Stochastically Dominant Distributional Reinforcement Learning Appendix

1. Experimental Details

1.1. Regression Comparison

This experiment compared empirical first and second moment estimates between quantile regression and solutions of a Wasserstein gradient flow. The distributions were parameterized with the same number of particles, which we varied for values of 5, 10, 20, and 50. Particles were trained on data from a five-component Gaussian mixture model of those sample sizes. We draw samples from each component with equal probability $c_i = 1/5$, using the means $\mu_i \in \{-5, -3, 0, 5, 6, 9\}$, and standard deviations $\sigma_i \in \{1, 2, 1, 2, 1, 0.5\}$, for $i = 1, \cdots, 5$. Models were evaluated on a separate draw of the same size as the training set. We computed the target values, y, using empirical estimates from 10^4 samples. The violin plots show the distribution of root mean square error $RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y - \hat{y})^2}$ samples between the targets and the estimates $\hat{y}$ over $N = 100$ trials.

1.2. Ablation Study

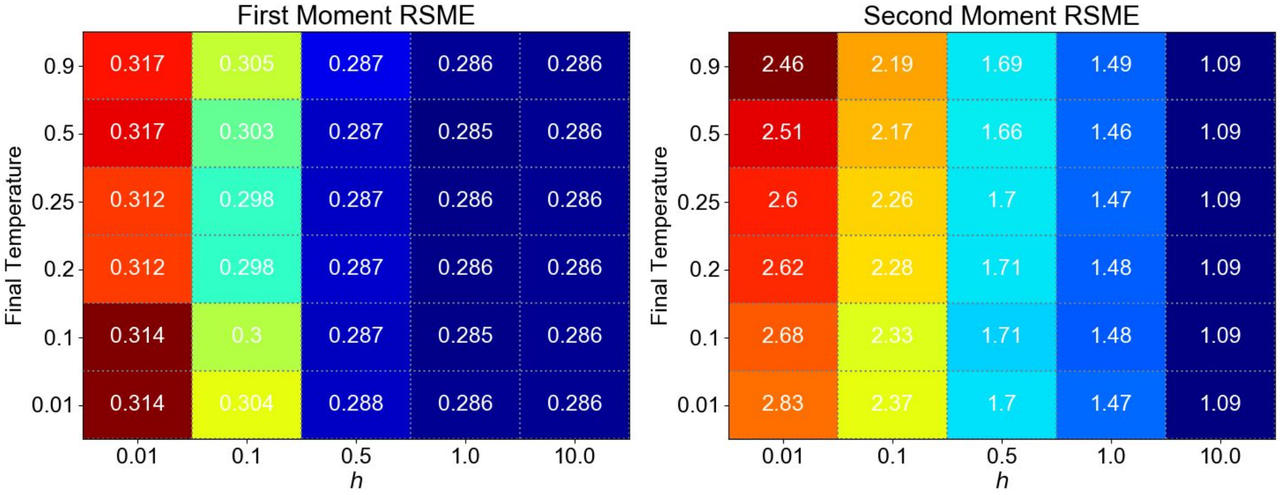


Figure 1. **Ablations:** Temperature β and step size h in proximal loss.

We ablated the minimum temperature $\beta^{-1} \in \{0.01, 0.1, 0.2, 0.25, 0.5, 0.9\}$ and step size $h \in \{0.01, 0.1, 0.5, 1., 10.\}$ over 50 trials. Data came from the five-component mixture model used in the Regression Comparison experiment. We report the root mean square error in the first and second moments with targets computed using 10^4 samples and empirical estimators.

