
Robust Reinforcement Learning in Finance: Modeling Market Impact with Elliptic Uncertainty Sets

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 In financial applications, reinforcement learning (RL) agents are commonly trained
2 on historical data, where their actions do not influence prices. However, during
3 deployment, these agents trade in live markets where their own transactions can
4 shift asset prices, a phenomenon known as market impact. This mismatch between
5 training and deployment environments can significantly degrade performance. Tra-
6 ditional robust RL approaches address this model misspecification by optimizing
7 the worst-case performance over a set of uncertainties, but typically rely on symmet-
8 ric structures that fail to capture the directional nature of market impact. To address
9 this issue, we develop a novel class of elliptic uncertainty sets. We establish both
10 implicit and explicit closed-form solutions for the worst-case uncertainty under
11 these sets, enabling efficient and tractable robust policy evaluation. Experiments
12 on single-asset and multi-asset trading tasks demonstrate that our method achieves
13 superior Sharpe ratio and remains robust under increasing trade volumes, offering
14 a more faithful and scalable approach to RL in financial markets.

15 1 Introduction

16 Reinforcement learning (RL) has emerged as a promising decision-making framework for quantitative
17 trading strategies [1, 2], including portfolio optimization [3–13], automatic trading [3, 14–19], market
18 making [20–24], and option hedging [25–30]. RL’s appeal in finance lies in its ability to learn
19 adaptive strategies directly from data, without strong market assumptions [1]. This makes it well-
20 suited for capturing complex market dynamics and aligning with the sequential nature of financial
21 decision-making.

22 One of the primary challenges in training a robust and consistently profitable RL agent lies in handling
23 the *market impact* [31–33]; that is, the influence of the agent’s own trades on asset prices in the
24 deployed environment. For example, when a trader buys or sells a large volume of an asset, it can
25 temporarily drive the price up or down, respectively (as illustrated in Figure 1). Typically, RL agents
26 are trained on historical market data where market impact is absent. However, during deployment,
27 the environment shifts from a passive historical setting to the real market, where the agent’s actions
28 actively affect prices. This discrepancy between training and deployment environments undermines
29 the optimality and robustness of the learned policy, and leads us to the central question in this paper:

30 *Q: Can we train RL agents on historical data while robustly accounting for
market impact during deployment?*

31 To address this central question, we adopt the framework of robust RL [34–51], which is designed
32 to handle *model misspecification*; that is, the mismatch between the training and deployment envi-

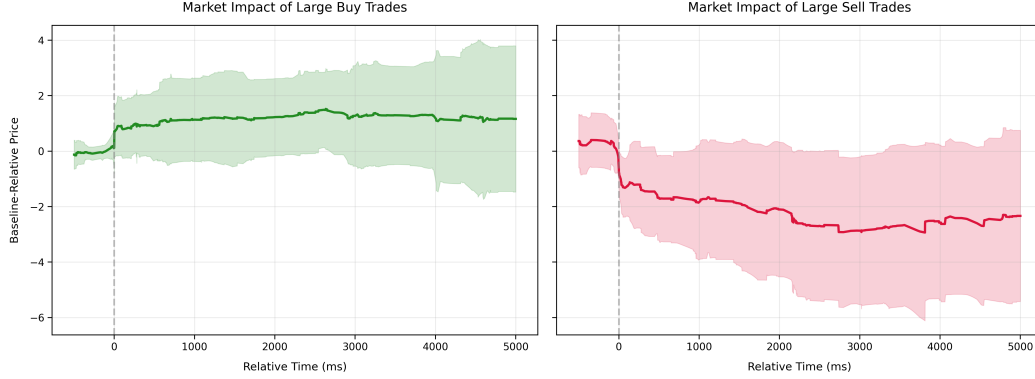


Figure 1: Market impact illustration using AMZN stock on June 21, 2012, based on 5-level LOBSTER data [52]. The left panel shows the price response to executing a buy order of 100 shares at the time $t = 0$ ms, which consumes ask-side liquidity within 1000 ms and induces an immediate upward shift in price. The right panel shows the analogous impact of a sell order. This plot indicates that the transition dynamics induced by trading are not symmetrically distributed around the nominal kernel.

ronments. The robust RL framework explicitly acknowledges that the real market environment may differ from the simulated training environment, and it aims to learn policies that are resilient under a range of perturbations.

However, existing robust RL approaches face a key limitations in financial applications; that is, traditional uncertainty sets are typically symmetric, which fail to capture the directional nature of market impact. Addressing this challenge motivates our threefold **contributions**:

- (1) We propose a novel class of uncertainty sets, *elliptic uncertainty sets* (Definition 3.2), which generalize traditional ℓ_p -norm balls to the ellipse-like structure. These sets better capture the empirically observed directional nature of market impact as illustrated in Figure 1.
- (2) On the theoretical side, we derive closed-form solutions for solving the worst-case transition kernel under the proposed uncertainty sets (Theorem 3.3). Furthermore, under certain conditions, we present the explicit solutions (Theorem 3.4). This development significantly broadens the scope of tractable robust RL problems beyond symmetric (ball-shaped) uncertainty sets, enabling more faithful representation of market impact.
- (3) We empirically evaluate our approach on real-world financial data using trade-level market impact simulations in Section 4. Experimental results demonstrate that our method consistently outperforms the standard single-asset intra-day trading strategy and existing RL baselines in terms of Sharpe ratio. Moreover, we validate the robustness of our method under increasing strategy volume, confirming its effectiveness in high-volume regimes.

1.1 Related Work

Existing Approaches to Handle Market Impact The most common approach for handling market impact is to simulate the electronic market more accurately. By applying high-fidelity market simulators, often built upon limit order book (LOB) dynamics [31, 32, 53–55], trade-level data [56, 57], or large-scale agent-based simulators [58–61], it captures more detailed market microstructure and reduces the gap between the simulated and the real trading environments. Prominent approaches to incorporating market impact into backtesting include agent-based simulation frameworks [58–61], data-driven LOB reconstruction models [62, 63, 55], and hybrid systems that integrate historical data replay with synthetic order flow generation [64, 65]. However, access to high-quality market data is often limited, and simulating market environments with agent-based systems remains prohibitively expensive. Therefore, a practical alternative is to train the agent directly on historical data without market impact while still encouraging it to account for potential worst-case scenarios.

Robust RL in Finance Robust RL, with its intrinsic ability to handle model misspecification, provides a natural framework for incorporating market impact considerations during training. Jaimungal

et al. [3] propose a robust reinforcement learning framework based on rank-dependent utility to address uncertainty in financial decision-making, demonstrating the effectiveness of robust RL in portfolio allocation, benchmark strategy optimization, and statistical arbitrage. Shi et al. [5] formulate portfolio optimization as a robust RL problem to enhance resilience. We et al. [30] extend robust risk-aware RL to manage the risks associated with path-dependent financial derivatives, showcasing its effectiveness in complex hedging scenarios. However, existing work primarily focuses on fully symmetric uncertainty sets, which fail to capture the directional characteristics of financial markets. Addressing this limitation is the central focus of our paper.

Modeling the Uncertainty Set in Robust RL The uncertainty set captures the discrepancy between training and deployment environments. However, robust RL becomes computationally intractable when the uncertainty set is highly irregular [34, 50, 45, 46]. To mitigate this issue, it is common to impose structural assumptions that enable tractable solutions. We highlight several representative structures, with further details deferred to Appendix A.1. The R -contamination model [37] defines an uncertainty set as a sphere of radius R centered around the nominal transition kernel, admitting analytical solutions for robust policy evaluation. The ℓ_p -norm uncertainty sets are also widely studied due to their closed-form solutions [45, 46]. The integral probability metric (IPM) and double-sampling uncertainty sets have been shown to allow efficient computation [49]. Other works include uncertainty sets based on the Wasserstein metric and f -divergence, which have also received considerable attention [66–68].

2 Preliminaries: Robust Reinforcement Learning

We focus on the discounted infinite-horizon Markov Decision Processes (MDPs) [69], formally defined as a five-tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$, where \mathcal{S} and \mathcal{A} denote the state and action spaces¹, respectively. The transition kernel $\mathbb{P}(s' | s, a)$ specifies the probability of transitioning to state s' from state s after taking action a . The reward function $r : \mathcal{S} \rightarrow [0, 1]$ assigns a bounded reward to each state, and the discount factor $\gamma \in (0, 1)$ models the agent’s preference for immediate rewards over future ones.

Instead of assuming a fixed transition kernel \mathbb{P} , we account for the effects of *model misspecification*. Let \mathbb{P}_0 denote the nominal transition kernel, and define the (s, a) -uncertainty set $\mathcal{U}_{s,a} \subset \mathbb{R}^{|\mathcal{S}|}$ at the point $(s, a) \in \mathcal{S} \times \mathcal{A}$ as

$$((s, a)\text{-Uncertainty Set}) \quad \mathcal{U}_{s,a} := \{u_{s,a} \in \mathbb{R}^{|\mathcal{S}|} \mid u_{s,a} \in C_{s,a}, u_{s,a}^\top \mathbf{1}_{|\mathcal{S}|} = 0\},$$

where $C_{s,a} \subset \mathbb{R}^{|\mathcal{S}|}$ is a convex set, $\mathbf{1}_{|\mathcal{S}|}$ is an all-one vector with the dimension $|\mathcal{S}|$ (for convenience, we will usually omit the subscript $|\mathcal{S}|$). The convex set is commonly chosen as a ball-shaped set (e.g. $C_{s,a} = B_p(\beta_{s,a}) := \{u_{s,a} \in \mathbb{R}^{|\mathcal{S}|} \mid \|u_{s,a}\|_p \leq \beta_{s,a}\}$ for the ℓ_p -norm uncertainty set), where $\beta_{s,a} \geq 0$ is a radius parameter that quantifies the allowable deviation. The zero-sum constraint $u_{s,a}^\top \mathbf{1}_{|\mathcal{S}|} = 0$ ensures the perturbed transition $\mathbb{P}(\cdot | s, a) + u_{s,a}$ remains a valid probability distribution. The *uncertainty set* \mathcal{U} and the *robust transition model* are then defined as

$$(\text{Uncertainty Set}) \quad \mathcal{U} := \bigtimes_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{U}_{s,a}, \quad (1)$$

$$(\text{Robust Transition Model}) \quad \mathcal{P}_{\mathcal{U}} := \left\{ \mathbb{P}_u := \mathbb{P}_0 + u \mid u \in \mathcal{U} \right\}, \quad (2)$$

respectively. Throughout this paper, we assume the parameter of uncertainty set is chosen appropriately such that all elements in the robust transition model is well-defined [45, 46]. Given these notations in place, we define the value function of the policy π with the uncertainty u as the value function with the transition probability $\mathbb{P}_u \in \mathcal{P}$:

$$V_u^\pi(s) := \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, \mathbb{P}_u, \pi \right],$$

¹For theoretical analysis, we restrict our attention to MDPs with finite state and action spaces. This assumption avoids the technical complications arising from continuous or hybrid spaces, which, although explored in some prior work [70–73], remain analytically open without imposing additional assumptions, especially for robust RL problems. Nevertheless, our empirical results extend to continuous settings, demonstrating the practical applicability of our approach beyond the theoretical scope.

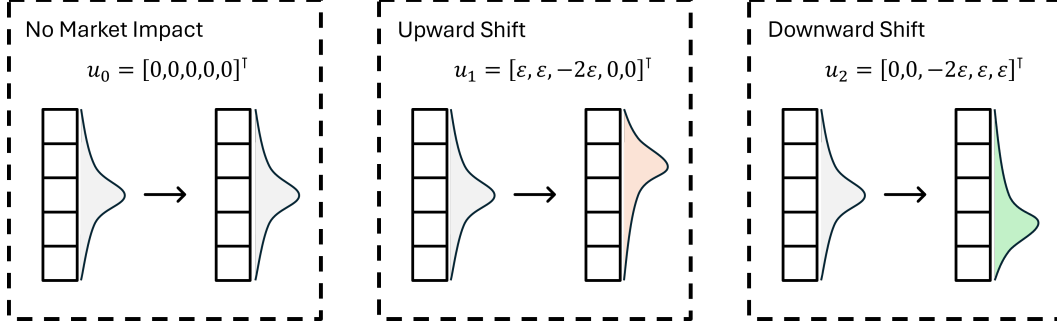


Figure 2: Illustration of the transition kernel in a simplified robust RL setting using an ℓ_∞ -norm uncertainty set. Only one of the upward (buy) or downward (sell) shift is plausible in a given scenario; however, the ball-shaped uncertainty set must include both due to its symmetric structure, which motivates us to propose an uncertainty set with non-symmetric structures.

104 The robust value function is the worst-case value function over all uncertainties; that is $V^\pi(s) :=$
 105 $\min_{u \in \mathcal{U}} V_u^\pi(s)$. Similarly, we can also define the robust Q-function and the robust advantage function
 106 as $Q^\pi(s, a) := \min_u \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a, \mathbb{P}_u, \pi]$ and $A^\pi(s, a) := Q^\pi(s, a) -$
 107 $V^\pi(s)$, respectively.

The goal of robust RL is to learn a parameterized policy π_θ that maximizes the worst-case value function $V^{\pi_\theta}(s_0)$, where s_0 denotes the initial state. A standard approach applies the robust policy gradient formula [34]:

$$\nabla_\theta V^\pi(s) = \mathbb{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta} [Q^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a|s)],$$

108 where d^{π_θ} is the stationary distribution induced by π_θ , and Q^{π_θ} denotes the robust Q-function. This
 109 formulation reduces robust RL to a gradient-based optimization problem, shifting the main challenge
 110 of robust RL to accurately evaluate the robust value function, which is our focus in Section 3.3.

111 3 Modeling Market Impact with Elliptic Uncertainty Sets

112 3.1 Limitations of Symmetric Uncertainty Sets

113 Robust MDPs offer an ideal framework to handle model misspecification by allowing the transition
 114 kernel to deviate within a prescribed uncertainty set. However, most existing formulations adopt
 115 symmetric structures, typically the ball defined by a specific norm, that treat all directions of
 116 perturbation equally. While mathematically convenient, these symmetric structures often fail to
 117 reflect the directional nature of real-world uncertainties.

