\pi)$ objective can be simplified to an equivalent objective $J'_{offline}(\pi)$ by setting $\alpha_1=\beta^2$ and $\alpha_2=\beta(1-\beta)Z$ where Z is the partition function for $e^{r(s,a,g)}$ over entire $\mathcal{S}\times\mathcal{A}\times\mathcal{G}$.

$$J'_{offline}(\pi) = \mathbb{E}_{\mathtt{Mix}_\beta(d,\rho)(s,a,g)}\left[\beta e^{r(s,a,g)} + (1-\beta)\rho(s,a,g).Z\right] \tag{25}$$

$$\begin{aligned}
J'_{offline}(\pi) &= \mathbb{E}_{\mathtt{Mix}_\beta(d,\rho)(s,a,g)}\left[\beta e^{r(s,a,g)} + (1-\beta)\rho(s,a,g).Z\right] && (26)\\
&= \beta^2\mathbb{E}_{d^\pi}\left[e^{r(s,a,g)}\right] + \beta(1-\beta)Z\mathbb{E}_{d^\pi}[\rho(s,a,g)] && (27)\\
&+ (1-\beta)\mathbb{E}_{d^O}\left[\beta e^{r(s,a,g)} + (1-\beta)\rho(s,a,g).Z\right]\beta && (28)\\
& && (29)\\
&= \beta^2\mathbb{E}_{d^\pi}\left[e^{r(s,a,g)}\right] + \beta(1-\beta)Z\mathbb{E}_{d^\pi}[\rho(s,a,g)] + C' && (30)\\
&= J_{offline}(\pi) + C' && (31)
\end{aligned}$$

Now that we have shown $J'_{offline}(\pi) \equiv J_{offline}(\pi)$ and hence solving the same optimization problem, we proceed to derive connections with mixture occupancy matching which follows through an application of Jensen's inequality:

$$\log J'_{offline}(\pi) = \log \mathbb{E}_{\texttt{Mix}_\beta(d,\rho)(s,a,g)}\left[\beta e^{r(s,a,g)} + (1-\beta)\rho(s,a,g).Z\right] \tag{32}$$

$$\geqslant \mathbb{E}_{\texttt{Mix}_\beta(d,\rho)(s,a,g)}\left[\log(\beta e^{r(s,a,g)} + (1-\beta)\rho(s,a,g).Z)\right] \tag{33}$$

$$\tag{34}$$

$$= \mathbb{E}_{\texttt{Mix}_\beta(d,\rho)(s,a,g)}[\log(\beta q(s,a,g) + (1-\beta)\rho(s,a,g))] + \log Z \tag{35}$$

$$= -D_{KL}[\texttt{Mix}_\beta(d,\rho)(s,a,g) \| \texttt{Mix}_\beta(q,\rho)(s,a,g)] - \mathcal{H}(\texttt{Mix}_\beta(d,\rho)(s,a,g)) + \log Z \tag{36}$$

For any f-divergence that upperbounds the KL divergence since $Z \geqslant 1$ we have:

$$\log J'_{offline}(\pi) + \frac{1}{\alpha}\mathcal{H}(\texttt{Mix}_\beta(d,\rho)(s,a,g)) \geqslant -\frac{1}{\alpha}D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g) \| \texttt{Mix}_\beta(q,\rho)(s,a,g)) \tag{37}$$

Further simplifying using Eq 31:

$$\log J_{offline}(\pi) + \frac{1}{\alpha}\mathcal{H}(\texttt{Mix}_\beta(d,\rho)(s,a,g) + C \geqslant -\frac{1}{\alpha}D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g) \| \texttt{Mix}_\beta(q,\rho)(s,a,g)) \tag{38}$$

□

Optimizing the mixture distribution matching objective of SMORe maximizes a variant of *offline/dataset regularized* GCRL objective where the entropy for distribution $\texttt{Mix}_\beta(d,\rho)(s,a,g)$ is jointly maximized. Therefore we have shown that the minimizing discrepancy of mixture distribution occupancy maximizes a lower bounds to an offline variant of maxent GCRL objective.

A.2 Convex Conjugates and f-divergences

We first review the basics of duality in reinforcement learning. Let $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ be a convex function. The convex conjugate $f^* : \mathbb{R}_+ \rightarrow \mathbb{R}$ of f is defined by:

$$f^*(y) = \sup_{x \in \mathbb{R}_+}[xy - f(x)]. \tag{39}$$

The convex conjugates have the important property that f^* is also convex and the convex conjugate of f^* retrieves back the original function f. We also note an important relation regarding f and f^*: $(f^*)' = (f')^{-1}$, where the $'$ notation denotes first derivative.

Going forward, we would be dealing extensively with f-divergences. Informally, f-divergences (Rényi, 1961) are a measure of distance between two probability distributions. Here's a more formal definition:

Let P and Q be two probability distributions over a space $\mathcal{Z}$ such that P is absolutely continuous with respect to Q [4]. For a function $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ that is a convex lower semi-continuous and $f(1) = 0$, the f-divergence of P from Q is

$$D_f(P \,||\, Q) = \mathbb{E}_{z\sim Q}\left[f\left(\frac{P(z)}{Q(z)}\right)\right]. \tag{40}$$

Table 3 lists some common f-divergences with their generator functions f and the conjugate functions f^*.

A.3 SMORe: Dual objective for Offline Goal conditioned reinforcement learning

In this section, we derive the dual objective for solving the multi-task occupancy problem formulation for GCRL. First, we derive the original variant of SMORe for the GCRL problem and later derive the action-free SMORe variant for the interested readers.

[4] Let z denote the random variable. For any measurable set $Z \subseteq \mathcal{Z}$, $Q(z \in Z) = 0$ implies $P(z \in Z) = 0$.

Divergence Name	Generator $f(x)$	Conjugate $f^*(y)$
KL (Reverse)	$x \log x$	$e^{(y-1)}$
Squared Hellinger	$(\sqrt{x}-1)^2$	$\frac{y}{1-y}$
Pearson χ^2	$(x-1)^2$	$y + \frac{y^2}{4}$
Total Variation	$\frac{1}{2}\lvert x-1\rvert$	y if $y \in [-\frac{1}{2}, \frac{1}{2}]$ otherwise ∞
Jensen-Shannon	$-(x+1)\log(\frac{x+1}{2}) + x\log x$	$-\log(2-e^y)$

Table 3: List of common f-divergences.