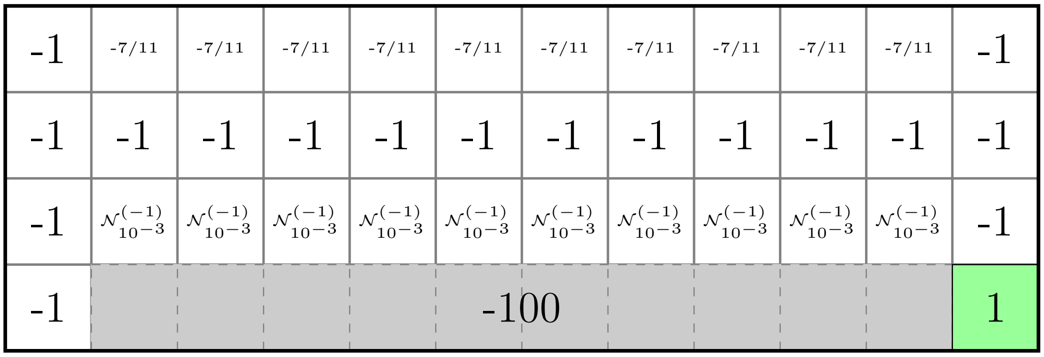
1.3. WGF Policy Evaluation

Here we perform policy evaluation on Monte Carlo (MC) returns from the optimal policy. The optimal policy was obtained by running Q-learning for 10^4 episodes with an $(\epsilon = 0.1)$-greedy behavior policy, $\gamma = 0.9$, learning rate $\alpha = 0.5$, and using an absorbing terminal state. MC returns were computed for each state from 200 rollouts of 200 time steps. We parameterized a discrete distribution with 200 particles initialized from a standard $\mathcal{N}(0, 1)$ Gaussian, then transported them using 100 gradient steps with a step size of 0.5. The proximal loss was annealed down from $\beta^{-1} = 1$ to 0.25 in minimum steps of 0.5; the proximal time step was set to $h = 1$. We report the curves of the proximal loss and the squared value error at each gradient step.

1.4. WGF in the Control Setting

This experiment used the OpenAI Gym (Brockman et al., 2016) environments MountainCar, CartPole, and LunarLander with discrete actions. We estimated particle locations using a two layer fully-connected neural network, each with 256 hidden units. We trained these networks with the WGF proximal loss and the quantile regression loss from (Dabney et al., 2017). Both models regressed 2 quantiles. We used the Adam optimizer (Kingma & Ba, 2015) with a step size of 10^{-3}. We used experience replay with batches of size 32 and a total capacity of 10^4. Agents explored with and $(\epsilon = 0.1)$-greedy policy, using $\gamma = 0.99$ until the absorbing state was reached. We report data for 5 independent trials.

1.5. Control in the Presence of Uncertainty



-1	-7/11	-7/11	-7/11	-7/11	-7/11	-7/11	-7/11	-7/11	-7/11	-7/11	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	$\mathcal{N}_{10^{-3}}^{(-1)}$	-1
-1	-100										1

Figure 2. Modified CliffWorld with multiple solutions.

This experiment used data generated in the CliffWalk environment (Sutton & Barto, 2018). The agent moves in four cardinal directions. We modified the reward function so they were assigned randomly according to Figure 2. Here $\mathcal{N}_{\sigma}^{(\mu)}$ denotes $\mathcal{N}(\mu, \sigma)$. For the random rewards, we clipped them to be within the interval $[-10, 10]$. Both the WGF and quantile regression agents used a tabular representation of 16 particles for each return distribution. Learning occurred with $\gamma = 1$, a horizon length of 500, and the same loss settings used in the policy evaluation experiment. However, the number of gradient steps was limited to 50, unless a tolerance of 10^{-8} was exceeded below first. We report data gathered from $M = 50$ independent trials. The 95% confidence intervals were computed using the standard t-distribution with $M - 1$ degrees of freedom.

2. Mathematical Proofs and References to Supporting Results

This section provides proofs to our main theoretical results. For our supporting results, we provide references to their original sources. We drop the superscript notation introduced in the main paper, used to denote single measures for state-action pairs. All the following results involving probability measure apply for single measures.