118 Symmetry in this context typically refers to invariance under the signed permutation group (see
 119 Appendix A.2 for details). A canonical example is the ℓ_p -norm ball:

$$B_p(\beta) := \{u \in \mathbb{R}^d \mid \|u\|_p \leq \beta\},$$

120 which satisfies the property that for any $u \in B_p(\beta)$, all signed permutations of u are also contained in
 121 the set. It enforces an implicit assumption of isotropic uncertainty, equally plausible in all directions,
 122 which often includes unrealistic perturbations. We demonstrate it in the following example:

Example 3.1 (Symmetric Sets Fail to Capture Directional Uncertainty). In financial markets, large buy or sell orders induce directional shifts in asset prices due to liquidity consumption (see Figure 1). In Figure 2, we consider the following classical ℓ_∞ -norm (s, a) -uncertainty set:

$$\mathcal{U}_{s,a} := \{u \in \mathbb{R}^d \mid \|u\|_\infty \leq 2\varepsilon, u^\top \mathbf{1} = 0\}.$$

This set includes, for example, the perturbation vectors

$$u_1 = [\varepsilon, \varepsilon, -2\varepsilon, 0, 0]^\top, \quad \text{and} \quad u_2 = [0, 0, -2\varepsilon, \varepsilon, \varepsilon]^\top.$$

123 Both u_1 and u_2 satisfy the norm and mean constraints, and since they are signed permutations of
 124 each other, they must either both belong to $\mathcal{U}_{s,a}$ or be excluded together. However, this symmetry

fails to reflect market realities: under a buy action, u_1 represents a plausible upward shift due to liquidity-driven market impact, while u_2 , corresponding to a downward shift, is implausible. Thus, the symmetric structure forces inclusion of perturbations that contradict the directional market impact, potentially leading to overly conservative or unrealistic robust policies.

This observation underscores the importance of developing a robust RL framework that can capture the directional nature of environment shifts observed in financial markets. To this end, we introduce a novel class of *elliptic uncertainty sets*, which generalize traditional norm-bounded sets by allowing non-symmetric perturbations, while retaining closed-form tractability under certain conditions.

3.2 The Elliptic Uncertainty Sets

The *elliptic uncertainty sets* generalize the classical ℓ_p -norm uncertainty set by incorporating directional non-symmetry. Formally, we have the following definition:

Definition 3.2 (Elliptic (s, a) -Uncertainty Set). For each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the *elliptic (s, a) -uncertainty set* is defined as

$$\mathcal{U}_{s,a} := \left\{ u \in \mathbb{R}^{|\mathcal{S}|} \mid \sum_{n=1}^N \|u - u_n^{s,a}\| \leq \beta_{s,a}, u^\top \mathbf{1}_{|\mathcal{S}|} = 0 \right\}, \quad (3)$$

where $\{u_n^{s,a}\}_{n=1}^N \in \mathbb{R}^{|\mathcal{S}|}$ are called the *foci* of the ellipse, $\beta_{s,a} \geq 0$ is the *uncertainty size*, and $\|\cdot\| : \mathbb{R}^d \rightarrow \mathbb{R}$ is an arbitrary norm. Particularly, when $\|\cdot\|$ is taken as the ℓ_p -norm ($p \in [1, +\infty]$), $\mathcal{U}_{s,a}$ is called the ℓ_p -*ellipse uncertainty set*.

Here the constraint $u^\top \mathbf{1}_{|\mathcal{S}|} = 0$ ensures the perturbed transition still defines a valid probability distribution. For convenience, we will omit the superscript in $u_n^{s,a}$ and the subscript in $\beta_{s,a}$ when the context is clear.

Connection to the Classical Ellipse Our definition draws directly from the geometric characterization of an ellipse: the set of points for which the sum of distances to multiple foci is bounded by a constant $\beta_{s,a}$. We show that it recovers classical structures as special cases with the following two concrete examples:

- (1) When $N = 1$ and $u_1 = 0$, the set Eq. (3) reduces to the ℓ_p -norm ball

$$B_p(\beta) := \{u \mid \|u\|_p \leq \beta\},$$

which aligns with the standard uncertainty set used in robust RL (e.g., [45, 46, 34]).

- (2) When $N = 2$ and $p = 2$, Eq. (3) becomes $\|u - u_1\|_2 + \|u - u_2\|_2 \leq \beta$. Defining the midpoint $\bar{u} := \frac{u_1 + u_2}{2}$, there exists a matrix A such that this constraint is equivalent to a classical quadratic form of an ellipse:

$$(u - \bar{u})^\top A (u - \bar{u}) \leq 1,$$

where the detailed derivation is given in Lemma B.10.

These examples demonstrate that our formulation significantly generalizes classical ellipses by allowing arbitrary norms and accommodating multiple foci, thereby enabling more flexible modeling of non-symmetric perturbations. However, such generality often comes at the cost of increased complexity. To address this concern, in the next section, we show that, under certain parameter choices, our proposed uncertainty set remains as trackable as traditional ℓ_p -norm uncertainty sets.

3.3 Solving the Worst-Case Uncertainty

Solving the worst-case uncertainty $u^* := \arg \min V_u^\pi(s_0)$ plays a crucial role in efficient robust policy evaluation. The ℓ_p -norm uncertainty set [45, 46] and the R -contamination model [37] are popular as the solution u^* is closed-form. When the uncertainty set is complicated, many existing robust RL methods require an external optimization loop to determine the worst-case transition probability, which can be impractical in real-world scenarios.

To address this issue, we derive an implicit solution (Theorem 3.3) for the worst-case uncertainty that avoids additional interaction with the environment. Under certain conditions, we further provide

an explicit closed-form solution (Theorem 3.4), enabling direct computation without the need for external solvers.

We start with recapping some backgrounds in robust TD learning. Let \mathcal{V} denote all functions mapping from the state space \mathcal{S} to the Euclidean space \mathbb{R} . Given that \mathcal{U} is an arbitrary uncertainty set, the robust Bellman operator associated with the policy π , $\mathcal{T}_{\mathcal{U}}^{\pi} : \mathcal{V} \rightarrow \mathcal{V}$, is defined as

$$\mathcal{T}_{\mathcal{U}}^{\pi} v(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left[r(s, a) + \gamma \min_{u \in \mathcal{U}_{s,a}} u^{\top} v + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}_0(s'|s, a) v(s') \right].$$

As shown by [74, 34], when $\gamma \in (0, 1)$, the robust Bellman operator is a contraction operator, which admits the unique fixed point as the robust value function $V^{\pi} \in \mathcal{V}$. When \mathcal{U} is given by the ℓ_p -ellipse uncertainty set (Eq. (3)), then

$$\mathcal{T}_{\mathcal{U}}^{\pi} v(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left[r(s, a) + \gamma \min_{\substack{\sum_n \|u - u_n\| \leq \beta \\ u^{\top} \mathbf{1} = 0}} u^{\top} v + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}_0(s'|s, a) v(s') \right].$$

It turns out that if we can solve the optimization problem

$$\min_{\substack{\sum_n \|u - u_n\| \leq \beta_{s,a} \\ u^{\top} \mathbf{1} = 0}} u^{\top} v, \quad (4)$$

the robust Bellman operator is just the standard Bellman operator (over the nominal transition probability) with a solved shift. As the result, we can simply apply the standard TD-learning with adding this correction term to solve the desired robust value function. In the following theorem, we present a general recipe of solving this optimization problem.

Theorem 3.3 (Implicit). *Let $d := |\mathcal{S}|$ be the cardinal of state space. Suppose that $\{u_n\}_{n=1}^N \subset \mathbb{R}^d$ for each $(s, a) \in \mathcal{S} \times \mathcal{A}$, the uncertainty size $\beta \geq 0$, and $v \in \mathbb{R}^d$. Then there exists λ^* and μ^* such that the optimization problem defined by Eq. (4) is solved by*

$$u^* = \arg \min_u [(v + \mu^* \mathbf{1})^{\top} u + \lambda^* \sum_{n=1}^N \|u - u_n\|_p].$$

Remark. We say a solution of the optimization problem is “implicit” if it can be represented as an equation in the form $g(u, v, \beta, \{u_n\}) = 0$. In this theorem, as the right-hand side

$$G(u, v, \beta, \{u_n\}) := (v + \mu^* \mathbf{1})^{\top} u + \lambda^* \sum_{n=1}^N \|u - u_n\|_p$$

is proper, convex, and coercive; we can surely re-write it as the sub-gradient form by letting

$$g(u, v, \beta, \{u_n\}) \in \partial_u G(u, v, \beta, \{u_n\}).$$

Then we obtain the implicit representation $g(u, v, \beta, \{u_n\}) = 0$. Moreover, we derive the formula of λ^* and μ^* beyond the existence; the full result is presented in Theorem B.12 with more details.

The implicit solution has already shown significant advances compared to some of existing robust RL methods which typically require to solve the worst-case transition probability using additional state-action sample generated from the agent-environment iteration. However, it is still (slightly) impractical to solve additional convex optimization problems in every iteration. Fortunately, under certain conditions, the solution can be “explicit”; that is, we can write the optimal solution in the form of $u^* = f(v, \beta, \{u_n\})$.

Theorem 3.4 (Explicit). *Let $d := |\mathcal{S}|$ be the cardinal of state space. Suppose that $\{u_n\}_{n=1}^N \subset \mathbb{R}^d$ for each $(s, a) \in \mathcal{S} \times \mathcal{A}$, the uncertainty size $\beta \geq 0$, and $v \in \mathbb{R}^d$. The optimization problem defined by Eq. (4) is explicitly solved in the following cases:*

(a) Let $N = 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then the minimizer of Eq. (4) is given by

$$u^* = u_1 + \beta \frac{\text{sign}(v + \mu^* \mathbf{1}) \odot |v + \mu^* \mathbf{1}|^{q-1}}{\|v + \mu^* \mathbf{1}\|_q^{q-1}}.$$

Here sign and $|\cdot|$ are coordinate-wise sign function and absolute value, respectively; \odot is the coordinate-wise product; and $\mu^* := \arg \min_{\mu \in \mathbb{R}} \|v + \mu \mathbf{1}\|_q$.

(b) Let $p = 1$ and $N = 2$. Suppose that $\beta > \|u_2 - u_1\|_1$. Then the minimizer of Eq. (4) is given by

$$u^* = \frac{u_1 + u_2}{2} - \frac{\beta - \|u_1 - u_2\|_1}{2} [\text{sign}(v + \mu^* \mathbf{1}) \odot \mathbb{I}_{\{|v + \mu^* \mathbf{1}| = 2\lambda^*\}}].$$

189 Here sign , $|\cdot|$, and \mathbb{I} are coordinate-wise sign function, absolute value, and indicator function,
 190 respectively; \odot is the coordinate-wise product; $\mu^* := -\frac{\max_j v_j + \min_j v_j}{2}$; and $\lambda^* :=$
 191 $-\frac{\max_j v_j - \min_j v_j}{4}$.

(c) Let $p = 2$ and $N = 2$. Suppose that $\beta > \|u_2 - u_1\|_2$. Then the minimizer of Eq. (4) is given by

$$u^* = \frac{u_1 + u_2}{2} - \frac{\sqrt{\beta^2 - \|u_2 - u_1\|_2^2}}{2} \frac{1}{\underline{\lambda}^*} \Omega^{-1}(v + \mu^* \mathbf{1}).$$

192 Here $\Omega := I - \frac{1}{\beta^2}(u_2 - u_1)(u_2 - u_1)^\top$, $\underline{\lambda}^* = \sqrt{(v + \mu^* \mathbf{1})^\top \Omega^{-1}(v + \mu^* \mathbf{1})}$, and $\mu^* =$
 193 $-\frac{v^\top \Omega^{-1} \mathbf{1}}{\mathbf{1}^\top \Omega^{-1} \mathbf{1}}$.

194 *Remark.* This result presents a clean form of the worst-case uncertainty u^* under certain conditions.
 195 Unlike the implicit case where it takes an additional root-finding algorithm to solve an approximated
 196 u^* , in the explicit case, if the current value function $v \in \mathcal{V}$ is given and the parameters of the
 197 uncertainty set (u_i and β) have been determined, the uncertainty u^* can be explicitly solved. The full
 198 proof is presented in Theorem B.13.

199 3.4 Robust TD Learning Algorithm

200 Given Theorem 3.3 and Theorem 3.4 in place, we immediately obtain the robust TD learning
 201 algorithm for robust policy evaluation. Given the current value function v , we can calculate u^* using
 202 the implicit and the explicit formula; assume the current state-action pair is given as (s, a, s') , then
 203 the updated value function is given by

$$v'(s) = v(s) + \eta \left(r(s, a) + \gamma \sum_{s'} (\mathbb{P}_0(s'|s, a) + u^*(s')) [v(s') + \gamma v(s')] - v(s) \right). \quad (5)$$

204 We can further simplify this update rule by using an unbiased estimator of that:

$$v'(s) = v(s) + \eta \gamma v^\top u^* + \eta (r(s, a) + \gamma v(s') - v(s)), \quad (6)$$

205 where $s' \sim \mathbb{P}_0(\cdot|s, a)$. The formulation leads us to Algorithm 1.

Algorithm 1: Robust Policy Evaluation

Input: The target policy π , the foci $\{p_i\}_{i=1}^N \subset \mathbb{R}^d$, and $\{\beta_{s,a}\}$ the uncertainty size

1 Sample the initial state s_0 from the initial distribution;
 2 Initialize the value function v_0 ;
 3 **for** $t = 0, 1, 2, \dots, T - 1$ **do**
 206 4 Sample the action $a_t \sim \pi(\cdot|s_t)$; Transition from s_t to $s_{t+1} \sim \mathbb{P}_0(\cdot|s_t, a_t)$;
 5 Calculate u^* using $\{p_i\}_{i=1}^N$, $\{\beta_{s,a}\}$, and v_t ;
 6 Robust TD-learning: $v'(s) = v(s) + \eta \gamma v^\top u^* + \eta (r(s, a) + \gamma v(s') - v(s))$;
 7 **end**

Output: The final value function v_T

207 *Remark.* The convergence of this algorithm, as well as its corresponding Actor-Critic-style policy
 208 gradient algorithm, follows directly by applying the standard proof routine from the robust RL
 209 literature (e.g. [49]). For completeness, we include the convergence result and full proof in the
 210 supplementary material.

211 4 Experiments

212 To validate our theoretical findings and demonstrate the practical effectiveness of robust RL framework
 213 in the environment with the market impact, we conduct experiments on two different tasks that are

Table 1: Performance comparison of different RL agents on selected assets under the simulated market impact from June 9 to December 9, 2022. Robust RL with ℓ_p -ellipse uncertainty set consistently achieves the highest Sharpe ratio.

