# Stochastic Dominance Driven First-Order Policy Optimization for Multi-Objective Reinforcement Learning

**Ege C. Kaya**[1] **Kadierdan Kaheman**[2] **Jason M Cloud**[2] **Abolfazl Hashemi**[1]

[1] Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
[2] Dolby Laboratories, San Francisco, CA, USA

## Abstract

We study multi-objective reinforcement learning with stochastic vector-valued returns and propose a policy improvement method based on multivariate $k$th-order stochastic dominance in the lower-orthant sense. Rather than scalarizing objectives or comparing only expected values, we compare entire return distributions using $k$th-order integrated cumulative distribution functions (cdfs). This yields a max-violation objective that measures the worst-case dominance gap of a candidate policy relative to the incumbent policy. Under regularity assumptions, we show that the Orthant Dominance Policy Optimization (ORDO) algorithm we introduce drives this objective to a point that certifies $\epsilon$-almost non-dominatedness, meaning that no policy in the class can improve the incumbent by more than $\epsilon$. We also propose smoothing approaches both for the inner maximization in ORDO, and for the hinge structure inherent in integrated cdfs, which improve stability and enable stable stochastic gradients in optimization.

## 1 INTRODUCTION

Reinforcement learning (RL) is often posed as an optimization problem over a scalar objective, typically the expected discounted return of a policy [Sutton and Barto, 2018, Agarwal et al., 2019]. In many domains, however, outcomes are naturally multivariate and uncertain [Vamplew et al., 2022, Bowling et al., 2023]. A robotic controller must find balance between speed, energy, and safety, whereas an autonomous driving policy must balance efficiency and comfort while avoiding rare but catastrophic failures. In such settings, focusing on expectations alone will often be inadequate. Two policies can exhibit the same expected performance while having extremely different variability and tail behavior. Distributional preferences, such as risk aversion or safety constraints, matter alongside mean performance.

Multi-objective reinforcement learning (MORL) formalizes these problems by letting the reward be vector-valued, in $\mathbb{R}^d$. The dominant practice is to reduce this vector to a scalar via a preference model, for instance by choosing weights and optimizing a weighted sum, or by selecting a utility and optimizing its expectation, and then to apply standard single-objective RL methods [Natarajan and Tadepalli, 2005, Van Moffaert et al., 2013, Abels et al., 2019]. This is workable when the preference is known and stable, and when the resulting scalar objective truly captures the preference that we care about. In many real problems, however, the trade-off across objectives is not a priori known, it may be negotiated after the fact, and it may change with context and constraints. Even if we are willing to commit to a particular trade-off across objectives, we still face the question of how uncertainty should be treated within and across objectives.

This motivates policy comparison principles that operate directly on *distributions* of multivariate returns. Stochastic dominance (SD) [Hadar and Russell, 1969, Shaked and Shanthikumar, 2007] is one of the classical principled ways of undertaking this in the univariate setting. Roughly speaking, $k$th-order SD compares random variables (RVs) by requiring a family of inequalities across thresholds, and it admits an expected-utility interpretation in which dominance implies higher expected utility for all utility functions in a broad class. The appeal of SD, at least from a modeling standpoint, is that it provides preference-robust guarantees without forcing one to pick a single risk measure or a single utility. At the same time, it is well known that SD is a partial order (so incomparability between pairs of RVs is common), and that checking SD involves a continuum of inequalities, leading to a search for computationally efficient procedures [Shaked and Shanthikumar, 2007].

Both of these issues become more pronounced in the multivariate setting. SD remains a partial order, now on joint distributions in $\mathbb{R}^d$, and the inequalities that must be en-

*Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence* (UAI 2026), PMLR 337:2828–2876.

forced across thresholds become high dimensional, making naive approaches even more challenging. Furthermore, in the univariate setting, several characterizations of SD involving views based on cdfs, utility functions, and lower sets align cleanly, whereas in $\mathbb{R}^d$, these equivalencies no longer hold. A practitioner must choose which notion of SD to operationalize, and this choice determines both what is being compared and what is algorithmically tractable. In this work, we adopt a tractable notion based on the *lower-orthant order* [Shaked and Shanthikumar, 2007, Kopa and Petrová, 2018], which is the natural multivariate analogue of cdf comparisons, since multivariate cdfs accumulate probability over lower orthants $L = \{x \in \mathbb{R}^d : x \leq m\}$ characterized by a single generator $m \in \mathbb{R}^d$. Throughout the work, whenever we mention any notion of dominance, it is to be taken in the lower-orthant sense, formally defined in Definition 3.2.

**Overview.** We compare policies via multivariate $k$th-order SD, which allows us to work with the integrated cdf form. We introduce the ORDO algorithm, which, at iteration $t$, treats a pointwise $k$th-order dominance gap functional $H_{k,t}(\theta, m)$ comparing the return distribution induced by a candidate policy $\pi_\theta$ to that of the incumbent policy $\pi_{\theta_t}$ at an orthant generator $m \in M$, where $M$ is the compact support of returns. These pointwise gaps are aggregated into the worst-case violation objective

$$\phi_{k,t}(\theta) := \max_{m \in M} H_{k,t}(\theta, m). \tag{1}$$

The maximization searches for the orthant in which the candidate policy is most dominated relative to the incumbent, and then an outer minimization over $\theta$ searches for a policy that reduces this worst-case violation. If we are able to find $\theta$ such that $\phi_{k,t}(\theta) \leq -\epsilon$, then the candidate dominates the incumbent by margin $\epsilon$ in the $k$th-order. If no such improvement exists within the policy class, we conclude that the incumbent is approximately non-dominated.

**Theory and algorithms.** On the theoretical side, we show that ORDO produces an iterate that certifies $\epsilon$-almost non-dominatedness in the nonconvex setting, in the sense that within the policy class there is no policy that can dominate the returned policy by more than $\epsilon$ in the $k$th-order, under regularity assumptions. We state the theory for $k \geq 2$ and our experiments use the practically most common second-order case, while the nonsmooth first-order case is left for future work. Because directly optimizing $\phi_{k,t}$ requires dealing with a hard inner maximization over $m$, we introduce Soft-ORDO, a log-sum-exp surrogate over a finite net of orthant generators, and separately smooth the hinge-polynomial kernels for stable rollout-based gradients.

**Contributions.** Our main contributions are:

- We formulate a tractable multivariate $k$th-order SD objective for MORL using the integrated cdf view for SD, and derive a worst-case dominance functional that supports iterative policy improvement.
- We introduce the ORDO algorithm and establish that under regularity assumptions, an $\epsilon$-almost non-dominated solution can be found in the nonconvex setting, thereby providing a principled certification notion compatible with the partial-order nature of SD.
- We introduce Soft-ORDO as a smooth alternative replacing maximization over orthant generators by a log-sum-exp over a finite net, and describe smoothing approaches for the hinge-polynomial structure of the integrated cdfs that enable stable rollout-based stochastic gradients.

## 2 RELATED WORK

SD is a classical framework for comparing uncertain outcomes while remaining robust to broad classes of non-decreasing utility functions [Dentcheva and Ruszczynski, 2003, Dentcheva and Ruszczyński, 2004, Shaked and Shanthikumar, 2007]. A common optimization viewpoint treats SD as a *constraint* relative to a reference distribution and optimizes a separate scalar objective, which induces min-max structures when enforcing dominance over a continuum of thresholds. Recent machine learning literature revisits this constrained perspective with scalable sample-based procedures and practical inner solvers [Dai et al., 2023]. In a different application domain, Farajzadeh et al. [2025] use pluralistic SD for imitation learning to match demonstrations in a preference-robust way.

Using SD as an *optimality criterion* is harder because SD is only a partial order and enforcing it theoretically involves infinitely many inequalities. Cen et al. [2026] propose a general learning framework which makes SD-based learning practical and amenable to first-order methods. Our work is closest in spirit to this line, but differs in two key ways: (i) our comparisons are inherently *multivariate*, because the joint dependence structure across objectives is part of the preference, and (ii) our guarantees target multi-objective $\epsilon$-almost non-dominatedness in a *nonconvex* policy class, rather than restricting to convex regimes.

Our approach also relates to distributional and risk-sensitive RL. Distributional RL (DRL) models return *distributions* rather than only expectations [Bellemare et al., 2017, 2023, Wiltzer et al., 2024], and risk-sensitive RL optimizes tail-sensitive objectives such as CVaR [Chow et al., 2015, Tamar et al., 2015]. SD provides a preference language compatible with many such criteria. For instance, Martin et al. [2020] study SD-based action selection rules within a DRL framework. In MORL, the dominant paradigm is scalarization and Pareto-style comparisons of expected return vectors [Roijers et al., 2013]. Closer to our motivation, Cai et al. [2023] formalize distributional preferences in MORL through learned families of diverse monotone utility functions, whereas we optimize a dominance-violation objective directly in a multivariate SD view and obtain a certifiably $\epsilon$-almost non-dominated solution.

Because dominance induces only a partial order, a substantial portion of literature studies refinements and relaxations of SD that reduce incomparability while preserving preference encoding. *Almost stochastic dominance* relaxations allow controlled violations under a large majority of utility functions [Leshno and Levy, 2002], and *generalized almost SD* extends this idea to broader settings [Tsetlin et al., 2015]. In a related direction, generalized SD (GSD) has been used as a decision-theoretic tool for comparing methods across multiple criteria without committing to weighted aggregates. Jansen et al. [2023] develop a framework to compare classifiers via generalized SD induced by preference systems, showing that the resulting dominance checks reduce to linear programs and can be equipped with statistical tests. Jansen et al. [2024] further introduce the GSD-front as a more discriminative alternative to the Pareto front for multicriteria benchmarking, together with estimators and statistical tests. These works share with ours the viewpoint that SD relations can yield sharper frontiers than Pareto analyses while avoiding arbitrary scalarization. Our focus is on sequential decision-making with vector-valued returns, and we pursue algorithmic and theoretical tools tailored to policy optimization, including $\epsilon$-almost non-dominatedness guarantees in nonconvex settings and smoothing techniques that stabilize the min-max optimization that arises from SD certification.

## 3 PRELIMINARIES

We consider an episodic Markov decision process (MDP) with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, and a vector-valued reward kernel $P_R : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d)$. A stochastic policy $\pi_\theta(a \mid s)$ is parameterized by $\theta \in \mathbb{R}^w$ and induces a trajectory density $p_\theta(\tau)$ over trajectories $\tau = (s_0, a_0, \ldots, s_T, a_T)$. For a discount factor $\gamma \in (0, 1)$, the random return vector is expressed by $R(\tau) := \sum_{t=0}^{T} \gamma^t R(s_t, a_t)$, where $\tau \sim p_\theta$.

We write $x \leq m$ for componentwise inequality in $\mathbb{R}^d$ and use this partial order for the entirety of the work. For a scalar $u \in \mathbb{R}$, we write $(u)_+ = \max\{u, 0\}$.

SD admits several characterizations, including cdf-based, utility function-based, and lower set-based views. Unlike the univariate setting, in the multivariate setting, these views are not generally equivalent. In the interest of tractability, we will adopt the cdf-based view of SD, which corresponds to the lower-orthant stochastic order as named by Shaked and Shanthikumar [2007]. We first define $k$th-order cdfs.

**Definition 3.1** ($k$th-order cdf)**.** Let $X$ be an RV with support in $\mathbb{R}^d$. The $k$th-order cdf $F_X^k$ of $X$ is defined recursively as

$$\begin{aligned} F_X^1(m) &:= \mathbb{P}(X \leq m), \\ F_X^k(m) &:= \int_{-\infty}^{m} F_X^{k-1}(\mu)\mathrm{d}\mu, \quad k \geq 2, \end{aligned} \tag{2}$$

for $m \in \mathbb{R}^d$.

We can now specify our notion of dominance.

**Definition 3.2** ($k$th-order SD in the multivariate setting)**.** Let $X$ and $Y$ be two RVs with compact support $M \subseteq \mathbb{R}^d$. Then, $X$ dominates $Y$ in the $k$th-order if, for all $m \in M$,

$$F_X^k(m) \leq F_Y^k(m). \tag{3}$$

The interpretation is clear: We say that $X$ dominates $Y$ in the $k$th-order if its $k$th-order cdf accumulates less density in the "less valuable orthants" of their support, as ranked by the usual componentwise partial order over $\mathbb{R}^d$. For true $k$th-order dominance, this must hold for any lower orthant.