Lemma 1. *Let $\tau \in (0, 1)$ and consider $\xi_\tau = F_X^{-1}(\tau)$. Then $F_X^{-2}(\tau) = \mathbf{E}[X \leq \xi_\tau]$.*

Proof. By conjugate duality,

$$\begin{aligned} F_X^{-2}(\tau) &= \tau\xi_\tau - F_X^{(2)}(x), \\ &= \tau\xi_\tau - \tau\mathbf{E}[X - \xi_\tau | X \leq \xi_\tau], \\ &= \tau\mathbf{E}[X | X \leq \xi_\tau], \\ &= \mathbf{E}[X \leq \xi_\tau]. \end{aligned}$$

□

Proposition 1. $Z^{(s,a_1)} \succeq_{(2)} Z^{(s,a_2)}$ *if, and only if* $\sum_{i=1}^{j} z_{a_1}^{[i]} \geq \sum_{i=1}^{j} z_{a_2}^{[i]},\ \forall\, j = 1, \cdots, N.$

Proof. We prove the result in the context of random returns. However, this holds for general random variables. We consider two random returns induced by the actions a_1 and a_2, respectively denoted $Z^{(s,a_1)}$, $Z^{(s,a_2)}$. Each return is approximated with a discrete Lagrangian measure

$$\mu^{(s,a_1)} \approx \frac{1}{N}\sum_{n=1}^{N}\delta_{z_{a_1}^{(n)}}, \qquad \mu^{(s,a_s)} \approx \frac{1}{N}\sum_{n=1}^{N}\delta_{z_{a_2}^{(n)}}.$$

Given that $Z^{(s,a_1)} \succeq_{(2)} Z^{(s,a_2)}$, we know by the definition that $F^{-2}_{Z^{(s,a_1)}}(\tau) \geq F^{-2}_{Z^{(s,a_2)}}(\tau)$ for all $\tau \in (0,1)$. Invoking Lemma **??** allows us to rewrite the definition with total expectations

$$\mathbf{E}[Z^{(s,a_1)} \leq \xi_{a_1}^{(\tau)}] \geq \mathbf{E}[Z^{(s,a_2)} \leq \xi_{a_2}^{(\tau)}],\ \forall\, \tau \in (0,1).$$

Denote the ordered coordinates of a return distribution to be $z^{[1]} \leq z^{[2]} \leq \cdots \leq z^{[N]}$. Then with particle sets from each measure, we have

$$\sum_{i=1}^{j} z_{a_1}^{[i]} \geq \sum_{i=1}^{j} z_{a_2}^{[i]},\ \forall\, j = 1, \cdots, N.$$

The other implication follows by normalizing the sums with $1/N$ and invoking Lemma **??** again to arrive at the definition. □

Proposition 2 (Fishburn (1980))**.** *Assume μ has two finite moments. Then $X \succeq_{(2)} Y$ implies $\mu_X^{(1)} \geq \mu_Y^{(1)}$ or $\mu_X^{(1)} = \mu_Y^{(1)}$ and $\mu_X^{(2)} \leq \mu_Y^{(2)}$, where $(\cdot)$ denotes a particular moment of the distribution μ.*

Proof. This result follows from Theorem 1 of Fishburn (1980), which proves an ordering dominance of any finite degree. □

Proposition 3. *Let $\{\mu_t\}_{t\in[0,1]}$ be an absolutely-continuous curve in $\mathcal{P}(\mathbb{R})$ with finite second-order moment. Then for $t \in [0,1]$, the vector field $\mathbf{v}_t = \nabla(\frac{\delta E}{\delta t}(\mu))$ defines a gradient flow on $\mathcal{P}(\mathbb{R})$ as $\partial_t \mu_t = -\nabla\cdot(\mu_t \mathbf{v}_t)$, where $\nabla \cdot \mathbf{u}$ is the divergence of some vector $\mathbf{u}$.*

Proof. See Ambrosio (2005), Theorem 8.3.1. □

Proposition 4. *Let $\mu_0 \in \mathcal{P}_2(\mathbb{R})$ have finite free energy $E(\mu_0) < \infty$, and for a given $h > 0$, let $\{\mu_t^{(h)}\}_{t=0}^{K}$ be the solution of the discrete-time variational problem, with measures restricted to $\mathcal{P}_2(\mathbb{R})$, the space with finite second moments. Then as $h \to 0$, $\mu_K^{(h)} \to \mu_T$, where μ_T is the unique solution of the Fokker-Plank equatio at $T = hK$.*

