Stochastically Dominant Distributional Reinforcement Learning

John D. Martin ¹  Michal Lyskawinski ¹  Xiaohu Li ²  Brendan Englot ¹

Abstract
We describe a new approach for mitigating risk in the Reinforcement Learning paradigm. Instead of reasoning about expected utility, we use second-order stochastic dominance (SSD) to directly compare the inherent risk of random returns induced by different actions. We frame the RL optimization within the space of probability measures to accommodate the SSD relation, treating Bellman’s equation as a potential energy functional. This brings us to Wasserstein gradient flows, for which the optimality and convergence are well understood. We propose a discrete-measure approximation algorithm called the Dominant Particle Agent (DPA), and we demonstrate how safety and performance are better balanced with DPA than with existing baselines.

1. Introduction
The behavior of Reinforcement Learning (RL) agents is dictated by the utility they prescribe to different outcomes. As rational agents operating in a stochastic world, they prefer outcomes with the greatest expected utility (von Neumann & Morgenstern, 1947). This is equivalent to choosing actions that lead to high sums of reward (Sutton & Barto, 1998).

Naturally, every decision comes with the possibility of an undesirable outcome. Rational agents avoid risky events by diminishing the utility of such actions in a concave manner (Fishburn, 1964; Borkar, 2002). The effect shifts preferences away from uncertainty and toward choices with greater predictability. Our goal for the paper is to remove the need for an explicit utility function and to reason about risk on a more intrinsic level. Specifically, we employ the second-order stochastic dominance (SSD) relation to directly compare the underlying risk of random returns induced by different actions.

The SSD relation is defined using distribution functions and compared over the continuum of their realizable values:

\[ X \succeq (2) Y \iff F_{X}^{(2)}(\alpha) \leq F_{Y}^{(2)}(\alpha) \quad \forall \alpha \in \mathbb{R}. \quad (1) \]

We say \( X \) stochastically dominates \( Y \) in the second order when their integrated CDFs, \( F_{X}^{(2)}(\alpha) = \int_{-\infty}^{\alpha} F(x) dx \), satisfy (1), and we denote it as \( X \succeq (2) Y \).

It is well known that stochastic dominance forms a partial order over the real random variables (Billingsley, 1986). But what is perhaps more subtle is how the second-order relation characterizes a random variable’s risk. The function \( F_{X}^{(2)} \) defines the frontier of what is known as the dispersion space (Figure 1). The space’s volume reflects the degree to which a random variable differs from its deterministic behavior - if it were simply a real number equal to its mean. Outcomes with large dispersion spaces are considered risky; hence risk-averse agents prefer \( X \) to \( Y \) when \( X \succeq (2) Y \).

This idea forms the basis of a dual expected utility theory, which is popular in economics and operations research (Dentcheva & Ruszczyński, 2006). Dispersion statistics in RL have been frequently estimated with the classic Markowitz mean-variance model (Markowitz, 1952; Sato et al., 2001; Tamar et al., 2013). Others have used quantile risk measures, such as Value at Risk (VaR) and its conditional variant (CVaR), to characterize tail effects of the return distribution (Morimura et al., 2010; Chow & Ghavamzadeh, 2014; Chow et al., 2017). Interestingly, we can show these measures are special cases of the SSD model.

We use the SSD relation to define a safe exploration policy for RL. This is made possible through the class of distributional algorithms described by Bellmeare et al. (2017). These algorithms model a distribution over the return, whose mean is the familiar value function, and use it to evaluate and optimize a policy (Hessel et al., 2018; Barth-Maron et al., 2018). Relevant methods that use particle (quantile) models have shown encouraging progress on empirical benchmarks (Dabney et al., 2017; 2018), but understanding their convergence has been more challenging.

Our analysis sheds light on one approach to convergence which has thus far been absent from distributional RL. By casting the optimization problem as free-energy minimiza-
tion in the space of probability measures, we show that solutions evolve as a Wasserstein Gradient Flow (WGF) (Ambrosio, 2005). Model updates in this framework have well-determined dynamics. And under certain conditions, we detail in Section 3, convergence can be better understood.

The full gradient flow problem is unfortunately intractable, as it requires the optimal transport of (infinite dimensional) probability mass through continuous time. We propose a finite-measure approximation algorithm using the discrete-time Jordan-Kinderlehrer-Otto (JKO) update scheme (Jordan et al., 1998). At each step, our method iteratively optimizes a set of particles, where the equilibrium corresponds to satisfying the distributional Bellman equation. In the paper, we derive the update procedure and integrate it into the RL pipeline. Our results show it is possible to simultaneously learn and balance risk.

2. Reinforcement Learning Preliminaries

Reinforcement Learning describes a sequential decision making problem, whereby an agent learns to act optimally from rewards collected after taking actions. Optimality is defined with respect to the random return:

$$Z_{s,a}^{(s,a)} = \sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \bigg| S_0 = s, A_0 = a.$$  (2)

It represents the outcome of a decision sequence as the total discounted reward obtained after taking action $a \in A$ in state $s \in S$, then following the policy $\pi \in \Pi$ thereafter. Polices are stationary distributions over actions, coming from the set $\Pi = \{\pi : S \to \mathcal{P}(A)\}$. Here, $\gamma \in [0,1)$ is a discount factor, and $R(S_t, A_t)$ is the real-valued random reward associated with state $S_t$ and action $A_t$.

The learning process evolves iteratively; at each time step, the agent selects an action based on its current state, then transitions to the next state and collects a reward. Formally, the interaction is modeled as a Markov Decision Process (MDP), associated with the transition kernel, $p : S \times A \to \mathcal{P}(\mathbb{R} \times S)$, defining a joint distribution over the reward and next state, given the current state-action pair. Here, $S$ and $A$ are measurable Borel subsets of complete and separable metric spaces, which we take as finite. And for each state $s \in S$, the set $A_s \subset A$ is a measurable set indicating the feasible actions from $s$.