Asset	Method	Final Value (\$)	Annualized Return (%)	Sharpe Ratio	Max Drawdown (%)
META	Momentum	96 334	-7.3%	-0.95	-4.2%
	Non-Robust RL	120 347	44.8%	1.74	-11.3%
	Robust RL (ℓ_p -Ball)	97 103	-5.7%	-0.28	-13.4%
	Robust RL (ℓ_p -Ellipse)	138 011	90.8%	2.48	-9.9%
MSFT	Momentum	105 163	10.6%	1.10	-5.3%
	Non-Robust RL	87 440	-23.5%	-1.75	-16.9%
	Robust RL (ℓ_p -Ball)	92 159	-15.1%	-0.82	-11.2%
	Robust RL (ℓ_p -Ellipse)	111 485	24.4%	1.20	-10.1%
SPY	Momentum	107 333	15.2%	1.69	-3.2%
	Non-Robust RL	91 947	-15.5%	-1.64	-11.9%
	Robust RL (ℓ_p -Ball)	100 560	1.1%	0.17	-6.4%
	Robust RL (ℓ_p -Ellipse)	109 272	19.4%	1.60	-5.8%

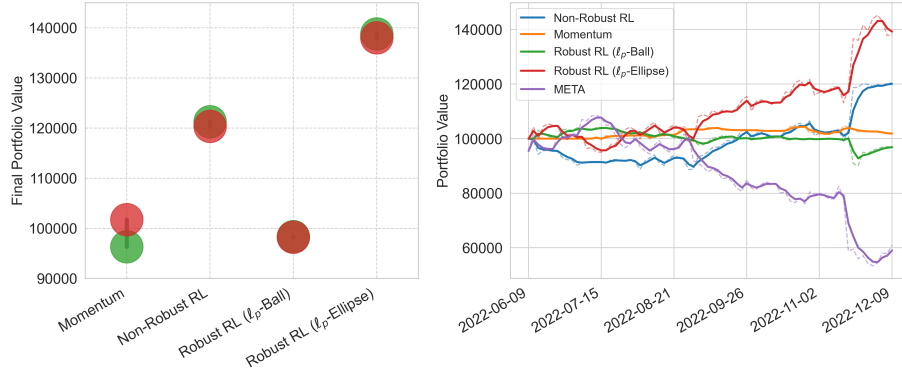


Figure 3: Performance comparison of trading strategies on the META stock from June 9 to December 9, 2022, under simulated market impact. The left panel compares the final portfolio values with (in red) and without market impact (in green), illustrating the robustness of each method to execution-related slippage. The right panel shows the cumulative returns over the evaluation period, tracking the performance of the four strategies in Table 1, alongside the baseline performance of the META stock.

214 closely tied to market impact: (1) minute-level single-asset strategy, and (2) large-volume portfolio
215 rebalancing. In minute-level trading, even small trade sizes can noticeably move prices, leading to
216 slippage. Similarly, large-scale portfolio rebalancing, often performed by large financial institutions,
217 can significantly affect asset prices due to the large order volumes involved.

218 4.1 Performance Comparison on Single-Asset Intra-Day Trading

219 We start with the single-asset minute-level trading. The non-RL baseline is chosen as the momentum
220 strategy [75], which is designed based on the empirical observation where assets that have performed
221 well in the recent past are more likely to continue performing well in the near future.

222 **Training and Evaluation of RL Agents** We implement a Gym-like RL environment [76, 77]
223 constructed on historical data, with full environment details provided in Appendix C.3. All RL agents
224 are trained on one year of earlier historical data (from May 9th, 2021 to May 9th, 2022 as the nominal
225 environment) without accounting for market impact. Their performances are then evaluated over
226 the period from June 9 to December 9, 2022, with the market impact included. To simulate market
227 impact, we reconstruct LOB dynamics using a short period of real trading orders and determine the
228 execution price via the volume-weighted average price (VWAP). A simple example illustrating this
229 estimation process is shown in Table 2, Appendix C.

230 **Results** As shown in Table 1, our proposed method consistently outperforms the momentum strategy,
231 the non-robust RL, and the symmetric robust RL baselines (based on ℓ_p -norm balls) in terms of the

232 risk-adjusted return (Sharpe Ratio). These experiments validate the following key understandings:
 233 (i) While robust RL with symmetric uncertainty sets significantly mitigates the effects of market
 234 impact (as illustrated in the left panel of Figure 3), it often produces overly conservative strategies
 235 that compromise profitability by taking implausible perturbations into consideration; (ii) the non-
 236 robust RL usually suffers greater risk exposure, resulting in the highest Max Drawdown among all
 237 methods; (iii) in contrast, the proposed ℓ_p -ellipse uncertainty set effectively captures the directional
 238 non-symmetry of market impact, allowing the agent to achieve a more favorable trade-off between
 239 robustness and return.

240 4.2 Robustness to the Market Impact Scaling in the Trading Volume

In this subsection, we show that a policy trained in a low-volume environment continues to mitigate market impact when transferred to portfolios with significantly larger volumes. We consider a multi-asset portfolio allocation task, modeling the realistic setting where large volumes are traded over short periods to maintain a low-variance portfolio. The same Gym-like RL environment and evaluation period from the previous experiment are used. We evaluate the robustness to the market impact using the relative portfolio gap, the normalized absolute difference in final portfolio value with and without market impact:

$$\text{Relative Port. Gap} = \frac{|\text{Port. Value with MI} - \text{Port. Value without MI}|}{\text{Initial Cash}},$$

241 where MI represents the market impact. Additional experimental details are provided in Appendix C.

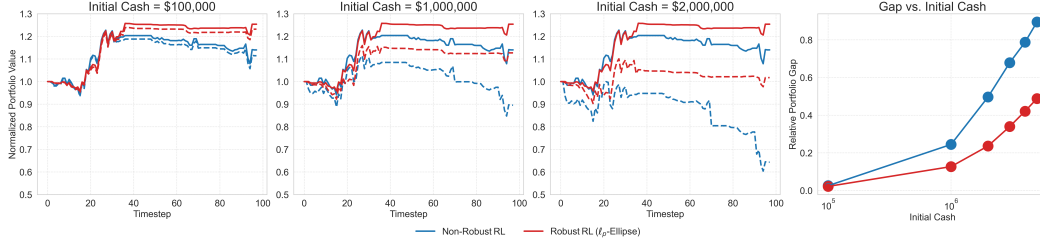


Figure 4: Robustness of RL agents to market impact under increasing trading volumes. The left three panels show normalized portfolio values over time across initial cash levels, with dashed and solid lines indicating performance with and without market impact, respectively. The right panel shows the relative portfolio gap, which increases sharply for the non-robust agent but remains small and stable for the robust RL agent with ℓ_p -elliptic uncertainty sets.

242 **Results** As shown in Figure 4, the robust RL agent with ℓ_p -ellipse uncertainty set consistently
 243 outperforms the non-robust RL method both in return and in mitigating the effects of market impact.
 244 While the performance gap is small at low volume, the non-robust agent degrades rapidly as volume
 245 increases, suffering from instability and larger drawdowns. In contrast, the robust agent remains
 246 stable and profitable even at high volume (~ 200 M), demonstrating strong scalability.

247 5 Conclusion & Broader Impact

248 This paper focuses on the market impact appearing in quantitative trading, where an agent’s actions
 249 affect prices. By modeling the training environment as the nominal transition kernel, the proposed
 250 novel ℓ_p -ellipse uncertainty sets better captures the non-symmetric nature of price responses compared
 251 to traditional symmetric sets. We established the theoretical tractability of this approach by deriving
 252 implicit and explicit closed-form solutions for robust policy evaluation within this framework,
 253 enabling efficient robust TD-learning algorithms that account for the market impact during training
 254 on the nominal historical environment. Experiments on real historical data demonstrated that our
 255 method significantly improves robustness and risk-adjusted returns over non-RL, non-robust RL, and
 256 symmetric robust RL baselines. This work broadens the applicability of tractable robust RL and
 257 offers a more faithful modeling approach for market impact. The broader impact involves potentially
 258 more stable and profitable automated trading strategies.

References

- [1] Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023.
- [2] Nikolaos Pippas, Cagatay Turkay, and Elliot A Ludvig. The evolution of reinforcement learning in quantitative finance. *arXiv preprint arXiv:2408.10932*, 2024.
- [3] Sebastian Jaimungal, Silvana M Pesenti, Ye Sheng Wang, and Hariom Tatsat. Robust risk-aware reinforcement learning. *SIAM Journal on Financial Mathematics*, 13(1):213–226, 2022.
- [4] Pengqian Yu, Joon Sern Lee, Ilya Kulyatin, Zekun Shi, and Sakyasingha Dasgupta. Model-based deep reinforcement learning for dynamic portfolio optimization. *arXiv preprint arXiv:1901.08740*, 2019.
- [5] Xiaochuan Shi, Yihua Zhou, and Lei Wu. Robust reinforcement learning for portfolio management via competition and cooperation strategies, 2023. ICLR 2024 Conference Withdrawn Submission.
- [6] Philip Ndikum and Serge Ndikum. Advancing investment frontiers: Industry-grade deep reinforcement learning for portfolio optimization. *arXiv preprint arXiv:2403.07916*, 2024.
- [7] Carlos Betancourt and Wen-Hui Chen. Deep reinforcement learning for portfolio management of markets with a dynamic number of assets. *Expert Systems with Applications*, 164:114002, 2021.
- [8] Amine Mohamed Aboussalah and Chi-Guhn Lee. Continuous control with stacked deep dynamic recurrent reinforcement learning for portfolio optimization. *Expert Systems with Applications*, 140:112891, 2020.
- [9] Min-Yuh Day, Ching-Ying Yang, and Yensen Ni. Portfolio dynamic trading strategies using deep reinforcement learning. *Soft Computing*, 28(15):8715–8730, 2024.
- [10] Yuh-Jong Hu and Shang-Jen Lin. Deep reinforcement learning for optimizing finance portfolio management. In *2019 amity international conference on artificial intelligence (AICAI)*, pages 14–20. IEEE, 2019.
- [11] Angelos Filos. Reinforcement learning for portfolio management. *arXiv preprint arXiv:1909.09571*, 2019.
- [12] WARAMETH NUIPIAN, PHAYUNG MEESAD, et al. *Dynamic Portfolio Management with Deep Reinforcement Learning*. PhD thesis, King Mongkut’s University of Technology North Bangkok, 2025.
- [13] Farzan Soleymani and Eric Paquet. Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder—deepbreath. *Expert Systems with Applications*, 156:113456, 2020.
- [14] Hyunmin Cho and Hyun Joon Shin. Trading strategies using reinforcement learning. *Journal of the Korea Academia-Industrial cooperation Society*, 22(1):123–130, 2021.
- [15] John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE transactions on neural Networks*, 12(4):875–889, 2001.
- [16] Yang Li, Wanshan Zheng, and Zibin Zheng. Deep robust reinforcement learning for practical algorithmic trading. *IEEE Access*, 7:108014–108022, 2019.
- [17] Ji-Heon Park, Jae-Hwan Kim, and Jun-Ho Huh. Deep reinforcement learning robots for algorithmic trading: Considering stock market conditions and us interest rates. *IEEE Access*, 12:20705–20725, 2024.
- [18] Yasmeen Ansari, Sadaf Yasmin, Sheneela Naz, Hira Zaffar, Zeeshan Ali, Jihoon Moon, and Seungmin Rho. A deep reinforcement learning-based decision support system for automated stock market trading. *IEEE Access*, 10:127469–127501, 2022.
- [19] Xing Wu, Haolei Chen, Jianjia Wang, Luigi Troiano, Vincenzo Loia, and Hamido Fujita. Adaptive stock trading strategies with deep reinforcement learning methods. *Information Sciences*, 538:142–158, 2020.
- [20] Olivier Guéant, Charles-Albert Lehalle, and Joaquin Fernandez-Tapia. Dealing with the inventory risk: a solution to the market making problem. *Mathematics and Financial Economics*, 7(4):477–507, 2013.
- [21] Joel Hasbrouck. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, Oxford, UK, 2007.
- [22] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deep reinforcement learning for trading. *arXiv preprint arXiv:1911.10107*, 2019.

- [23] Kyung Hyun Park, Hyeong Jin Kim, and Woo Chang Kim. Deep reinforcement learning for limit order book-based market making. *Expert Systems with Applications*, 169:114338, 2021.
- [24] Pierre Casgrain, Anirudh Kulkarni, and Nicholas Watters. Learning to trade with continuous action spaces: Application to market making. *arXiv preprint arXiv:2303.08603*, 2023.
- [25] Petter N Kolm, Sebastian Krügel, and Sergiy V Zadorozhnyi. Reinforcement learning for optimal hedging. *The Journal of Trading*, 14(4):4–17, 2019.
- [26] Hans Buehler, Lukas Gonon, Josef Teichmann, and Ben Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019.
- [27] H Cao, Y Wang, and Y Zhang. Risk-averse reinforcement learning for optimal option hedging. *Journal of Computational Finance*, 24(2):1–31, 2020.
- [28] W L Chan and R O Shelton. Can machine learning improve delta hedging? *Journal of Derivatives*, 9(1):39–56, 2001.
- [29] Z Ning and Y K Kwok. Q-learning for option pricing and hedging with transaction costs. *Applied Economics*, 52(55):6033–6048, 2020.
- [30] David Wu and Sebastian Jaimungal. Robust risk-aware option hedging. *Applied Mathematical Finance*, 30(3):153–174, 2023.
- [31] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.
- [32] Anna A Obizhaeva and Jiang Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial markets*, 16(1):1–32, 2013.
- [33] Bence Tóth, Zoltán Eisler, and J-P Bouchaud. The square-root impace law also holds for option markets. *Wilmott*, 2016(85):70–73, 2016.
- [34] Yan Li, Guanghui Lan, and Tuo Zhao. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022.
- [35] Yike Li, Yunzhe Tian, Endong Tong, Wenjia Niu, and Jiqiang Liu. Robust reinforcement learning via progressive task sequence. In *IJCAI*, pages 455–463, 2023.
- [36] Guanlin Liu, Zhihan Zhou, Han Liu, and Lifeng Lai. Efficient action robust reinforcement learning with probabilistic policy execution uncertainty. *Transactions on Machine Learning Research*, 2024.
- [37] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- [38] Shangding Gu, Laixi Shi, Muning Wen, Ming Jin, Eric Mazumdar, Yuejie Chi, Adam Wierman, and Costas Spanos. Robust gymnasium: A unified modular benchmark for robust reinforcement learning. In *International Conference on Learning Representations*, 2025.
- [39] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. *arXiv preprint arXiv:1703.02702*, 2017.
- [40] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- [41] Minghong Fang, Xilong Wang, and Neil Zhenqiang Gong. Provably robust federated reinforcement learning. *arXiv preprint arXiv:2502.08123*, 2025.
- [42] Wei Shen, Xiaoying Zhang, Yuanshun Yao, Rui Zheng, Hongyi Guo, and Yang Liu. Robust rlhf with noisy rewards. In *International Conference on Learning Representations*, 2025. Withdrawn submission.
- [43] Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371):1–40, 2023.
- [44] Pierre Clavier, Laixi Shi, Erwan Le Pennec, Eric Mazumdar, Adam Wierman, and Matthieu Geist. Near-optimal distributionally robust reinforcement learning with general l_p norms. *Advances in Neural Information Processing Systems*, 37:1750–1810, 2024.
- [45] Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. An efficient solution to s-rectangular robust markov decision processes. *arXiv preprint arXiv:2301.13642*, 2023.