Theorem 1. *The dual problem to the primal occupancy matching objective (Equation 6) is given by:*

$$\max_{\pi_g} \min_{S} \beta(1-\gamma)\mathbb{E}_{d_0,\pi_g}[S(s,a,g)] + \mathbb{E}_{Mix_\beta(q,\rho)}[f^*(\gamma P^{\pi_g} S(s,a,g) - S(s,a,g))] \qquad (7)$$

$$- (1-\beta)\mathbb{E}_\rho[\gamma P^{\pi_g} S(s,a,g) - S(s,a,g)],$$

where f^* is conjugate function of f and S is the Lagrange dual variable defined as $S : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}$. Moreover, as strong duality holds from Slater's conditions the primal and dual share the same optimal solution π_g^* for any offline transition distribution ρ.

Proof. Recall that: $\mathtt{Mix}_\beta(d,\rho)(s,a,g) := \beta d(s,a,g) + (1-\beta)\rho(s,a,g)$ and $\mathtt{Mix}_\beta(q,\rho)(s,a,g) := \beta q(s,a,g) + (1-\beta)\rho(s,a,g)$. $\mathtt{Mix}_\beta(d,\rho)(s,a,g)$ denotes the mixture between the current agent's joint-goal visitation distribution with an offline transition dataset potentially suboptimal and $\mathtt{Mix}_\beta(q,\rho)(s,a,g)$ is the mixture between the expert's visitation distribution with arbitrary experience from the offline transition dataset. Minimizing the divergence between these visitation distributions still solves the occupancy problem, i.e $d^{\pi_g} = q$ when q is achievable. We start with the primal formulation from Eq 6 for mixture divergence regularization:

$$\max_{d(s,a,g)\geqslant 0, \pi(a|s)} -D_f(\mathtt{Mix}_\beta(d,\rho)(s,a,g) \,||\, \mathtt{Mix}_\beta(q,\rho)(s,a,g))$$

$$\text{s.t } d(s,a,g) = (1-\gamma)\rho_0(s,g).\pi(a|s,g) + \gamma\pi(a|s,g)\sum_{s',a'} d(s',a',g)p(s|s',a').$$

Applying Lagrangian duality and convex conjugate (39) to this problem, we can convert it to an unconstrained problem with dual variables $S(s,a,g)$ defined for all $s, a \in \mathcal{S} \times \mathcal{A} \times \mathcal{G}$:

$$\max_{\pi, d\geqslant 0} \min_{S(s,a,g)} -D_f(\mathtt{Mix}_\beta(d,\rho)(s,a,g) \,||\, \mathtt{Mix}_\beta(q,\rho)(s,a,g))$$

$$+ \sum_{s,a,g} S(s,a,g)\left((1-\gamma)d_0(s,g).\pi(a|s,g) + \gamma\sum_{s',a'} d(s',a',g)p(s|s',a')\pi(a|s,g) - d(s,a,g)\right) \qquad (41)$$

$$= \max_{\pi, d\geqslant 0} \min_{S(s,a,g)} (1-\gamma)\mathbb{E}_{d_0(s,g),\pi(a|s,g)}[S(s,a,g)]$$

$$+ \mathbb{E}_{s,a,g\sim d}\left[\gamma\sum_{s',a'} p(s'|s,a)\pi(a'|s')S(s',a',g) - S(s,a,g)\right] \qquad (42)$$

$$- D_f(\mathtt{Mix}_\beta(d,\rho)(s,a,g) \,||\, \mathtt{Mix}_\beta(q,\rho)(s,a,g)) \qquad (43)$$

$$= \max_{\pi, d\geqslant 0} \min_{S(s,a,g)} \beta(1-\gamma)\mathbb{E}_{d_0(s,g),\pi(a|s,g)}[S(s,a,g)]$$

$$+ \beta\mathbb{E}_{s,a,g\sim d}\left[\gamma\sum_{s',a'} p(s'|s,a)\pi(a'|s')S(s',a',g) - S(s,a,g)\right] \qquad (44)$$

$$
\begin{aligned}
&+(1-\beta)\mathbb{E}_{s,a,g\sim\rho}\left[\gamma\sum_{s',a'}p(s'|s,a)\pi(a'|s')S(s',a',g)-S(s,a,g)\right] \\
&-(1-\beta)\mathbb{E}_{s,a,g\sim\rho}\left[\gamma\sum_{s',a'}p(s'|s,a)\pi(a'|s',g)S(s',a',g)-S(s,a,g)\right] && (45)\\
&-D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g)\,||\,\texttt{Mix}_\beta(q,\rho)(s,a,g)) && (46)
\end{aligned}
$$

Now using the fact that strong duality holds in this problem we can swap the inner max and min resulting in:

$$
\begin{aligned}
&=\max_{\pi}\min_{S(s,a,g)}\max_{\texttt{Mix}_\beta(d,\rho)(s,a,g)\geqslant 0}\beta(1-\gamma)\mathbb{E}_{d_0(s,g),\pi(a|s,g)}[S(s,a,g)] \\
&+\beta\mathbb{E}_{s,a,g\sim d}\left[\gamma\sum_{s',a'}p(s'|s,a)\pi(a'|s')S(s',a',g)-S(s,a,g)\right] \\
&+(1-\beta)\mathbb{E}_{s,a,g\sim\rho}\left[\gamma\sum_{s',a'}p(s'|s,a)\pi(a'|s')S(s',a',g)-S(s,a,g)\right] \\
&-(1-\beta)\mathbb{E}_{s,a,g\sim\rho}\left[\gamma\sum_{s',a'}p(s'|s,a)\pi(a'|s',g)S(s',a',g)-S(s,a,g)\right] && (47)\\
&-D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g)\,||\,\texttt{Mix}_\beta(q,\rho)(s,a,g)) && (48)\\
& && (49)
\end{aligned}
$$