The agent starts with no knowledge of the return. Its task is to posit a representation of (2) and use transition observations to improve its estimates for better decision making.

2.1. Bellman’s Equation

The expected return obeys a recursive decomposition, known as Bellman’s equation (Bellman, 1957). For a general policy $\pi \in \Pi$, the equation is

$$E_{\mu}[Z_{s,a}^{(s,a)}] = Q_{\pi}(s,a) = r(s,a) + \gamma E_{p,\pi}[Q_{\pi}^{(s,a)}].$$  (3)

Here, the measure $\mu \in \mathcal{P}(\mathbb{R})^{(s,a)}$ captures all possible realizations of the return for each state-action pair. The decomposition is called Bellman’s optimality equation when actions are prescribed as $a^* \in \arg \max_{a \in A_s} E_{\mu}[Z_{s,a}^{(s,a)}]$:  

$$Q_{\pi}^{(s,a)} = r(s,a) + \gamma E_{p,\pi}[Q_{\pi}^{(s,a^*)}].$$  (4)

Viewed as a functional operator, (4) is known to contract to a unique fixed point, $Q_{\pi}^*$: the optimal value, corresponding to the set of optimal policies $\Pi_* = \{\pi_* \in \Pi : E_{\pi_*}[Q_{\pi_*}^{(s,a)}] = Q_{\pi_*}^{(s,a)} \quad \forall s \in S\}$. The formula establishes a connection between the observed reward, $r(s,a)$, and latent return, giving rise to many solution methods (Szepesvári, 2010). Typically $Q$ is directly modeled, and the underlying representation is updated to minimize the estimation error (Bellman error) formed with (3) or (4).

2.2. Distributional Bellman Operators

Instead of learning the expected return, the class of distributional RL methods model the return distribution, $\mu$ (Bellemare et al., 2017). The return distribution has been shown to satisfy a distributional variant of Bellman’s equation: $\mu^{(s,a)} = T^\pi \mu^{(s,a)}$, for all $(s,a) \in S \times A$, and where $T^\pi$ is the distributional Bellman operator:

$$T^\pi \mu^{(s,a)} = \int_\mathbb{R} \sum_{(s',a')} f_{s,a}^{(r,\gamma)} \mu^{(s,a')} \pi(a'|s') \rho(dr, s'|s,a).$$  (5)
Embedded here is the measurable mapping reflecting the Bellman action: \( f^{(r,\gamma)}(x) = r + \gamma x \), and the corresponding push-forward measure \( f^\#\mu(x) = \mu(f(x)) = \mu(A) \), for all Borel measurable sets \( A \). Just as the standard Bellman equation (3) is the focus of standard value-based RL, the distributional Bellman operator (5) plays the central role in distributional RL; it motivates algorithms which attempt to represent \( \mu \) and approximate it by repeated application of the update \( \mu_{t+1} \leftarrow T^\gamma \mu_t \), for steps \( t = 0, 1, \ldots \). The optimal version of the operator is obtained when actions are selected to maximize the expected value: \( \mathcal{T}\mu(x,a) = \int_{\mathbb{R}} \sum_{(s',a') \in S \times A} f^{(r,\gamma)}(s,a) p(ds',s|s,a) \). (6)

3. Reducing Risk in Distributional RL

Risk-neutral agents rank the expected return for optimal control (6). Performance is their only objective; thus the return distribution’s shape does not influence decision making. This can be problematic in high-stakes environments. Consider the example in Fig. 3. The agent is faced with two competing outcomes, \( Z_{a_1} \) and \( Z_{a_2} \). Here, \( Z_{a_2} \) has a larger mean, but it also has a much greater likelihood of realizing the least optimal outcome. This illustrates that distributional features such as the modality, tail size, spread, and dispersion are all critically important for mitigating risk. In the remainder of this section we describe a method to learn return distributions which formally permit agents to select safe actions with the SSD relation.

![Figure 3. Risky action: \( E[Z_{a_1}] < E[Z_{a_2}] \) but \( Z_{a_1} \succeq_{(2)} Z_{a_2} \)](image)

How is the safe action determined? At each state \( s \), the agent must evaluate the risk of every action in \( \mathcal{A}_s \). The action that dominates all the other choices in the second order has the lowest risk, and is considered safest. In precise terms, the safest action is the singular element of the set
\[
\{ a_\ast \in \mathcal{A}_s : Z^{(x,a_\ast)} \succeq_{(2)} Z^{(x,a')} \} \quad \forall a' \in \mathcal{A}_s \setminus \{ a_\ast \}.
\]
The safest action is not always guaranteed to exist. Indeed, it is quite possible that a given variety of return distributions have no valid dominance ordering (Fig. 2). The set above will be empty in these cases, but fortunately the agent can always default back to the standard approach of ranking with expectations. We propose an agent in Section 4 that adopts this contingency rule.

Given how conceptually straightforward it is to replace expectations with the SSD relation, it is natural to wonder if one can take any distributional RL algorithm and proceed with our approach. This is not advised unless certain technical requirements are met on the learned distributions.

**Assumption 1.** The first two moments of the learned return distributions are finite.

Under this assumption, an ordering holds on the moments.

**Lemma 1.** When Assumption 1 is valid, then \( X \succeq_{(2)} Y \) if, and only if \( \mu_X^{(1)} \geq \mu_Y^{(1)} \) or \( \mu_X^{(1)} = \mu_Y^{(1)} \) and \( \mu_X^{(2)} \leq \mu_Y^{(2)} \), where \( (\cdot) \) denotes a particular moment.

This result demands that any distributional representation of (2) emulates the first two moments correctly. Otherwise, SSD orderings may be invalid. Through the lens of Lemma 8, it becomes clear that risk-averse decision makers should consider algorithms that converge, at minimum, in the first two moments. As we will now describe, this is equivalent to convergence in the second-order Wasserstein distance.