The tractability of this formulation, however, is still in question. To obtain a more approachable form, we first use a simple $d$-fold application of Fubini's theorem and induction to express the $k$th-order cdfs in a product-of-ReLUs form [Mosler, 1984, Denuit and Mesfioui, 2010]:

$$\begin{aligned} F_X^k(m) &= \frac{1}{(k-1)!^d} \mathbb{E}\left[\prod_{i=1}^{d} (m_i - X_i)_+^{k-1}\right] \\ &= c_k \mathbb{E}[h(X, m)], \end{aligned} \tag{4}$$

where we define $c_k := (k-1)!^{-d}$ and

$$h_k(x, m) := \prod_{i=1}^{d} (m_i - x_i)_+^{k-1}. \tag{5}$$

This motivates the definition of the following quantity, which, given two RVs $X$ and $Y$ with support in $\mathbb{R}^d$, relates the worst unilateral $k$th-order *dominance violation* of $X$ over $Y$ [Cen et al., 2026]:

$$\begin{aligned} \max_{m \in M} \mathcal{L}_k(X, Y, m) &:= \max_{m \in M} \left(F_X^k(m) - F_Y^k(m)\right) \\ &= c_k \max_{m \in M} \left(\mathbb{E}\left[h_k(X, m)\right] - \mathbb{E}\left[h_k(Y, m)\right]\right). \end{aligned} \tag{6}$$

$\max_{m \in M} \mathcal{L}_k(X, Y, m)$ gives us a measure of how much $X$ *fails to dominate* $Y$ in the $k$th-order at the worst orthant generator, by comparing the densities accumulated by the $k$th-order cdfs of the two RVs over the worst lower orthant. If this quantity is nonpositive, $X$ does not fail to dominate $Y$. Otherwise, the quantity itself tells us how badly the dominance relation is violated over the highest-offending lower orthant. Another step to simplify our undertaking is to limit our search to a class of RVs $X_\theta$ parameterized by $\theta \in \mathbb{R}^w$. This will allow us to treat the problem as a first-order optimization problem over the parameter space. Analysis and minimization of this dominance violation functional will be the main object of the rest of our inquiry.

Since in the practical RL setting, we will soon have to ditch the population quantities and work with sample rollouts of

the environment, we define the empirical estimate $\hat{\mathcal{L}}_k$

$$\begin{aligned}&\hat{\mathcal{L}}_k(R(\tau), R(\tau_t), m)\\ &:= \frac{c_k}{N}\sum_{j=1}^{N}\left(\prod_{i=1}^{d}(m_i - R_{j,i}(\tau))_+^{k-1}\right.\\ &\left.-\prod_{i=1}^{d}(m_i - R_{j,i}(\tau_t))_+^{k-1}\right),\end{aligned} \tag{7}$$

where $X_{j,i}$ indicates the $i$th component of the $j$th sample.

To end the section, we define the notion of optimality that we would like to achieve.

**Definition 3.3** ($k$th-order $\epsilon$-almost non-dominatedness)**.** Let $X_\theta$ be an RV with compact support $M \subseteq \mathbb{R}^d$. $X_\theta$ is $k$th-order $\epsilon$-almost non-dominated if for any other RV $X_{\theta'}$ with support $M$,

$$\max_{m\in M} \mathcal{L}_k(X_{\theta'}, X_\theta, m) \geq -\epsilon. \tag{8}$$

That is, there exists no other RV $Y_\theta$ with support $M$ such that $X_{\theta'}$ dominates $X_\theta$ by more than $\epsilon$ in the $k$th order.

## 4 ORDO

Let us now pull back into the setting of MORL and begin proposing an algorithmic way of obtaining a stochastically non-dominated policy. Let us consider a class of policies parameterized by a vector $\theta \in \mathbb{R}^w$. Assume we initialize parameters $\theta_0$ arbitrarily and thus start with an initial policy $\pi_{\theta_0}$. We would like to leverage the dominance violation functional of the previous section within an optimization framework to arrive at a non-dominated policy. Let us consider an iterative approach, as guided by first-order optimization methods. At each iteration $t+1$, we have an incumbent policy $\pi_{\theta_t}$, and we want to solve

$$\min_{\theta\in\mathbb{R}^w}\max_{m\in M} \mathcal{L}_k(R(\tau), R(\tau_t), m), \tag{9}$$

where $\tau \sim p_\theta$, $\tau_t \sim p_{\theta_t}$, the trajectory densities induced by policies $\pi_\theta$ and $\pi_{\theta_t}$, respectively.

There are multiple ways to view (9), each useful in a different context. One way is to view it, at any time $t$, as a function of the decision parameter $\theta$, i.e.,

$$\min_{\theta\in\mathbb{R}^w}\max_{m\in M} H_{k,t}(\theta, m) = \min_{\theta\in\mathbb{R}^w} \phi_{k,t}(\theta). \tag{10}$$

This view will enable us to argue for the use of iterative first-order methods to solve the problem in the ensuing analysis. The use of first-order methods will naturally mean that being able to compute $\nabla_\theta\phi_k(\theta)$ is essential. However, because $\phi_k(\theta)$ is in reality a pointwise maximum over a family of functions, differentiability is not straightforward. We need to invoke an envelope theorem in the effect of those stated by Danskin [2012] and Bertsekas [1997]. For an extensive analysis of the dominance-violation functional, culminating in the applicability of Danskin's envelope theorem (TheoremB.9) and enabling the use of gradient methods, we refer the reader to Appendix B.

Motivated by the analysis, we present Algorithm 1, ORDO (**Or**thant **D**ominance Policy **O**ptimization). ORDO implements a nested policy-improvement scheme driven by the maximum dominance-violation objective $\phi_{k,t}(\theta) = \max_{m\in M} H_{k,t}(\theta, m)$, which measures the worst $k$th-order dominance gap of a candidate policy $\pi_\theta$ relative to the incumbent $\pi_{\theta_t}$. At outer iteration $t$, the incumbent policy's parameter $\theta_t$ is held fixed and the inner loop runs a stochastic first-order method over $\theta$, starting from $\theta_{t,0} = \theta_t$. Each inner iteration $\bar{t}$ collects a batch of rollouts from the incumbent to form reference returns $\{R_{t,i}\}_{i=1}^N$, and independently, rollouts from the current candidate $\pi_{\theta_{t,\bar{t}}}$. These candidate rollouts are split into two roles. First, a batch $\{R^m_{t,\bar{t},i}\}_{i=1}^N$ is used to approximately solve the inner maximization in $\max_{m\in M} \hat{\mathcal{L}}_k(R^m_{t,\bar{t}}, R_t, m)$, producing an (approximate) worst-offending orthant generator $\hat{m}$ which highlights where the candidate is most dominated by the incumbent. Second, a separate batch $\{R_{t,\bar{t},i}\}_{i=1}^N$ is used to form a REINFORCE-style [Williams, 1992, Sutton et al., 1999] stochastic subgradient $g_{t,\bar{t}}$ of $\phi_{k,t}$ at $\theta_{t,\bar{t}}$ evaluated at $\hat{m}$, and to update $\theta_{t,\bar{t}+1} = \theta_{t,\bar{t}} - \eta_{\bar{t}} g_{t,\bar{t}}$. If, at any inner iterate, the empirical violation certificate satisfies $\hat{\mathcal{L}}_k(R_{t,\bar{t}}, R_t, \hat{m}) \leq -\epsilon/2$, then the candidate is accepted as an improving update, we set $\theta_{t+1} = \theta_{t,\bar{t}+1}$, and the next outer iteration begins with this new incumbent. If no such improvement is found within $\bar{T}$ inner steps, the algorithm terminates and returns $\theta_t$, which is $\epsilon$-almost non-dominated within the policy class under $k$th-order SD.

As it is written, the execution of ORDO (Algorithm 1) is still nontrivial, specifically due to line 9. There is no closed-form expression or obvious way to produce a maximizer $m^*$ to $\hat{\mathcal{L}}_k$ for any value of $k$ or $d$. Although $d = 1$, corresponding to the univariate SD case, admits some practical algorithms for $k \leq 3$ [Dai et al., 2023, Cen et al., 2026], such algorithms do not translate well into the multivariate setting in terms of time and space complexity. One possible approach is to leverage the fact that the inner problem can be expressed as the maximization of the difference of two convex functions. In Appendix C, we justify the use of a first-order method for the solution of this subproblem, using principles of *difference-of-convex (DC) optimization*. An optional stronger solver evaluates a finite net, refines its top-$K$ points, and retains the best objective, at additional computational cost.

We now summarize the main convergence guarantee for ORDO. At a high level, the analysis treats each inner loop as a biased stochastic subgradient method applied to the weakly convex max-function $\phi_{k,t}$, with bias from the inexact inner maximizer. The sharpness and tube conditions in Appendix

**Algorithm 1** ORDO (**Or**thant **D**ominance Policy **O**ptimization)

**Require:** Number of maximum outer and inner iterations $T, \bar{T}$, progress threshold $\epsilon > 0$, initialization $\theta_0$.

**for** $t = 0, \dots, T-1$ **do**
  $\theta_{t,0} \leftarrow \theta_t$
  **for** $\bar{t} = 0, \dots, \bar{T}-1$ **do**
    Sample $\{\tau_{t,i}\}_{i=1}^N \sim p_{\theta_t}(\tau)$ by rolling out MDP with policy $\pi_{\theta_t}$
    $\{R_{t,i}\}_{i=1}^N \leftarrow \{R(\tau_{t,i})\}_{i=1}^N$
    Sample $\{\tau_{t,\bar{t},i}\}_{i=1}^N, \{\tau^m_{t,\bar{t},i}\}_{i=1}^N \sim p_{\theta_{t,\bar{t}}}(\tau)$ by rolling out MDP with policy $\pi_{\theta_{t,\bar{t}}}$
    $\{R_{t,\bar{t},i}\}_{i=1}^N \leftarrow \{R(\tau_{t,\bar{t},i})\}_{i=1}^N$
    $\{R^m_{t,\bar{t},i}\}_{i=1}^N \leftarrow \{R(\tau^m_{t,\bar{t},i})\}_{i=1}^N$
    Find $\hat{m} \in \arg\max_{m \in M} \hat{\mathcal{L}}_k(R^m_{t,\bar{t}}, R_t, m)$
    $g_{t,\bar{t}} \leftarrow \frac{c_k}{N} \sum_{i=1}^N h_k(R_{t,\bar{t},i}, \hat{m}) \nabla_\theta \log p_{\theta_{t,\bar{t}}}(\tau_{t,\bar{t},i})$
    $\theta_{t,\bar{t}+1} \leftarrow \theta_{t,\bar{t}} - \eta_{\bar{t}} g_{t,\bar{t}}$
    **if** $\hat{\mathcal{L}}_k(R_{t,\bar{t}}, R_t, \hat{m}) \le -\epsilon/2$ **then**
      $\theta_{t+1} \leftarrow \theta_{t,\bar{t}+1}$
      **break**
    **end if**
  **end for**
  **if** $\theta_t$ has not been updated **then**
    **return** $\theta_t$
  **end if**
**end for**

D present a set of sufficient conditions when incumbent-relative dominance improvement can be certified. Under bounded subgradients and sufficiently small bias, either the inner loop certifies an $\epsilon$ improvement or the incumbent is already $\epsilon$-almost non-dominated. Chaining these steps yields $O(\epsilon^{-2})$ iteration complexity. The full proof, including the concentration event that keeps the inner iterates inside the sharpness tube is deferred to Appendix D.

**Theorem 4.1** (Convergence of ORDO)**.** *Fix $\epsilon > 0$ and integer $k \ge 2$ and consider Algorithm 1 (ORDO). For each outer iteration $t$, assume*

$$\phi_{k,t}(\theta) := \max_{m \in M} H_{k,t}(\theta, m), \quad \Theta_t^* := \arg\min_{\theta \in \mathbb{R}^w} \phi_{k,t}(\theta). \tag{11}$$

*Under Assumptions B.1-B.2, Appendix D shows that $\phi_{k,t}$ is $\rho$-weakly convex (Corollary D.7) and that the stochastic directions $g_{t,\bar{t}}$ produced by line 10 of the algorithm are uniformly bounded by $G_{\max}$ (Proposition D.4) and have a controlled bias $\beta_N$ due to the inexact inner maximization (Proposition D.3).*

*Assume the regularity condition of Assumption D.9 holds with sharpness constant $\kappa_{\text{sh}} > 0$ and that $\beta_N < \kappa_{\text{sh}}/2$. Let the inner-loop step size be*

$$\eta_{\bar{t}} \equiv \min\left\{ \frac{(\kappa_{\text{sh}} - 2\beta_N)\epsilon}{4G_{\max}^3}, \frac{(\kappa_{\text{sh}} - 2\beta_N)\kappa_{\text{sh}}}{4G_{\max}^2 \rho} \right\}. \tag{12}$$

*If the inner loop at outer iteration $t$ is started inside the sharpness tube, meaning that $\text{dist}(\theta_{t,0}, \Theta_t^*) \le \kappa_{\text{sh}}/(2\rho)$, then Lemma D.10 implies that, with high probability, all inner iterates $\{\theta_{t,\bar{t}}\}_{\bar{t}=0}^{\bar{T}}$ remain inside the tube where the sharpness geometry is effective.*

*On the intersection of these high-probability tube events over all executed outer iterations, ORDO terminates after $O(\epsilon^{-2})$ total inner-loop updates and returns a parameter $\theta_{\text{out}}$ such that, for every $\theta \in \mathbb{R}^w$,*

$$\mathbb{E}\left[ \max_{m \in M} \mathcal{L}_k\left(R(\tau_\theta), R(\tau_{\theta_{\text{out}}}), m\right) \right] \ge -\epsilon, \tag{13}$$

*where the expectation is over the rollouts and any randomness in the inner maximization routine. In particular, $\pi_{\theta_{\text{out}}}$ is kth-order $\epsilon$-almost non-dominated in expectation within the policy class.*

### 4.1 SOFT-ORDO

In addition to the ORDO objective which is dependent on a hard-maximization inner objective, we also consider a smoothed variant, which we refer to as Soft-ORDO. The key idea is to replace the inner maximization over orthant generators by a smooth log-sum-exp surrogate. This yields a differentiable approximation to the ORDO objective and potentially leads to better-behaved gradients in optimization, while still encouraging reduction in dominance violations.