Proof. See Jordan et al. (1998), Theorem 5.1. □

Proposition 5. *Let $\{\mu_t^{(h)}\}_{t=0}^{K}$ be the solution of the discrete-time JKO variational problem, with measures restricted to $\mathcal{P}_2(\mathbb{R})$, the space with finite second moments. Then $E(\mu_t)$ is a decreasing function of time.*

Proof. We show that the free-energy $E(\mu) = F(\mu) + \beta^{-1}H(\mu)$ is a Lyapunov functional for the Fokker-Planck (FP) equation. Following the approach of (Markowich & Villani, 1999), we consider the change of variables $\mu_t = h_t e^{-U}$, where we let $\beta = 1$ without loss of generality. With this, FP is equivalent to

$$\partial_t h_t = \Delta h_t - \nabla U \cdot \nabla h_t. \tag{1}$$

Whenever ϕ is a convex function, one can check the following is a Lyapunov functional for (1), and equivalently FP:

$$\int \phi(h_t)e^{-U}dz = \int \phi(\mu_t e^{U})e^{-U}dz.$$

Differentiating with respect to time shows

$$\frac{d}{dt}\int \phi(h_t)e^{-U}dz = -\int \phi''(h_t)|\nabla h_t|^2 e^{-U}dz < 0.$$

Now consider $\phi(h_t) = h_t \log(h_t) - h_t + 1$. With the identity $\int (h_t - 1)e^{-U}dz = 0$, we find

$$\int \phi(h_t)e^{-U}dz = \int \mu_t \log\left(\frac{\mu_t}{e^{-U}}\right)dz = \int \mu_t(U + \log \mu_t)dz = E(\mu).$$

Thus, the free-energy functional is a Lyapunov function for the Fokker-Planck equation, and $E(\mu_t)$ is a decreasing function of time. In the low-energy state the optimal distributional Bellman equation is satisfied with pure Brownian motion. □

Theorem 1. *If* $\mathcal{T}\mu = \mu$*, then* $\text{Prox}^{\mathcal{W}}_{hE}(\mu) = \mu$ *as* $\beta \to \infty$.

Proof. Let $d(\mu, \nu)$ be some distributional distance between measures μ and ν, such as the supremal k-Wasserstein $= \sup_{s,a} \mathcal{W}_k(\mu, \nu)$. Furthermore, suppose $\mu^* = \mathcal{T}\mu^*$ is the fixed point of the optimal distributional Bellman operator $\mathcal{T}$. We consider the proximal operator

$$\text{Prox}^{\mathcal{W}}_{hE}(\mu_k) = \arg\min_{\mu} \mathcal{W}_2^2(\mu, \mu_k) + 2hE(\mu).$$

It follows that $\mu^* = \mathcal{T}\mu^*$ and

$$\begin{aligned} d(\mathcal{T}\mu^*, \mu^*) &\leq d(\text{Prox}^{\mathcal{W}}_{hE}(\mu^*), \mu^*) = d\bigg(\arg\min_{\mu} \mathcal{W}_2^2(\mu, \mu^*) + 2h \underbrace{E(\mu)}_{0 \text{ as } \beta \to \infty}, \mu^*\bigg), \\ &\leq d\bigg(\arg\min_{\mu} \mathcal{W}_2^2(\mu, \mu^*) = \mu^*, \mu^*\bigg) \\ &\leq 0 \end{aligned}$$

Distance is non-negative, so it must be that $\text{Prox}^{\mathcal{W}}_{hE}(\mu^*) = \mathcal{T}\mu^* = \mu^*$. □

3. Expanded Background

3.1. Euclidean Gradient Flows

Suppose we have a smooth function $F \colon \mathbb{R}^d \to \mathbb{R}$ and an initial point $\mathbf{x}_0 \in \mathbb{R}^d$. The gradient flow of $F(\mathbf{x})$ is defined as the solution to the differential equation $\frac{d\mathbf{x}}{d\tau} = -\nabla F(\mathbf{x}(\tau))$, $\tau > 0$, and $\mathbf{x}(0) = \mathbf{x}_0$. This has a unique solution if ∇F is Lipschitz continuous.