### 3.1. Wasserstein Convergence

In the simple case with two univariate measures \( \mu, \nu \in \mathcal{P}(\mathbb{R}) \), the \( k \)-th order Wasserstein distance is defined as
\[
W_k(\mu, \nu) = \inf_{\gamma \in \mathcal{P}_k(\mu, \nu)} \left\{ \int_{\mathbb{R}^2} |x - y|^k d\gamma(x,y) \right\}^{1/k},
\]
where \( \mathcal{P}_k(\mu, \nu) \) is the set of all joint distributions with marginals \( \mu \) and \( \nu \) having \( k \) finite moments. The Wasserstein distance describes an optimal transport problem, where one seeks to transform \( \mu \) to \( \nu \) with minimum cost (Villani, 2008). The cost here is \( |x - y|^k \). If \( \mu \) is absolutely continuous with respect to the Lesbesgue measure, then there is an optimal plan from \( \mu \to \nu \), given as a mapping \( T: \mathbb{R} \to \mathbb{R} \), pushing \( \mu \) onto \( \nu \) with minimum cost: \( T_\# \mu = \nu \).

The \( W_k \) distance is appealing as a distributional learning objective, because its convergence implies convergence in the first \( k \) moments (Villani, 2008).

**Lemma 2.** Let \( \mu_t, \mu \in \mathcal{P}(\mathbb{R}^d) \) and \( k \geq 1 \), then \( W_k(\mu_t, \mu) \to 0 \) as \( t \to \infty \) if, and only if \( \mu_t \to \mu \) and \( \mu_t^{(k)} \to \mu^{(k)} \), where \( (\cdot) \) denotes the \( k \)-th moment.

Bellemare et al (2017) first showed the distributional Bellman operator contracts in the supremal Wasserstein distance. They proposed a discrete-measure approximation algorithm (CDRL) using a fixed mesh in probability space and later showed their approximation converges in \( W_2 \) (Rowland et al., 2018). This implies CDRL distributions can support...
3.2. Distributional RL as free-energy minimization

We propose an alternative way to learn the return distribution while minimizing $W_2$. We cast the RL problem as a free-energy minimization in terms of the functional

$$E(\mu) = F(\mu) + \beta^{-1}H(\mu). \quad (7)$$

Here, $F$ is the potential, $H$ is the entropy, and $\beta$ an inverse temperature parameter.

The potential energy defines what it means for a distribution to be optimal. We choose the low-energy equilibrium to coincide with minimum expected Bellman error, formed from (6). When the mapping $T$ reaches a fixed point, $T\mu^{(s,a)} = \mu^{(s,a)}$ for some $(s,a)$, energy is minimized. Given a transition sample $(s,a,r,s')$, we compute the distributional targets $Tz^{(s,a)}$, which denote realizations of $T\mu^{(s,a)}$, and define Bellman’s potential energy as

$$F(\mu) = \frac{1}{2} \int \left( Tz^{(s,a)} - z^{(s,a)} \right)^2 d\mu = \int U(z) d\mu. \quad (8)$$

Many RL inference frameworks enjoy the benefits of energy-based models (Haarnoja et al., 2017; Zhang et al., 2018). Given the functional form of our potential, we can describe the optimal probability measure in closed form.

**Lemma 3.** The minimizer of (7) is the Gibbs density,

$$\mu_*(z) = \mathcal{Z}^{-1} \exp\{-\beta U(z)\}, \quad (9)$$

where $\mathcal{Z} = \int \exp\{-\beta U(z)\} dz$.

3.3. The Fokker-Planck Equation

We would like to understand the convergence behavior of return distributions as the free-energy is minimized. Systems of this nature are typically modeled as continuous-time stochastic diffusion processes, where the distributions $\{\mu_t\}_{t \in [0,1]}$ evolve over a smooth manifold of probability measures from $\mathcal{P}_2(\mathbb{R})$. The second-order Wasserstein distance, $W_2$, endows $\mathcal{P}_2(\mathbb{R})$ with a Riemannian geometry (Ambrosio, 2005), allowing convergence to be precisely analyzed. The dynamics of $\mu_t$ obey the diffusive partial differential equation called the Fokker-Planck equation:

$$\partial_t \mu_t = \nabla \cdot \left( \mu_t \nabla \frac{\delta E}{\delta \mu_t} \right). \quad (10)$$

Here, we denote the sub-gradient with respect to time $\partial_t$, and the first variation (Gâteaux derivative) of free energy $\frac{\delta E}{\delta \mu_t}$. The Fokker-Planck equation plays a central role in statistical physics, chemistry, and biology. In the analysis of optimization procedures, (10) dictates the solution path, or gradient flow, of $\mu$ as it evolves over the manifold of probability measures. This is captured in the following lemma.

**Lemma 4.** Let $\{\mu_t\}_{t \in [0,1]}$ be an absolutely-continuous curve in $\mathcal{P}_2(\mathbb{R})$. Then for $t \in [0,1]$, the vector field $v_t = \nabla \left( \frac{\delta E}{\delta \mu_t} \right)$ defines a gradient flow on $\mathcal{P}_2(\mathbb{R})$ as $\partial_t \mu_t = -\nabla \cdot (\mu_t v_t)$, where $\nabla \cdot \mathbf{u}$ is the divergence of some vector $\mathbf{u}$.

The free-energy functional, $E$, intuitively characterizes the diffusion and thus, the optimization landscape of the RL problem. Convergence to an optimal point can be guaranteed provided $E$ is convex. By inspection, we know this is the case for the return variable $z$ in (7).

**Lemma 5.** The energy functional (7) with potential (8) is convex in the return variable, $z$.

A normalized solution to the Fokker-Planck equation de-
scribes the probability density for an Itô advection-diffusion process, whose dynamics are governed by the stochastic differential equation:

$$dZ_t = -\nabla U(Z_t)dt + \sqrt{2}\beta^{-1/2}dW_t,$$  \hspace{1cm} (11)

with $\nabla U$ representing the system’s drift, and $W_t$ the standard Brownian motion. The Gibbs measure is the unique invariant measure for the Markov return update process $Z_t$.