Soft-ORDO is useful both algorithmically and analytically. It avoids dependence on exact inner maximizers at each update and provides a natural smoothing parameter that interpolates between a softer aggregate objective and the hard-max ORDO objective in the low-temperature limit. We defer the complete treatment to Appendix E and only state the main convergence theorem informally.

**Theorem 4.2** (Convergence of Soft-ORDO (informal))**.** *Fix $\epsilon > 0$ and integer $k \ge 2$. With appropriate step and batch sizes, Soft-ORDO produces an iterate $\theta$ with $\|\nabla \tilde{\phi}_{k,t}(\theta)\| \le \epsilon$ in expectation in $O(\epsilon^{-2})$ iterations. Further, this iterate is also $(\epsilon+\epsilon')$-Pareto stationary for the original objective $\phi_{k,t}$, in the sense that there is no direction $d$ that simultaneously decreases all coordinate objectives by more than $\epsilon + \epsilon'$ at $\theta$, where $\epsilon'$ is exactly the approximation error induced by smoothing and discretizing the support $M$.*

## 5 SMOOTHING THE OBJECTIVES VIA SMOOTH-RELU

This section concerns a different smoothing layer from Soft-ORDO: Soft-ORDO smooths the hard maximum over orthant generators, whereas here we smooth the ReLU/hinge terms inside the integrated cdf kernel. We use smoothed approximations such as softplus [Dugas et al., 2000] or

GELU [Hendrycks, 2016] to alleviate discontinuity or non-smoothness in rollout-based gradients. In particular, we resort to the widely accepted formalism of Gaussian smoothing [Nesterov and Spokoiny, 2017]. While this smoothing is standard, we demonstrate that it induces a strict ordering hierarchy for our specific $k$-th order integrated cdf objective. We start by defining smooth-ReLU formally.

**Definition 5.1** (Smooth-ReLU via Gaussian smoothing)**.** Let $\phi_\mu$ and $\Phi_\mu$ denote the probability density function (pdf) and cdf, respectively, of a zero-mean Gaussian variable with variance $\mu^2$. Let $Z \sim \mathcal{N}(0, \mu^2 I_d)$.

We define the *smooth-ReLU* function, denoted as $\zeta_\mu$, via the convolution:

$$\zeta_\mu(x) := \int_{-\infty}^{\infty} (x - z)_+ \phi_\mu(z) \mathrm{d}z = \mathbb{E}_{Z_i}[(x - Z_i)_+] \quad (14)$$

$$= x\Phi_\mu(x) + \mu^2 \phi_\mu(x). \quad (15)$$

Crucially, due to the infinite support of the Gaussian pdf, $\zeta_\mu(x) > 0$ for all $x \in \mathbb{R}$.

We now define two approximations to the cdf that leverage Gaussian smoothing via two distinct yet similar approaches.

**Definition 5.2** (Gaussian-smoothed objectives)**.** We define the *approximation* $\tilde{F}_k^{\text{approx}}$ and the *true convolution* $\tilde{F}_k^{\text{true}}$ of the $k$-th order integrated cdf at a reference point $m$ as:

$$\tilde{F}_k^{\text{approx}}(m) := c_k \mathbb{E}_R \left[ \prod_{i=1}^{d} (\zeta_\mu(m_i - R_i))^{k-1} \right],$$

$$\tilde{F}_k^{\text{true}}(m) := c_k \mathbb{E}_R \left[ \mathbb{E}_Z \left[ \prod_{i=1}^{d} (m_i - R_i - Z_i)_+^{k-1} \right] \right]. \quad (16)$$

The key result of the section is that we can use Gaussian-smoothed counterparts of our objectives for $k \geq 2$, resulting in alternative but valid notions of $k$th-order SD while enjoying a smoother and friendlier landscape for gradient-based optimization in practice. However, Gaussian smoothing does not resolve the difficulty of the $k = 1$ case: the resulting ordering can be indeterminate.

**Proposition 5.3** (Comparison of smoothing approximations)**.** *For any $m \in M$ and $\mu > 0$, the relationship between the approximation and the true convolution depends on the order $k$:*

1. *For $k = 2$, the approximation is exact: $\tilde{F}_2^{approx}(m) = \tilde{F}_2^{true}(m)$.*
2. *For integer $k > 2$, the approximation strictly lower bounds the true convolution: $\tilde{F}_k^{approx}(m) < \tilde{F}_k^{true}(m)$.*
3. *For $k = 1$, using the convention $0^0 = 0$ and $x^0 = 1$ for $x > 0$, the approximation degenerates to a constant upper bound: $1 \equiv \tilde{F}_1^{approx}(m) > \tilde{F}_1^{true}(m)$.*

We now explore the relationships between the two approximations we explored above and the true cdf.

**Proposition 5.4** (Ordering of exact and smoothed objectives)**.** *Let $F_R^k(m)$ be the exact $k$-th order cdf objective, and let $\tilde{F}_k^{true}(m)$ and $\tilde{F}_k^{approx}(m)$ be the smoothed variations defined in Definition 5.2. For any $m \in M$ and $\mu > 0$, the following ordering relations hold almost everywhere (with respect to the measure of $R$):*

1. ***For** $k = 2$ **(Overestimation):** Both approximations are strictly greater than the exact objective, and they are equal to each other:*

$$\tilde{F}_2^{true}(m) = \tilde{F}_2^{approx}(m) > F_R^2(m). \quad (17)$$

2. ***For integers** $k > 2$ **(Strict overestimation chain):** Both approximations strictly overestimate the exact objective, with the true convolution providing a looser bound than the approximation:*

$$\tilde{F}_k^{true}(m) > \tilde{F}_k^{approx}(m) > F_R^k(m). \quad (18)$$

3. ***For** $k = 1$ **(Indeterminate ordering):** The approximation $\tilde{F}_1^{approx}(m)$ is a trivial upper bound (constant 1). However, the true convolution $\tilde{F}_1^{true}(m)$ admits no global ordering with respect to $F_R^1(m)$; it overestimates $F_R^1(m)$ in regions of low probability density and underestimates it in regions of high probability density.*

For a complete analysis and proofs of these propositions, we refer the interested reader to Appendix F.

## 6 EXPERIMENTS

**Benchmarks.** We evaluate our approach on five multi-objective benchmarks from MO-Gymnasium [Felten et al., 2023]: DeepSeaTreasure, FruitTree, MountainCar, Hopper, and HalfCheetah. These environments span both discrete and continuous control. Following the protocol of Cai et al. [2023], we use the two-objective setting and retain only the first two reward coordinates when environments expose more than two reward dimensions.

**Policy-set construction.** For each environment, we train a portfolio of $N = 20$ policies independently. To obtain a diverse set of policies without coupling multiple runs, we associate each policy $\pi_i$ with a preference vector $w_i \in \Delta_2$ and define a standard scalarized utility $u_{w_i}(\tau) = w_i^\top R(\tau)$ for trajectory return $R(\tau) \in \mathbb{R}^2$. Each policy is initially trained to be competitive for its scalarization, while still enabling later SD-driven refinement. All reported metrics are computed on the resulting set of $N$ policies.

**PPO and ORDO mixture.** Our theory analyzes pure ORDO. Our empirical implementation is a practical PPO-ORDO surrogate: PPO supplies the same kind of stable

exploration and policy-gradient machinery used by strong MORL baselines, while ORDO supplies late-stage dominance refinement. This comparison is fairer since the benchmark methods include PPO-based methods such as DP-MORL. Concretely, for a policy parameter $\theta$, the update rule is a convex combination of gradients:

$$g_t = (1-\alpha_t)g_t^{\mathrm{PPO}} + \alpha_t g_t^{\mathrm{ORDO}}, \quad \theta_{t+1} = \theta_t + \eta_t g_t, \tag{19}$$

where $g_t^{\mathrm{PPO}}$ is the policy-gradient component obtained from a PPO-style clipped surrogate objective for maximizing $\mathbb{E}[u_{w_i}(\tau)]$, and $g_t^{\mathrm{ORDO}}$ is the policy-gradient estimator induced by the ORDO dominance-violation objective. The mixing weight $\alpha_t \in [0,1]$ is scheduled so that the early updates emphasize PPO to obtain stable exploration and diverse behaviors, while later updates increase the influence of ORDO to refine dominance properties of the learned policy set and reach non-dominated policies. Although this mixed update is not covered by our pure ORDO theorem, it provides a viable practical surrogate. It is also compatible with the dominance objective, in that SD improvements imply improvements in expected utility under any nondecreasing utility function, including linear scalarizations $u_w$, so the PPO component does not conflict with later dominance refinement.

**Inner maximization and smoothing.** To compute the ORDO signal, we approximately solve the inner maximization over orthant generators by projected gradient ascent. The reproducibility default initializes at the return-box midpoint. We also provide a finite-net top-$K$ option that evaluates a Cartesian net in two dimensions or a Sobol net in higher dimensions, includes observed return vectors, and refines the strongest starts. We use $k = 2$ throughout and a softplus hinge surrogate for stability. As shown in Table 1, the finite-net top-$K$ solver matches the finite-grid check in all 72 DeepSeaTreasure diagnostics and is within 5% in 48 of 72 Hopper diagnostics, with approximately one second average runtime.

Table 1: **Finite-net top-$K$ inner-max diagnostics.** "Matched" counts gaps at most $10^{-4}$, the percentage columns allow relative error up to the stated threshold. Runtime includes net evaluation and refinement.

| Environment | Checks | Matched | Within 1% | Within 5% | Mean time (s) |
|---|---|---|---|---|---|
| DeepSeaTreasure | 72 | 72 | 72 | 72 | 0.92 |
| Hopper | 72 | 33 | 38 | 48 | 1.15 |

**Training budget and schedules.** Each policy is trained for a fixed budget of $10^7$ environment transitions. We use PPO hyperparemeters aligned with standard Stable Baselines3 [Raffin et al., 2021] conventions (e.g., clipped surrogate, GAE, gradient clipping). The exact schedules and hyperparameters in all runs are reported in Appendix G.

**Evaluation metrics and protocol.** We evaluate the learned policy portfolio with the four metrics used by Cai et al. [2023]: expected utility (EU), hypervolume (HV), constraint satisfaction (Constr.) and variance objective (Var.). Each trained policy is evaluated for 40 episodes. Metrics are computed from undiscounted episodic returns, and then aggregated over the policy set. See Appendix G for the exact definitions of the metrics, and reference-point conventions used in computing HV.

**Results and discussion.** Tables 2 and 3 report the four benchmark metrics. For ORDO, we report mean $\pm$ standard deviation across three independently trained portfolios that preserve the same 20 preference assignments and environment-specific hyperparameters. Among the directly comparable EU, Constr., and Var. entries, ORDO has the highest displayed mean in 9 of 15 environment-metric pairs. It leads all three metrics on Hopper and HalfCheetah, leads EU and Var. on DeepSeaTreasure, and leads Constr. on FruitTree, while MountainCar remains challenging. The baseline values are copied point estimates without replicate uncertainty, so these comparisons are descriptive. Figure 1 visualizes the empirical return samples from one learned portfolio.

**Inner maximization.** The finite-net top-$K$ inner solver matches the finite-grid check in all 72 DeepSeaTreasure diagnostics and is within 5% in 48 of 72 Hopper diagnostic see Table 1 and Appendix G for further discussion.

## 7 CONCLUSION

We introduced ORDO, a first-order policy optimization algorithm for MORL based on multivariate $k$th-order SD in the lower-orthant sense. By optimizing a worst-case dominance-violation objective, ORDO provides a natural $\epsilon$-almost non-dominatedness certificate within a policy class under regularity assumptions, aligning optimization with the partial-order structure induced by SD. We further developed smooth surrogates for the hinge-polynomial structure inherent in integrated cdf objectives, yielding more stable optimization and facilitating practical rollout-based gradient estimators.

Empirically, we instantiated ORDO as a dominance-refinement signal layered on top of a PPO backbone. Across three independently trained portfolios, it is strongest on Hopper and HalfCheetah and obtains competitive but mixed results on the remaining benchmarks. Promising directions include improving the inner maximization routine, studying the interaction between dominance refinement and the PPO schedule, and developing sharper theory and practical surrogates for the challenging $k = 1$ case.

### Acknowledgements

This work was supported in part by NSF under grants DMS-2502560 and CNS-2313109.