Exact solutions are typically intractable. A standard numerical method, called the Minimizing Movement Scheme (MMS) (Gobbino, 1999), evolves $\mathbf{x}$ iteratively for small steps along the gradient of F at the current point $\mathbf{x}_k$. The next point is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \nabla F(\mathbf{x}_{k+1})h,$$

for the step size h. Determining $\mathbf{x}_{k+1}$ is equivalent to solving the optimization problem

$$\mathbf{x}_{k+1} \in \arg\min_{\mathbf{x}} F(\mathbf{x}) + \frac{||\mathbf{x} - \mathbf{x}_k||_2^2}{2h}.$$

Where the squared Euclidean norm is denoted $|| \cdot ||_2^2$. Convergence of the sequence $\{\mathbf{x}_k\}$ to the exact solution has been established for this method, (Ambrosio, 2005).

3.2. Sinkhorn's Algorithm

We describe how the Kantorovich problem can be made tractable through entropy regularization, then present an algorithm for approximating the $\mathcal{W}_2^2$ distance. The key message is that including entropy reduces the original Optimal Transport problem to one of matrix scaling. Sinkhorn's algorithm can be applied for this purpose to admit unique solutions.

The optimal value of the Kantorovich problem is the exact $\mathcal{W}_2^2$ distance. Given probability measures $\alpha = \sum_{i=1}^{N} \alpha_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{M} \beta_j \delta_{y_j}$, the problem is to compute a minimum-cost mapping, π, defined as a non-negative matrix on the product space of atoms $\{x_1, \cdots, x_N\} \times \{y_1, \cdots, y_M\}$. Denoting the cost to move x_i to y_j as $C_{ij} = ||x_i - y_j||^2$, we have

$$\mathcal{W}_2^2(\alpha, \beta) = \min_{\pi \in \mathbb{R}_{\geq 0}^{N \times M}} \langle \pi, C \rangle = \sum_{ij} \pi_{ij} C_{ij}, \tag{2}$$

$$\text{such that } \pi \mathbf{1}_M = \alpha,\ \pi^\top \mathbf{1}_N = \beta. \tag{3}$$

This approach constitutes a linear program, which unfortunately scales cubically in the number of atoms. We can reduce the complexity by considering an entropically regularized version of the problem. Let ε be a regularization parameter. The new problem is written in terms of the generalized Kullback Leibler (KL) divergence:

$$\mathcal{W}_2^2(\alpha, \beta) \approx \mathcal{W}_\varepsilon(\alpha, \beta) = \min_{\pi \in \mathbb{R}_{\geq 0}^{N \times M}} \langle \pi, C \rangle + \varepsilon \mathsf{KL}(\pi || \alpha \otimes \beta), \tag{4}$$

$$= \sum_{i,j} \pi_{ij} C_{ij} + \varepsilon \sum_{i,j} [\pi_{ij} \log \frac{\pi_{ij}}{\alpha_i \beta_j} - \pi_{ij} + \alpha_i \beta_j], \tag{5}$$

$$\text{such that } \pi \mathbf{1}_M = \alpha,\ \pi^\top \mathbf{1}_N = \beta. \tag{6}$$

The value of $\mathcal{W}_\varepsilon(\alpha, \beta)$ occurs necessarily at the critical point of the constrained objective function

$$\begin{aligned} L_\varepsilon = \sum_{i,j} \pi_{ij} C_{ij} + \varepsilon \sum_{i,j} [\pi_{ij} \log \frac{\pi_{ij}}{\alpha_i \beta_j} - \pi_{ij} + \alpha_i \beta_j] \\ - \sum_i f_i \left(\sum_j \pi_{ij} - \alpha_i \right) - \sum_j g_j \left(\sum_i \pi_{ij} - \beta_j \right), \end{aligned} \tag{7}$$

$$\frac{\partial L_\varepsilon}{\partial \pi_{ij}} = 0 \Longrightarrow \forall\, i, j,\ C_{ij} + \varepsilon \log \frac{\pi^*_{ij}}{\alpha_i \beta_j} = f^*_i + g^*_j. \tag{8}$$