### 3.4. Discrete Time Solutions

We adopt an iterative procedure (Jordan et al., 1998) to approximately solve the gradient flow problem (10). The method discretizes time in steps of $h$ and applies the proximal operator

$$\text{Prox}_{hE}^W(\mu_k) = \arg \min_{\mu} W^2_2(\mu, \mu_k) + 2hE(\mu). \hspace{1cm} (12)$$

For every step $k$, the operator generates a path of distributions $\{\mu_i\}_{i=1}^K$ such that $\mu_{k+1} = \text{Prox}_{hE}^W(\mu_k)$ is equivalent to $\mu_k$. In contrast with the distributional Bellman operator (6), the proximal operator regulates Bellman minimization directly with $W_2$. And because of Lemma 8, this method obtains the unique solution to (10). The following general result is due to Jordan et al. (1998):

**Lemma 6.** Let $\mu_0 \in \mathcal{P}_2(\mathbb{R})$ have finite free energy $E(\mu_0) < \infty$, and for a given $h > 0$, let $\{\mu_i\}_{i=1}^K$ be the solution of the discrete-time variational problem (12), with measures restricted to $\mathcal{P}_2(\mathbb{R})$, the space with finite second moments. Then as $h \to 0$, $\mu_i \to \mu_T$, where $\mu_T$ is the unique solution of (10) at $T = hK$.

Furthermore, one can evaluate the free-energy (8) over the solution sequence and observe it becomes a decreasing function of time (a Lyapunov function). The result implies that the expected distributional Bellman error is minimized when using the JKO approach.

**Proposition 1.** Let $\{\mu_k\}_{k=0}^K$ be the solution of the discrete-time variational problem (12), with measures restricted to $\mathcal{P}_2(\mathbb{R})$, the space with finite second moments. Then $E(\mu_k)$ is a decreasing function of time.

### 4. Discrete Measure Solutions

Our goal is to numerically compute the steady-state distribution of the diffusion (11). Direct application of (12) is infeasible, because the $\{\mu_k\}$ are infinite-dimensional objects. We propose a discrete-measure approximation of $\mu$ using a Lagrangian (particle-based) discretization, where the measure is supported on $N$ equally-likely diracs $\mu_k \approx \frac{1}{N} \sum_{i=1}^N \delta_{z_i}$.

Given an initial set of particles at some state-action pair $z(s, a) = \{z^{(1)}, \cdots, z^{(N)}\}$, we evolve them forward in time with steps of $h$ to obtain the solution at $t + h$. We apply a finite number of gradient steps to approximate the convergence limit $T = hK$.

Following several prior works (Cuturi, 2013; Peyré, 2015; Zhang et al., 2018), we lump the entropy term from $E$ into $W_2^2$. Let $\mu = \sum_{i=1}^N \mu_i \delta_{z_i}$ and $\nu = \sum_{j=1}^N \nu_j \delta_{y_j}$ be finite distributions. Then the entropic-regulated $W_2^2$ distance is compactly written

$$W_\beta(\mu, \nu) = \inf_{P \in \mathbb{R}^N_+ \times N} \langle P, C \rangle + \beta KL(P\|\nu),$$

s.t. $\sum_{j=1}^N P_{ij} = \mu_i$, $\sum_{i=1}^N P_{ij} = \nu_j$.

Here, $\langle P, C \rangle$ denotes the Frobenius norm between the joint $P$ and the square Euclidean cost $C_{ij} = (x_i - y_j)^2$, and $KL(P\|\nu) = \sum_{i,j} P_{ij} \log(P_{ij}/\mu_i\nu_j) - P_{ij} + \mu_i\nu_j$. The entropic term promotes numerical stability by acting as a barrier function in the positive octant. Moreover, entropy’s connection to the Kullback-Libeler (KL) divergence enables closed form solutions to common marginal projection problems in Optimal Transport (Peyré, 2015). JKO stepping under this new distance is denoted

$$\text{Prox}_{hF}^W(\mu_k) = \arg \min_{\mu \in \mathcal{P}_2(\mu_k)} W_\beta(\mu, \mu_k) + 2hF(\mu). \hspace{1cm} (13)$$

One can compute the entropic-regularized distance, $W_\beta$, using Sinkhorn’s algorithm (Sinkhorn, 1967). The algorithm applies kernel products and point-wise division to
perform coordinate ascent on the dual maximization problem (Peyré & Cuturi, 2018; Cuturi & Doucet, 2014). One can also consider parametric models for the particles, and apply auto-differentiation to back-propagate gradient information through the Sinkhorn procedure into the model. Our experiments use this approach with a tabular model (Fig. 4). A study of more sophisticated particle models is left to future work.

4.1. SSD action selection with discrete measures

We want to compare discrete measures comprised of \( N \) equally-likely diracs for SSD action selection. Given the measures of two random returns, the SSD relation (1) checks the integrated CDFs, \( F^{(2)}(\alpha) \), for every \( \alpha \in \mathbb{R} \). An equivalent comparison can be made with cumulative quantile functions, \( F^{-2}(\tau) = \int_0^\tau F^{-1}(t) \, dt \); where,

\[
X \succeq_{(2)} Y \iff F_X^{-2}(\tau) \geq F_Y^{-2}(\tau) \forall \tau \in (0, 1).
\]

(14)

Here we assume that \( F_Y^{-2}(0) = 0 \), and \( F_Y^{-2}(1) = \infty \). Notice that \( F_X^{-2}(\tau)/\tau = E[X|X \leq \xi(\tau)] \) is the Conditional Value at Risk for level \( \tau \). Thus, the SSD condition can be interpreted as a continuum of CVaR constraints for risk levels \( \tau \in (0, 1) \). This connection suggest an efficient way to compute the cumulative quantile function \( F^{-2} \).