Table 2: **Standard MORL metrics (higher is better).** ORDO reports mean ± standard deviation across three independently trained $N = 20$ policy portfolios. Other values are copied point estimates from Cai et al. [2023].

| Environment | Hopper | | HalfCheetah | | MountainCar | | DeepSeaTreasure | | FruitTree | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | EU | HV | EU | HV | EU | HV | EU | HV | EU | HV |
| GPI-PD | 2.37 K | 5.03 M | 412.62 | 1.08 M | −55.44 | 7.37 K | 5.93 | 9.75 | **6.98** | 0.14 |
| GPI-LS | 1.40 K | 1.45 M | 98.31 | 2.84 M | −37.37 | 8.05 K | 5.04 | 9.36 | 3.84 | 1.49 |
| OLS | 175.07 | 3.30 K | −580.38 | 1.42 M | −470.00 | 2.45 | 4.64 | 10.28 | 6.27 | 7.49 |
| PGMORL | 300.19 | 32.62 K | 111.30 | 383.68 K | −429.48 | 94.45 | – | – | – | – |
| DPMORL | 3.49 K | 12.15 M | 1.19 K | **8.59 M** | **−29.89** | 8.66 K | 6.70 | 8.68 | 6.89 | 16.39 |
| **PPO-ORDO (ours)** | **3.68 ± 0.08 K** | **13.08 ± 0.62 M** | **2.47 ± 0.18 K** | 1.77 ± 0.07 M | −56.98 ± 0.12 | **875.81 ± 0.43 K** | **7.26 ± 0.00** | **2.16 ± 0.01 K** | 6.64 ± 0.43 | **39.56 ± 0.48** |

Table 3: **Distributional MORL metrics (higher is better).** ORDO reports mean ± standard deviation across three independently trained $N = 20$ policy portfolios. Other values are copied point estimates from Cai et al. [2023].

| Environment | Hopper | | HalfCheetah | | MountainCar | | DeepSeaTreasure | | FruitTree | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Constr. | Var. | Constr. | Var. | Constr. | Var. | Constr. | Var. | Constr. | Var. |
| GPI-PD | 0.47 | 979.74 | 0.64 | 83.49 | **1.00** | −31.15 | 0.85 | 2.59 | 0.65 | **3.42** |
| GPI-LS | 0.25 | 607.47 | 0.60 | 51.67 | **1.00** | −22.73 | 0.85 | 1.99 | 0.33 | 1.35 |
| OLS | 0.05 | 75.48 | 0.43 | −311.27 | 0.05 | −244.21 | 0.80 | 1.86 | 0.50 | 2.98 |
| PGMORL | 0.05 | 98.47 | 0.50 | 45.48 | 0.11 | −249.69 | – | – | – | – |
| DPMORL | 0.76 | 1.65 K | 0.82 | 431.26 | **1.00** | **−16.32** | **0.90** | 2.54 | 0.67 | 3.21 |
| **PPO-ORDO (ours)** | **0.92 ± 0.02** | **1.71 ± 0.08 K** | **0.89 ± 0.03** | **1.07 ± 0.07 K** | 0.91 ± 0.07 | −28.17 ± 0.04 | 0.87 ± 0.21 | **3.38 ± 0.00** | **0.83 ± 0.11** | 3.18 ± 0.18 |

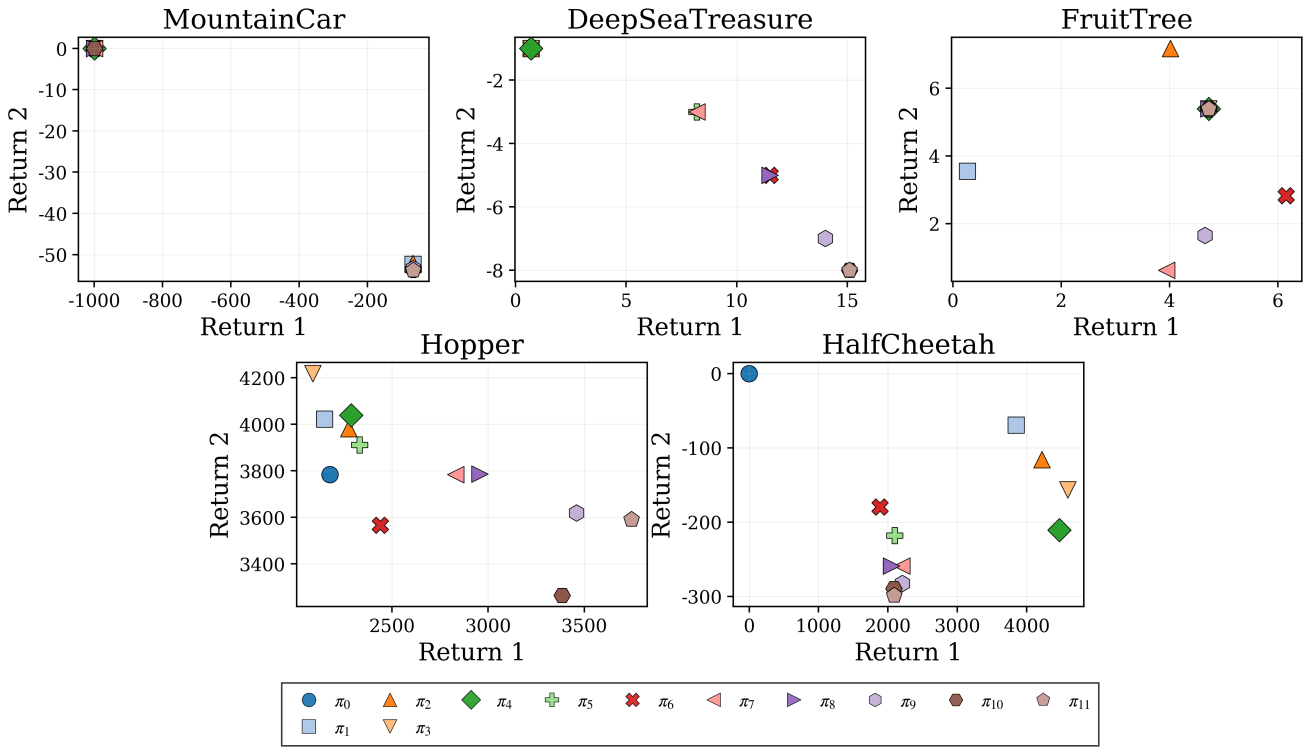


Figure 1: **Return distributions of learned policy portfolios.** For each environment, we evaluate the portfolio of $N = 20$ learned policies $\{\pi_i\}_{i=0}^{19}$ and plot the resulting empirical *undiscounted* episodic return vectors $(R_1, R_2)$. For readability, the figure shows a 12-policy subset of each learned 20-policy portfolio so that overlapping policies and trade-offs remain visible. Each marker denotes a different policy $\pi_i$. Multiple points per marker come from repeated evaluation episodes (one point per episode). The plots visualize portfolio diversity and show how different policies concentrate return mass in different regions of the bi-objective return space.

## References


Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. In *International conference on machine learning*, pages 11–20. PMLR, 2019.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32: 96, 2019.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. The MIT Press, 05 2023. ISBN 9780262374026. doi: 10.7551/mitpress/14207.001.0001. URL https://doi.org/10.7551/mitpress/14207.001.0001.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, pages 3003–3020. PMLR, 2023.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Xin-Qiang Cai, Pushi Zhang, Li Zhao, Jiang Bian, Masashi Sugiyama, and Ashley Llorens. Distributional pareto-optimal multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 36:15593–15613, 2023.

Shicong Cen, Jincheng Mei, Hanjun Dai, Dale Schuurmans, Yuejie Chi, and Bo Dai. Beyond expectations: Learning with stochastic dominance made practical. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=ebyPKXsweD.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1):79–127, 2006.

Hanjun Dai, Yuan Xue, Niao He, Yixin Wang, Na Li, Dale Schuurmans, and Bo Dai. Learning to optimize with stochastic dominance constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 8991–9009. PMLR, 2023.

John M Danskin. *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media, 2012.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

Darinka Dentcheva and Andrzej Ruszczynski. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization*, 14(2):548–566, 2003.

Darinka Dentcheva and Andrzej Ruszczyński. Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints. *Mathematical Programming*, 99(2):329–350, 2004.

Michel M Denuit and Mhamed Mesfioui. Generalized increasing convex and directionally convex orders. *Journal of Applied Probability*, 47(1):264–276, 2010.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000.

Ali Farajzadeh, Danyal Saeed, Syed M Abbas, Rushit N Shah, Aadirupa Saha, and Brian D Ziebart. Imitation beyond expectation using pluralistic stochastic dominance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C. da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

Josef Hadar and William R Russell. Rules for ordering uncertain prospects. *The American economic review*, 59 (1):25–34, 1969.

Juha Heinonen. *Lectures on Lipschitz analysis*. Number 100. University of Jyväskylä, 2005.

D Hendrycks. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Kihyuk Hong and Ambuj Tewari. A primal-dual algorithm for offline constrained reinforcement learning with linear mdps. In *Proceedings of the 41st International Conference on Machine Learning*, pages 18711–18737, 2024.

Christoph Jansen, Malte Nalenz, Georg Schollmeyer, and Thomas Augustin. Statistical comparisons of classifiers by generalized stochastic dominance. *Journal of Machine Learning Research*, 24(231):1–37, 2023.

Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin. Statistical multicriteria benchmarking via the gsd-front. *Advances in Neural Information Processing Systems*, 37:98143–98179, 2024.

Milos Kopa and Barbora Petrová. Strong and weak multivariate first-order stochastic dominance. *Available at SSRN 3144058*, 2018.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Preventing reward hacking with occupancy measure regularization. 2023.

Moshe Leshno and Haim Levy. Preferred by "all" and preferred by "most" decision makers: Almost stochastic dominance. *Management Science*, 48(8):1074–1085, 2002.

Sergio López-Rivera, Pedro Pérez-Aros, and Emilio Vilches. A projected variable smoothing for weakly convex optimization and supremum functions. *Journal of Optimization Theory and Applications*, 208(3):115, 2026.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

John Martin, Michal Lyskawinski, Xiaohu Li, and Brendan Englot. Stochastically dominant distributional reinforcement learning. In *International conference on machine learning*, pages 6745–6754. PMLR, 2020.

Washim U Mondal and Vaneet Aggarwal. Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3097–3105. PMLR, 2024.

KC Mosler. Stochastic dominance decision rules when the attributes are utility independent. *Management Science*, 30(11):1311–1322, 1984.

Sriraam Natarajan and Prasad Tadepalli. Dynamic preferences in multi-criteria reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 601–608, 2005.

Edward Nelson. Probability theory and euclidean field theory. In *Constructive quantum field theory*, pages 94–124. Springer, 2007.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2):255–283, 2015.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Martin Schetzen. *The Volterra and Wiener theories of nonlinear systems*. Krieger Publishing Co., Inc., 2006.

Moshe Shaked and J George Shanthikumar. *Stochastic orders*. Springer, 2007.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Pham Dinh Tao and LT Hoai An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997.

Ilia Tsetlin, Robert L Winkler, Rachel J Huang, and Larry Y Tzeng. Generalized almost stochastic dominance. *Operations Research*, 63(2):363–377, 2015.

Peter Vamplew, Benjamin J Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M Roijers, Conor F Hayes, Fredrik Heintz, Patrick Mannion, Pieter JK Libin, et al. Scalar reward is not enough: A response to silver,

singh, precup and sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2):41, 2022.

Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*, pages 191–199. IEEE, 2013.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Edward CF Wilson. A practical guide to value of information analysis. *Pharmacoeconomics*, 33(2):105–121, 2015.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of multivariate distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 37:101297–101336, 2024.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlxIJBFDr.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure (cccp). *Advances in neural information processing systems*, 14, 2001.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

# First-Order Stochastic Dominance Optimization in Multi-Objective Reinforcement Learning (Supplementary Material)

**Ege C. Kaya**[1] **Kadierdan Kaheman**[2] **Jason M Cloud**[2] **Abolfazl Hashemi**[1]

[1] Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
[2] Dolby Laboratories, San Francisco, CA, USA

## A NOTATION GUIDE

Table 4: Recurring notation used in the main text, algorithms, theorem statements, and experiments.

| Symbol | Meaning |
|---|---|
| **MDP and trajectory notation** | |
| $\mathcal{S}, \mathcal{A}$ | State and action spaces. |
| $P_R$ | Vector-valued reward kernel. |
| $d$ | Number of reward objectives. |
| $w$ | Dimension of the policy parameter vector. |
| $s, a$ | State and action. |
| $\theta \in \mathbb{R}^w$ | Candidate policy parameter. |
| $\pi_\theta$ | Stochastic policy parameterized by $\theta$. |
| $\tau$ | Trajectory $(s_0, a_0, \ldots, s_T, a_T)$. |
| $p_\theta(\tau)$ | Trajectory density induced by $\pi_\theta$. |
| $T$ | Maximum number of outer ORDO iterations. Also used as a trajectory horizon when clear from context. |
| $\gamma$ | Discount factor. |
| $R(s, a)$ | Random vector reward at state-action pair $(s, a)$. |
| $R(\tau)$ | Discounted vector return (cumulative reward) of trajectory $\tau$. |
| $X_\theta$ | Random return vector induced by candidate policy $\pi_\theta$. |
| $Y$ | Reference or incumbent random return vector. |
| $N$ | Number of rollout samples in an empirical batch. |
| $B_p$ | Uniform bound on trajectory score norms. |
| **Stochastic dominance and integrated-cdf notation** | |
| $x \leq m$ | Componentwise inequality, defining a lower orthant. |
| $(u)_+$ | Positive part of the argument: $\max\{u, 0\}$. |
| $k$ | Order of stochastic dominance. The theory is stated for $k \geq 2$. |
| $m \in \mathbb{R}^d$ | Lower-orthant generator or threshold vector. |
| $M$ | Compact return-support box over which orthants are searched. |
| $D$ | Coordinatewise diameter bound for $M$. |
| $F_X^k(m)$ | $k$th-order integrated cdf of random vector $X$ evaluated at $m$. |
| $h_k(x, m)$ | Hinge-polynomial kernel $\prod_i (m_i - x_i)_+^{k-1}$. |
| $c_k$ | Normalization constant $((k-1)!)^{-d}$. |
| $\succeq_k$ | Lower-orthant $k$th-order stochastic-dominance relation. |
| $\mathcal{L}_k(X, Y, m)$ | Pointwise dominance gap $F_X^k(m) - F_Y^k(m)$. |
| $\widehat{\mathcal{L}}_k$ | Rollout-based empirical estimate of $\mathcal{L}_k$. |
| $V_k(X, Y)$ | Worst-orthant gap $\max_{m \in M} \mathcal{L}_k(X, Y, m)$, used in supplementary discussion. |
| **ORDO and Soft-ORDO algorithm notation** | |
| $t$ | Outer ORDO iteration index. |