The last line of (8) shows that the entropically-regularized solution is characterized by two vectors $f^* \in \mathbb{R}^N$, $g^* \in \mathbb{R}^M$. With the following definitions

$$u_i = \exp(f^*_i / \varepsilon), \qquad v_j = \exp(g^*_j / \varepsilon), \qquad K_{ij} = \exp(-C_{ij} / \varepsilon), \tag{9}$$

we can write the optimal transport plan as $\pi^* = \mathbf{diag}(\alpha_i u_i) K \mathbf{diag}(v_j \beta_j)$. And the approximate Wasserstein distance can be computed simply as

$$\mathcal{W}_\varepsilon(\alpha, \beta) = \langle \pi^*, C \rangle + \varepsilon \mathsf{KL}(\pi^* || \alpha \otimes \beta) = \sum_{ij} (f^*_i + g^*_j) = \langle f^*, \alpha \rangle + \langle g^*, \beta \rangle$$

We mentioned that Optimal Transport reduces to positive matrix scaling. Indeed, using the vectors u and v, Sinkhorn's algorithm provides a way to iteratively scale K such that the unique solution is π^*. Initialize $u^{(0)} = \mathbf{1}_N$, and $v^{(0)} = \mathbf{1}_M$, then perform the following iterations for all i, j

$$\begin{aligned} v_j^{(1)} &= \frac{1}{[K^\top (\alpha \odot u^{(0)})]_j}, & u_i^{(1)} &= \frac{1}{[K(\beta \odot v^{(1)})]_i}, \\ &\vdots & &\vdots \\ v_j^{(n+1)} &= \frac{1}{[K^\top (\alpha \odot u^{(n)})]_j}, & u_i^{(n+1)} &= \frac{1}{[K(\beta \odot v^{(n+1)})]_i}. \end{aligned} \tag{10}$$

Sinkhorn's algorithm performs coordinate ascent with f and g to maximize the dual maximization problem

$$\mathcal{W}_\varepsilon(\alpha, \beta) = \max_{f \in \mathbb{R}^N, g \in \mathbb{R}^M} \langle f, \alpha \rangle + \langle g, \beta \rangle - \varepsilon \langle \alpha \otimes \beta, \exp\{(f \oplus g - C)/\varepsilon\} - 1 \rangle. \tag{11}$$

Each update consists of kernel products, $K^\top(\alpha \odot u)$ and $K(\beta \odot v)$, and point-wise divisions. We describe this procedure in Algorithm 1, using computations in the log domain to numerically stabilize the updates. The log updates derive from (9) and (12):

$$\begin{aligned} \log v_j &= -\log \sum_i K_{ij}\alpha_i u_i & \log u_i &= -\log \sum_j K_{ij}\beta_j v_j, \\ g_j &= -\varepsilon \log \sum_i \exp\{(-C_{ij} + f_i)/\varepsilon + \log \alpha_i\} & f_i &= -\varepsilon \log \sum_j \exp\{(-C_{ij} + g_j)/\varepsilon + \log \beta_j\}. \end{aligned} \tag{12}$$

The Sinkhorn iterations typically loop until convergence. In practice, we choose a decreasing temperature sequence $\{\varepsilon_n\}$ with which to bound the number of iterations.

Algorithm 1 Sinkhorn's Algorithm in the log domain for $\mathcal{W}_2^2$

```
input: Source and target measures $\alpha = \sum_{i=1}^N \alpha_i \delta_{x_i}$, $\beta = \sum_{j=1}^M \beta_j \delta_{y_j}$, Annealing temperature sequence $\{\varepsilon_n\}$
# Initialize dual variables
$i \in \{1, \cdots, N\}, j \in \{1, \cdots, M\}$
$f_i \leftarrow 0, g_j \leftarrow 0 \; \forall \, i, j$
# Perform coordinate ascent in the log domain
for $\varepsilon \in \{\varepsilon_n\}$ do
    $C_{ij} = \frac{1}{2\varepsilon}||x_i - y_j||^2 \; \forall \, i, j$
    $g_j^{(n+1)} \leftarrow -\varepsilon \log \sum_i \exp\{(-C_{ij} + f_i^{(n)})/\varepsilon + \log \alpha_i\} \; \forall \, j$
    $f_i^{(n+1)} \leftarrow -\varepsilon \log \sum_j \exp\{(-C_{ij} + g_j^{(n+1)})/\varepsilon + \log \beta_j\} \; \forall \, i$
end for
# Return the entropic-regularized OT distance
output: $\langle f, \alpha \rangle + \langle g, \beta \rangle$
```