**Lemma 7**. Let \( \tau \in (0, 1) \) and consider \( \xi(\tau) = F_X^{-1}(\tau) \). Then \( F_X^{-2}(\tau) = E[X \leq \xi(\tau)] \).

Lemma 7 makes it possible to compare total expectations on subsets of the return space, instead of dealing with probability integrals over an unbounded domain. Computations simplify even further when we consider finite sets of returns.

We use \( z[i] \) to denote the ordered coordinates of a return distribution: \( z[1] \leq z[2] \leq \cdots \leq z[N] \). Then given particle sets for two returns which were induced by actions \( a_1 \) and \( a_2 \), we have \( Z^{(s,a_1)} \geq_{(2)} Z^{(s,a_2)} \iff E[Z^{(s,a_1)} \leq \xi(\tau)] \geq E[Z^{(s,a_2)} \leq \xi(\tau)], \forall \tau \in (0, 1) \);

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

(15)

We propose a new behavior policy for RL that uses (15) to find the safest action; the one whose induced return dominates all others in SSD. The policy is implemented by checking (15) for all \( (|A|^2) \) pairs with quantiles derived empirically from the particles. When dominance cannot be established (Fig. 2c), our policy returns the greedy action. Values are straightforward to compute with equally-likely particles: \( Q^{(s,a)} = \frac{1}{N} \sum_{i=1}^N z[i] \).

4.2. The Dominant Particle Agent

We introduce the Dominant Particle Agent (DPA); a reinforcement learner that tries to recover latent return distribu-

4.2. The Dominant Particle Agent

We introduce the Dominant Particle Agent (DPA); a reinforcement learner that tries to recover latent return distribu-

4.2. The Dominant Particle Agent

We introduce the Dominant Particle Agent (DPA); a reinforcement learner that tries to recover latent return distribu-

Algorithm 1 Dominant Particle Agent (Tabular)

1: Initialize particles \( z(s, a) = \{z[i]_i\}_{i=1}^N \forall (s, a) \in S \times A \)
2: for \( t = 1, 2, \cdots \) do
3: \# Explore with the SSD behavior policy
4: \( a_t \leftarrow B_{s_t, z} \)
5: \( s_{t+1}, r_t \sim p(\cdot|s_t, a_t) \)
6: \# Exploit with the greedy target policy
7: \( a^* \leftarrow \arg \max_{a \in A} \tau_t(z_{t+1}, a) \}
8: \( T_t(1) \leftarrow r_t + \gamma z(t+1, a^*) \forall i = 1, \cdots, N \)
9: \# Update particles with proximal step
10: \( z(s_t, a_t) \leftarrow \arg \min_{z'} W_{hF}^\beta(z(s_t, a_t), z'; T_t) \)
11: end for

The proximal loss is computed in Alg. 2 using the expected Bellman error (scaled \( L_2 \) loss) and Sinkhorn algorithm for the entropic-regulated second-order Wasserstein distance.

Algorithm 2 Proximal Loss

1: input: Source, argument, and target particles \( z_0, z, T z \)
2: \( F(z) \leftarrow \frac{1}{N} \sum_{i=1}^N [T z(i) - z(i)]^2 \)
3: \( W_\beta(z, z_0) \leftarrow \text{Sinkhorn}(\beta, z, z_0) \)
4: output: \( W_{hF}^\beta = W_\beta(z, z_0) + 2hF(z) \# \text{JKO loss} \)

5. Connections with Related Work

**Quantile Regression RL**: The QR-DQN agent trains a convolutional neural network to output return particles corresponding to a fixed grid of quantiles (Dabney et al., 2017). The grid is precisely aligned to minimize the distance with its targets. The IQN agent (Dabney et al., 2018) extends the QR-DQN Huber loss to train a sampling network that mimics the target distribution. These methods perform impressively well in complex Atari environments, where the state-space is high-dimensional. We also model the return distribution with particle sets. Though, our experiments use tabular models instead of deep parametric networks. Our primary concern in this paper has been analytical. We sought a theoretical treatment for particle-based distributional RL methods that permit the application of an SSD behavior policy.

**Wasserstein Gradient Flows in RL**: DP-WGF (Zhang et al., 2018) models stochastic policy inference as free-
Figure 5. Safe learning in the CliffWalk domain: The left plot shows the DPA estimated histograms of the target densities. Convergence of the proximal loss and squared value error for each action are shown in the top two plots. The right plot arranges outcome-risk diagrams in descending order with respect to their dispersion space size. Although moving right is optimal, results indicate $a = \text{up}$ is the safest decision. Indeed, this is the action DPA takes.

energy minimization. They too apply the JKO scheme to derive learning algorithms. And in the same way our formalism leads to a convergent algorithm for return distributions, so too does their approach for stochastic policies. DP-WGF couches their training procedure within the soft-$Q$ learning paradigm (Haarnoja et al., 2017; 2018). These algorithms train a deep parametric model to sample from a target Gibbs density using Stein Variational Gradient Descent (Liu & Wang, 2016). In contrast, our method fixes a set of particles and adopts the more traditional Sinkhorn algorithm from optimal transport to compute the Wasserstein distance. We complete the JKO step using gradient methods and auto-differentiation.

Wasserstein minimization: The Wasserstein distance has become as a compelling objective in Machine Learning (Arjovsky et al., 2017; Chen et al., 2018). Its introduction to deep RL (Bellemare et al., 2017) has spurred research that focuses on improving empirical performance on Atari benchmarks (Barth-Maron et al., 2018; Hessel et al., 2018). Though the exact reason for observed improvements remain an open area of research, evidence suggests the distributional model improves generalization (Imani & White, 2018). Our paper addresses particle models as they apply within Wasserstein Gradient Flows.

Variational RL: Approximate inference has a well-established history in RL (Dayan & Hinton, 1997; Ziebart et al., 2008; Toussaint, 2009; Neumann, 2011). Recent work draws connections between the actor-critic framework and expectation-maximization (Fellows et al., 2018). The implications of a return-based WGF actor-critic would be interesting to explore in future work.