We can now apply the convex conjugate (Eq. (39)) definition to obtain a closed form for the inner maximization problem simplifying to:

$$
\begin{aligned}
&\max_{\pi(a|s,g)}\min_{S(s,a,g)}\beta(1-\gamma)\mathbb{E}_{d_0(s,g),\pi(a|s,g)}[S(s,a,g)] \\
&+\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[f^*(\gamma\sum_{s',a'}p(s'|s,a,g)\pi(a'|s')S(s',a',g)-S(s,a,g))\right] \\
&-(1-\beta)\mathbb{E}_{s,a,g\sim\rho}\left[\gamma\sum_{s',a'}p(s'|s,a,g)\pi(a'|s')S(s',a',g)-S(s,a,g)\right] && (50)
\end{aligned}
$$

This completes our derivation of the SMORe objective. Since strong duality holds (objective convex, constraints linear and feasible), SMORe and the primal mixture occupancy matching share the same global optima π^*_g. □

A.4 Action-free SMORe: Dual-V objective for offline goal conditioned reinforcement learning

The primal problem in Equation 6 is over-constrained. The objective determines the visitation distribution d uniquely under a fixed policy. It turns out we can further relax this constraint to get an objective that results in the same optimal solution (Agarwal et al., 2019) π^*_g by rewriting our primal formulation as:

$$
\begin{aligned}
&\max_{d(s,a,g)\geqslant 0}-D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g)\,||\,\texttt{Mix}_\beta(q,\rho)(s,a,g)) \\
&\text{s.t}\ \sum_a d(s,a,g)=(1-\gamma)\rho_0(s,g)+\gamma\sum_{s',a'}d(s',a',g)p(s|s',a'). && (51)
\end{aligned}
$$

Theorem 2. *Let $y(s,a,g)=\gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}[S(s',g)]-S(s,g)$. The action-free dual problem to the multi-task mixture occupancy matching objective (Equation 51) is given by:*

$$
\begin{aligned}
&\min_{S(s,g)}\beta(1-\gamma)\mathbb{E}_{d_0(s,g)}[S(s,g)] \\
&+\mathbb{E}_{s,a,g\sim Mix_\beta(q,\rho)(s,a,g)}\left[\max\left(0,(f')^{-1}\left(y(s,a,g)\right)\right)y(s,a,g)-f\left(\max\left(0,(f')^{-1}\left(y(s,a,g)\right)\right)\right)\right] \\
&\qquad-(1-\beta)\mathbb{E}_{s,a,g\sim\rho}\left[\gamma\sum_{s'}p(s'|s,a)S(s',g)-S(s,g)\right]
\end{aligned}
$$

where S is the lagrange dual variable defined as $S : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$. Moreover, strong duality holds from Slater's conditions the primal and dual share the same optimal solution π_g^* for any offline transition distribution d^O.

Proof. Proceeding as before and applying Lagrangian duality and convex conjugate (39) to this problem, we can convert it to an unconstrained problem with dual variables $S(s, g)$ defined for all $s, g \in \mathcal{S} \times \mathcal{G}$:

$$
\begin{aligned}
&\max_{d \geqslant 0} \min_{S(s,g)} -D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g) \,||\, \texttt{Mix}_\beta(q,\rho)(s,a,g)) \\
&+ \sum_{s,g} S(s,g) \left((1-\gamma) d_0(s,g) + \gamma \sum_{s',a',g} d(s',a',g) p(s|s',a',g) - \sum_a d(s,a,g) \right) \qquad (52) \\
&= \max_{d \geqslant 0} \min_{S(s,g)} (1-\gamma) \mathbb{E}_{d_0(s,g)}[S(s,g)] \\
&+ \mathbb{E}_{s,a,g \sim d} \left[\gamma \sum_{s'} p(s'|s,a) \pi(a'|s') S(s',g) - S(s,g) \right] \qquad (53) \\
&- D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g) \,||\, \texttt{Mix}_\beta(q,\rho)(s,a,g)) \qquad (54)
\end{aligned}
$$

$$
\begin{aligned}
&= \max_{d \geqslant 0} \min_{S(s,g)} \beta(1-\gamma) \mathbb{E}_{d_0(s,g)}[S(s,g)] \\
&+ \beta \mathbb{E}_{s,a,g \sim d} \left[\gamma \sum_{s'} p(s'|s,a) S(s',g) - S(s,g) \right] \\
&+ (1-\beta) \mathbb{E}_{s,a,g \sim d^O} \left[\gamma \sum_{s'} p(s'|s,a) S(s',g) - S(s,g) \right] \\
&- (1-\beta) \mathbb{E}_{s,a,g \sim d^O} \left[\gamma \sum_{s'} p(s'|s,a) S(s',g) - S(s,g) \right] \qquad (55) \\
&- D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g) \,||\, \texttt{Mix}_\beta(q,\rho)(s,a,g)) \qquad (56)
\end{aligned}
$$

Now using the fact that strong duality holds in this problem we can swap the inner max and min resulting in:

$$
\begin{aligned}
&= \min_{S(s,g)} \max_{\texttt{Mix}_\beta(d,\rho)(s,a,g) \geqslant 0} \beta(1-\gamma) \mathbb{E}_{d_0(s,g)}[S(s,g)] \\
&+ \beta \mathbb{E}_{s,a,g \sim d} \left[\gamma \sum_{s'} p(s'|s,a) S(s',g) - S(s,g) \right] \\
&+ (1-\beta) \mathbb{E}_{s,a,g \sim d^O} \left[\gamma \sum_{s'} p(s'|s,a) S(s',g) - S(s,g) \right] \\
&- (1-\beta) \mathbb{E}_{s,a,g \sim d^O} \left[\gamma \sum_{s'} p(s'|s,a) S(s',g) - S(s,g) \right] \qquad (57) \\
&- D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g) \,||\, \texttt{Mix}_\beta(q,\rho)(s,a,g)) \qquad (58)
\end{aligned}
$$

Unlike previous case where constraints uniquely define a valid d for any given π, in this case we need to take into account the hidden constraint $d \geqslant 0$ or equivalently $\texttt{Mix}_\beta(d,\rho)(s,a,g) \geqslant 0$. To incorporate the non-negativity constraints we consider the inner maximization separately and derive a closed-form solution that adheres to the non-negativity constraints. Recall $y(s,a,g) = \mathbb{E}_{s' \sim p(s,a)}[S(s',g)] - S(s,g)$.

