

## A Appendix

### A.1 Gradient of divergences w.r.t. the policy parameters

We derive the expressions for  $\nabla_{\theta_j} D_{\text{JS}}$  and  $\nabla_{\theta_j} D_{\text{KLS}}$  mentioned in Equation 4.  $\theta_j$  denotes the parameters for  $\pi_j$ . The distribution ratio,  $\zeta_{ij} = \rho_i / \rho_j$ , depends on  $\theta_j$  through  $\rho_j$ . A bar above a symbol signifies that it is a constant w.r.t.  $\theta_j$ ; for instance, while  $\rho_j$  depends on  $\theta_j$ ,  $\bar{\rho}_j$  does not. The derivation uses the property that the expectation of the score function estimator is 0.

**Jenson-Shannon divergence.**

$$D_{\text{JS}}(\rho_i, \rho_j) = \frac{1}{2} \mathbb{E}_{\rho_i} \log \frac{\rho_i}{\rho_i + \rho_j} + \frac{1}{2} \mathbb{E}_{\rho_j} \log \frac{\rho_j}{\rho_i + \rho_j} + \log 2$$

Differentiating with the product rule,

$$\begin{aligned} \nabla_{\theta_j} 2D_{\text{JS}} &= -\mathbb{E}_{\rho_i} \nabla_{\theta_j} \log[\rho_i + \rho_j] + \underbrace{\mathbb{E}_{\rho_j} \nabla_{\theta_j} \log[\rho_j]}_{\substack{=0 \\ \text{Exp. score function}}} + \nabla_{\theta_j} \mathbb{E}_{\rho_j} \log[\bar{\rho}_j] - \mathbb{E}_{\rho_j} \nabla_{\theta_j} \log[\rho_i + \rho_j] - \nabla_{\theta_j} \mathbb{E}_{\rho_j} \log[\rho_i + \bar{\rho}_j] \\ &= -\underbrace{\mathbb{E}_{\rho_i + \rho_j} \nabla_{\theta_j} \log[\rho_i + \rho_j]}_{\substack{=0 \\ \text{Exp. score function}}} + \nabla_{\theta_j} \mathbb{E}_{\rho_j} \log[\bar{\rho}_j] - \nabla_{\theta_j} \mathbb{E}_{\rho_j} \log[\rho_i + \bar{\rho}_j] \\ \nabla_{\theta_j} D_{\text{JS}} &= - (1/2) \nabla_{\theta_j} \mathbb{E}_{\rho_j} \log[1 + \bar{\zeta}_{ij}] \end{aligned}$$

**Symmetric Kullback-Leibler divergence.**

$$D_{\text{KLS}}(\rho_i, \rho_j) = \mathbb{E}_{\rho_i} \log \frac{\rho_i}{\rho_j} - \mathbb{E}_{\rho_j} \log \frac{\rho_i}{\rho_j}$$

Differentiating with the product rule,

$$\nabla_{\theta_j} D_{\text{KLS}} = -\mathbb{E}_{\rho_i} \nabla_{\theta_j} \log[\rho_j] - \nabla_{\theta_j} \mathbb{E}_{\rho_j} \log \bar{\zeta}_{ij} + \underbrace{\mathbb{E}_{\rho_j} \nabla_{\theta_j} \log[\rho_j]}_{\substack{=0 \\ \text{Exp. score function}}}$$

For the first term, interchanging the gradient and the expectation, we can write:

$$\mathbb{E}_{\rho_i} \nabla_{\theta_j} \log[\rho_j] = \sum_{(s,a)} \rho_i \frac{\nabla_{\theta_j} \rho_j}{\bar{\rho}_j} = \sum_{(s,a)} \bar{\zeta}_{ij} \nabla_{\theta_j} \rho_j = \nabla_{\theta_j} \mathbb{E}_{\rho_j} [\bar{\zeta}_{ij}]$$

Therefore,

$$\nabla_{\theta_j} D_{\text{KLS}} = \nabla_{\theta_j} \mathbb{E}_{\rho_j} [-\bar{\zeta}_{ij} - \log \bar{\zeta}_{ij}]$$

### A.2 DualDICE min-max objective with Fenchel conjugates

We start with the *DualDICE* objective from Section 3.3:

$$J(\nu) = \frac{1}{2} \mathbb{E}_{(s,a) \sim \rho_j} [(\nu - \mathcal{B}^{\pi_i} \nu)(s, a)^2] - (1 - \gamma) \mathbb{E}_{\substack{s_0 \sim \mu_0 \\ a_0 \sim \pi_i(s_0)}} [\nu(s_0, a_0)]$$

Fenchel duality provides that  $\frac{1}{2}x^2 = \max_g gx - \frac{1}{2}g^2$  for a scalar  $g \in \mathbb{R}$ . Nachum et al. [18] rewrite the quadratic (first) term in the objective using this maximization and use the interchangeability principle [34] to replace the inner max over scalar  $g$  to a max over functions  $g : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ . Given the definition of the  $\mathcal{B}^{\pi_i}$  operator, this yields the min-max objective:

$$\begin{aligned} \min_{\nu} \max_g J(\nu, g) &= \mathbb{E}_{(s,a) \sim \rho_j, s' \sim p(\cdot|s,a)} \left[ (\nu(s, a) - \gamma \nu(s', a')) g(s, a) - \frac{g(s, a)^2}{2} \right] \\ &\quad - (1 - \gamma) \mathbb{E}_{\substack{s_0 \sim \mu_0 \\ a_0 \sim \pi_i(s_0)}} [\nu(s_0, a_0)] \end{aligned}$$

The distribution ratio is obtained from the saddle-point solution  $(\nu^*, g^*)$  using the following equivalence,  $\zeta_{ij}(s, a) = g^*(s, a) = (\nu^* - \mathcal{B}^{\pi_i} \nu^*)(s, a)$ .