*Table 4, continued*

| Symbol | Meaning |
|---|---|
| $\bar{t}$ | Inner policy-update index for a fixed incumbent. |
| $\bar{T}$ | Maximum number of inner policy updates. |
| $\theta_t$ | Incumbent parameter at outer iteration $t$. |
| $\theta_{t,\bar{t}}$ | Candidate parameter at outer iteration $t$ and inner iteration $\bar{t}$. |
| $R_{t,i}$ | Return from incumbent rollout $i$at outer iteration $t$. |
| $R^m_{t,\bar{t},i}$ | Candidate return used to solve the inner maximization. |
| $H_{k,t}(\theta, m)$ | Pointwise candidate-versus-incumbent dominance gap. |
| $\phi_{k,t}(\theta)$ | Worst-orthant violation $\max_{m\in M} H_{k,t}(\theta, m)$. |
| $\widehat{m}$ | Approximate maximizer returned by the inner orthant search. |
| $g_{t,\bar{t}}$ | Stochastic policy-gradient or subgradient estimate. |
| $\eta_{\bar{t}}$ | Inner-loop step size. |
| $\epsilon$ | Required progress margin or target certificate tolerance. |
| $M_J = \{m_j\}_{j=1}^J$ | Finite $\alpha_{\text{net}}$-net of orthant generators used by Soft-ORDO. |
| $J$ | Number of finite-net elements. |
| $\alpha_{\text{net}}$ | Covering radius of $M_J$. |
| $q_j(\theta)$ | Softmax weight assigned to finite-net element $m_j$. |
| $\kappa$ | Soft-ORDO log-sum-exp temperature. |
| $\widetilde{\phi}_{k,t}$ | Soft-ORDO finite-net log-sum-exp objective. |
| $\widetilde{L}$ | Smoothness constant of $\widetilde{\phi}_{k,t}$. |
| $\epsilon'$ | Error transferred from finite-net and softmax approximation. |
| **PPO-ORDO implementation notation** | |
| $g_t^{\text{PPO}}$ | Gradient from the PPO clipped-surrogate update. |
| $g_t^{\text{ORDO}}$ | Gradient induced by the ORDO dominance-violation objective. |
| $\alpha_t^{\text{mix}}$ | Scheduled weight on $g_t^{\text{ORDO}}$ in the practical PPO-ORDO mixture. |
| $1-\alpha_t^{\text{mix}}$ | Scheduled weight on $g_t^{\text{PPO}}$. |
| $w_i \in \Delta_2$ | Scalarization preference assigned to portfolio policy $i$. |
| $u_{w_i}(\tau)$ | Scalarized utility $w_i^\top R(\tau)$ used by PPO. |
| $\Pi$ | Learned policy portfolio. |
| $\lvert\Pi\rvert$ | Number of independently trained policies in a portfolio. |
| $\pi_{\text{ref}}$ | Periodically refreshed reference snapshot used by PPO-ORDO. |
| $\mathcal{L}_{\text{PPO}}$ | PPO clipped surrogate with value and entropy terms. |
| **Convergence assumptions, constants, and certificates** | |
| $\Theta_t^*$ | Minimizer set $\arg\min_\theta \phi_{k,t}(\theta)$. |
| $\phi_{k,t}^*$ | Minimum value of $\phi_{k,t}$. |
| $\text{dist}(\theta, \Theta_t^*)$ | Euclidean distance from $\theta$ to the minimizer set. |
| $\rho$ | Weak-convexity constant of $\phi_{k,t}$. |
| $\kappa_{\text{sh}}$ | Sharpness constant. |
| $\kappa_{\text{sh}}/(2\rho)$ | Inner radius of the assumed sharpness tube. |
| $\kappa_{\text{sh}}/\rho$ | Outer radius in the concentration lemma. |
| $G_{\max}$ | Uniform bound on stochastic ORDO directions. |
| $\beta_N$ | Bias bound caused by an inexact empirical inner maximizer. |
| $B, G$ | Bounds on the policy score and its Lipschitz constant. |
| $\delta_c$ | Irreducible inner-solver bias term used in the hard-max analysis. |
| $\Delta(\theta)$ | Gap between maximal and submaximal finite-net values. |
| $\delta_\Delta$ | Positive margin separating near-maximal from submaximal net points. |
| $L, L_m$ | Lipschitz constants in the orthant variable and its policy gradient. |
| $\delta_{\text{val}}$ | Value approximation error $L\alpha_{\text{net}} + \kappa \log J$. |
| **Smoothing notation** | |
| $\mu$ | Gaussian smooth-ReLU scale used for the hinge-polynomial kernel. |
| $\phi_\mu, \Phi_\mu$ | Density and cdf of a centered Gaussian with variance $\mu^2$. |
| $\zeta_\mu$ | Gaussian smooth-ReLU, the convolution of ReLU with $\phi_\mu$. |
| $s$ | Logistic smoothing scale for the softplus analogue. |
| $\zeta_s^{\text{sp}}$ | Softplus/logistic smooth-ReLU. |
| $\widetilde{F}_k^{\text{true}}$ | True Gaussian convolution of the integrated cdf. |
| $\widetilde{F}_k^{\text{approx}}$ | Tractable product-of-smoothed-hinges approximation. |
| $\widetilde{\mathcal{L}}_k^\mu$ | Symmetrically smoothed dominance-gap objective. |
| $\Lambda$ | Total-variation distance between candidate and incumbent trajectory laws. |

*Table 4, continued*

| Symbol | Meaning |
|---|---|
| $\Delta_\theta$ | Parameter displacement $\|\theta - \theta_t\|$ controlling smoothing bias. |

## B ANALYSIS OF THE DOMINANCE VIOLATION FUNCTIONAL

We now start the analysis of the dominance violation functional, viewed as a function of the parameter $\theta$. To simplify the crowded notation, we will drop the $t$ subscript for much of the analysis, since the following results will hold for any time step $t$. We express

$$H_k(\theta, m) = c_k \cdot \left(\mathbb{E}\left[h(R(\tau), m)\right] - \mathbb{E}\left[h(R(\tau_t), m)\right]\right), \tag{20}$$

It will sometimes be informative to analyze $h_k$ to draw conclusions about the dominance violation functional $H_k$. This analysis depends on the value of $k$, the order of stochastic dominance we would like to pursue. In fact, only two main cases arise. When $k = 1$, with the standard convention that $0^0 = 0$, we have

$$h_1(x, m) = \prod_{i=1}^{d} \mathbf{1}_{\{x_i \le m_i\}}(x_i). \tag{21}$$

This points to $H_1(\theta, m)$ not even being continuous in the loose sense, let alone Lipschitz or differentiable, which greatly hinders further analysis. Our only hope of differentiability will be through the use of a smooth approximation of the indicator function, the subject of Section 5.

When $k \ge 2$, we are in the setting where $H_k(\theta, m)$ is a continuous function, and further analysis is possible. The following analysis pertains to this case.

Recall that

$$F_R^k(m) = c_k \mathbb{E}\left[\prod_{i=1}^{d} (m_i - R_i)_+^{k-1}\right], \quad c_k = \frac{1}{(k-1)!^d}. \tag{22}$$

We start with two standard assumptions. We assume, componentwise bounded rewards, which implies a normalized, compact support for the return distribution without loss of generality. The following assumption formalizes this.

**Assumption B.1** (Boundedness of reward)**.** For every state-action pair $(s, a)$ and every $i \in [d]$,

$$R_i(s, a) \in [0, (1-\gamma)(k-1)!^{1/(k-1)}]. \tag{23}$$

In particular, Assumption B.1 implies that for any trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T)$, the return $R(\tau) := \sum_{t=0}^{T} \gamma^t R(s_t, a_t)$ has compact support $[0, (k-1)!^{1/(k-1)}]^d =: M$, with diameter $D := (k-1)!^{1/(k-1)}$. All dominance comparisons in the paper are enforced over the compact domain $M$. This restriction avoids the numerical subtleties that arise when enforcing SD over an unbounded domain, and is commonplace in related literature [Leshno and Levy, 2002, Dentcheva and Ruszczynski, 2003, Laidlaw et al., 2023, Cai et al., 2023, Hong and Tewari, 2024, Cen et al., 2026].

We also impose the following common regularity condition on the policy functions, standardly adopted in the theoretical analysis of most policy gradient methods [Pirotta et al., 2015, Xu et al., 2020, Zhang et al., 2020, Mondal and Aggarwal, 2024].

**Assumption B.2** (Boundedness and Lipschitz continuity of score function)**.** The score function $\nabla_\theta \log \pi_\theta$ is bounded and Lipschitz continuous in $\theta$. That is, there exist constants $B, G > 0$ such that for every state-action pair $(s, a)$ and all $\theta, \theta' \in \mathbb{R}^w$,

$$\begin{aligned} \|\nabla_\theta \log \pi_\theta(a \mid s)\| &\le B, \\ \|\nabla_\theta \log \pi_\theta(a \mid s) - \nabla_\theta \log \pi_{\theta'}(a \mid s)\| &\le G \|\theta - \theta'\|. \end{aligned} \tag{24}$$

*Remark* B.3. We note that the condition $\|\nabla_\theta \log \pi_\theta(a \mid s)\| \le B$, from Assumption B.2, for policy $\pi_\theta$ also implies boundedness of the score function for the induced trajectory density $p_\theta$. It suffices to note that for a trajectory $\tau = (s_0, a_0, \ldots, s_T, a_T)$,

$$\|\nabla_\theta \log p_\theta(\tau)\| = \left\| \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right\| \le TB. \tag{25}$$

In particular, in the episodic setting, where the length of any trajectory is at most $H$, $\|\nabla_\theta \log p_\theta(\tau)\| \leq HB =: B_p$. Similarly, the smoothness also extends to the trajectory density through the simple use of triangle inequality:

$$\begin{aligned}\mathbb{E}\left[\|\nabla_\theta \log p_\theta(\tau) - \nabla_\theta \log p_{\theta'}(\tau)\|\right] &\leq \mathbb{E}\left[\left\|\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) - \nabla_\theta \log \pi_{\theta'}(a_t \mid s_t)\right\|\right] \\ &\leq \mathbb{E}\left[\sum_{t=0}^{T} \|\nabla_\theta \log \pi_\theta(a_t \mid s_t) - \nabla_\theta \log \pi_{\theta'}(a_t \mid s_t)\|\right] \\ &\leq HG\|\theta - \theta'\|.\end{aligned} \tag{26}$$

We also name $G_p := HG$ for convenience.

**Proposition B.4** (Boundedness of $H_k$)**.** *For all $m \in M$ and $\theta \in \mathbb{R}^w$, $|H_k(\theta, m)| \leq 2$.*

*Proof.* For any $i \in [d]$ and $m, x \in M$, $0 \leq (m_i - x_i)_+^{k-1} \leq D$. Hence, $h_k(x, m) \leq D^{(k-1)d}$. We conclude

$$\begin{aligned}|H_k(\theta, m)| &= c_k \left|\mathbb{E}\left[h_k(R(\tau), m)\right] - \mathbb{E}\left[h_k(R(\tau_t), m)\right]\right| \\ &\leq c_k \left(\mathbb{E}\left[h_k(R(\tau), m)\right] + \mathbb{E}\left[h_k(R(\tau_t), m)\right]\right) \\ &\leq 2c_k D^{(k-1)d} \\ &= 2.\end{aligned} \tag{27}$$

□

**Proposition B.5** (Lipschitz continuity of $H_k$ in $m$)**.** *$H_k$ is globally Lipschitz continuous in $m$ in the $\ell^\infty$ metric with Lipschitz constant*

$$L := 2d(k-1)D^{-1}. \tag{28}$$

*That is, for any $m, m' \in M$,*

$$|H_k(\theta, m) - H_k(\theta, m')| \leq L\|m - m'\|. \tag{29}$$

*Proof.* Let us first begin by analyzing $\varphi_k : \mathbb{R} \to \mathbb{R}$, $\varphi_k(x) := (x)_+^{k-1}$. When $k = 2$, $\varphi_k$ is continuous and continuously differentiable everywhere except at $x = 0$. Its derivative, where it exists, is given by

$$\varphi_2'(x) = \begin{cases} 1, & x > 0, \\ 0, & x < 0. \end{cases} \tag{30}$$

Then, for all $t \in [-D, D] \setminus \{0\}$, $|\varphi_2'(x)| \leq 1$.