Risk Modeling: Lemma 8 describes a moment ordering imposed by the SSD. This reveals a connection to the popular Markowitz mean-variance risk model (Markowitz, 1952), which often attempts to reduce variance. Although conceptually appealing, variance minimization can sometimes produce safe outcomes that are stochastically dominated by a feasible alternative (Ogryczak & Ruszczynski, 2002). DPA considers a comprehensive measure of dispersion, which includes as special cases, the VaR and CVaR metrics (Rockafellar & Uryasev, 2000; Morimura et al., 2010; Chow & Ghavamzadeh, 2014; Chow et al., 2017).

6. Experiments

We have introduced DPA, a reinforcement learner that uses information from the return distribution to mitigate risk. In this section we verify several prior assertions. Namely, we show the optimization procedure in Alg. 1 minimizes the proximal loss (13) and recovers the latent return distribution. We also quantify the agent’s ability to mitigate risk in a policy learning setting. In doing so, we draw comparisons with other baselines.

6.1. Solving the Gradient Flow Problem

Proposition 1 argues that repeated application of the proximal step (12) produces a decreasing function of time, implying that a Wasserstein-regulated free energy is minimized at convergence. Here, we verify this is indeed the case by learning a non-trivial return distribution from Monte Carlo (MC) targets.

The problem is set within the CliffWalk domain, shown in Figure 5. The transition dynamics follow those in Sutton & Barto (1998). Additionally, we include a ten-percent chance of taking a random action at each step. The optimal policy, starting from $s_0$, moves up, then along the cliff, until moving down brings the agent to the goal.
We show the convergence of the proximal loss and the mean square value error in Figure 5. The target distribution was obtained by rolling out a SARSA policy 200 times from $s_0$. As we can see, the estimated distribution accurately captures the target’s features: the near certainty of walking off the cliff when moving right, the added chance of doing the same when choosing left or down, and finally the most profitable choice, moving up. In all experiments, we fix $\beta^{-1} = 0.5$ and $h=0.05$.

6.2. Learning in the Presence of Risk

This experiment tasks the agent to learn a full CliffWalk policy from transition data. We study how optimality and risk are balanced throughout this process by measuring the return learned by DPA and the total number of times the agent falls off the cliff. These metrics are evaluated with three different behavior criteria: greedy, CVaR, and SSD. The standard greedy policy selects actions to maximize the expected return. The CVaR policy prefers actions with the largest five-percent lower tail return. And the SSD policy prefers actions with the smallest dispersion space.

Results are shown in Fig. 6. Learning evolves quickly in all cases: within 100 episodes. We were surprised to find each agent performed approximately the same, given the risk-sensitive policies prioritize optimality second. With high levels of environment stochasticity, however, the greedy agent may initially undervalue states adjacent to the cliff. This would cause it to avoid the cliff until sufficient episodes have passed to recognize the benefit of the lower path.

The total number of times each agent fell from the cliff provides an indication of risk awareness. DPA-SSD exhibited the most risk-sensitive behavior, followed by DPA-CVaR, then DPA-Greedy. This was the expected outcome, since SSD decisions are based on a conjunction of CVaR comparisons (15), and the greedy approach is risk-neutral.

7. Discussion

This paper argues for the use of stochastic dominance as a means to compare the inherent risk of random returns. Our experiments show that choosing the least disperse action avoids the pathology of greedy and CVaR policies, offering a more complete understanding of the agent’s uncertain condition. Before closing, we discuss several important points to clarify our contributions and design decisions.

Why optimize? If the target particle set is known a priori, then why not simply assign them to their target value, as the distributional Bellman operator (6) would suggest? In the WGF framework, completing a proximal step is not the same as swapping estimates for targets. The procedure is more regulated, because optimization must balance energy minimization and a sense of nearness in probability space. The benefit is that learned distributions will have two correct moments.

Optimal Policy Our method strives to learn the optimal target policy using data gathered from the SSD behavior policy. However, we did not prove this procedure is guaranteed to recover the optimal policy. We expect the SSD policy will return greedy actions when the environment is deterministic and contains few solution paths. For environments with greater solution diversity, SSD actions will most-likely produce a suboptimal policy that allows the agent to act safely.

Scaling to large state spaces Adopting the Lagrangian discretization requires a set of particles for every element in $S \times A$. In large spaces this is impractical. It is possible to use a parameterized model that outputs the particles. The model parameters can be updated by back propagating gradient information through the Sinkhorn algorithm. This approach lends itself nicely to methods that learn a generative model of the target distribution (Zhang et al., 2018).

7.1. Conclusion

To guarantee feasible application of the SSD policy, we introduced DPA, a new particle-based method for recovering the latent return distribution. DPA leverages the theoretical scaffolding of Wasserstein gradient flows to guarantee model updates converge to a unique optimum described by the Bellman potential energy functional. In closing, we believe our paper highlights the importance that the SSD relation can bring to risk-sensitive RL.
References

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

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Barth-Maron, G., Hoffman, M., Budden, D., Dabney, W., Horgan, D., TB, D., Muldal, A., Heess, N., and Lillicrap, T. Distributed distributional policy gradients. In *International Conference on Learning Representations (ICLR)*, 2018.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Bellman, R. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

Billingsley, P. *Probability and Measure*. John Wiley and Sons, second edition, 1986.

Borkar, V. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 2002.

Chen, C., Zhang, R., Wang, W., Li, B., and Chen, L. A unified particle-optimization framework for scalable bayesian sampling. *CoRR*, 2018.

Chow, Y. and Ghavamzadeh, M. Algorithms for cvar optimization in mdps. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, 2014.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18, 2017.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, 2013.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Dayan, P. and Hinton, G. E. Using expectation-maximization for reinforcement learning. *Neural Computation*, 1997.