### A.3 Optimality in the Donsker-Varadhan representation

The Donsker-Varadhan representation [26] of the KL-divergence is given by:

$$D_{\text{KL}}(\rho_i || \rho_j) = \sup_{x: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}} \mathbb{E}_{(s,a) \sim \rho_i} [x(s,a)] - \log \mathbb{E}_{(s,a) \sim \rho_j} [e^{x(s,a)}]$$

The optimality is achieved at  $x^*(s,a) = \log \zeta_{ij}(s,a) + C$ , for some constant  $C \in \mathbb{R}$ .

*Proof.* We begin with a re-write of the expression inside the supremum:

$$\begin{aligned} & \mathbb{E}_{(s,a) \sim \rho_i} \left( \log \left[ e^{x(s,a)} \cdot \frac{\rho_i}{\rho_j} \cdot \frac{\rho_j}{\rho_i} \right] \right) - \log \mathbb{E}_{(s,a) \sim \rho_j} [e^{x(s,a)}] \\ &= \underbrace{\mathbb{E}_{(s,a) \sim \rho_i} \left[ \log \frac{\rho_i}{\rho_j} \right]}_{\text{KL}} + \mathbb{E}_{(s,a) \sim \rho_i} \left( \log \left[ \frac{\rho_j}{\rho_i} \cdot e^{x(s,a)} \right] \right) - \log \mathbb{E}_{(s,a) \sim \rho_j} [e^{x(s,a)}] \\ &\leq D_{\text{KL}}(\rho_i || \rho_j) + \log \mathbb{E}_{(s,a) \sim \rho_i} \left[ \frac{\rho_j}{\rho_i} \cdot e^{x(s,a)} \right] - \log \mathbb{E}_{(s,a) \sim \rho_j} [e^{x(s,a)}] \quad (\text{Jensen's inequality}) \\ &= D_{\text{KL}}(\rho_i || \rho_j) + \log \mathbb{E}_{(s,a) \sim \rho_j} [e^{x(s,a)}] - \log \mathbb{E}_{(s,a) \sim \rho_j} [e^{x(s,a)}] \\ &= D_{\text{KL}}(\rho_i || \rho_j) \end{aligned}$$

Therefore, this expression is upper bounded by  $D_{\text{KL}}(\rho_i || \rho_j)$ . To complete the proof, we show that this upper bound is indeed achieved when  $x(s,a) = \log \zeta_{ij}(s,a) + C$ , for some constant  $C \in \mathbb{R}$ . Inserting this into the expression, we get:

$$\begin{aligned} & \mathbb{E}_{(s,a) \sim \rho_i} [\log \zeta_{ij}(s,a) + C] - \log \mathbb{E}_{(s,a) \sim \rho_j} [e^{\log \zeta_{ij}(s,a) + C}] \\ &= D_{\text{KL}}(\rho_i || \rho_j) + C - \log \left( e^C \underbrace{\mathbb{E}_{(s,a) \sim \rho_j} [\zeta_{ij}(s,a)]}_{=1} \right) \\ &= D_{\text{KL}}(\rho_i || \rho_j) \end{aligned}$$

□

### A.4 GenDICE min-max objective with Fenchel conjugates

We start with the *GenDICE* objective from Section 3.3 that minimizes the  $f$ -divergence between the quantities on the two sides of the Bellman flow constraint, along with a penalty regularization:

$$J(\theta) = D_f(\mathcal{T}_{(\pi_i, \rho_j)} \circ \zeta_\theta || \rho_j \cdot \zeta_\theta) + \frac{\lambda}{2} (\mathbb{E}_{\rho_j} [\zeta_\theta] - 1)^2$$

where  $\mathcal{T}_{(\pi_i, \rho_j)}$  is an operator such that:

$$(\mathcal{T}_{(\pi_i, \rho_j)} \circ \zeta_\theta)(s', a') := (1 - \gamma) \mu_0(s') \pi_i(a' | s') + \gamma \int \pi_i(a' | s') p(s' | s, a) \zeta_\theta(s, a) \rho_j(s, a) ds da$$