- [46] Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Y Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. *Advances in Neural Information Processing Systems*, 36:59477–59501, 2023.
- [47] Runyu Zhang, Yang Hu, and Na Li. Soft robust MDPs and risk-sensitive MDPs: Equivalence, policy gradient, and sample complexity. In *The Twelfth International Conference on Learning Representations*, 2024.
- [48] Zifan Wu, Chao Yu, Chen Chen, Jianye Hao, and Hankz Hankui Zhuo. Plan to predict: Learning an uncertainty-foreseeing model for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35:15849–15861, 2022.
- [49] Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, PR Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation. *Advances in neural information processing systems*, 36:97–133, 2023.
- [50] Qiuhaio Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. In *International Conference on Machine Learning*, pages 35763–35797. PMLR, 2023.
- [51] Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Policy optimization for robust average reward mdps. *Advances in Neural Information Processing Systems*, 37:17348–17372, 2024.
- [52] Ruihong Huang and Tomas Polak. Lobster: Limit order book reconstruction system. *Available at SSRN 1977207*, 2011.
- [53] Jean-Philippe Bouchaud, Julien Kockelkoren, and Zoltan Eisler. The price impact of order book events: Market orders, limit orders and cancellations. *Quantitative Finance*, 9(3):283–297, 2009.
- [54] Jim Gatheral, Alexander Schied, and Aleksey Slynko. Transient linear price impact and fredholm integral equations. *Mathematical Finance*, 22(3):445–474, 2012.
- [55] Leonardo Berti, Bardh Prenkaj, and Paola Velardi. Trades: Generating realistic market simulations with diffusion models. *arXiv preprint arXiv:2502.07071*, 2025.
- [56] Robert Almgren, Chee Thum, Emmanuel Hauptmann, and Hong Li. Direct estimation of equity market impact. *Risk*, 18(7):58–62, 2005.
- [57] Anastasia Bugaenko. Empirical study of market impact conditional on order-flow imbalance. *arXiv preprint arXiv:2004.08290*, 2020.
- [58] Tianlang He, Keyan Lu, Chang Xu, and Jiang Bian. Multi-agent reinforcement learning in a realistic limit order book market simulation. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 464–472, 2020.
- [59] Andrew Todd, Peter Beling, William Scherer, and Steve Y. Yang. Agent-based financial markets: A review of the methodology and domain. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–5. IEEE, 2016.
- [60] Shen Gao, Yuntao Wen, Minghang Zhu, Jianing Wei, Yuhang Cheng, Qunzi Zhang, and Shuo Shang. Simulating financial market via large language model based agents. *arXiv preprint arXiv:2406.19966*, 2024.
- [61] Junjie Li, Yang Liu, Weiqing Liu, Shikai Fang, Lewen Wang, Chang Xu, and Jiang Bian. Mars: a financial market simulation engine powered by generative foundation model. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [62] Martin D. Gould, Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.
- [63] Avraam Tsantekidis, Nikolaos Passalis, Anastasios Tefas, Juho Kannianen, Moncef Gabbouj, and Alexandros Iosifidis. Forecasting stock prices from limit order book using convolutional neural networks. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, volume 1, pages 7–12. IEEE, 2017.
- [64] Antoine Ragel. *Reinforcement Learning for Systematic Market Making Strategies*. PhD thesis, Université Paris-Saclay, 2024. Available at <https://theses.hal.science/tel-04913317v1>.
- [65] Zhenglong Li, Vincent Tam, and Kwan L. Yeung. Developing a multi-agent and self-adaptive framework with deep reinforcement learning for dynamic portfolio risk management. *arXiv preprint arXiv:2402.00515*, 2024.

- 403 [66] Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian
404 Zhang, and Jun Wang. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*, 2019.
- 405 [67] Yufei Kuang, Miao Lu, Jie Wang, Qi Zhou, Bin Li, and Houqiang Li. Learning robust policy against
406 disturbance in transition dynamics via state-conservative policy optimization. In *Proceedings of the AAAI*
407 *Conference on Artificial Intelligence*, volume 36–7, pages 7247–7254, 2022.
- 408 [68] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally
409 robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.
- 410 [69] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning. *Journal of Cognitive Neuroscience*,
411 11(1):126–134, 1999.
- 412 [70] Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space:
413 Theory and algorithms. *Journal of Machine Learning Research*, 23(275):1–50, 2022.
- 414 [71] Huaqing Xiong, Tengyu Xu, Lin Zhao, Yingbin Liang, and Wei Zhang. Deterministic policy gradient:
415 Convergence analysis. In *Uncertainty in Artificial Intelligence*, pages 2159–2169. PMLR, 2022.
- 416 [72] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Networked multi-agent reinforcement learning in
417 continuous spaces. In *2018 IEEE conference on decision and control (CDC)*, pages 2771–2776. IEEE,
418 2018.
- 419 [73] Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical
420 foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and*
421 *Autonomous Systems*, 6(1):123–158, 2023.
- 422 [74] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of*
423 *Operations Research*, 38(1):153–183, 2013.
- 424 [75] Carlo Zarattini, Andrew Aziz, and Andrea Barbon. Beat the market: An effective intraday momentum
425 strategy for s&p500 etf (spy). *Available at SSRN 4824172*, 2024.
- 426 [76] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and
427 Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.
- 428 [77] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel
429 Goulao, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for
430 reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.
- 431 [78] Stephen Boyd and Lieven Vandenbergh. *Convex Optimization*. Cambridge University Press, 2004.
- 432 [79] Stephen Roman. *Advanced Linear Algebra*. Graduate Texts in Mathematics. Springer, third edition, 2008.
- 433 [80] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*,
434 1(3):127–239, 2014.
- 435 [81] Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and*
436 *Examples*. Springer, 2 edition, 2006.
- 437 [82] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
438 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Appendix

Table of Contents

443	A Backgrounds	14
444	A.1 Common Uncertainty Sets in the Literature	14
445	A.2 The Signed Permutation Group	15
446	B Worst-Case Uncertainty under ℓ_p-Ellipse Uncertainty Sets	16
447	B.1 Supporting Lemmas	16
448	B.2 Implicit Solution	23
449	B.3 Explicit Solution	24
450	C Experiment Details	27
451	C.1 Hardware and System Environment	27
452	C.2 Task Descriptions	27
453	C.3 Reinforcement Learning Framework	28
454	C.4 Parameter Details	29
455	C.5 Omitted Visualization	29
456	C.6 Other Implementation Details	29
457	D Limitations	30

A Backgrounds

A.1 Common Uncertainty Sets in the Literature

When evaluating the robust Bellman operator

$$\mathcal{T}_{\mathcal{U}}^{\pi} v(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left[r(s, a) + \gamma \min_{\mathbb{P} \in \mathcal{P}_{s,a}} \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, a) v(s') \right], \quad (7)$$

the uncertainty set $\mathcal{U} = \{\mathcal{P}_{s,a}\}_{(s,a) \in \mathcal{S} \times \mathcal{A}}$ plays a crucial role. Certain structures in \mathcal{U} enable efficient robust policy evaluation. Below, we summarize several widely adopted constructions.

f -divergence The f -divergence family [67, 68] generalizes statistical distances between distributions using a convex function $f : (0, \infty) \rightarrow \mathbb{R}$ with $f(1) = 0$. For distributions \mathbb{P} and \mathbb{P}_0 such that $\mathbb{P} \ll \mathbb{P}_0$, the f -divergence is defined as

$$D_f(\mathbb{P} \parallel \mathbb{P}_0) := \int_{\mathcal{S}} \mathbb{P}_0(s) f\left(\frac{\mathbb{P}(s)}{\mathbb{P}_0(s)}\right) ds.$$

Special cases include the Kullback-Leibler divergence ($f(t) = t \log t$), total variation distance ($f(t) = \frac{1}{2}|t - 1|$), and χ^2 -divergence ($f(t) = (t - 1)^2$). In robust RL, the f -divergence ball around the nominal transition kernel $\mathbb{P}_0(\cdot|s, a)$ yields the uncertainty set

$$\mathcal{P}_{s,a} = \left\{ \mathbb{P}(\cdot|s, a) \in \Delta^{|\mathcal{S}|} \mid D_f(\mathbb{P}(\cdot|s, a) \parallel \mathbb{P}_0(\cdot|s, a)) \leq \beta_{s,a} \right\}.$$

The inner minimization in Eq. (7) becomes a distributionally robust optimization problem over \mathbb{P}_0 . As the result, the robust policy evaluation under the KL-divergence often requires to repeatedly solve an additional convex program.

475 **R -contamination Model** The R -contamination model [37] assumes that the true transition kernel
 476 lies within a convex mixture of the nominal model \mathbb{P}_0 and an arbitrary distribution \mathbb{P}_1 :

$$\mathcal{P}_{s,a} = \left\{ (1-R)\mathbb{P}_0(\cdot|s,a) + R\mathbb{P}_1(\cdot|s,a) \mid \mathbb{P}_1(\cdot|s,a) \in \Delta^{|\mathcal{S}|} \right\},$$

477 where $R \in [0, 1]$ quantifies the contamination level. This model leads to closed-form solutions for the
 478 robust Bellman operator, with the worst-case distribution taking mass at the minimum of the value
 479 function v . As a result, this setup enables efficient and model-free learning algorithms, including
 480 robust variants of Q-learning, TD learning, and policy gradients. It is particularly well-suited for
 481 online learning, where \mathbb{P}_0 evolves with the observed data.

482 **ℓ_p -norm** These sets constrain the deviation from the nominal model $\mathbb{P}_0(\cdot|s,a)$ using the ℓ_p -norm:

$$\mathcal{P}_{s,a}^{(p)} = \left\{ \mathbb{P}(\cdot|s,a) \in \Delta^{|\mathcal{S}|} \mid \|\mathbb{P}(\cdot|s,a) - \mathbb{P}_0(\cdot|s,a)\|_p \leq \beta_{s,a} \right\}.$$

483 When $p = 1$, the constraint corresponds to total variation distance, while $p = \infty$ bounds the largest
 484 single-coordinate deviation. These sets are commonly used due to their interpretability and explicit
 485 analytical solution given in [45, 46]. However, their axis-aligned geometry can lead to overly
 486 conservative policies in high dimensions.

487 **Integral Probability Metric (IPM)** The IPM measures the discrepancy between distributions
 488 through expectations over a function class \mathcal{F} :

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{P}_0) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{P}_0}[f]|.$$

489 The corresponding uncertainty sets are:

$$\mathcal{P}_{s,a} = \left\{ \mathbb{P}(\cdot|s,a) \in \Delta^{|\mathcal{S}|} \mid d_{\mathcal{F}}(\mathbb{P}(\cdot|s,a), \mathbb{P}_0(\cdot|s,a)) \leq \beta_{s,a} \right\}.$$

490 The IPM-based uncertainty sets are particularly useful when the state space is extremely large or
 491 continuous, as explicitly solve the minimization problem in Eq. (7) does not requires to access values
 492 at all states [49].

493 **Wasserstein Distance** The Wasserstein distance [66], grounded in optimal transport theory, ac-
 494 counts for the geometry of the state space. Given a cost function $d : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}_+$ and $p \geq 1$, the
 495 p -Wasserstein distance between \mathbb{P} and \mathbb{P}_0 is

$$W_p(\mathbb{P}, \mathbb{P}_0) := \left(\inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{P}_0)} \int_{\mathcal{S} \times \mathcal{S}} d(s, s')^p d\gamma(s, s') \right)^{1/p},$$

496 where $\Gamma(\mathbb{P}, \mathbb{P}_0)$ denotes the set of joint distributions (couplings) with marginals \mathbb{P} and \mathbb{P}_0 . The
 497 uncertainty set is then

$$\mathcal{P}_{s,a} = \left\{ \mathbb{P}(\cdot|s,a) \in \Delta^{|\mathcal{S}|} \mid W_p(\mathbb{P}(\cdot|s,a), \mathbb{P}_0(\cdot|s,a)) \leq \beta_{s,a} \right\}.$$

498 Despite their strong theoretical properties, solving the inner minimization often requires dual formu-
 499 lations or approximation techniques.

500 **General Uncertainty Sets** There are also many techniques developed to handle the situation
 501 where the uncertainty set is general. For example, However, [50] proposes a bilevel approach that
 502 iteratively solves the worst-case transition kernel to approximate the robust value function. However,
 503 as demonstrated in [34, 50, 74], solving robust RL problems in the general case is NP-hard.

504 A.2 The Signed Permutation Group

505 The **signed permutation group** plays a central role in characterizing the symmetry structure of
 506 uncertainty sets in our robust RL framework. Informally, this group consists of all matrices in
 507 $\mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ satisfying the following conditions:

- 508 1. Each entry is either 0, 1, or -1 .

509 2. Each row and each column contains exactly one nonzero entry.

510 In other words, every element of this group is a matrix obtained by permuting the standard basis
 511 vectors of $\mathbb{R}^{|S|}$ and possibly flipping their signs. Each such matrix can be expressed as the product
 512 DP , where D is a diagonal matrix with diagonal entries in $\{\pm 1\}$, and P is a permutation matrix
 513 representing an element of the symmetric group $S_{|S|}$. This leads to the following formal definition:

514 **Definition A.1** (Signed Permutation Group). Let $S_{|S|}$ denote the permutation group over $|S|$ elements,
 515 and let $(\mathbb{Z}_2)^{|S|}$ be the direct sum of $|S|$ copies of the cyclic group of order 2. Then the *signed*
 516 *permutation group*, denoted by $\text{Signed}(S_{|S|})$, is the semidirect product:

$$\text{Signed}(S_{|S|}) \cong (\mathbb{Z}_2)^{|S|} \rtimes S_{|S|},$$

517 where the action of $S_{|S|}$ on $(\mathbb{Z}_2)^{|S|}$ is given by permuting the order.