$$\max_{\texttt{Mix}_\beta(d,\rho)(s,a,g)\geqslant 0} \mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(d,\rho)(s,a,g)}\left[\gamma\sum_{s'} p(s'|s,a)S(s',g) - S(s,g)\right] \\ - D_f(\texttt{Mix}_\beta(d,\rho)(s,a,g) \mid\mid \texttt{Mix}_\beta(q,\rho)(s,a,g))$$

We can now construct the Lagrangian dual to incorporate the constraint $\texttt{Mix}_\beta(d,\rho)(s,a,g)\geqslant 0$ in its equivalent form $w(s,a,g)\geqslant 0$ and obtain the following where $w \triangleq \frac{\texttt{Mix}_\beta(d,\rho)(s,a,g)}{\texttt{Mix}_\beta(q,\rho)(s,a,g)}$:

$$\max_{w(s,a,g)}\max_{\lambda\geqslant 0}\mathbb{E}_{s,a\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}[w(s,a,g)y(s,a,g)] - \mathbb{E}_{\texttt{Mix}_\beta(q,\rho)(s,a,g)}[f(w(s,a,g))] + \sum_{s,a,g}\lambda(w(s,a,g)-0) \tag{59}$$

Since strong duality holds, we can use the KKT constraints to find the solutions $w^*(s,a,g)$ and $\lambda^*(s,a,g)$.

1. **Primal feasibility**: $w^*(s,a,g)\geqslant 0 \ \ \forall\ s,a$
2. **Dual feasibility**: $\lambda^*\geqslant 0 \ \ \forall\ s,a$
3. **Stationarity**: $\texttt{Mix}_\beta(q,\rho)(s,a,g)(-f'(w^*(s,a,g)) + y(s,a,g) + \lambda^*(s,a,g)) = 0 \ \ \forall\ s,a$
4. **Complementary Slackness**: $(w^*(s,a,g)-0)\lambda^*(s,a,g) = 0 \ \ \forall\ s,a$

Using stationarity we have the following:

$$f'(w^*(s,a,g)) = y(s,a,g) + \lambda^*(s,a,g) \ \ \forall\ s,a,g \tag{60}$$

Now using complementary slackness, only two cases are possible $w^*(s,a,g)\geqslant 0$ or $\lambda^*(s,a,g)\geqslant 0$. Combining both cases we arrive at the following solution for this constrained optimization:

$$w^*(s,a) = \max\left(0, {f'}^{-1}(y(s,a,g))\right) \tag{61}$$

Using the optimal closed-form solution (w^*) for $\texttt{Mix}_\beta(d,\rho)(s,a,g)$ of the inner optimization in Eq. (57) we obtain

$$\min_{S(s,a)} \beta(1-\gamma)\mathbb{E}_{d_0(s)}[S(s,g)] \\ + \mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[\max\left(0,(f')^{-1}\left(y(s,a,g)\right)\right)y(s,a,g) - \alpha f\left(\max\left(0,(f')^{-1}\left(y(s,a,g)\right)\right)\right)\right] \\ - (1-\alpha)\mathbb{E}_{s,a\sim\rho}\left[\gamma\sum_{s'}p(s'|s,a)\pi(a'|s')S(s',g) - S(s,g)\right] \tag{62}$$

For deterministic dynamics, this reduces to the action-free SMORe objective:

$$\min_{S(s,a)} \beta(1-\gamma)\mathbb{E}_{d_0(s)}[S(s,g)] \\ + \mathbb{E}_{s,a\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[\max\left(0,(f')^{-1}\left(y(s,a,g)\right)\right)y(s,a,g) - f\left(\max\left(0,(f')^{-1}\left(y(s,a,g)\right)\right)\right)\right] \\ - (1-\beta)\mathbb{E}_{s,a\sim\rho}\left[\gamma S(s',g) - S(s,g)\right] \tag{63}$$

where $y(s,a,g) = \gamma S(s',g) - S(s,g)$.

Note that we no longer need actions in the offline dataset to learn an optimal goal conditioned score function. This score function can be used to learn presentation in action-free datasets as well as for transfer of value function across differing action-modalities where agents share the same observation space (eg. images as observations).

□

B SMORe ALGORITHMIC DETAILS

B.1 SMORe WITH COMMON f-DIVERGENCES

a. KL divergence

We consider the reverse KL divergence and start with the general SMORe objective:

$$\max_{\pi_g}\min_{S}\beta(1-\gamma)\mathbb{E}_{d_0,\pi_g}[S(s,a,g)]+\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}[f^*(\gamma P^{\pi_g}S(s,a,g)-S(s,a,g))] \\ -(1-\beta)\mathbb{E}_{s,a,g\sim\rho}[\gamma P^{\pi_g}S(s,a,g)-S(s,a,g)] \tag{64}$$

Plugging in the conjugate f^* for reverse KL divergence we get:

$$\max_{\pi_g}\min_{S}\beta(1-\gamma)\mathbb{E}_{d_0,\pi_g}[S(s,a,g)]+\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[e^{(\gamma P^{\pi_g}S(s,a,g)-S(s,a,g))}\right] \\ -(1-\beta)\mathbb{E}_{s,a,g\sim\rho}[\gamma P^{\pi_g}S(s,a,g)-S(s,a,g)] \tag{65}$$

Using the telescoping sum for the last term in the objective above, we can simplify it as follows:

$$\max_{\pi_g}\min_{S}\beta(1-\gamma)\mathbb{E}_{d_0,\pi_g}[S(s,a,g)]+\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[e^{(\gamma P^{\pi_g}S(s,a,g)-S(s,a,g))}\right] \\ +(1-\beta)\mathbb{E}_{s,g\sim d_0,a\sim\rho(\cdot|s,g)}[S(s,a,g)] \tag{66}$$