4. Supporting Results

4.1. Proof that the Gibbs measure minimizes free energy.

Remark 1. *Let $E(\mu) = F(\mu) + \beta^{-1}H(\mu)$, with $F(\mu) = \int U(z)d\mu$. The minimizer is the Gibbs density,*

$$\mu_*(z) = Z^{-1} \exp\{-\beta\psi(z)\},$$

where $\psi(z) = U(z) + \int_0^1 \lambda(\tau)S(z, \tau)d\tau$, and $Z = \int \exp\{-\beta\psi(z)\}dz$.

Proof. We set the functional derivative, or the first variation, of E to zero and solve for μ. The derivatives are

$$\frac{\delta F}{\delta \mu} = U(z), \qquad \frac{\delta H}{\delta \mu} = \log(\mu) + 1.$$

Solving for μ_* emits a proportionality, which can be normalized as described:

$$U(z) + \beta^{-1}(\log(\mu_*) + 1) = 0 \implies \mu_* \propto \exp\{-\beta\psi(z)\}$$

□

4.2. On the prevalence of multiple solutions

We are concerned with settings where the agent must select between multiple equivalently-valued actions in a way that minimizes uncertainty. Figure 3 shows the number of times these events occurred during the Control in the Presence of Uncertainty experiment. We present this data to support the claim that multiple solutions occur often enough in our experiment to merit a policy for selecting among the options.

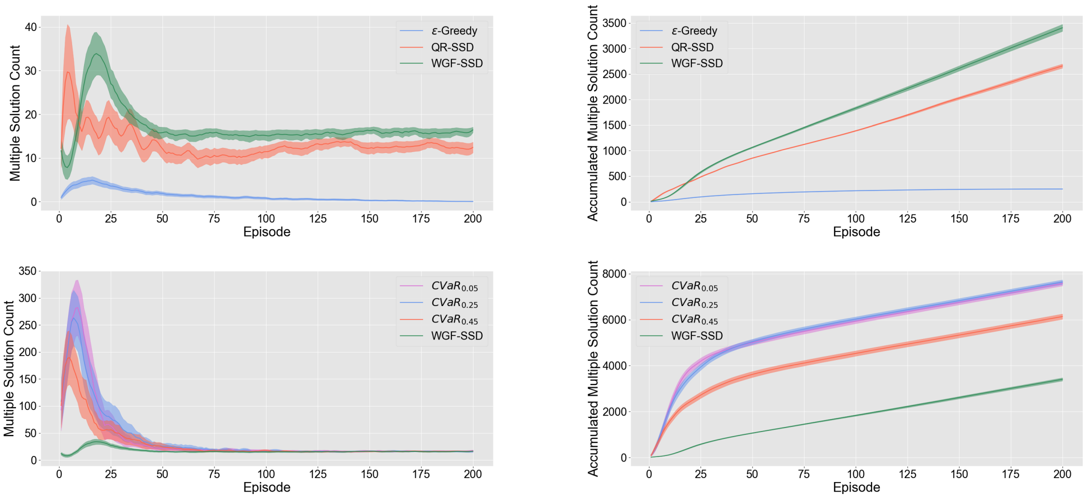


Figure 3. Frequency of multiple-solution events that occurred during the Control in the Presence of Uncertainty experiment.

References

Ambrosio, L. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zurich, 2005.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym, 2016.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. Distributional Reinforcement Learning with Quantile Regression. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

Fishburn, P. C. Stochastic dominance and moments of distributions. *Mathematics of Operations Research*, 5(1):94–100, 1980.

Gobbino, M. Minimizing movements and evolution problems in Euclidean spaces. *Annali di Matematica Pura ed Applicata*, 176(1):29–48, 1999.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, (ICLR)*, 2015.

Markowich, P. and Villani, C. On the trend to equilibrium for the fokker-planck equation: An interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*, 1999.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.