When $k \geq 3$, $\varphi_k$ is continuously differentiable everywhere, with derivative

$$\varphi_k'(x) = \begin{cases} (k-1)x^{k-2}, & t > 0, \\ 0, & t \leq 0. \end{cases} \tag{31}$$

Then, for all $x \in [-D, D]$, $|\varphi_k'(x)| \leq (k-1)D^{k-2}$. In particular, for any $k \geq 2$, $\varphi_k$ is a continuous, piecewise continuously differentiable function with bounded derivative in each piece. Hence, $\varphi_k$ is globally Lipschitz continuous on $[-D, D]$. Then, for any $m, m' \in M$, we have

$$\begin{aligned}|\varphi_k(m_i - x_i) - \varphi_k(m_i' - x_i)| &\leq (k-1)D^{k-2}\,|m_i - m_i'| \\ &\leq (k-1)D^{k-2}\,\|m - m'\|.\end{aligned} \tag{32}$$

We now move on to comparing the products

$$\begin{aligned}h_k(x, m) &= \prod_{i=1}^{d} \varphi_k(m_i - x_i), \\ h_k(x, m') &= \prod_{i=1}^{d} \varphi_k(m_i' - x_i).\end{aligned} \tag{33}$$

Let us define $A_i := \varphi_k(m_i - x_i)$, $A'_i := \varphi_k(m'_i - x_i)$ for convenience. Then,

$$\begin{aligned} |h_k(x,m) - h_k(x,m')| &= \left| \prod_{i=1}^{d} A_i - \prod_{i=1}^{d} A'_i \right| \\ &= \left| \sum_{j=1}^{d} (A_j - A'_j) \left( \prod_{i<j} A_i \right) \left( \prod_{i>j} A'_i \right) \right| \\ &\leq \sum_{j=1}^{d} |A_j - A'_j| \left( \prod_{i<j} |A_i| \right) \left( \prod_{i>j} |A'_i| \right). \end{aligned} \tag{34}$$

By Proposition B.4, we know that for all $i \in [d]$, $A_i, A'_i \leq D^{k-1}$. Then, for every term in the sum in (34),

$$\left( \prod_{i<j} |A_i| \right) \left( \prod_{i>j} |A'_i| \right) \leq D^{(k-1)(d-1)}. \tag{35}$$

Using this and (32) in (34), we have

$$\begin{aligned} |h_k(x,m) - h_k(x,m')| &\leq \sum_{j=1}^{d} (k-1) D^{k-2} \|m - m'\| \cdot D^{(k-1)(d-1)} \\ &= d(k-1) D^{d(k-1)-1} \|m - m'\|. \end{aligned} \tag{36}$$

Now, because this holds for every $x \in M$, it is also the case that

$$\begin{aligned} \left|\Delta_{R(\tau)}\right| &:= |\mathbb{E}\left[h_k(R(\tau), m)\right] - \mathbb{E}\left[h_k(R(\tau), m')\right]| \\ &\leq d(k-1) D^{d(k-1)-1} \|m - m'\|, \end{aligned} \tag{37}$$

and that

$$\begin{aligned} \left|\Delta_{R(\tau_t)}\right| &:= |\mathbb{E}\left[h_k(R(\tau_t), m)\right] - \mathbb{E}\left[h_k(R(\tau_t), m')\right]| \\ &\leq d(k-1) D^{d(k-1)-1} \|m - m'\|. \end{aligned} \tag{38}$$

Therefore,

$$\begin{aligned} |H_k(\theta,m) - H_k(\theta,m')| &= c_k \left|\mathbb{E}\left[h_k(R(\tau), m)\right] - \mathbb{E}\left[h_k(R(\tau_t), m)\right] - \mathbb{E}\left[h_k(R(\tau), m')\right] + \mathbb{E}\left[h_k(R(\tau_t), m')\right]\right| \\ &\leq c_k \left( \left|\Delta_{R(\tau)}\right| + \left|\Delta_{R(\tau_t)}\right| \right) \\ &\leq 2c_k d(k-1) D^{d(k-1)-1} \|m - m'\| \\ &= 2d(k-1) D^{-1} \|m - m'\|. \end{aligned} \tag{39}$$

□

**Corollary B.6** (Existence of maximizers)**.** *$\phi_k(\theta) = \max_{m \in M} H_k(\theta, m)$ is attained and $\arg\max_{m \in M} H_k(\theta, m)$ is a nonempty compact set.*

**Proposition B.7** (Gradient of $H_k$ in $\theta$)**.** *Let $m \in M$. The function $\theta \mapsto H_k(\theta, m)$ is differentiable, and*

$$\nabla_\theta H_k(\theta, m) = c_k \mathbb{E}\left[h_k(R(\tau), m) \nabla_\theta \log p_\theta(\tau)\right]. \tag{40}$$

*Proof.* We have

$$\begin{aligned} \nabla_\theta c_k \mathbb{E}\left[h_k(R(\tau), m)\right] &= \nabla_\theta c_k \int h_k(R(\tau), m) p_\theta(\tau) \mathrm{d}\tau \\ &= c_k \int h_k(R(\tau), m) \nabla_\theta p_\theta(\tau) \mathrm{d}\tau \\ &= c_k \int h_k(R(\tau), m) \nabla_\theta \log p_\theta(\tau) p_\theta(\tau) \mathrm{d}\tau \\ &= c_k \mathbb{E}\left[h_k(R(\tau), m) \nabla_\theta \log p_\theta(\tau)\right], \end{aligned} \tag{41}$$

where we have leveraged the boundedness of $\|\nabla_\theta \log p_\theta(\tau)\|$ by Remark B.3 and of $h_k$, as seen in Proposition B.4, to invoke the dominated convergence theorem for the result. Yet, this is precisely the gradient $\nabla_\theta H_k(\theta, m)$, since the $R(\tau_t)$ term in $H_k(\theta, m)$ is independent of $\theta$. □

**Proposition B.8** (Joint continuity of $\nabla_\theta H_k(\theta, m)$)**.** *$\nabla_\theta H_k(\theta, m)$ is jointly continuous in $(\theta, m)$.*

*Proof.* Let $(\theta_n, m_n) \to (\theta, m)$. Consider

$$\begin{aligned} \psi_n(\tau) &:= c_k h_k(R(\tau), m_n) \nabla_{\theta_n} \log p_{\theta_n}(\tau), \\ \psi(\tau) &:= c_k h_k(R(\tau), m) \nabla_\theta \log p_\theta(\tau). \end{aligned} \tag{42}$$

For each $n$, $|\psi_n(\tau)| \le D^{(k-1)d} g(\tau)$, and $\psi_n(\tau) \to \psi(\tau)$ pointwise. Then, the dominated convergence theorem yields

$$c_k \mathbb{E}\left[\psi_n\right] \to c_k \mathbb{E}\left[\psi\right]. \tag{43}$$

That is, $c_k \mathbb{E}[\psi(\tau)] := \nabla_\theta H_k(\theta, m)$ is jointly continuous in $(\theta, m)$. □

With this, we have now stated every condition that is required for Danskin's theorem, which follows.

**Theorem B.9** (Danskin's theorem Danskin [2012])**.** *Let $M$ be a nonempty compact topological space, and $f : \mathbb{R}^w \times M \to \mathbb{R}$ be such that $\theta \mapsto f(\theta, m)$ is differentiable for every $m \in M$, and $\nabla_\theta f(\theta, m)$ is continuous on $\mathbb{R}^w \times M$. Let $M^*(\theta) := \arg\max_{m \in M} f(\theta, m)$.*

*Then, the corresponding max-function $\nu(\theta) := \max_{m \in M} f(\theta, m)$ is locally Lipschitz continuous, directionally differentiable, and for any direction $u$, its directional derivative satisfies*

$$D_u \nu(\theta) = \max_{m \in M^*(\theta)} u^\top \nabla_\theta f(\theta, m). \tag{44}$$

*In particular, if for some $\theta \in \mathbb{R}^w$, the set $M^*(\theta) = \{m^*\}$, then the max-function is differentiable at $\theta$ and*

$$\nabla \nu(\theta) = \nabla_\theta f(\theta, m^*). \tag{45}$$

*Furthermore, the subdifferential of $\nu(\theta)$ is given by*

$$\partial \nu(\theta) = \operatorname{conv}\left\{\nabla_\theta f(\theta, m^*) : m^* \in M^*(\theta)\right\}. \tag{46}$$

*Remark* B.10. A locally Lipschitz continuous is almost everywhere differentiable, by Rademacher's theorem Heinonen [2005]. The max-function $\nu(\theta)$ of Theorem B.9 is therefore also almost everywhere differentiable.

**Corollary B.11** (Valid descent directions Madry et al. [2018])**.** *Let $\bar{m}$ be such that $\bar{m} \in M$ and is a maximizer for $\max_{m \in M} H_k(\theta, m)$. Then, as long as it is nonzero, $-\nabla_\theta H(\theta, \bar{m})$ is a valid descent direction for $\phi_k(\theta) = \max_{m \in M} H_k(\theta, m)$.*

*Proof.* Applying Theorem B.9 to $H_k(\theta, m)$, we see that the directional derivative in direction $\bar{u} := \nabla_\theta H_k(\theta, \bar{m})$ satisfies

$$D_{\bar{u}} \phi(\theta) = \max_{m \in M^*(\theta)} \bar{u}^\top \nabla_\theta H_k(\theta, m) = \bar{u}^\top \bar{u} = \|\nabla_\theta H_k(\theta, \bar{m})\|_2^2 \ge 0, \tag{47}$$

Therefore, $D_{-\bar{u}} \phi(\theta) \le 0$, with strict inequality if the gradient is nonzero. □

## C DIFFERENCE-OF-CONVEX OPTIMIZATION AND CCCP

In this section, we will justify and motivate a first-order method for the solution of the subproblem

$$\max_{m \in M} \frac{1}{N} \sum_{i=1}^{N} \left( h_k(R_{t,\bar{t},i}, m) - h_k(R_{t,i}, m) \right). \tag{48}$$

We first state the following result:

**Proposition C.1.** *$h_k(x, m) := \prod_{i=1}^{d} (m_i - x_i)_+^{k-1}$ can be decomposed into the difference of two convex functions in $m$. We write that $h_k(x, m)$ **is DC in** $m$.*

*Proof.* We use the polarization identity for the product of $d$ factors, as presented, for instance, by Schetzen [2006] and Nelson [2007] to write:

$$h_k(x, m) = \frac{1}{d!} \sum_{\mathcal{S} \subseteq [d]} (-1)^{d-|\mathcal{S}|} \left( \sum_{i \in \mathcal{S}} (m_i - x_i)_+^{k-1} \right)^d. \tag{49}$$

Firstly, for any $x$, the function $m \mapsto (m_i - x_i)_+^{k-1}$ is convex, as it is simply the composition of the nondecreasing convex function $t \mapsto t^{k-1}$ with the composition of the convex function $t \mapsto t_+$ and the affine function $(m_i - x_i)$. A nonnegative weighted sum of such functions, therefore, remains convex. Finally, once again composing this with the nondecreasing convex function $t \mapsto t^d$ preserves convexity. The term $(\sum_{i \in \mathcal{S}} (m_i - x_i)_+^{k-1})^d$ is therefore convex in $m$, which means that $h_k$ is a sum of such convex functions of $m$, some of them having coefficient 1 and the others having coefficient $-1$. We can then conclude:

$$h_k(x, m) = h_k^+(x, m) - h_k^-(x, m), \tag{50}$$

where

$$h_k^+(x, m) := \sum_{\substack{\mathcal{S} \subseteq [d] \\ d-|\mathcal{S}| \text{ even}}} \left( \sum_{i \in \mathcal{S}} (m_i - x_i)_+^{k-1} \right)^d, \tag{51}$$

and

$$h_k^-(x, m) := \sum_{\substack{\mathcal{S} \subseteq [d] \\ d-|\mathcal{S}| \text{ odd}}} \left( \sum_{i \in \mathcal{S}} (m_i - x_i)_+^{k-1} \right)^d, \tag{52}$$

where both $h_k^+(x, m)$ and $h_k^-(x, m)$ are convex in $m$. □

**Corollary C.2.**

$$\hat{\mathcal{L}}_k(X, Y, m) := \frac{1}{N} \sum_{i=1}^{N} \left( h_k(X_i, m) - h_k(Y_i, m) \right) \tag{53}$$

*is DC in $m$. Similarly to Proposition C.1, we write $\hat{\mathcal{L}}_k(X, Y, m) = \hat{\mathcal{L}}_k^+(X, Y, m) - \hat{\mathcal{L}}_k^-(X, Y, m)$, where*

$$\hat{\mathcal{L}}_k^+(X, Y, m) := \frac{1}{N} \sum_{i=1}^{N} h_k^+(X_i, m) + h_k^-(Y_i, m), \tag{54}$$

*and*

$$\hat{\mathcal{L}}_k^-(X, Y, m) := \frac{1}{N} \sum_{i=1}^{N} h_k^-(X_i, m) + h_k^+(Y_i, m). \tag{55}$$