With the initial state distribution d_0 set to the offline dataset distribution ρ, and Since our initial state distribution is the same as offline data distribution, we get:

$$\max_{\pi_g}\min_{S}\beta(1-\gamma)\mathbb{E}_{\rho,\pi_g}[S(s,a,g)]+\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[e^{(\gamma P^{\pi_g}S(s,a,g)-S(s,a,g))}\right] \\ +(1-\beta)\mathbb{E}_{\rho}[S(s,a,g)] \tag{67}$$

Collecting terms together we get:

$$\max_{\pi_g}\min_{Q}\mathbb{E}_{\rho}[\mathbb{E}_{a\sim\pi}[\beta(1-\gamma)S(s,a,g)]+\mathbb{E}_{a\sim\rho}[(1-\beta)S(s,a,g)]] \\ +\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[e^{(\gamma P^{\pi_g}S(s,a,g)-S(s,a,g))}\right] \tag{68}$$

The objective for SMORe with reverse KL divergence pushes down the "score" of offline dataset transitions selectively (without pushing down score of the goal-transition distribution) while minimizing the term resembling bellman regularization that also encourages increasing score at the mixture dataset jointly over the offline dataset as well as the goal transition distribution.

b. Pearson chi-squared divergence

We consider the Pearson χ^2 and start with the general SMORe objective:

$$\max_{\pi_g}\min_{S}\beta(1-\gamma)\mathbb{E}_{d_0,\pi_g}[S(s,a,g)]+\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}[f^*(\gamma P^{\pi_g}S(s,a,g)-S(s,a,g))] \\ -(1-\beta)\mathbb{E}_{s,a,g\sim\rho}[\gamma P^{\pi_g}S(s,a,g)-S(s,a,g)] \tag{69}$$

With the initial state distribution d_0 set to the offline dataset distribution ρ, and plugging in the conjugate f^* for Pearson χ^2 divergence we get:

$$\max_{\pi_g}\min_{S}\beta(1-\gamma)\mathbb{E}_{d_0,\pi_g}[S(s,a,g)]+0.25\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}[(\gamma P^{\pi_g}S(s,a,g)-S(s,a,g))^2] \\ +\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}[(\gamma P^{\pi_g}S(s,a,g)-S(s,a,g))]-(1-\beta)\mathbb{E}_{s,a,g\sim\rho}[\gamma P^{\pi_g}S(s,a,g)-S(s,a,g)] \tag{70}$$

Using the fact that $\texttt{Mix}_\beta(q,\rho)(s,a,g) = \beta q(s,a,g) + (1-\beta)\rho(s,a,g)$, we can further simplify the above equation to:

$$\max_{\pi_g}\min_S \beta(1-\gamma)\mathbb{E}_{d_0,\pi_g}[S(s,a,g)] + 0.25\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[(\gamma P^{\pi_g}S(s,a,g) - S(s,a,g))^2\right] + \beta\mathbb{E}_{s,a,g\sim q}[(\gamma P^{\pi_g}S(s,a,g) - S(s,a,g))] \quad (71)$$

Collecting terms together we get:

$$\max_{\pi_g}\min_S \beta(1-\gamma)\mathbb{E}_{\rho,\pi_g}[S(s,a,g)] + \beta\mathbb{E}_{s,g\sim q,a\sim\pi_g}[\gamma P^{\pi_g}S(s,a,g)] - \beta\mathbb{E}_{s,a,g\sim q}[S(s,a,g)] + 0.25\mathbb{E}_{s,a,g\sim\texttt{Mix}_\beta(q,\rho)(s,a,g)}\left[(\gamma P^{\pi_g}S(s,a,g) - S(s,a,g))^2\right] \quad (72)$$

Observing the equation above, we note that the first two terms decrease score at offline data distribution as well as the goal transition distribution when actions are sampled according to the policy π_g. Simultaneously the third term pushes score up for the $\{s,a,g\}$ tuples that are sampled from goal transition distribution. Finally the last term encouraged enforces a bellman regularization enforcing smoothness is the scores of neighbouring states.

C SMORe EXPERIMENTAL DETAILS

C.1 Tasks with observations as states

Environments: For the offline GCRL experiments we consider the benchmark used in prior work GoFar and extend it with locomotion tasks. For the manipulations tasks we consider the Fetch environment and a dextrous shadow hand environment. Fetch environments (Plappert et al., 2018) consists of a manipulator with seven degrees of freedom along with a parallel gripper. The set of environments get a sparse reward of 1 when the goal is within 5 cm and 0 otherwise. The action space is 4 dimensional (3 dimension cartesian control + 1 dimension gripper control). The shadow hand is 24 DOF manipulator with 20-dimensional action space. The goal is 15-dimension specifying the position for each of the five fingers. The tolerance for goal reaching is 1 cm. For the locomotion environments, the task is to achieve a particular velocity in the x direction and stay at the velocity. For HalfCheetah, the target velocity is set to 11.0 and for Ant the target velocity is 5.0. For locomotion environments, the tolerance for goal reaching if 0.5. The MuJoCo environments used in this work are licensed under CC BY 4.0.

Offline Datasets: We use existing datasets from the offline GCRL benchmark used in (Ma et al., 2022) for all manipulation tasks except Reacher, SawyerReach, and SawyerDoor. For Reacher, SawyerReach, and SawyerDoor we use existing datasets from (Yang et al., 2022). These datasets are comprised on x% random data and (100-x)% expert data depending on the coverage over goals reached in individual datasets. We create our own datasets for locomotion by using 'random/medium/medium-replay' data as our offline (suboptimal) data combined with 30 trajectories from corresponding 'expert' datasets. The datasets used from D4RL are licensed under Apache 2.0.

Baselines: To benchmark and analyze the performance of our proposed methods for offline imitation learning with suboptimal data, we consider the following representative baselines in this work: GoFAR (Ma et al., 2022), WGCSL (Yang et al., 2022), GCSL (Ghosh et al., 2019), and Actionable Models (Chebotar et al., 2021), Contrastive RL (Eysenbach et al., 2020) and GC-IQL Kostrikov et al. (2021). GoFAR is a dual occupancy matching approach to GCRL that formulates it as a weighted regression problem. WGCSL and GSCL use goal-conditioned behavior cloning with goal relabelling as the base algorithms and WGCL uses weights to learn improved policy over GCSL. Actionable models uses conservative learning with goal chaining to learn goal-reaching behaviours using offline datasets. Contrastive RL treats GCRL as a classification problem - contrastive goals that are achieved in trajectory from random goals. Finally, GC-IQL extends the single task offline RL algorithm IQL to GCRL.