Dentcheva, D. and Ruszczyński, A. Inverse stochastic dominance constraints and rank dependent expected utility theory. *Mathematical Programming*, 2006.

Fellows, M., Mahajan, A., Rudner, T., and Whiteson, S. Virel: A variational inference framework for reinforcement learning. *CoRR*, abs/1811.01132, 2018.

Fishburn, P. *Decision and Value Theory*. Wiley, 1964.

Fishburn, P. Stochastic dominance and moments of distributions. *Math. Operations Research*, 1980.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M. G., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Imani, E. and White, M. Improving regression performance with distributional losses. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 1998.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems* 29, 2016.

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.
Markowitz, H. Portfolio selection. *Journal of Finance*, 1952.

Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., and Tanaka, T. Parametric return density estimation for reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

Neumann, G. Variational inference for policy search in changing situations. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011.

Ogryczak, W. and Ruszczynski, A. Dual stochastic dominance and related mean-risk models. *SIAM J. on Optimization*, 2002.

Peyré, G. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 2015.

Peyré, G. and Cuturi, M. Computational optimal transport. 2018. URL https://arxiv.org/abs/1803.00567.

Rockafellar, R. and Uryasev, S. Optimization of conditional value-at-risk. *Journal of Risk*, 2000.

Rowland, M., Bellemare, M., Dabney, W., Munos, R., and Teh, Y. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Sato, M., Kimura, H., and Kobayashi, S. Td algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 2001.

Sinkhorn, R. Diagonal equivalence to matrices with prescribed row and column sums. In *The American Mathematical Monthly*, 1967.

Sutton, R. and Barto, A. *Introduction to Reinforcement Learning*. MIT Press, 1st edition, 1998.

Szepesvári, C. *Algorithms for Reinforcement Learning*. Morgan & Claypool, 2010.

Tamar, A., Castro, D. D., and Mannor, S. Temporal difference methods for the variance of the reward to go. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Toussaint, M. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 2009.

Villani, C. *Optimal Transport: Old and New*. Springer, 2008.

von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, 1947.

Zhang, R., Chen, C., Li, C., and Carin, L. Policy optimization as Wasserstein gradient flows. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Ziebart, B., Maas, A., Bagnell, A., and Dey, A. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 2008.
8. Proofs

Lemma 1. When distributions of $X$ and $Y$ have finite first and second moments, then $X \succeq (2) Y$ if, and only if $\mu_{X}^{(1)} \geq \mu_{Y}^{(1)}$ or $\mu_{X}^{(1)} = \mu_{Y}^{(1)}$ and $\mu_{X}^{(2)} \leq \mu_{Y}^{(2)}$.

Proof. See Theorem 1 of (Fishburn, 1980).

Lemma 2. Let $\mu_{t}, \mu \in P(\mathbb{R}^d)$ and $k \geq 1$, then $W_k(\mu_{t}, \mu) \to 0$ as $t \to \infty$ if, and only if $\mu_{t} \to \mu$ and $\mu_{t}^{(k)} \to \mu^{(k)}$.

Proof. See (Villani, 2008).

Lemma 3. Let $E(\mu) = F(\mu) + \beta^{-1}H(\mu)$, with $F(\mu) = \int U(z)d\mu$. The minimizer is the Gibbs density,

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

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 $\mu$. The derivatives are

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

Solving for $\mu_\star$ emits a proportionality, which can be normalized as described:

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

Lemma 4. Let $\{\mu_t\}_{t \in [0, 1]}$ be an absolutely-continuous curve in $P(\mathbb{R})$ with finite second-order moment. Then for $t \in [0, 1]$, the vector field $v_t = \nabla(\frac{\delta E}{\delta \mu}(\mu))$ defines a gradient flow on $P(\mathbb{R})$ as $\partial_t \mu_t = -\nabla \cdot (\mu_t v_t)$, where $\nabla \cdot \mathbf{u}$ is the divergence of some vector $\mathbf{u}$.

Proof. See (Ambrosio, 2005) Theorem 8.3.1.

Lemma 5. The energy functional $E(\mu) = F(\mu) + \beta^{-1}H(\mu)$, with $F(\mu) = \int U(z)d\mu$ and $U(z) = \frac{1}{2}(Tz - z)^2$ is convex in the return variable, $z$.

Proof. By inspection.

Lemma 6. Let $\mu_0 \in P_2(\mathbb{R})$ 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 $P_2(\mathbb{R})$, the space with finite second moments. Then as $h \to 0$, $\mu_t^{(h)} \to \mu_T$, where $\mu_T$ is the unique solution of the Fokker-Plank equation at $T = hK$. 
Proof. See (Jordan et al., 1998) Theorem 5.1.

**Proposition 1.** 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

\[
\partial_t h_t = \Delta h_t - \nabla U \cdot \nabla h_t.
\]  

(17)

Whenever \( \phi \) is a convex function of \( \mathbb{R} \), one can check the following is a Lyapunov functional for (17), and equivalently the FP equation:

\[
\int \phi(h_t) e^{-U} \, dz = \int \phi(\mu_t e^U) e^{-U} \, dz
\]

Indeed

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

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

\[
\int \phi(h_t) e^{-U} \, dz = \int \mu_t \log \left( \frac{h_t}{e^{-U}} \right) \, dz
\]

\[
= \int \mu_t (U + \log \mu_t) \, dz = E(\mu).
\]

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, when there is only pure Brownian motion, the optimal distributional Bellman equation is satisfied.

**Lemma 7.** Let \( \tau \in (0, 1) \) and consider \( \xi_\tau = F_X^{-1}(\tau) \). Then \( F_X^{-2}(\tau) = \mathbb{E}[X \leq \xi_\tau] \).