518 In this work, we define the “symmetry” of sets as the invariance under the group action induced by
 519 the signed permutation group. Specifically, we say a set $B \subset \mathbb{R}^{|S|}$ is *symmetric under a group action*
 520 by G if $g \cdot B \subseteq B$ for all $g \in G$ (or $G \cdot B := \{g \cdot b\}_{g \in G, b \in B} \subseteq B$). This notion of symmetry leads to
 521 the following structural property of ℓ_p -norm balls:

522 **Proposition A.2.** Let $\mathcal{B} := \{B_p(\beta)\}_{p \geq 1, \beta \geq 0}$ be the family of ℓ_p -norm balls, where $B_p(\beta) := \{u \in$
 523 $\mathbb{R}^d \mid \|u\|_p \leq \beta\}$. If there exists a group G such that all elements in \mathcal{B} are symmetric under the group
 524 action by G , then G must be isomorphic to a subgroup of $\text{Signed}(S_d)$.

525 *Proof.* All elements in \mathcal{B} are symmetric under the group action by G ; that is, for every $p \geq 1$ and
 526 every $\beta \geq 0$,

$$g(B_p(\beta)) = B_p(\beta) \quad \forall g \in G.$$

527 Therefore, the act $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a norm-preserving bijection; by the Mazur–Ulam theorem, it must
 528 be affine. Then as it preserves 0, it must be linear.

529 In particular, taking $p = 1$ and $\beta = 1$, each $g \in G$ is an (invertible) linear isometry of the 1-norm unit
 530 ball

$$B_1(1) = \{u \in \mathbb{R}^d : \|u\|_1 \leq 1\},$$

531 whose extreme points are exactly

$$\{\pm e_1, \pm e_2, \dots, \pm e_d\}.$$

532 Because a linear automorphism of a polytope must permute its extreme points, for each i and each
 533 $g \in G$ there must exist a sign $\varepsilon_i \in \{\pm 1\}$ and an index $\sigma(i) \in \{1, \dots, d\}$ such that

$$g(e_i) = \varepsilon_i e_{\sigma(i)}.$$

534 Thus in the standard basis g is represented by a *signed permutation matrix*:

$$g = DP,$$

535 where $D = \text{diag}(\varepsilon_1, \dots, \varepsilon_d)$ and P is the permutation matrix corresponding to $\sigma \in S_d$. Hence every
 536 $g \in G$ lies in the signed permutation group $\text{Signed}(S_d)$. In other words $G \subseteq \text{Signed}(S_d)$, which
 537 equivalently shows G is isomorphic to a subgroup of $\text{Signed}(S_d)$. \square

538 This result provides a useful insight in designing the ℓ_p -ellipse set: the signed permutation group
 539 is the *largest group* under which all ℓ_p -norm balls are symmetric. Consequently, to construct a set
 540 with less symmetry than the standard ℓ_p -norm balls (as we aim to do with our ℓ_p -ellipse sets), it is
 541 necessary to enlarge the family \mathcal{B} to include non-ball shapes.

542 B Worst-Case Uncertainty under ℓ_p -Ellipse Uncertainty Sets

543 B.1 Supporting Lemmas

Definition B.1 (Minkowski sum). Given two sets $A, B \subseteq \mathbb{R}^d$, the Minkowski sum $+$: $2^{\mathbb{R}^d} \times 2^{\mathbb{R}^d} \rightarrow$
 $2^{\mathbb{R}^d}$ is defined as

$$A + B = \{a + b \mid a \in A, b \in B\}.$$

Definition B.2 (Fenchel conjugate [78]). Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function over \mathbb{R}^d . Its Fenchel conjugate is denoted as $g^* : \mathbb{R}^d \rightarrow \mathbb{R}$ and is defined as

$$g^*(y) := \sup_{x \in \mathbb{R}^d} \{y^\top x - g(x)\}.$$

544 We include the following famous Hölder's inequality without providing the proof.

545 **Lemma B.3** (Hölder's inequality). Let $p, q \in [1, +\infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. For every $f, g \in \mathbb{R}^d$

$$f^\top g \leq \|f\|_p \|g\|_q.$$

546 Moreover, equality holds if and only if

$$g = 0 \quad \text{or} \quad \frac{g}{\|g\|_q} \in J_p(f),$$

547 where $J_p(f)$ denotes any ℓ_p -unit vector that attains the maximum inner product with f :

$$J_p(f) = \arg \max_{\|u\|_p=1} f^\top u. \quad (8)$$

548 *Proof.* The proof can be found in [79]. □

549 **Lemma B.4.** For any $x, y \in \mathbb{R}^d$ and radii $r, s \geq 0$, the Minkowski sum of the two ℓ_p -norm balls
550 ($p \geq 1$)

$$B_p(x, r) = \{u \in \mathbb{R}^d \mid \|u - x\| \leq r\}, \quad B_p(y, s) = \{v \in \mathbb{R}^d \mid \|v - y\| \leq s\}$$

551 is again a ball, namely

$$B_p(x, r) + B_p(y, s) = B_p(x + y, r + s).$$

552 *Proof.* For convenience, we omit p at the subscript in this proof. It suffices to show two inclusions.

553 • $B(x, r) + B(y, s) \subseteq B(x + y, r + s)$.

554 Take any

$$z = u + v, \quad u \in B(x, r), v \in B(y, s).$$

555 Then by the triangle inequality,

$$\|z - (x + y)\| = \|(u - x) + (v - y)\| \leq \|u - x\| + \|v - y\| \leq r + s.$$

556 Hence $z \in B(x + y, r + s)$, proving the first inclusion.

557 • $B(x + y, r + s) \subseteq B(x, r) + B(y, s)$.

558 Let $z \in B(x + y, r + s)$, so $\|z - (x + y)\| \leq r + s$. Set

$$\alpha = \frac{r}{r + s}, \quad \beta = \frac{s}{r + s} \quad (\text{if } r + s = 0 \text{ then } r = s = 0 \text{ and the statement is trivial}).$$

559 Define $u = x + \alpha(z - (x + y))$ and $v = y + \beta(z - (x + y))$. Then

$$u + v = x + y + (\alpha + \beta)(z - (x + y)) = x + y + z - (x + y) = z,$$

560 and

$$\begin{aligned} \|u - x\| &= \alpha \|z - (x + y)\| \leq \frac{r}{r + s} (r + s) = r, \\ \|v - y\| &= \beta \|z - (x + y)\| \leq \frac{s}{r + s} (r + s) = s. \end{aligned}$$

561 Thus $u \in B(x, r)$ and $v \in B(y, s)$, so $z = u + v \in B(x, r) + B(y, s)$. This proves the
562 reverse inclusion.

563 Combining (1) and (2) gives the desired equality $B(x, r) + B(y, s) = B(x + y, r + s)$. □

Lemma B.5. Let $g(u) = \sum_{i=1}^N \|u - u_i\|_p$. Then the Fenchel conjugate of $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is

$$g^*(u) = \begin{cases} \inf_{\substack{\sum_i y_i = y \\ \|y_i\|_q \leq 1}} \sum_i y_i^\top u_i, & \|y\|_q \leq 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Proof. The key is that $g(u) = \sum_{i=1}^N \|u - u_i\|_p$ is a sum of N “shifted-norms,” and the Fenchel conjugate (Definition B.2) of a sum is the infimal convolution of the conjugates. We proceed in two steps.

- Let $f(x) = \|x\|_p$ be a single ℓ_p -norm mapping and q be the dual of p satisfying $\frac{1}{p} + \frac{1}{q} = 1$. By [78], it is standard that

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{y^\top x - \|x\|_p\} = \begin{cases} 0, & \|y\|_q \leq 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Then we consider its “shift” by u_i . Let $f_i(u) := \|u - u_i\|_p$. By the translation rule for Fenchel conjugates [80, 81],

$$f_i^*(y) = \sup_u \{y^\top u - \|u - u_i\|_p\} = \sup_t \underbrace{\{y^\top (t + u_i) - \|t\|_p\}}_{t = u - u_i} = y^\top u_i + f^*(y).$$

Hence, $f_i^*(y) = \begin{cases} y^\top u_i, & \|y\|_q \leq 1, \\ +\infty, & \text{otherwise.} \end{cases}$

- As the Fenchel conjugate of $\|u - u_i\|_p$ has been evaluated, the Fenchel conjugate of their sum is given by

$$g^*(y) = \begin{cases} \inf_{\substack{\sum_i y_i = y \\ \|y_i\|_q \leq 1}} \sum_i y_i^\top u_i, & \|y\|_q \leq 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Applying this infimal convolution requires each component $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is proper, convex, and lower semicontinuous, which is automatically satisfied by the ℓ_p -norm.

□

Lemma B.6. Let $p_o, \lambda_o, \tilde{\omega}_o \in \mathbb{R}^d$ and $\gamma_o \geq 0$ be given constants. Let $\frac{1}{p} + \frac{1}{q} = 1$. Then the optimization problem

$$\vartheta_o = \inf_{w: \|w\|_q \leq C} (p_o + \lambda_o)^\top w$$

has the unique minimizer given by

$$\begin{aligned} w_* &= -C \frac{J_q(p_o + \lambda_o)}{\|p_o + \lambda_o\|_p} \\ &= -C \frac{\text{sign}(p_o + \lambda_o) \odot |p_o + \lambda_o|^{p-1}}{\|p_o + \lambda_o\|_p^{p-1}}, \end{aligned}$$

where \odot represent the coordinate-wise product and $|\cdot|$ is the coordinate-wise absolute value. The optimal value is solved as

$$\vartheta_o = -C \|p_o + \lambda_o\|_p.$$

Proof. By Hölder’s inequality,

$$(p_o + \lambda_o)^\top w \geq -\|p_o + \lambda_o\|_p \|w\|_q.$$

To make the Hölder’s inequality achieve the equality, we choose $w = -t J_q(p_o + \lambda_o)$ for some t , where J_q is the q -unit vector defined in Eq. (8). Then by letting $\|w\|_q = C$, we obtain the final result. □

579 **Lemma B.7.** Suppose that $v \in \mathbb{R}^d$, $\{u_i\}_{i=1}^N \subset \mathbb{R}^d$, and $\mu \geq 0$, and the norm exponent $\frac{1}{p} + \frac{1}{q} = 1$
 580 (for $p \geq 1$). Let $w = v + \mu \mathbf{1}$ and $C := \|w\|_q$. Then the optimization problem

$$\begin{aligned} \min_{\{w_i\}_{i=1}^N \subset \mathbb{R}^d} \quad & \sum_{i=1}^N w_i^\top u_i \\ \text{s.t.} \quad & \sum_{i=1}^N w_i = w \\ & \|w_i\|_q \leq C \end{aligned}$$

is feasible and solves the minimizer

$$w_{i,*} = -\|v + \mu \mathbf{1}\|_q \frac{\text{sign}(u_i + \tilde{\lambda}) \odot |u_i + \tilde{\lambda}|^{p-1}}{\|u_i + \tilde{\lambda}\|_p^{p-1}},$$

581 for $i = 1, 2, \dots, N$, where $\tilde{\lambda}^* \in \mathbb{R}^d$ is given by

$$\tilde{\lambda}^* = \arg \min_{\tilde{\lambda}} \left\{ \tilde{\lambda}^\top w + \|w\|_q \sum_{i=1}^N \|u_i + \tilde{\lambda}\|_p \right\}. \quad (*)$$

582 *Proof.* We consider the constrained Lagrangian function

$$\tilde{L}(\{w_i\}, \tilde{\lambda}) = \sum_{i=1}^N \left[u_i^\top w_i + \tilde{\lambda}^\top w_i \right] - \tilde{\lambda}^\top w.$$

where $\|w_i\|_q \leq C := \|w\|_q$. Let the dual function $\vartheta(\tilde{\lambda}) := \inf_{\{w_i\}} \tilde{L}(\{w_i\}, \tilde{\lambda})$. Define

$$\vartheta_i(\tilde{\lambda}) = \inf_{w_i} \left(u_i + \tilde{\lambda} \right)^\top w_i.$$

583 By Lemma B.6, it solves

$$\begin{aligned} w_{i,*} &= -C \frac{J_q(u_i + \lambda)}{\|u_i + \lambda\|_p} \\ &= -C \frac{\text{sign}(u_i + \tilde{\lambda}) \odot |u_i + \tilde{\lambda}|^{p-1}}{\|u_i + \tilde{\lambda}\|_p^{p-1}} \end{aligned}$$

584 As the result,

$$\max_{\tilde{\lambda}} \vartheta(\tilde{\lambda}) = \min_{\tilde{\lambda}} \left\{ \tilde{\lambda}^\top w + C \sum_{i=1}^N \|u_i + \tilde{\lambda}\|_p \right\}$$

Recall that $w = v + \mu \mathbf{1} \in \mathbb{R}^d$ is a given vector. Denote

$$\tilde{\lambda}^* = \arg \min_{\tilde{\lambda}} \left\{ \tilde{\lambda}^\top w + C \sum_{i=1}^N \|u_i + \tilde{\lambda}\|_p \right\}.$$

585 Put it back to $w_{i,*}$, we obtain the final result. □

Lemma B.8. Suppose that $\{u_i\}_{i=1}^N \subset \mathbb{R}^d$, $\beta \geq 0$, and $v \in \mathbb{R}^d$. Let

$$\varphi(\lambda, \mu) = -\lambda\beta + \inf_u \left[(v + \mu \mathbf{1})^\top u + \lambda \sum_{i=1}^N \|u - u_i\|_p \right],$$

586 where $p \in [1, +\infty]$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\sup_{\lambda \geq 0} \varphi(\lambda, \mu) = -\|v + \mu \mathbf{1}\|_q \beta + \sum_{i=1}^N w_{i,*}^\top u_i,$$

where

$$w_{i,*} := -\|v + \mu \mathbf{1}\| \frac{\text{sign}(u_i + \tilde{\lambda}_*) \odot |u_i + \tilde{\lambda}_*|^{p-1}}{\|u_i + \tilde{\lambda}_*\|_p^{p-1}}, \quad \tilde{\lambda}_* = \arg \min_{\tilde{\lambda}} \left\{ \tilde{\lambda}^\top w + C \sum_{i=1}^N \|u_i + \tilde{\lambda}\|_p \right\}.$$

Consequently, the optimal (λ^*, μ^*) of achieving $\sup_{\lambda, \mu} \varphi(\lambda, \mu)$ is given by

$$\mu^* = \arg \max_{\mu} \left\{ -\|v + \mu \mathbf{1}\|_q \beta + \sum_{i=1}^N w_{i,*}^\top u_i \right\}, \quad \text{and} \quad \lambda^* = \|v + \mu^* \mathbf{1}\|_q.$$