The open-source implementations of the baselines GoFAR, WGCSL, GCSL, Actionable models, Contrastive RL and IQL are provided by the authors (Ma et al., 2022) and employed in our experiments. We use the hyperparameters provided by the authors, which are consistent with those used in the original GoFAR paper, for all the MuJoCo locomotion and manipulation environments. We implement contrastive learning using the code from Contrastive RL repository. GC-IQL is implemented using code from author's implementation found here.

Architecture and Hyperparameters For the baselines, we use tuned hyperparameters from previous works that were tuned on the same set of tasks and datasets. Implementation for `SMORe` shares the same network architecture as baselines. GoFAR additionally requires training a discriminator. For all experiments, all methods are trained for 10 seeds with each training run. Fetch manipulation (except Push) tasks are trained for a maximum of 400k minibatch updates of size 512 whereas all other environments training is done for 1M minibatch updates. The expectile parameter τ was searched over [0.65 ,0.7,0.8,0.85]. For the results shown in table 1, Fetch and Sawyer environments use $\tau = 0.8$, Locomotion and Adroit hand environments use $\tau = 0.7$. In general, the HER ratio is searched over [0.2,0.5,0.8,1.0] for all methods and the best one was selected. HER ratio of 0.8 gave best performance across all tasks for `SMORe`.

The architectures and hyperparameters for all methods are reported in Table 4.

Hyperparameter	**Value**
Policy updates n_{pol}	1
Policy learning rate	3e-4
Value learning rate	3e-4
MLP layers	(256,256)
LR decay schedule	cosine
Discount factor	0.99
LR decay schedule	cosine
Batch Size	512
Mixture ratio β	0.5
Policy temperature (α)	3.0

Table 4: Hyperparameters for `SMORe`.

C.2 Tasks with observations as images

Hyperparameters	Values
batch size	2048
number of training epochs	300
number of training iterations per epoch	1000
Horizon	400
image encoder architecture	3-layer CNN
policy network architecture	(1024, 4) MLP
critic/score network architecture	(1024, 4) MLP
weight initialization for final layers of critic and policy	$\textsc{Unif}[-10^{-12}, 10^{-12}]$
policy std deviation	0.15
representation dimension	16
data augmentation	random cropping
discount factor	0.99
learning rate	3e-4

Table 5: Hyperparameters for image-observation GCRL from Zheng et al. (2023).

Tasks and dataset Our experiments use a suite of simulated goal-conditioned control tasks based on prior work Zheng et al. (2023). The observations and goals are $48 \times 48 \times 3$ RGB images. The evaluation tasks require multi-trajectory stitching whereas the dataset contains trajectories solving only parts of the evaluation tasks.

Task	Behavior cloning		Contrastive RL	RL+sparse reward	
	WGCSL	GCSL	CRL	AM	IQL
Reacher	15.30 ±0.58	14.01 ±0.36	16.62 ±2.09	23.68±0.58	8.86 ± 0.61
SawyerReach	14.06 ±0.08	12.05±1.23	23.03±1.17	23.37±2.29	36.19 ± 0.01
SawyerDoor	16.79±0.75	18.29±0.94	12.26 ±3.94	16.63 ±0.76	29.31± 0.88
FetchPick	6.87± 0.77	6.54 ± 1.85	0.21± 0.29	0.45 ± 0.32	15.24± 1.27
FetchPush	10.62± 0.98	12.38 ± 1.10	3.60 ± 0.59	2.74 ± 0.70	19.95 ± 1.94
FetchSlide	2.62± 1.15	2.03 ± 0.01	0.41± 0.03	0.31 ± 0.31	3.25 ± 1.02

Table 6: Discounted Return for the offline GCRL benchmark with 5% expert data. Results are averaged over 10 seeds.

Task	Behavior cloning		Contrastive RL	RL+sparse reward	
	WGCSL	GCSL	CRL	AM	IQL
Reacher	13.03±0.56	12.17 ±0.8	19.63 ±3.09	24.78±0.23	4.44± 0.70
SawyerReach	11.455±1.37	11.34±1.18	25.35 ±0.8	25.19±0.61	35.73 ± 0.22
SawyerDoor	16.79±0.29	13.20±0.53	14.78 ±5.29	16.59 ±1.39	16.87 ± 4.21
FetchPick	4.39± 1.35	4.99 ± 0.11	0.21± 0.29	0.24± 0.27	11.79 ± 1.78
FetchPush	8.01± 1.96	8.04± 0.34	3.60± 0.59	2.02± 0.48	19.66 ± 1.69
FetchSlide	2.33± 0.23	2.37 ± 0.83	0.44 ± 0.016	0.45± 0.44	1.83± 1.31

Table 7: Discounted Return for the offline GCRL benchmark with 2.5% expert data. Results are averaged over 10 seeds.

In the simulation, we employed an offline manipulation dataset comprising near-optimal examples of basic action sequences, including the initiation of drawer movement, the displacement of blocks, and the grasping of items. The demonstrations exhibit variations in length, ranging between 50 to 100 horizon, while the offline dataset contains a total of 250,000 state transitions in its entirety. It is important to note that the offline trajectories do not depict a complete progression from the initial condition to the final objective. For the purposes of evaluation, we consider 4 tasks similar to Zheng et al. (2023), against which we compare the success rates in realizing these specified objectives.

Baseline and `SMORe` implementations. We use the open-source implementation of Stable-contrastive RL to use the same design decisions and implement `SMORe`, GC-IQL on that codebase. We use the same hyperparameters as the stable-contrastive RL implementation for the shared hyperparameters. The hyperparameters for `SMORe` were kept the same as in Table 4.

D Additional experiments

D.1 Results on offline GCRL benchmark with varying expert coverage in offline dataset

We ablate the effect of dataset quality on the performance of an offline GCRL method in this sections. Table 6, 7, 8 show performance of all methods with 5%, 2.5% and 1% expert data in the offline dataset respectively.