Proof. By conjugate duality,

\[
F_X^{-2}(\tau) = \tau \xi_\tau - F_X^{-2}(x),
\]

\[
= \tau \xi_\tau - \tau \mathbb{E}[X - \xi_\tau | X \leq \xi_\tau],
\]

\[
= \tau \mathbb{E}[X | X \leq \xi_\tau],
\]

\[
= \mathbb{E}[X \leq \xi_\tau].
\]

\[
\Box
\]

9. Sinkhorn’s Algorithm

We describe how the Kantorovich problem can be made tractable through entropy regularization, then present an algorithm for approximating the \( W_2^2 \) 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 \( W_2^2 \) 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, \( \pi \), defined as non-negative matrix on the product space of atoms \( \{x_1, \ldots, x_N\} \times \{y_1, \ldots, y_M\} \). Denoting the cost to move \( x_i \) to \( y_j \) as \( C_{ij} = ||x_i - y_j||^2 \), we have

\[
W_2^2(\alpha, \beta) = \min_{\pi \in \mathbb{R}^{N \times M}} \langle \pi, C \rangle = \sum_{ij} \pi_{ij} C_{ij},
\]

such that \( \pi 1_N = \alpha, \pi^T 1_M = \beta \).
We mentioned that Optimal Transport reduces to positive matrix scaling. Indeed, using the vectors 

\[ \pi \in S_{n \times n}^+ \]

with which to bound the number of iterations.

The Sinkhorn iterations typically loop until convergence. In practice, we choose a decreasing temperature sequence 

\[ \{ \epsilon_n \} \]

such that \[ \pi \mathbf{1}_M = \alpha, \pi^\top \mathbf{1}_N = \beta. \]

The value of \( W_\epsilon(\alpha, \beta) \) occurs necessarily at the critical point of the constrained objective function

\[
L_\epsilon = \sum_{i,j} \pi_{ij} C_{ij} + \epsilon \sum_{i,j} [\pi_{ij} \log \frac{\pi_{ij}}{\alpha_i \beta_j} - \pi_{ij} + \alpha_i \beta_j] - \sum_i f_i \left( \sum_j \pi_{ij} - \alpha_i \right) - \sum_j g_j \left( \sum_i \pi_{ij} - \beta_j \right),
\]

\[
\frac{\partial L_\epsilon}{\partial \pi_{ij}} = 0 \implies \forall i, j, C_{ij} + \epsilon \log \frac{\pi_{ij}}{\alpha_i \beta_j} = f_i^* + g_j^*.
\]

The last line of (23) 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

\[
u_i = \exp \left( \frac{f_i^*}{\epsilon} \right), \quad v_j = \exp \left( \frac{g_j^*}{\epsilon} \right), \quad K_{ij} = \exp \left( -\frac{C_{ij}}{\epsilon} \right),
\]

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

\[
W_\epsilon(\alpha, \beta) = \langle \pi^*, C \rangle + \epsilon \text{KL}(\pi^* || \alpha \otimes \beta) = \sum_{i,j} (f_i^* + g_j^*) = \langle f^*, \alpha \rangle + \langle g^*, \beta \rangle.
\]

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 \( \pi^* \). Initialize \( u^{(0)} = \mathbf{1}_N \), and \( v^{(0)} = \mathbf{1}_M \), then perform the following iterations for all \( i, j \)

\[
v_j^{(1)} = \frac{1}{[K^\top(\alpha \otimes u^{(0)})]_j}, \quad u_i^{(1)} = \frac{1}{[K(\beta \otimes v^{(1)})]_i}, \quad \vdots \quad u_i^{(n+1)} = \frac{1}{[K(\beta \otimes v^{(n+1)})]_i}, \quad v_j^{(n+1)} = \frac{1}{[K^\top(\alpha \otimes u^{(n)})]_j},
\]

Sinkhorn’s algorithm performs coordinate ascent with \( f \) and \( g \) to maximize the dual maximization problem

\[
W_\epsilon(\alpha, \beta) = \max_{f \in \mathbb{R}^N, g \in \mathbb{R}^M} \langle f, \alpha \rangle + \langle g, \beta \rangle - \epsilon (\alpha \otimes \beta, \exp \{ (f + g - C) / \epsilon \} - 1).
\]

Each update consists of kernel products, \( K^\top(\alpha \otimes u) \) and \( K(\beta \otimes v) \), and point-wise divisions. We describe this procedure in Algorithm 3, using computations in the log domain to numerically stabilize the updates. The log updates derive from (24) and (27):

\[
\log v_j = -\log \sum_i K_{ij} \alpha_i u_i, \quad \log u_i = -\log \sum_j K_{ij} \beta_j v_j,
\]

\[
g_j = -\epsilon \log \sum_i \exp \{ (-C_{ij} + f_i) / \epsilon + \log \alpha_i \}, \quad f_i = -\epsilon \log \sum_j \exp \{ (-C_{ij} + g_j) / \epsilon + \log \beta_j \}.
\]
Algorithm 3 Sinkhorn’s Algorithm in the log domain for $W_2^2$

1: **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 $\{\varepsilon_n\}$
2: # Initialize dual variables
3: $i \in \{1, \cdots, N\}$, $j \in \{1, \cdots, M\}$
4: $f_i \leftarrow 0$, $g_j \leftarrow 0 \; \forall \; i, j$
5: # Perform coordinate ascent in the log domain
6: **for** $\varepsilon \in \{\varepsilon_n\}$ **do**
7: $C_{ij} = \frac{1}{2\varepsilon}||x_i - y_j||^2 \; \forall \; i, j$
8: $g_{j}^{(n+1)} \leftarrow -\varepsilon \log \sum_i \exp\{(-C_{ij} + f_{i}^{(n)})/\varepsilon + \log \alpha_i\} \; \forall \; j$
9: $f_{i}^{(n+1)} \leftarrow -\varepsilon \log \sum_j \exp\{(-C_{ij} + g_{j}^{(n+1)})/\varepsilon + \log \beta_j\} \; \forall \; i$
10: **end for**
11: # Return the entropic-regularized OT distance
12: **output:** $\langle f, \alpha \rangle + \langle g, \beta \rangle$