Let  $g : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  be a function. The  $f$ -divergence could be substituted with its variational representation [35] which involves the Fenchel conjugate ( $f^*$ ) of the  $f$  function in  $D_f$ :

$$D_f(\mathcal{T}_{(\pi_i, \rho_j)} \circ \zeta_\theta || \rho_j \cdot \zeta_\theta) = \max_g \mathbb{E}_{\mathcal{T}_{(\pi_i, \rho_j)} \circ \zeta_\theta} [g(s, a)] - \mathbb{E}_{\rho_j \cdot \zeta_\theta} [f^*(g(s, a))]$$

This expression can be simplified by using the definition of the  $\mathcal{T}_{(\pi_i, \rho_j)}$  operator. Furthermore, since Fenchel duality provides that  $\frac{1}{2}x^2 = \max_u ux - \frac{1}{2}u^2$ , the quadratic penalty regularization can also be written in form of a max over a scalar variable  $u \in \mathbb{R}$ . This yields the min-max objective:

$$\begin{aligned} \min_{\theta} \max_{g, u} J(\theta, g, u) &= (1 - \gamma) \mathbb{E}_{\mu_0(s) \pi_i(a|s)} [g(s, a)] + \gamma \mathbb{E}_{(s,a) \sim \rho_j, s' \sim p(\cdot | s, a)} [\zeta_\theta(s, a) g(s', a')] \\ &\quad - \mathbb{E}_{(s,a) \sim \rho_j} [\zeta_\theta(s, a) f^*(g(s, a))] + \lambda (\mathbb{E}_{\rho_j} [u \zeta_\theta(s, a) - u] - \frac{u^2}{2}) \end{aligned}$$

For a practical instantiation, Zhang et al. [19] suggest the  $\chi^2$  divergence, which is an  $f$ -divergence with  $f(x) = (x - 1)^2$  and  $f^*(x) = x + \frac{x^2}{4}$

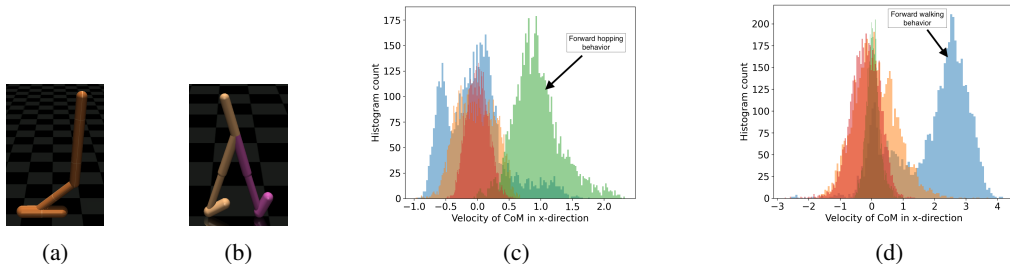


Figure 4: (a)-(b) Rendering of the Hopper and Walker tasks, respectively; (c)-(d) Center-of-mass velocity histograms for Hopper and Walker, respectively, when trained with QD-*GenDICE*-JS. Arrows point to the emergence of locomotion in one of the policies from the ensemble.

## A.5 Further Experiments

**Diversity helps in the absence of environmental rewards.** Designing a task-relevant reward function is typically laborious and error-prone. In the absence of an external reward signal, the diversity objective alone has been previously demonstrated to lead to useful skills [6]. We test the efficacy of our method in this setting using the Hopper and Walker tasks from Gym (Figures 4a- 4b) but modify the code to return a zero reward for each timestep. Thus, the gradient from the quality-enforcing component in Equation 2 is absent and the QD ensemble is trained only to maximize diversity. After training, we generate a few trajectories with the constituent policies and plot histograms with the velocity of the center-of-mass of the bot on the  $x$ -axis and the respective counts on the  $y$ -axis (Figures 4c- 4d). Both tasks are learned with QD-*GenDICE*-JS and each policy is colored differently. We note that the hopping (respectively walking) behavior *emerges* even in the absence of Gym rewards, suggesting that diversity is a strong signal for learning interesting skills.

## A.6 Hyperparameters

Hyper-parameter	Value
Kernel temperatures	$k_{JS}(0.5), k_{KLS}(1.0)$
<i>GenDICE</i> penalty coefficient	10.
Policy network	2 hidden layers, 64 hidden dim, tanh
RL algorithm	PPO (clip=0.2), lr=1e-4
$\gamma$ (Discount), $\lambda$ (GAE [36])	0.99, 0.95

The networks trained in each of the distribution ratio estimation methods (Section 3.3) are as follows:

- **NCE:** For each policy  $\pi_i$ , an estimator for its stationary distribution  $\rho_i$ . These are {2 hidden layers, 100 hidden dim, tanh} networks.
- **DualDICE:** For each  $\zeta_{ij}$  estimation, a network for the  $\nu$  function and a network for the  $g$  function (Appendix A.2). These are {2 hidden layers, 100 hidden dim, tanh} networks.
- **ValueDICE:** For each  $\zeta_{ij}$  estimation, a network for the  $\nu$  function. These are {2 hidden layers, 100 hidden dim, tanh} networks.
- **GenDICE:** For each  $\zeta_{ij}$  estimation, a network for the  $\zeta_\theta$  function and a network for the  $g$  function (Appendix A.4). These are {2 hidden layers, 100 hidden dim, tanh} networks.