D.2 Success Rate and Final distance to goal on Manipulation tasks

Table 10 and Table 11 reports the success rate and final distance to goal metrics on manipulation tasks.

D.3 Robustness of mixture distribution parameter β

We find that `SMORe` is quite robust to the mixture distribution parameter β except in the environment FetchPush where $\beta = 0.5$ is the most performant. Table 9 shows this result empirically.

Task	Behavior cloning		Contrastive RL	RL+sparse reward	
	WGCSL	GCSL	CRL	AM	IQL
Reacher	13.56±0.69	12.27 ±1.45	17.94±3.71	24.89±0.34	4.28 ± 0.92
SawyerReach	10.71 ±0.69	11.79±1.46	25.61±0.39	25.54±0.95	31.31 ± 2.08
SawyerDoor	15.18 ±0.81	11.89±1.51	10.26±4.61	18.04±1.8	17.11 ± 4.45
FetchPick	1.89 ± 1.22	3.30 ± 0.66	0.42 ± 0.29	0.41 ± 0.22	7.90 ± 1.22
FetchPush	6.44 ± 3.64	6.43 ± 0.56	1.69 ± 1.56	2.63± 3.04	7.11 ± 2.60
FetchSlide	1.77 ± 0.24	1.11± 0.26	0.0 ± 0.0	0.10 ± 0.11	0.80 ± 0.48

Table 8: Discounted Return for the offline GCRL benchmark with 1% expert data. Results are averaged over 10 seeds.

D.4 How much does HER contribute to the performance improvements of SMORe?

GoFAR demonstrated improved performance without relying on HER. The authors also demonstrated that HER is detrimental to GoFAR's performance. In this section, we aim to conduct a similar stufy and see how much HER contributed to SMORe's performance. Table 12 shows that HER gives SMORe a small performance boost and show that SMORe is still able to outperform GoFAR without HER.

D.5 Comparison to variants of GoFAR

GoFAR formulates GCRL as an occupancy matching problem, but it is also suggested that using a discriminator is optional. Without a discriminator, GoFAR reduces to a sparse reward RL problem. Table 13 shows that GoFAR achieves poor performance when a reward function is substituted in place on an discriminator. We also study if the performance benefits we obtain are due to the offline learning strategy we used from IQL. We modify GoFAR with discriminator reward to use expectile loss for value learning and AWR for policy learning. Results in Table 13 shows that no performance gains were observed.

D.6 Offline GCRL with purely suboptimal data

In this experiment, we study offline GCRL from purely suboptimal datasets. Except FetchReach, these datasets provide very sparse coverage of goals expected to reach in evaluation. Table 14 shows the robustness of SMORe even in the setting of poor quality offline data.

D.7 Comparison with in-sample learning methods

In-sample learning methods perform value improvement using bellman backups without OOD action sampling. This makes them a particularly suitable candidate for offline setting. We compare against a number of recent in-sample learning methods, IQL (Kostrikov et al., 2021), SQL/EQL Xu et al. (2023) and XQL (Garg et al., 2023). Table 15 compares SMORe to in-sample learning methods adapted to GCRL.

D.8 Ablating components of SMORe for offline setting

In offline setting, it is well known that bellman backups suffer from overestimation and results in poor policy performance. We validate the utility of the components used in this work in Table 16 : expectile loss function and constrained policy optimization with AWR.

Task	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.8$	$\beta = 0.9$
FetchReach	35.08 ± 0.54	36.57 ± 0.20	36.59± 0.30	36.30± 0.30
FetchPick	26.47± 0.34	27.04± 0.81	27.43± 0.97	27.89 ± 1.19
FetchPush	26.83 ± 1.21	16.20± 1.11	11.50± 1.19	13.85± 5.53
FetchSlide	4.99± 0.40	3.76± 0.75	3.43± 2.4	4.10± 1.20

Table 9: Discounted Return for the offline GCRL benchmark with varying mixture coefficients in offline dataset. Results are averaged over 10 seeds.

Task	Occupancy Matching		Behavior cloning		Contrastive RL	RL+sparse reward	
	SMORe	GoFAR	WGCSL	GCSL	CRL	AM	IQL
Reacher	0.875±0.07	0.90±0.01	0.97±0.014	0.92 ±0.08	0.76±0.74	1.0±0.1	0.26 ± 0.06
SawyerReach	0.98±0.014	0.75±0.04	1.0±0.0	0.98±0.02	0.98±0.018	1.0±0.1	0.81 ± 0.01
SawyerDoor	0.875±0.038	0.5±0.12	0.78 ±0.10	0.5±0. 12	0.22±0.11	0.3±0.11	0.84 ± 0.06
FetchReach	1.0± 0.0	1.0 ± 0.0	1.0± 0.0	0.98 ± 0.05	1.0± 0.0	1.0± 1.0	1.0 ± 0.0
FetchPick	0.925 ± 0.045	0.84 ± 0.09	0.54± 0.16	0.54 ± 0.20	0.42 ± 0.29	0.78 ± 0.15	0.86 ± 0.11
FetchPush	0.90± 0.07	0.88± 0.09	0.76± 0.12	0.72 ± 0.15	0.06± 0.03	0.67± 0.14	0.65 ± 0.052
FetchSlide	0.315± 0.07	0.18 ± 0.12	0.18± 0.14	0.17± 0.13	0.0 ± 0.0	0.11± 0.09	0.26± 0.057
HandReach	0.47± 0.11	0.40 ± 0.20	0.25± 0.23	0.047± 0.10	0.0± 0.0	0.0 ± 0.0	0.0 ± 0.0

Table 10: Success Rate for the offline GCRL benchmark with 10% expert data. Results are averaged over 10 seeds.