**Algorithm 2** CCCP

**Require:** DC function $f(m) := f^+(m) - f^-(m)$, number of iterations $T$, stopping threshold $\epsilon$

Initialize $m_0 \in M$ arbitrarily
**for** $t = 0, \ldots, T-1$ **do**
  Pick $g_t \in \partial f^+(m_t)$
  $m_{t+1} \leftarrow \arg\max_{m \in M} \langle g_t, m \rangle - f^-(m_t)$
**end for**
**if** $\|m_{t+1} - m_t\| \leq \epsilon$ **then**
  **break**
**end if**
**return** $m_{t+1}$

We now intend to leverage the results of Corollary C.2 and Proposition B.5 to use first-order methods in the solution of Problem (48). In particular, the discrete iterative concave-convex procedure (CCCP) algorithm (Algorithm 2) Tao and An [1997], Yuille and Rangarajan [2001] can be used to maximize $\hat{\mathcal{L}}_k$ in $m$, with the following guarantee to converge to a first-order stationary point:

**Theorem C.3** (Ascent, convergence, and stationarity of CCCP Tao and An [1997], Yuille and Rangarajan [2001])**.** *CCCP (Algorithm 2) satisfies*

1. ***Monotone ascent.*** *The sequence of values $f(m_t)$ is nondecreasing, i.e., $f(m_{t+1}) \geq f(m_t)$ for all $t$. The inequality is strict whenever the subproblem is solved exactly and $m_{t+1} \neq m_t$.*
2. ***Convergence.*** *If $f$ is bounded above on $M$, then $f(m_t)$ converges and the sequence of iterates $\{m_t\}$ is bounded.*
3. ***Stationarity.*** *Every limit point $m^*$ of $m_t$ satisfies the DC stationarity condition $\mathbf{0} \in \partial g(m^*) - \partial h(m^*) + N_M(m^*)$, where $N_M(m^*)$ is the normal cone of $M$ at $m^*$.*

In intuitive terms, CCCP can be thought of as performing minorization-maximization by linearizing the convex part of a DC function to obtain a concave lower bound, which is subsequently maximized using principles of convex optimization.

## D ANALYSIS OF ORDO

The analysis of ORDO depends on the solution of the maximization subproblem within ORDO, expressed in line 9 of the algorithm. As evident from the results of Theorem C.3, and as can be anticipated from the nonconvex nature of the problem, we can only realistically expect a guarantee to converge to a stationary point, and not necessarily to a true maximizer of $\hat{\mathcal{L}}_k(R_{t,\bar{t}}, R_t, m)$. The problem is further multiplied by the fact that the function we are maximizing is $\hat{\mathcal{L}}_k$, the empirical estimate of the function we actually want to maximize, which is $\mathcal{L}_k$.

We will abstract away the suboptimality of our maximizer estimate $\hat{m}$ with the following assumption:

**Assumption D.1** (Suboptimality bound on $\hat{m}$)**.** At any pair of iteration indices $(t, \bar{t})$, line 8 of Algorithm 1 returns a solution $\hat{m}$ which satisfies

$$\mathbb{E}\left[\operatorname{dist}\left(\hat{m}, M^*(\theta_{t,\bar{t}})\right)^2 \mid \theta_{t,\bar{t}}\right] \leq \frac{\delta_m}{N} + \delta_c, \tag{56}$$

for some constants $\delta_m, \delta_c \geq 0$.

To make proper use of Assumption D.1, we also prove the following result.

**Proposition D.2** (Lipschitz continuity of $\nabla_\theta H_k$ in $m$)**.** *For each $\theta \in \mathbb{R}^w$, $m \mapsto \nabla_\theta H_k(\theta, m)$ is $L_m$-Lipschitz continuous, with $L_m := B_p d(k-1)D^{-1}$, that is, for any $m, m' \in M$*

$$\|\nabla_\theta H_k(\theta, m) - \nabla_\theta H_k(\theta, m')\| \leq L_m \|m - m'\|. \tag{57}$$

*Proof.* Simply noting that

$$\nabla_\theta H_k(\theta, m) - \nabla_\theta H_k(\theta, m') = c_k \mathbb{E}\left[\left(h_k(R(\tau), m) \; - \; h_k(R(\tau), m')\right) \nabla_\theta \log p_\theta(\tau)\right] \tag{58}$$

and using Assumption B.2 (through Remark B.3) to bound $|\nabla_\theta \log p_\theta(\tau)| \leq B_p$ and (36) to bound

$$|h_k(R(\tau), m) - h_k(R(\tau), m')| \leq d(k-1)D^{d(k-1)-1} \|m - m'\|, \tag{59}$$

we have

$$\|\nabla_\theta H_k(\theta, m) - \nabla_\theta H_k(\theta, m')\| \leq L_m \|m - m'\|. \tag{60}$$

with $L_m := B_p d(k-1)D^{-1}$. □

Now, Assumption D.1 and Proposition D.2 together allow us to bound the bias of our stochastic subgradient, defined in line 10 of Algorithm 1.

**Proposition D.3** (Bounded bias of $g_{t,\bar{t}}$)**.** *At any pair of iteration indices $(t, \bar{t})$, let the random variable $m^*(\theta_{t,\bar{t}})$ denote the closest maximizer to $\hat{m}$, i.e., $m^*(\theta_{t,\bar{t}}) \in \arg\min_{m \in M^*(\theta_{t,\bar{t}})} \|m - \hat{m}\|$. Let $v(\theta_{t,\bar{t}}) := \mathbb{E}[\nabla_\theta H_k(\theta_{t,\bar{t}}, m^*(\theta_{t,\bar{t}})) \mid \theta_{t,\bar{t}}]$. Then, the gradient estimator $g_{t,\bar{t}}$ given by line 9 of Algorithm 1 is an estimator of a subgradient of $\phi_k(\theta_{t,\bar{t}})$ with bounded bias, satisfying*

$$\left\|\mathbb{E}\left[g_{t,\bar{t}} \mid \theta_{t,\bar{t}}\right] - v(\theta_{t,\bar{t}})\right\|^2 \leq L_m^2 \left(\frac{\delta_m}{N} + \delta_c\right) := \beta_N^2. \tag{61}$$

*Proof.* First, note that $v(\theta_{t,\bar{t}}) \in \operatorname{conv}\{\nabla_\theta H_k(\theta_{t,\bar{t}}, m) : m \in M^*(\theta_{t,\bar{t}})\}$, so it is a valid subgradient of $\phi_k$ at $\theta_{t,\bar{t}}$ by Theorem B.9. Now, we write, using Proposition D.2,

$$\begin{aligned}
\left\|\mathbb{E}\left[g_{t,\bar{t}} \mid \theta_{t,\bar{t}}\right] - v(\theta_{t,\bar{t}})\right\|^2 &\leq \left\|\mathbb{E}\left[\nabla_\theta H_k(\theta, \hat{m}) - \nabla_\theta H_k(\theta, m^*(\theta_{t,\bar{t}})) \mid \theta_{t,\bar{t}}\right]\right\|^2 \\
&\leq \mathbb{E}\left[\left\|\nabla_\theta H_k(\theta, \hat{m}) - \nabla_\theta H_k(\theta, m^*(\theta_{t,\bar{t}}))\right\|^2 \mid \theta_{t,\bar{t}}\right] \\
&\leq L_m^2 \mathbb{E}\left[\left\|\hat{m} - m^*(\theta_{t,\bar{t}})\right\|^2 \mid \theta_{t,\bar{t}}\right].
\end{aligned} \tag{62}$$

But, by construction, $\|\hat{m} - m^*(\theta_{t,\bar{t}})\| = \mathrm{dist}(\hat{\mathrm{m}}, \mathrm{M}^*(\theta_{\mathrm{t},\bar{\mathrm{t}}}))$. Hence,

$$\begin{aligned}\left\|\mathbb{E}\left[g_{t,\bar{t}} \mid \theta_{t,\bar{t}}\right] - v(\theta_{t,\bar{t}})\right\|^2 &\leq L_m^2 \mathbb{E}\left[\mathrm{dist}(\hat{m}, M^*(\theta_{t,\bar{t}}))^2 \mid \theta_{t,\bar{t}}\right] \\ &\leq L_m^2 \left(\frac{\delta_m}{N} + \delta_c\right)\end{aligned} \tag{63}$$

by Assumption D.1. □

Finally, we state the following elementary results that generalizes the nice property of first-order methods, namely, that the updates are bounded by the learning rate.

**Proposition D.4** (Boundedness of stochastic gradients)**.** *Consider the stochastic gradient estimator defined in Algorithm 1, line 9:*

$$g_{t,\bar{t}} := \frac{c_k}{N} \sum_{i=1}^{N} h_k(R_{t,\bar{t},i}, \hat{m}) \nabla_\theta \log p_{\theta_{t,\bar{t}}}(\tau_{t,\bar{t},i}). \tag{64}$$

*Under Assumption B.1 and Assumption B.2, the norm of the gradient estimator is uniformly bounded almost surely by:*

$$\|g_{t,\bar{t}}\| \leq G_{\max} := B_p. \tag{65}$$

*Proof.* By the triangle inequality on the sum and the scaling property of norms:

$$\|g_{t,\bar{t}}\| \leq \frac{c_k}{N} \sum_{i=1}^{N} |h_k(R_{t,\bar{t},i}, \hat{m})| \cdot \|\nabla_\theta \log p_{\theta_{t,\bar{t}}}(\tau_{t,\bar{t},i})\|. \tag{66}$$

From Proposition B.4, the kernel $h_k$ is bounded by $D^{(k-1)d}$. From Remark B.3, the trajectory score function is bounded by $B_p = HB$. Substituting these uniform bounds:

$$\|g_{t,\bar{t}}\| \leq \frac{c_k}{N} \sum_{i=1}^{N} (D^{(k-1)d})(B_p) = c_k D^{(k-1)d} B_p = B_p. \tag{67}$$

□

With Algorithm 1 now fully specified, two natural questions arise: Is it guaranteed to converge to and return a satisfactory solution? If yes, what conclusion can be drawn about its rate of convergence? To be able to answer these questions, we first state the following proposition.

**Proposition D.5** ($\rho$-smoothness of $H_k$ in $\theta$)**.** *For each $m \in M$, $\theta \mapsto H_k(\theta, m)$ is $\rho$-smooth, with $\rho := G_p + B_p^2$, that is, for any $\theta, \theta' \in \mathbb{R}^w$*

$$\|\nabla_\theta H_k(\theta, m) - \nabla_\theta H_k(\theta', m)\| \leq \rho \|\theta - \theta'\|. \tag{68}$$

*Proof.* For notational convenience, let us define $S_\theta(\tau) := \nabla_\theta \log p_\theta(\tau)$ as a shorthand for the score function. We write

$$\Delta := \nabla_\theta H_k(\theta, m) - \nabla_\theta H_k(\theta', m) = c_k \left(\mathbb{E}_{p_\theta}\left[h_k(R(\tau), m) S_\theta(\tau)\right] - \mathbb{E}_{p_{\theta'}}\left[h_k(R(\tau), m) S_{\theta'}(\tau)\right]\right). \tag{69}$$

Adding and subtracting the "mixed term" $c_k \mathbb{E}_{p_\theta}\left[h_k(R(\tau), m) S_{\theta'}(\tau)\right]$, we get

$$\Delta = c_k \mathbb{E}_{p_\theta}\left[h_k(R(\tau), m)(S_\theta - S_{\theta'})(\tau)\right] + c_k \left(\mathbb{E}_{p_\theta}\left[h_k(R(\tau), m) S_{\theta'}(\tau)\right] - \mathbb{E}_{p_{\theta'}}\left[h_k(R(\tau), m) S_{\theta'}(\tau)\right]\right). \tag{70}$$

Assumption B.2 (through Remark B.3), Proposition B.4, and Jensen's inequality allow us to upper-bound the first term above as follows:

$$\|c_k \mathbb{E}_{p_\theta}\left[h_k(R(\tau), m)(S_\theta - S_{\theta'})(\tau)\right]\| \leq G_p \|\theta - \theta'\|. \tag{71}$$

The second term requires a more sophisticated approach. We would like to proceed by deriving an upper bound on $\|p_\theta - p_{\theta'}\|_{\mathrm{TV}}$, the total variation distance between the two trajectory densities. By the fundamental theorem of calculus, we write

$$\log \frac{p_\theta(\tau)}{p_{\theta'}(\tau)} = \int_0^1 \langle S_{\theta_\lambda}(\tau), \theta - \theta' \rangle \, \mathrm{d}\lambda, \tag{72}$$

where $\theta_\lambda := \theta' + \lambda(\theta - \theta')$. Hence,

$$\left| \log \frac{p_\theta(\tau)}{p_{\theta'}(\tau)} \right| \leq \int_0^1 \|S_{\theta_\lambda}\| \, \mathrm{d}\lambda \, \|\theta - \theta'\| \leq B_p \, \|\theta - \theta'\| \, . \tag{73}$$

This bound on the log-likelihood ratio allows us to write

$$\|p_\theta - p_{\theta'}\|_{\mathrm{TV}} \leq \frac{1}{2} B_p \, \|\theta - \theta'\| \, . \tag{74}$$