587 *Proof.* We take the transformation $w := v + \mu \mathbf{1}$. Then

$$\varphi(\lambda, \mu) = -\lambda \beta + \inf_u \left[w^\top u + \lambda \sum_{i=1}^N \|u - u_i\|_p \right].$$

588 For any convex g , by the definition of Fenchel conjugate (Definition B.2), we have

$$\begin{aligned} \inf_u [w^\top u + \lambda g(u)] &= -\lambda \underbrace{g^*(-w/\lambda)}_{= \sup_y \{(-w/\lambda)^\top y - g(y)\}}. \\ &= \sup_y \{(-w/\lambda)^\top y - g(y)\} \end{aligned}$$

Here $g(u) = \sum_i \|u - u_i\|_p$. Its conjugate is given by Lemma B.5,

$$g^*(z) = \begin{cases} \inf_{\sum_i z_i = z, \|z_i\|_q \leq 1} \sum_i z_i^\top u_i, & \|z\|_q \leq 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

589 where $\frac{1}{p} + \frac{1}{q} = 1$. We put it back to $\inf_u [w^\top u + \lambda g(u)] = -\lambda g^*(-w/\lambda)$, which leads to

$$\inf_u [w^\top u + \lambda g(u)] = -\lambda \inf_{\substack{\sum_i z_i = -w/\lambda \\ \|z_i\|_q \leq 1}} \sum_i z_i^\top u_i,$$

590 where $\|w/\lambda\|_q \leq 1$. As the result, we take $w_i = -\lambda z_i$ to obtain

$$\inf_u [w^\top u + \lambda g(u)] = \begin{cases} \inf_{\sum_i w_i = w, \|w_i\|_q \leq \lambda} \sum_{i=1}^N w_i^\top u_i, & \|w\|_q \leq \lambda, \\ -\infty, & \text{else.} \end{cases}$$

591 Thus, the full dual becomes

$$\varphi(\lambda, \mu) = \begin{cases} -\lambda \beta + \inf_{\sum_i w_i = v + \mu \mathbf{1}, \|w_i\|_q \leq \lambda} \sum_{i=1}^N w_i^\top u_i, & \|w\|_q \leq \lambda, \\ -\infty, & \text{else.} \end{cases}$$

592 For a fixed μ , we need $\lambda \geq \|w\|_q$ to keep φ finite, and $\varphi(\lambda, \mu)$ is decreasing in λ . Hence the best
593 choice is

$$\lambda^* = \|w\|_q = \|v + \mu \mathbf{1}\|_q,$$

594 giving

$$\sup_{\lambda \geq 0} \varphi(\lambda, \mu) = -\|v + \mu \mathbf{1}\|_q \beta + \inf_{\sum_i w_i = v + \mu \mathbf{1}, \|w_i\|_q \leq \|v + \mu \mathbf{1}\|_q} \sum_{i=1}^N w_i^\top u_i.$$

595 It leads to another optimization problem $\inf_{\sum_i w_i = v + \mu \mathbf{1}, \|w_i\|_q \leq \|v + \mu \mathbf{1}\|_q} \sum_{i=1}^N w_i^\top u_i$. We construct
596 another Lagrangian function to solve it. Denote $w_{i,*}$ as the minimizer given by Lemma B.7. Then
597 we obtain

$$\max_{\lambda \geq 0} \varphi(\lambda, \mu) = -\|v + \mu \mathbf{1}\|_q \beta + \sum_{i=1}^N w_{i,*}^\top u_i.$$

It recovers the optimal dual variable z is given by

$$z_* = -\sum_{i=1}^N \frac{w_{i,*}}{\lambda^*} = \sum_{i=1}^N \frac{\text{sign}(u_i + \tilde{\lambda}_*) \odot |u_i + \tilde{\lambda}_*|^{p-1}}{\|u_i + \tilde{\lambda}_*\|_p^{p-1}},$$

where

$$\tilde{\lambda}_* = \arg \min_{\tilde{\lambda}} \left\{ \tilde{\lambda}^\top w + C \sum_{i=1}^N \|u_i + \tilde{\lambda}\|_p \right\}.$$

598 □

Lemma B.9. Suppose that $\{u_i\}_{i=1}^2 \subset \mathbb{R}$, $\lambda > 0$, $\beta \geq 0$, $w \in \mathbb{R}$ is non-zero, $q \in [1, +\infty]$, and $\frac{1}{p} + \frac{1}{q} = 1$. Define

$$\varphi = \min_{u \in \mathbb{R}} \left\{ uw + \lambda \sum_{i=1}^2 |u - u_i| \right\}.$$

Then when $|w| \leq 2\lambda$, the problem is solved as

$$\varphi = w \frac{u_1 + u_2}{2} + \left(\lambda - \frac{|w|}{2} \right) |u_1 - u_2|.$$

599 *Proof.* We start from the general case. Define $f(u) = uw + \lambda \sum_{i=1}^N |u - u_i|$. If $|w| \leq \lambda N$, then the
600 sub-gradient is given by

$$\partial f(u) \ni w + \lambda \sum_{i=1}^N \text{sign}(u - u_i)$$

where $\text{sign}(t) := \begin{cases} -1, & t < 0, \\ 0, & t = 0, \\ 1, & t > 0. \end{cases}$ Write $\{u_i\}_{i=1}^N \subset \mathbb{R}$ in the increasing order:

$$u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(N)}.$$

Define $k^* := \lceil \frac{N-w/\lambda}{2} \rceil$. Then

$$u^* = u_{(k^*)}$$

is the explicit solution. To prove it, we consider $u \in (u_{(k^*)}, u_{(k^*+1)})$; it is larger than exactly k^* u_i 's. That is,

$$\partial f(u) = w + \lambda(k^* - (N - k^*)) \geq 0.$$

601 Whenever $u < u_{(k^*)}$, the sign of sub-gradient becomes negative. As the result, $f(u)$ is decreasing
602 when $u < u_{(k^*)}$ then increasing when $u > u_{(k^*)}$. Now we set $N = 2$. The problem gives

$$u^* = \frac{u_1 + u_2 - \text{sign}(w)|u_1 - u_2|}{2}.$$

603 When $w \geq 0$, $\text{sign}(w) = +1$ and

$$u^* = \frac{u_1 + u_2 - |u_1 - u_2|}{2} = \min\{u_1, u_2\} = u_{(1)}.$$

604 When $w < 0$, $\text{sign}(w) = -1$ and

$$u^* = \frac{u_1 + u_2 + |u_1 - u_2|}{2} = \max\{u_1, u_2\} = u_{(2)}.$$

605 Therefore, this formula recovers the original general case solution. Putting it back to φ solves this
606 problem. \square

607 The following lemma connects the sum-of-distance description to the quadratic form of an ellipse.

Lemma B.10. Suppose that $\{u_i\}_{i=1}^2 \subset \mathbb{R}^d$, $\beta \geq 0$, and $\beta > \|u_1 - u_2\|_2$. The ellipse set is given by

$$\mathcal{E} := \{u \mid \|u - u_1\|_2 + \|u - u_2\|_2 \leq \beta\}.$$

Then there exists a matrix A such that

$$\mathcal{E} = \{u \mid (u - \bar{u})^\top A (u - \bar{u}) \leq 1\},$$

where $\bar{u} := \frac{u_1 + u_2}{2}$. More explicitly, the matrix A has the form

$$A = \frac{4}{\beta^2 - \|u_2 - u_1\|_2^2} I - \frac{4}{\beta^2(\beta^2 - \|u_2 - u_1\|_2^2)} (u_2 - u_1)(u_2 - u_1)^\top.$$

608 *Proof.* Define

$$\bar{u} = \frac{u_1 + u_2}{2}, \quad f = \frac{\|u_2 - u_1\|_2}{2}, \quad a = \frac{\beta}{2}, \quad b = \sqrt{a^2 - f^2}, \quad e = \frac{u_2 - u_1}{\|u_2 - u_1\|_2},$$

609 and decompose each $u \in \mathbb{R}^d$ by

$$x = u - \bar{u}.$$

610 Then $u_1 = \bar{u} - fe$, $u_2 = \bar{u} + fe$, and

$$\|u - u_1\|_2 + \|u - u_2\|_2 = \|x + fe\|_2 + \|x - fe\|_2.$$

611 Hence

$$\|u - u_1\|_2 + \|u - u_2\|_2 \leq \beta \iff \|x + fe\|_2 + \|x - fe\|_2 \leq 2a.$$

612 Now decompose x into

$$\alpha = e^\top x, \quad \xi^2 = \|x\|_2^2 - \alpha^2,$$

613 so that

$$\|x \pm fe\|_2 = \sqrt{(\alpha \pm f)^2 + \xi^2}.$$

614 The inequality $\sqrt{(\alpha + f)^2 + \xi^2} + \sqrt{(\alpha - f)^2 + \xi^2} \leq 2a$ is equivalent, after two squarings, to

$$\frac{\alpha^2}{a^2} + \frac{\xi^2}{b^2} \leq 1, \quad \text{where } b^2 = a^2 - f^2.$$

615 Finally, observe that

$$\alpha^2 = x^\top (ee^\top)x, \quad \xi^2 = x^\top (I - ee^\top)x,$$

616 so

$$\frac{\alpha^2}{a^2} + \frac{\xi^2}{b^2} = x^\top \left(\frac{1}{a^2} ee^\top + \frac{1}{b^2} (I - ee^\top) \right) x.$$

Setting $A = \frac{1}{a^2} ee^\top + \frac{1}{b^2} (I - ee^\top)$ implies $(u - \bar{u})^\top A (u - \bar{u}) \leq 1$. It exactly characterizes $\{u \mid \|u - u_1\|_2 + \|u - u_2\|_2 \leq \beta\}$. Simplifying the form of A leads to

$$A = \frac{4}{\beta^2 - \|u_2 - u_1\|_2^2} I - \frac{4}{\beta^2(\beta^2 - \|u_2 - u_1\|_2^2)} (u_2 - u_1)(u_2 - u_1)^\top.$$

617 This completes the proof. \square

618 **Lemma B.11.** Let $v \in \mathbb{R}^d$, let $A \in \mathbb{R}^{d \times d}$ be symmetric positive definite, and let $\bar{p} \in \mathbb{R}^d$. Define

$$F(\mu) = -\sqrt{(v + \mu \mathbf{1})^\top A^{-1} (v + \mu \mathbf{1})} + (v + \mu \mathbf{1})^\top \bar{p}, \quad \mu \in \mathbb{R}.$$

619 Further set

$$\alpha = \mathbf{1}^\top A^{-1} \mathbf{1}, \quad \beta = v^\top A^{-1} \mathbf{1}, \quad \gamma = v^\top A^{-1} v, \quad \delta = \mathbf{1}^\top \bar{p}.$$

620 If $\delta^2 < \alpha$, then F attains a unique maximizer

$$\begin{aligned} \mu^* &= -\frac{\beta}{\alpha} + \frac{\delta}{\alpha \sqrt{\alpha - \delta^2}} \sqrt{\alpha \gamma - \beta^2} = \mu^* \\ &= -\frac{v^\top A^{-1} \mathbf{1}}{\mathbf{1}^\top A^{-1} \mathbf{1}} + \frac{\mathbf{1}^\top \bar{p}}{\mathbf{1}^\top A^{-1} \mathbf{1} \sqrt{\mathbf{1}^\top A^{-1} \mathbf{1} - (\mathbf{1}^\top \bar{p})^2}} \sqrt{(\mathbf{1}^\top A^{-1} \mathbf{1})(v^\top A^{-1} v) - (v^\top A^{-1} \mathbf{1})^2}. \end{aligned}$$

621 *Proof.* We begin by computing the derivative of

$$F(\mu) = -\sqrt{(v + \mu \mathbf{1})^\top A^{-1} (v + \mu \mathbf{1})} + (v + \mu \mathbf{1})^\top \bar{p}.$$

622 Using the notation $\alpha = \mathbf{1}^\top A^{-1} \mathbf{1}$, $\beta = v^\top A^{-1} \mathbf{1}$, $\gamma = v^\top A^{-1} v$, and $\delta = \mathbf{1}^\top \bar{p}$, one checks

$$(v + \mu \mathbf{1})^\top A^{-1} (v + \mu \mathbf{1}) = \alpha \mu^2 + 2\beta \mu + \gamma,$$

623 so

$$F'(\mu) = -\frac{\alpha \mu + \beta}{\sqrt{\alpha \mu^2 + 2\beta \mu + \gamma}} + \delta.$$

624 Setting $F'(\mu) = 0$ gives the stationarity condition

$$\alpha\mu + \beta = \delta\sqrt{\alpha\mu^2 + 2\beta\mu + \gamma},$$

625 which upon squaring yields the quadratic equation

$$(\alpha^2 - \alpha\delta^2)\mu^2 + 2\beta(\alpha - \delta^2)\mu + (\beta^2 - \delta^2\gamma) = 0.$$

626 Let $\delta^2 < \alpha$. Here $\alpha - \delta^2 > 0$, so dividing by $\alpha - \delta^2$ gives

$$\alpha\mu^2 + 2\beta\mu + \frac{\beta^2 - \delta^2\gamma}{\alpha - \delta^2} = 0,$$

627 whose two roots are

$$\mu = -\frac{\beta}{\alpha} \pm \frac{\delta}{\alpha\sqrt{\alpha - \delta^2}}\sqrt{\alpha\gamma - \beta^2}.$$

628 One verifies by inspecting $\lim_{\mu \rightarrow \pm\infty} F'(\mu) = \delta \mp \sqrt{\alpha}$ that exactly the “+” choice yields a change of
629 sign from + to −, and hence is the unique global maximizer. \square

630 B.2 Implicit Solution

631 In this subsection, we recap and prove the full version of [Theorem 3.3](#).

632 **Theorem B.12.** *Let $d := |S|$ be the cardinal of state space and $\frac{1}{p} + \frac{1}{q} = 1$ for $p, q \in [1, +\infty]$.
633 Suppose that $\{u_i\}_{i=1}^N \subset \mathbb{R}^d$ for each $(s, a) \in \mathcal{S} \times \mathcal{A}$, the uncertainty size $\beta \geq 0$, and $v \in \mathbb{R}^d$. The
634 solution of*

$$\begin{aligned} \min_{u \in \mathbb{R}^d} \quad & v^\top u \\ \text{s.t.} \quad & \sum_{i=1}^N \|u - u_i\|_p \leq \beta, \\ & \mathbf{1}^\top u = 0. \end{aligned} \tag{9}$$

is given by

$$u^* = \arg \min_u [(v + \mu^* \mathbf{1})^\top u + \lambda^* \sum_{i=1}^N \|u - u_i\|_p],$$