Task	Occupancy Matching		Behavior cloning		Contrastive RL	RL+sparse reward	
	SMORe	GoFAR	WGCSL	GCSL	CRL	AM	IQL
Reacher	0.02±0.01	0.03±0.01	0.011±0.01	0.016 ±0.00	0.05±0.03	0.013±0.00	0.12± 0.005
SawyerReach	0.008±0.004	0.04±0.00	0.004±0.00	0.00±0.00	0.01±0.01	0.01 ±0.00	0.053 ± 0.004
SawyerDoor	0.02±0.029	0.18±0.00	0.011±0.00	0.017±0.01	0.14±0.07	0.06 ±0.01	0.019 ± 0.01
FetchReach	0.004± 0.0012	0.018± 0.003	0.007± 0.0043	0.008 ± 0.008	0.007 ± 0.001	0.007 ± 0.001	0.002± 0.001
FetchPick	0.04± 0.018	0.036 ± 0.013	0.094± 0.043	0.108± 0.06	0.25 ± 0.025	0.04± 0.02	0.04± 0.012
FetchPush	0.03± 0.003	0.033± 0.008	0.041± 0.02	0.042± 0.018	0.15± 0.036	0.07±0.039	0.05± 0.006
FetchSlide	0.09± 0.012	0.12 ± 0.02	0.173± 0.04	0.204± 0.051	0.42± 0.05	0.198± 0.059	0.09± 0.013
HandReach	0.039± 0.0108	0.024± 0.009	0.035 ± 0.012	0.038± 0.013	0.04 ± 0.005	0.037 ±0.004	0.08 ± 0.005

Table 11: Final distance to goal for the offline GCRL benchmark with 10% expert data. Results are averaged over 10 seeds.

Task	SMORe	SMORe w/o HER	GoFAR
FetchReach	35.08 ± 0.54	34.86± 1.03	28.2±0.61
SawyerReach	37.67± 0.12	37.34± 0.36	15.34±0.64
SawyerDoor	31.48± 0.46	31.53± 0.62	18.94±0.01
FetchPick	26.47± 0.34	25.72± 3.88	19.7±2.57
FetchPush	26.83 ± 1.21	25.62± 1.67	18.2±3.00
FetchSlide	4.99± 0.40	4.09± 0.33	2.47±1.44

Table 12: Performance gains using HER (resampling ratio=0.8) on `SMORe`. All other hyperparameters are kept the same between SMORe and SMORe w/o HER.

Task	SMORe	GoFAR (discriminator)	GoFAR (sparse reward)	r-GoFAR (sparse reward)	GoFAR (expectile loss+AWR)
FetchReach	35.08 ± 0.54	28.2±0.61	26.1±1.14	0.30±0.43	26.90±0.41
SawyerReach	37.67± 0.12	15.34±0.64	—	0.34±0.33	16.17±3.02
SawyerDoor	31.48± 0.46	18.94±0.01	—	10.36±3.27	22.47±1.13
FetchPick	26.47± 0.34	19.7±2.57	17.4±1.78	0.25± 0.02	18.46± 2.72
FetchPush	26.83 ± 1.21	18.2±3.00	17.4±2.67	4.23±3.96	17.39±5.44
FetchSlide	4.99± 0.40	2.47±1.44	5.13±4.05	0.29±0.03	3.59±2.30

Table 13: Ablation comparison with GoFAR: a) a sparse binary reward is used in place of a learned discriminator in GoFAR b) The policy and value update is replaced by AWR and expectile loss respectively. The main difference with SMORe remains the use of discriminator for training. — indicates these environments were not considered in (Ma et al., 2022). Our reproduced results (denoted by r-) with the official code base for binary results did not match up to the reported results.

Task	SMORe	GoFAR	WGCSL	GC-IQL
FetchReach	35.01± 0.47	27.37±0.4	21.65±0.61	23.72±1.18
SawyerReach	36.26± 0.93	5.89±1.36	7.27±1.14	33.08±0.81
SawyerDoor	20.28± 2.65	15.33±1.30	13.81±2.72	16.05±4.97
FetchPick	0.61± 0.5	0.0±0.0	0.0±0.0	1.31±1.86
FetchPush	6.39± 0.68	4.23±3.96	4.27±3.9	2.63±1.68
FetchSlide	0.42±0.01	0.059±0.08	0.93±0.69	0.75±0.58

Table 14: Discounted Return for the offline GCRL benchmark with 0% expert data. Results are averaged over 10 seeds.

Task	SMORe	GC-IQL	GC-SQL	GC-EQL	GC-XQL
FetchReach	35.08 ± 0.54	34.43± 1.00	35.67±0.70	29.23±0.2	33.94±0.49
SawyerReach	37.67± 0.12	35.18± 0.29	37.10±0.24	30.19±1.66	32.88±2.85
SawyerDoor	31.48± 0.46	25.52± 1.45	27.96±0.45	3.57±3.51	5.85±4.21
FetchPick	26.47± 0.34	16.8± 3.10	18.35±6.67	1.31±1.86	1.31±1.82
FetchPush	26.83 ± 1.21	22.40± 0.74	17.19±2.56	2.64±1.30	3.79±0.21
FetchSlide	4.99± 0.40	4.80± 1.59	4.68±3.32	0.06±0.08	0.36±0.52

Table 15: Comparison of SMORe to in-sample RL methods - IQL (Kostrikov et al., 2021),SQL/EQL (Xu et al., 2023), XQL (Garg et al., 2023) that learn from sparse rewards. — in EQL denotes the learning diverged. We observed IQL to be the most stable alternative compared to SQL, EQL and XQL. SQL, EQL and XQL were implemented using author's official codebase

Task	SMORe	SMORe (w/o AWR)	SMORe (w/o AWR and Expectile loss)
FetchReach	35.08 ± 0.54	0.30±0.29	0.10±0.13
SawyerReach	36.26 ± 0.93	29.31±0.53	29.64±0.62
SawyerDoor	20.28 ± 2.65	5.06±0.52	2.11±1.59
FetchPick	26.47± 0.34	1.79±0.65	1.77±1.51
FetchPush	26.83 ± 1.21	4.60±2.51	2.69±1.01
FetchSlide	4.99± 0.40	0.22±0.33	0.50±0.02

Table 16: Ablating practical components of SMORe. Without adapting for offline setting we consider in this work by using in-sample maximization or constrained policy optimization using AWR the performance degrades as expected. Without in-sample-maximization value function explodes in the offline setting and using policy that maximizes Q can often select OOD actions leading to poor performance.