635 where

$$\begin{cases} \mu^* = \arg \max_{\mu} \left\{ -\|v + \mu \mathbf{1}\|_q \beta + \sum_{i=1}^N w_{i,*}^\top u_i \right\}, \\ \lambda^* = \|v + \mu^* \mathbf{1}\|_q, \\ \tilde{\lambda}_*(\mu) = \arg \min_{\tilde{\lambda}} \left\{ \tilde{\lambda}^\top (v + \mu \mathbf{1}) + \|v + \mu \mathbf{1}\|_q \sum_{i=1}^N \|u_i + \tilde{\lambda}\|_p \right\}, \\ w_{i,*}(\mu) = -\|v + \mu \mathbf{1}\|_q \frac{\text{sign}(u_i + \tilde{\lambda}_*) \odot |u_i + \tilde{\lambda}_*|^{p-1}}{\|u_i + \tilde{\lambda}_*\|_p^{p-1}}, \end{cases} \tag{10}$$

636 *Remark* (The procedure of solving u^*). To obtain u^* , it suffices to solve μ^* and λ^* as v and all u_i 's
637 have been given. The first step is to solve

$$\begin{cases} \tilde{\lambda}_*(\mu) = \arg \min_{\tilde{\lambda}} \left\{ \tilde{\lambda}^\top (v + \mu \mathbf{1}) + \|v + \mu \mathbf{1}\|_q \sum_{i=1}^N \|u_i + \tilde{\lambda}\|_p \right\}, \\ w_{i,*}(\mu) = -\|v + \mu \mathbf{1}\|_q \frac{\text{sign}(u_i + \tilde{\lambda}_*) \odot |u_i + \tilde{\lambda}_*|^{p-1}}{\|u_i + \tilde{\lambda}_*\|_p^{p-1}}, \end{cases}$$

638 for $i = 1, 2, \dots, N$. Both variables depend on the variable μ and other values are known. The next
639 step is to solve

$$\begin{cases} \mu^* = \arg \max_{\mu} \left\{ -\|v + \mu \mathbf{1}\|_q \beta + \sum_{i=1}^N w_{i,*}^\top u_i \right\}, \\ \lambda^* = \|v + \mu^* \mathbf{1}\|_q. \end{cases}$$

640 Once (μ^*, λ^*) is solved, the primal variable μ^* is obtained immediately.

641 *Proof.* Our goal is to solve the following constrained optimization problem:

$$\begin{aligned} \min_{u \in \mathbb{R}^d} \quad & v^\top u \\ \text{s.t.} \quad & \sum_{i=1}^N \|u - u_i\|_p \leq \beta, \\ & \mathbf{1}^\top u = 0. \end{aligned}$$

642 As it is a constrained optimization problem, the standard approach of solving this problem is using
643 Lagrangian multipliers. we introduce the Lagrangian multipliers $\lambda \geq 0$ for the inequality and $\mu \in \mathbb{R}$
644 for the equality. The Lagrangian function is

$$L(u, \lambda, \mu) = v^\top u + \lambda \left(\sum_{i=1}^N \|u - u_i\|_p - \beta \right) + \mu (\mathbf{1}^\top u),$$

with $\lambda \geq 0, \mu \in \mathbb{R}$. Because this optimization problem is a standard convex optimization problem with satisfying the Slater's condition, we have the strong duality

$$\underbrace{\inf_u \sup_{\lambda, \mu} L(u, \lambda, \mu)}_{\text{original opt. prob.}} = \sup_{\lambda, \mu} \inf_u L(u, \lambda, \mu).$$

645 Then we turn the original optimization problem into solving its dual-form problem. We let the dual
646 function be $\varphi(\lambda, \mu) := \inf_u L(u, \lambda, \mu)$. Then

$$\varphi(\lambda, \mu) = -\lambda\beta + \inf_u \left[(v + \mu\mathbf{1})^\top u + \lambda \sum_{i=1}^N \|u - u_i\|_p \right].$$

647 The above formulation plays the crucial role in our proof. For the implicit solution, we will follow
648 [Lemma B.8](#) to complete the remaining calculation. For the explicit solution, this dual form can be
649 significantly simplified in some cases.

650 By [Lemma B.8](#), the dual form can be simplified as

$$\max_{\lambda, \mu} \varphi(\lambda, \mu) = -\|v + \mu^* \mathbf{1}\|_q \beta + \sum_{i=1}^N w_{i,*}^\top u_i,$$

where

$$w_{i,*} := -\|v + \mu \mathbf{1}\| \frac{\text{sign}(u_i + \tilde{\lambda}_*) \odot |u_i + \tilde{\lambda}_*|^{p-1}}{\|u_i + \tilde{\lambda}_*\|_p^{p-1}}, \quad \tilde{\lambda}_* = \arg \min_{\tilde{\lambda}} \left\{ \tilde{\lambda}^\top w + C \sum_{i=1}^N \|u_i + \tilde{\lambda}\|_p \right\}.$$

and

$$\mu^* = \arg \max_{\mu} \left\{ -\|v + \mu \mathbf{1}\|_q \beta + \sum_{i=1}^N w_{i,*}^\top u_i \right\}, \quad \lambda^* = \|v + \mu^* \mathbf{1}\|_q.$$

The optimal primary variable is given as

$$u^* = \arg \min_u \left[(v + \mu^* \mathbf{1})^\top u + \lambda^* \sum_{i=1}^N \|u - u_i\|_p \right].$$

651

□

652 B.3 Explicit Solution

653 In this subsection, we recap and prove the full version of [Theorem 3.4](#).

654 **Theorem B.13.** Let $d := |\mathcal{S}|$ be the cardinal of state space. Suppose that $\{u_i\}_{i=1}^N \subset \mathbb{R}^d$ for each
655 $(s, a) \in \mathcal{S} \times \mathcal{A}$, the uncertainty size $\beta \geq 0$, and $v \in \mathbb{R}^d$. The optimization problem defined by [Eq. \(4\)](#)
656 is explicitly solved in the following cases:

(a) Let $N = 1$. then

$$\mu^* = u_1 + \beta \frac{\text{sign}(v + \mu^* \mathbf{1}) \odot |v + \mu^* \mathbf{1}|^{q-1}}{\|v + \mu^* \mathbf{1}\|_q^{q-1}}.$$

(b) Let $p = 1$ and $N = 2$. Define $\bar{u} = \frac{u_1 + u_2}{2}$,

$$\mu^* = -\frac{v_{\max} + v_{\min}}{2} \quad \text{and} \quad \lambda^* = \frac{1}{2} \|v + \mathbf{1} \mu^*\|_\infty = \frac{v_{\max} - v_{\min}}{4}.$$

Suppose that $\beta > \|u_2 - u_1\|_1$. Then the explicit solution to the optimization problem Eq. (4) is given as

$$u^* = \bar{u} - \frac{\beta - \|u_1 - u_2\|_1}{2} [\text{sign}(v + \mu^* \mathbf{1}) \odot \mathbb{I}_{\{|v + \mu^* \mathbf{1}| = 2\lambda^*\}}].$$

657 (c) Let $p = 2$ and $N = 2$. Define $\bar{u} = \frac{u_1 + u_2}{2}$,

$$\Omega := \Omega(u_1, u_2, \beta) = \left[I - \frac{1}{\beta^2} (u_2 - u_1)(u_2 - u_1)^\top \right], \quad (11)$$

$$\underline{\lambda}^* = \sqrt{(v + \mu^* \mathbf{1})^\top \Omega^{-1} (v + \mu^* \mathbf{1})}, \text{ and } \mu^* = -\frac{v^\top \Omega^{-1} \mathbf{1}}{\mathbf{1}^\top \Omega^{-1} \mathbf{1}}. \quad (12)$$

Suppose that $\beta > \|u_2 - u_1\|_2$. Then the explicit solution to the optimization problem Eq. (4) is given as

$$u^* = \bar{u} - \frac{\sqrt{\beta^2 - \|u_2 - u_1\|_2^2}}{2} \frac{1}{\underline{\lambda}^*} \Omega^{-1} (v + \mu^* \mathbf{1}).$$

658 *Proof.* We follow the standard routine used in proving [Theorem B.12](#).

659 (a) When $N = 1$ the objective optimization problem is given by

$$\begin{aligned} \min_{u \in \mathbb{R}^d} \quad & v^\top u \\ \text{s.t.} \quad & \|u - u_1\|_p \leq 1, \\ & \mathbf{1}^\top u = 0. \end{aligned}$$

660 We take the transformation $u' = u - u_1$. It still satisfies $\mathbf{1}^\top u' = 0$. Then the problem become

$$\begin{aligned} \min_{u' \in \mathbb{R}^d} \quad & v^\top u' \\ \text{s.t.} \quad & \|u'\|_p \leq 1, \\ & \mathbf{1}^\top u' = 0. \end{aligned}$$

This transformation has turned this problem into the standard ℓ_p -norm structure, which has been explicitly solved in [\[45, 46\]](#). The optimal u' is given for arbitrary $p \geq 1$ as

$$u'_* = \beta \frac{\text{sign}(v + \mu^* \mathbf{1}) |v + \mu^* \mathbf{1}|^{q-1}}{\|v + \mu^* \mathbf{1}\|_q^{q-1}},$$

where $\mu^* = \arg \min_{\mu \in \mathbb{R}} \|v + \mu \mathbf{1}\|_q$. As the result,

$$u^* = u'_* + u_1 = u_1 + \beta \frac{\text{sign}(v + \mu^* \mathbf{1}) |v + \mu^* \mathbf{1}|^{q-1}}{\|v + \mu^* \mathbf{1}\|_q^{q-1}}.$$

(b) When $p = 1$, we define the order

$$u_{(1),j} \leq u_{(2),j} \leq \dots \leq u_{(N),j},$$

661

We follow the same routine as [Theorem B.12](#) and derive the dual function $\varphi(\lambda, \mu)$:

$$\begin{aligned}\varphi(\lambda, \mu) &= -\lambda\beta + \inf_u \left[(v + \mu \mathbf{1})^\top u + \lambda \sum_{i=1}^N \|u - u_i\|_1 \right] \\ &\stackrel{(i)}{=} -\lambda\beta + \inf_u \left[\sum_{j=1}^d (v_j + \mu) u_j + \lambda \sum_{i=1}^N \sum_{j=1}^d |u_j - u_{ij}| \right] \\ &= -\lambda\beta + \sum_{j=1}^d \inf_{u_j} \left[(v_j + \mu) u_j + \lambda \sum_{i=1}^N |u_j - u_{ij}| \right].\end{aligned}$$

662

where (i) decomposes the ℓ_1 -norm by coordinates. When $N = 2$, by [Lemma B.9](#), the dual function is solved as

663

$$\begin{aligned}\varphi(\lambda, \mu) &= -\lambda\beta + \sum_{j=1}^d \inf_{u_j} \left[(v_j + \mu) u_j + \lambda \sum_{i=1}^N |u_j - u_{ij}| \right] \\ &= -\lambda\beta + \sum_{j=1}^d \left[(v_j + \mu) \frac{u_{1j} + u_{2j}}{2} + \left(\lambda - \frac{|v_j + \mu|}{2} \right) |u_{1j} - u_{2j}| \right] \\ &= -\lambda\beta + (v + \mathbf{1}\mu)^\top \bar{u} + [2\lambda\mathbf{1} - (v + \mathbf{1}\mu)]^\top \left| \frac{u_1 - u_2}{2} \right|.\end{aligned}$$

664

As the smaller λ is, the larger $\varphi(\lambda, \mu)$ is. It achieves the supremum at $\lambda^* = \max_j \frac{v_j + \mu}{2} = \frac{1}{2} \|v + \mathbf{1}\mu\|_\infty$. Then we solve

665

$$\begin{aligned}\sup_{\lambda} \varphi(\lambda, \mu) \\ &= -\frac{1}{2} \|v + \mathbf{1}\mu\|_\infty \beta + (v + \mathbf{1}\mu)^\top \bar{u} + \|v + \mathbf{1}\mu\|_\infty \left\| \frac{u_1 - u_2}{2} \right\|_1 - (v + \mathbf{1}\mu)^\top \left| \frac{u_1 - u_2}{2} \right|.\end{aligned}$$

Then we have

$$\sup_{\mu} \sup_{\lambda} \varphi(\mu, \lambda) = -\frac{\beta - \|u_2 - u_1\|_1}{2} \inf_{\mu} \left[\|v + \mu \mathbf{1}\|_\infty + \frac{(v + \mu \mathbf{1})^\top |u_1 - u_2|}{\beta - \|u_2 - u_1\|_1} \right] + v^\top \bar{u}.$$

As $\beta - \|u_2 - u_1\|_1 > 0$, it solves

$$\mu^* = -\frac{v_{\max} + v_{\min}}{2} \quad \text{and} \quad \lambda^* = \frac{1}{2} \|v + \mathbf{1}\mu^*\|_\infty = \frac{v_{\max} - v_{\min}}{4}.$$

Now we consider the KKT condition of the original Lagrangian function. We solve

$$\partial_{u_j} L(u, \lambda, \mu) = \lambda \text{sign}(u_j - u_{1j}) + \lambda \text{sign}(u_j - u_{2j}) + v_j + \mu \ni 0.$$

For inactive coordinate, the optimal value is attained for arbitrary $u_j \in (u_{(1)j}, u_{(2)j})$; in these cases, we simply take $u_j = \frac{u_{1j} + u_{2j}}{2}$. There are exactly two active coordinates matching the corner-case condition $|v_j + \mu^*| = 2\lambda^*$: $j^* = \arg \max v_j$ and $j_* = \arg \min v_j$. In these cases we take u_j as $\frac{u_{1j} + u_{2j}}{2}$ subtracting a drift. It finally solves

$$u^* = \bar{u} - \frac{\beta - \|u_1 - u_2\|_1}{2} \left[\text{sign}\left(v - \frac{v_{\max} + v_{\min}}{2}\right) \odot \mathbb{I}_{\{|v - \frac{v_{\max} + v_{\min}}{2}| = \frac{v_{\max} - v_{\min}}{2}\}} \right].$$

666

The magnitude coefficient is used to ensure that u^* belongs to $\|u^*\|_1 \leq \beta$.

(c) By [Lemma B.10](#), there exists a semi-positive definite matrix A such that

$$\{u \mid \|u - u_1\|_2 + \|u - u_2\|_2 \leq \beta\} = \{u \mid (u - \bar{u})^\top A(u - \bar{u}) \leq 1\},$$

667

where $\bar{u} := \frac{u_1 + u_2}{2}$. As the result, the objective optimization problem can be simplified as

$$\begin{aligned}\min_{u \in \mathbb{R}^d} \quad & v^\top u \\ \text{s.t.} \quad & (u - \bar{u})^\top A(u - \bar{u}) \leq 1, \\ & \mathbf{1}^\top u = 0.\end{aligned}$$

668 We follow the same routine as [Theorem B.12](#) and derive the dual function $\varphi(\lambda, \mu)$:

$$\begin{aligned}\varphi(\lambda, \mu) &= -\lambda + \inf_u \left[(v + \mu \mathbf{1})^\top u + \lambda(u - \bar{u})^\top A(u - \bar{u}) \right] \\ &= -\lambda + (v + \mu \mathbf{1})^\top \bar{u} - \frac{1}{4\lambda} (v + \mu \mathbf{1})^\top A^{-1} (v + \mu \mathbf{1}),\end{aligned}$$

669 where the infimum is attained at $u^* = \bar{u} - \frac{1}{2\lambda} A^{-1} (v + \mu \mathbf{1})$. Then

$$\begin{aligned}\sup_{\lambda, \mu} \varphi(\lambda, \mu) &= \sup_{\lambda, \mu} \left\{ -\lambda + (v + \mu \mathbf{1})^\top \bar{u} - \frac{1}{4\lambda} (v + \mu \mathbf{1})^\top A^{-1} (v + \mu \mathbf{1}) \right\} \\ &\stackrel{(i)}{=} \sup_{\mu} \left\{ -\sqrt{(v + \mu \mathbf{1})^\top A^{-1} (v + \mu \mathbf{1})} + (v + \mu \mathbf{1})^\top \bar{u} \right\}\end{aligned}$$

where (i) applies the optimal choice $\lambda^*(\mu) = \frac{1}{2} \sqrt{(v + \mu \mathbf{1})^\top A^{-1} (v + \mu \mathbf{1})}$ and μ^* is given by [Lemma B.11](#) to solve this maximization problem. As the result,

$$u^* = \bar{u} - \frac{1}{2\lambda^*} A^{-1} (v + \mu^* \mathbf{1})$$

670 where $\lambda^* = \frac{1}{2} \sqrt{(v + \mu^* \mathbf{1})^\top A^{-1} (v + \mu^* \mathbf{1})}$ and

$$\mu^* = -\frac{v^\top A^{-1} \mathbf{1}}{\mathbf{1}^\top A^{-1} \mathbf{1}} + \frac{\mathbf{1}^\top \bar{p}}{\mathbf{1}^\top A^{-1} \mathbf{1} \sqrt{\mathbf{1}^\top A^{-1} \mathbf{1} - (\mathbf{1}^\top \bar{p})^2}} \sqrt{(\mathbf{1}^\top A^{-1} \mathbf{1})(v^\top A^{-1} v) - (v^\top A^{-1} \mathbf{1})^2}.$$

Here, the matrix A is determined by $\{u_1, u_2\}_{i=1}^2$ and the vector v in [Lemma B.10](#). We recall that $\mathbf{1}^\top \bar{u} = 0$. Therefore, μ^* is further simplified as

$$\mu^* = -\frac{v^\top A^{-1} \mathbf{1}}{\mathbf{1}^\top A^{-1} \mathbf{1}}.$$

671 It completes the proof in the part (c).

672

□

673 C Experiment Details

674 In this section, we include the omitted details of [Section 4](#). All source codes and hyper-parameter
675 settings are available in the supplementary materials.

676 C.1 Hardware and System Environment

677 We conducted our experiments on a laptop running Windows 11 Home. The device is equipped with
678 32GB of RAM, 1TB SSD, an AMD Ryzen 9 7940HS processor and a NVIDIA GeForce RTX 4070
679 Laptop GPU. Our implementation was tested using Python version 3.10.10. Additional dependencies
680 are listed in the supplementary `requirements.txt` file.

681 The actual hardware requirement for running our implementation is significantly lower than the
682 specification listed above. Most experiments can be reproduced on consumer-grade machines with
683 8–16GB RAM and any CUDA-compatible NVIDIA GPU.

684 C.2 Task Descriptions

685 **Multi-Asset Portfolio Rebalancing** This task captures the setting where an institutional investor
686 reallocates capital across multiple assets over discrete time intervals to maintain a desired risk-return
687 profile. It reflects strategic portfolio management, such as end-of-day rebalancing or tactical asset
688 allocation. The task emphasizes robustness to market impact under varying capital scales, which
689 is critical for large-volume institutional strategies. In our experiment, we select five representative
690 ETFs, SPY, TLT, GLD, EFA, and VNQ, and use historical data from May 9, 2021 to May 9, 2022 for
691 training. Evaluation is conducted on the out-of-sample period from June 9 to December 9, 2022.

Table 2: Example of ask-side execution from an AMZN order book snapshot (21 June 2012). This table shows how our simulator determines the execution price for a market buy order. Instead of using the last trade price, the simulator consumes liquidity by filling shares at each ask level in order, starting from the best price. It calculates the VWAP based on the prices and quantities filled, and uses this VWAP as the final executed price.

Level	Price (USD)	Depth (sh)	Exec. 100 sh	Exec. 1 000 sh
Level 1	223.95	100	100	100
Level 2	223.99	100	0	100
Level 3	224.00	220	0	220
Level 4	224.25	100	0	100
Level 5	224.40	547	0	480
VWAP paid	—	—	223.95	224.21

Single-Asset Intra-Day Trading This task models the decision-making process of an agent that repeatedly buys or sells a single asset over short time intervals, such as minutes or seconds. It reflects the setting of mid-frequency or algorithmic trading, where even small trades can shift market prices. The goal is to maximize the risk-adjusted cumulative return while accounting for execution slippage, making it a standard benchmark for evaluating RL-based trading strategies. We follow the exactly same training and evaluation period as the multi-asset portfolio rebalancing task. In our experiments, we use minute-level historical data for the assets including META, MSFT, and SPY. The observation includes stock price, trading volume, implied volatility, and current portfolio return. During evaluate, we reconstruct the LOB dynamics using the tick-level trade data; the execution price is then simulated as illustrated in Table 2. Same as the multi-asset portfolio rebalancing experiment, the training data covers the period from May 9th, 2021 to May 9th, 2022, and the evaluation period is from June 9 to December 9, 2022. All data are accessed via the Polygon.io Stock Market API.

C.3 Reinforcement Learning Framework

We implemented a Gym-like RL trading environment [76, 77] using the historical data including the stock price, trading volume, the implied volatility, and the current portfolio return. Instead of using a single timestep data, we set `lookback_period` = 30 to account for 30 previous timestep; as the result, the observed state contains `lookback_period` \times 4 = 120 dimension. Given the state described above, we assign the transition probability from the current state s_t to the next state s_{t+1} as the probability 1, independent with the action. The action is implemented as a single float value that determines the position size relative to the maximum possible position size based on the available capital. The reward is a Sharpe-like ratio that consists of two components:

$$\text{reward} = \frac{\text{current_return}}{\text{current_volatility} + \varepsilon} - \delta \times \frac{|\text{current_position} - \text{previous_position}|}{\text{max_shares}}.$$

where $\varepsilon > 0$ is a small constant to ensure numerical stability, and δ is a tunable coefficient controlling the strength of the transaction cost penalty. The first term encourages higher risk-adjusted return, while the second term discourages frequent or drastic position shifts, which aligns with practical trading considerations under market impact or slippage. We also include 0.1% transaction cost throughout the experiment.

Uncertainty Restriction to Execution Prices Instead of perturbing the whole transition probability, we restrict the perturbation on the execution price. The formulation of restriction is given as follows: we denote the state s_t as two components (p_t, f_t) , and we hope perturb the transition on p_t only. We re-write the transition probability using the conditional probability

$$\begin{aligned} \mathbb{P}(s_{t+1} \mid s_t, a_t) &= \mathbb{P}(p_{t+1}, f_{t+1} \mid s_t, a_t) \\ &= \mathbb{P}(p_{t+1} \mid s_t, a_t) \mathbb{P}(f_{t+1} \mid p_{t+1}, s_t, a_t). \end{aligned}$$

Then we set the uncertainty u as the perturbation on the kernel $\mathbb{P}(p_{t+1} \mid s_t, a_t)$ instead of the whole transition kernel $\mathbb{P}(s_{t+1} \mid s_t, a_t)$.

Momentum Strategy The momentum strategy is a classical intra-day minute-level algorithmic trading method that exploits intraday time-series momentum to generate trading signals. We follow the classical implementation presented by [75]. The core idea of the momentum strategy is based on the empirical observation where assets that have performed well in the recent past are more likely to continue performing well in the near future, while poorly performing assets are likely to continue underperforming.

C.4 Parameter Details

Detailed descriptions of all parameters are provided in the separate supplementary materials.

C.5 Omitted Visualization

In Figure 3, we have visualized the return curve of the META stock. In this subsection, we include the visualization of the other two assets.

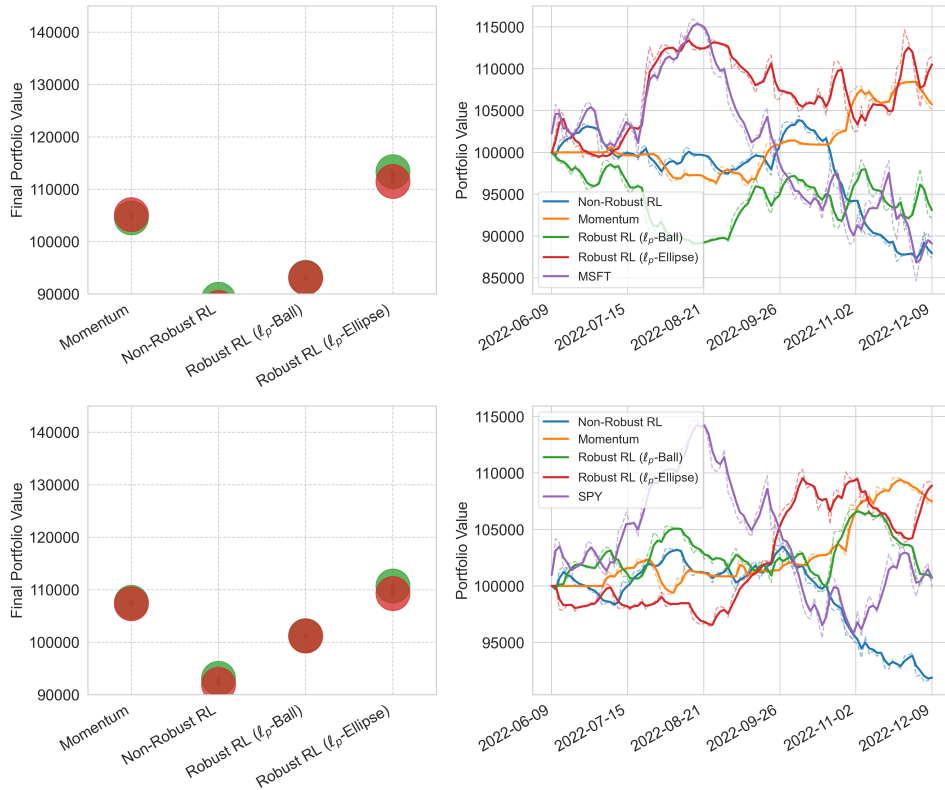


Figure 5: Performance comparison of trading strategies on the MSFT stock and SPY ETF.

C.6 Other Implementation Details

In this subsection, we discuss the practical implementation considerations and provide the details for reproducing our experiments. The source codes are also provided in the supplementary material.

RL Algorithm and Value Function Approximations We use the standard Robust Actor-Critic algorithm [49] to train the RL agent with replacing its IPM uncertainty set or doubly-sampling uncertainty set with the ℓ_p -ellipse uncertainty set and the ℓ_p -norm uncertainty set. We adopt an Actor-Critic architecture with a shared feature extractor and separate heads for the actor and critic. The shared backbone consists of three fully connected layers with ReLU activations, mapping the (flattened) input state to a high-level representation. The actor head outputs the vector with the same dimension as the action through a two-layer MLP followed by a sigmoid/tanh activation, depending

737 on the underlying task. The critic head predicts the state value using a similar two-layer MLP. All
 738 linear layers are orthogonally initialized to promote stable training. When updating the Actor network,
 739 we apply the classical PPO algorithm [82].

740 **Discretization of Execution Prices** To apply the robust RL framework, we apply the discretization
 741 to the execution price. We introduce two additional hyper-parameter to control the discretization
 742 level: $2N + 1$ represents the number of total discretization in the execution prices and δ represents
 743 the strength of each level. Given the execution price p , we have $2N + 1$ potential execution prices
 744 $[p - N\delta, \dots, p - \delta, p, p + \delta, \dots, p + N\delta]$ in total. This discretization approach allows us to directly
 745 apply the closed-form solution to solve the optimal u^* .

746 D Limitations

747 Despite the promising results, this work has several limitations. First, the theoretical analysis largely
 748 relies on the assumption of finite state and action spaces for tractability, which is primarily due to
 749 the limited development of deep learning theory. Second, the market impact simulation during the
 750 evaluation stage, while based on trade-level data, remains an approximation and may lack accuracy
 751 for extreme volumes. In fact, when the volume is sufficiently high, it often triggers momentum-based
 752 strategies deployed by other institutions in the market, resulting in higher market impact than reflected
 753 VWAP. Third, although our experiments demonstrate the robustness of the ℓ_p -ellipse uncertainty set
 754 under increasing volumes, we do not test its robustness under increasing trading frequency. Designing
 755 experiments in this setting is more challenging, as the behavior of RL agents varies significantly
 756 across different time scales.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Yes, we have clearly listed the main contributions in the introduction section.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have discussed the limitation in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We have provided the full proof in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include the source code and reproducing instructions in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include the source code and reproducing instructions in the supplementary material. The data is accessed using the third-party API.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We explicitly specify the details in the code and supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our evaluation environment is deterministic; that is, there is no randomness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Included in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed this code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have included this in the conclusion section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The source code is for reviewing purpose only and is not considered as the released new asset.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

1066 **16. Declaration of LLM usage**

1067 Question: Does the paper describe the usage of LLMs if it is an important, original, or

1068 non-standard component of the core methods in this research? Note that if the LLM is used

1069 only for writing, editing, or formatting purposes and does not impact the core methodology,

1070 scientific rigorousness, or originality of the research, declaration is not required.

1071 Answer: [NA]

1072 Justification: The core method development in this research does not involve LLMs as any

1073 important, original, or non-standard components.

1074 Guidelines:

1075 • The answer NA means that the core method development in this research does not

1076 involve LLMs as any important, original, or non-standard components.

1077 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)

1078 for what should or should not be described.

1079