We now turn to upper-bounding the second term in (70). By Proposition B.4, Assumption B.2 (through Remark B.3), and (74),

$$\begin{aligned} \|c_k(\mathbb{E}_{p_\theta}\left[h_k(R(\tau), m) S_{\theta'}(\tau)\right] - \mathbb{E}_{p_{\theta'}}\left[h_k(R(\tau), m) S_{\theta'}(\tau)\right])\| &\leq 2c_k \sup_\tau \|h_k(R(\tau), m) S_{\theta'}(\tau)\| \\ &\leq 2B_p \, \|p_\theta - p_{\theta'}\|_{\mathrm{TV}} \\ &\leq B_p^2 \, \|\theta - \theta'\| \, . \end{aligned} \tag{75}$$

Now, combining (71) and (75) in (70),

$$\Delta \leq (G_p + B_p^2) \, \|\theta - \theta'\| \, . \tag{76}$$

Since this upper bound is independent of $m$, letting $\rho = G_p + B_p^2$, $\theta \mapsto H_k(\theta, m)$ is $\rho$-smooth uniformly in $m$. □

**Lemma D.6** (Pointwise maximum of $\rho$-smooth functions [López-Rivera et al., 2026])**.** *Let $f(x, y)$ be a function such that $x \mapsto f(x, y)$ is $\rho$-smooth uniformly in $y$. Then, the max-function $g(x) := \sup_y f(x, y)$ is $\rho$-weakly convex.*

**Corollary D.7** ($\rho$-weak convexity of $\phi_k$)**.** *$\phi_k(\theta) := \max_{m \in M} H_k(\theta, m)$ is $\rho$-weakly convex.*

*Proof.* By Proposition D.5, $\theta \mapsto H_k(\theta, m)$ is $\rho$-smooth uniformly in $m$. Applying Lemma D.6, $\phi_k(\theta)$ is $\rho$-weakly convex. □

With these last insights, a clearer view of our optimization scheme is shaping up. At each outer iteration $t$ of Algorithm 1, we are in fact minimizing a $\rho$-weakly convex function $\phi_{k,t}$ through a stochastic subgradient method. As shown by Proposition D.3, our stochastic subgradients are unfortunately biased because of the potential suboptimality of the solution $\hat{m}$ of the inner maximization routine, but luckily, part of this bias is controlled in $O(N)$. The next natural question that arises is whether we can draw a similar conclusion on the variance of our stochastic subgradients. The next result affirms.

**Proposition D.8** (Bounded variance of $g_{t,\bar{t}}$)**.** *Let $\mu_{t,\bar{t}} := \mathbb{E}[g_{t,\bar{t}} \mid \theta_{t,\bar{t}}]$ at any pair of iteration indices $(t, \bar{t})$ of Algorithm 1. Then,*

$$\mathbb{E}\left[ \left\| g_{t,\bar{t}} - \mu_{t,\bar{t}} \right\|^2 \mid \theta_{t,\bar{t}} \right] \leq \frac{4B_p^2}{N}. \tag{77}$$

*Proof.* Let

$$g_{t,\bar{t},i} := h_k(R_{t,\bar{t},i}, \hat{m}) \nabla_\theta \log p_{\theta_{t,\bar{t}}}(\tau_{t,\bar{t},i}), \tag{78}$$

so that $g_{t,\bar{t}} := \sum_{i=1}^N g_{t,\bar{t},i}$. Firstly, by Remark B.3 and Proposition B.4, for each $i \in [N]$,

$$\begin{aligned} \left\| g_{t,\bar{t},i} \right\| &\leq B_p, \\ \left\| \mathbb{E}\left[ g_{t,\bar{t},i} \mid \theta_{t,\bar{t}} \right] \right\| &\leq B_p. \end{aligned} \tag{79}$$

Then,

$$\mathbb{E}\left[\left\|g_{t,\bar{t}} - \mu_{t,\bar{t}}\right\|^2 \mid \theta_{t,\bar{t}}\right]. \tag{80}$$

Hence,

$$\mathbb{E}\left[\left\|g_{t,\bar{t}} - \mu_{t,\bar{t}}\right\|^2 \mid \theta_{t,\bar{t}}\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} g_{t,\bar{t},i} - \mathbb{E}\left[g_{t,\bar{t},i} \mid \theta_{t,\bar{t}}\right]\right\|^2 \mid \theta_{t,\bar{t}}\right] \leq \frac{4B_p^2}{N}. \tag{81}$$

□

We have now set the stage for the optimization problem at hand. At each outer iteration $t$, we will attempt to minimize a weakly convex function $\phi_{k,t}$ through a stochastic subgradient method, where our first-order oracle has bounded bias and variance scaling in $O(1/N)$.[1] Variants of this problem have been extensively studied in the literature [Davis and Drusvyatskiy, 2018, Davis et al., 2018, Davis and Drusvyatskiy, 2019], and our analysis will adopt an assumption from, and follow similar lines as that of Davis and Drusvyatskiy [2018]. Here, we state the assumption:

**Assumption D.9** ($\kappa_{\text{sh}}$-sharpness of $\phi_{k,t}$ [Davis and Drusvyatskiy, 2019])**.** For all $t \in \{0, \ldots, T-1\}$, the set of minimizers $\Theta_t^* := \arg\min_{\theta \in \mathbb{R}^w} \phi_{k,t}(\theta)$ is nonempty, and the inequality

$$\phi_{k,t}(\theta) - \phi_{k,t}^* \geq \kappa_{\text{sh}} \cdot \text{dist}(\theta, \Theta_t^*) \tag{82}$$

hold for all $\theta \in \mathbb{R}^w$.

We emphasize that we must assume such a global regularity condition in order to be able to make a global qualification about the solution returned by the algorithm, such as $\epsilon$-non-dominatedness. Sharpness is one such sufficient condition. Although we cannot claim that it may hold for every MORL problem, it is nevertheless structurally plausible for a worst-orthant objective: when the active violating orthant is nondegenerate, the pointwise maximum creates a margin or kink rather than a flat averaged loss. This is especially transparent for finite-net and Soft-ORDO objectives, which are maxima or log-sum-exp surrogates over finitely many orthant generators.

Before we state the theorem on the convergence of Algorithm 1 (ORDO), we must first state a useful lemma.

**Lemma D.10** (Concentration near minimizer)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a $\rho$-weakly convex, $\kappa_{\text{sh}}$-sharp function, and $\mathcal{X}^* := \arg\min_{x \in \mathcal{X}} f(x)$. Consider the biased stochastic subgradient update*

$$x_{t+1} = x_t - \alpha g_t, \quad g_t = v_t + b_t, \tag{83}$$

*where $v_t \in \partial f(x_t)$, $\|b_t\| \leq \beta < \kappa_{\text{sh}}/2$ and $\|g_t\| \leq G$ for all $t$, almost surely. Choose the step size*

$$\alpha = \min\left\{\frac{(\kappa_{\text{sh}} - 2\beta)\kappa_{\text{sh}}}{4G^2\rho}, \frac{\kappa_{\text{sh}}}{4G\rho}\right\}. \tag{84}$$

*Furthermore, suppose we initialize $x_0$ such that $\|x_0 - x^*\| \leq \kappa_{\text{sh}}/2\rho$ for some $x^* \in \mathcal{X}^*$. Then,*

$$\mathbb{P}\left(\max_{0 \leq t \leq T} \|x_t - x^*\| \leq \frac{\kappa_{\text{sh}}}{\rho}\right) \geq 1 - \exp\left(-\frac{\Delta^2}{2TB(\alpha)^2}\right), \tag{85}$$

*where $\Delta := \frac{7\kappa_{\text{sh}}^2}{16\rho^2}$ and*

$$B(\alpha) := \alpha G\left(\frac{2\kappa_{\text{sh}}}{\rho} + \alpha G\right) + (\kappa_{\text{sh}} - 2\beta)\frac{\kappa_{\text{sh}}}{4\rho}\alpha. \tag{86}$$

*That is, the sequence of iterates found by the algorithm remains within $\kappa_{\text{sh}}/\rho$ of $x^*$ with high probability.*

[1] Here, we readopt the $t$ subscript we dropped earlier, because the context of Algorithm 1 (ORDO) will require to distinguish between $\phi_{k,t}$ at different outer iterations $t$ of the algorithm.

*Proof.* Firstly, we will show that for any $x_1^*, x_2^* \in \mathcal{X}^*$ such that $x_1^* \neq x_2^*$, $\|x_1^* - x_2^*\| \geq 4\kappa_{\text{sh}}/\rho$. Let $\delta := \|x_1^* - x_2^*\|$, $\bar{x} := \frac{1}{2}(x_1^* + x_2^*)$, and $f^* := \min_{x \in \mathcal{X}} f(x)$. By $\kappa_{\text{sh}}$-sharpness,

$$f(\bar{x}) - f^* \geq \kappa_{\text{sh}} \cdot \text{dist}(\bar{x}, \mathcal{X}^*) \geq \kappa_{\text{sh}} \cdot \frac{\delta}{2}. \tag{87}$$

By $\rho$-weak convexity of $f$, we know that $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex. Then,

$$\begin{aligned} f(\bar{x}) + \frac{\rho}{2}\|\bar{x}\|^2 &\leq \frac{1}{2}\left(f(x_1^*) + \frac{\rho}{2}\|x_1^*\|^2 + f(x_2^*) + \frac{\rho}{2}\|x_2^*\|^2\right) \\ &= f^* + \frac{\rho}{4}\left(\|x_1^*\|^2 + \|x_2^*\|^2\right). \end{aligned} \tag{88}$$

Hence,

$$f(\bar{x}) \leq f^* + \frac{\rho}{4}\left(\|x_1^*\|^2 - 2\|\bar{x}\|^2 + \|x_2^*\|^2\right). \tag{89}$$

Now let $x_1^* - x_2^* =: 2u$. We have

$$\begin{aligned} \|x_1^*\|^2 + \|x_2^*\|^2 &= \|\bar{x} + u\|^2 + \|\bar{x} - u\|^2 \\ &= 2\|\bar{x}\|^2 + 2\|u\|^2, \end{aligned} \tag{90}$$

or alternatively,

$$\|x_1^*\|^2 + \|x_2^*\|^2 - 2\|\bar{x}\|^2 = 2\|u\|^2 = 2 \cdot \frac{\|x_1^* - x_2^*\|^2}{4} = \frac{\delta^2}{2}. \tag{91}$$

Using this in (89),

$$f(\bar{x}) \leq f^* + \frac{\rho}{8}\delta^2. \tag{92}$$

Finally, once again using $\kappa_{\text{sh}}$-sharpness,

$$\kappa_{\text{sh}} \cdot \frac{\delta}{2} \leq f(\bar{x}) - f^* \leq \frac{\rho}{8}\delta^2. \tag{93}$$

Rearranging, we have the result

$$\delta \geq \frac{4\kappa_{\text{sh}}}{\rho}. \tag{94}$$

In particular, this means that any point $x \in \mathcal{X}$ with $\|x - x^*\| \leq \kappa_{\text{sh}}/\rho$ is strictly closer to $x^*$ than to any other minimizer, i.e., $\text{dist}(x, \mathcal{X}^*) = \|x - x^*\|$.

Now, for any $x \in \mathcal{X}$ such that $\|x - x^*\| \leq \kappa_{\text{sh}}/\rho$, and $v \in \partial f(x)$, we have by $\rho$-weak convexity and $\kappa_{\text{sh}}$-sharpness,

$$\begin{aligned} \langle v, x - x^* \rangle &\geq f(x) - f(x^*) - \frac{\rho}{2}\|x^* - x\|^2 \\ &\geq \kappa_{\text{sh}} \cdot \|x^* - x\| - \frac{\rho}{2}\|x^* - x\|^2. \end{aligned} \tag{95}$$

Since $\|x - x^*\| \leq \kappa_{\text{sh}}/\rho$,

$$\|x - x^*\|^2 \leq \frac{\kappa_{\text{sh}}}{\rho}\|x - x^*\|, \tag{96}$$

hence

$$\langle v, x - x^* \rangle \geq \kappa_{\text{sh}} \cdot \|x - x^*\| - \frac{\kappa_{\text{sh}}}{2}\|x - x^*\| = \frac{\kappa_{\text{sh}}}{2}\|x - x^*\|. \tag{97}$$

We will now construct an annulus which **(i)** the sequence of iterates must visit for at least one iteration before leaving the desired neighborhood, and **(ii)** has an inward drift property that pushes the iterates, on expectation, uniformly towards $x^*$. We choose the inner and outer radii of this annulus to be

$$r_{\text{in}} := \frac{\kappa_{\text{sh}}}{2\rho}, \qquad r_{\text{out}} := \frac{\kappa_{\text{sh}}}{\rho}. \tag{98}$$

Let $d_t := x_t - x^*$, and $R_t := \|d_t\|$. We also define the the *time of first entry into the annulus*

$$\sigma := \inf\{t \geq 0 : R_t \geq r_{\text{in}}\}, \tag{99}$$

and the *time of first exit outside the annulus*

$$\tau := \inf\{t \geq 0 : R_t \geq r_{\text{out}}\}. \tag{100}$$

Let us begin with property **(i)**. Firstly,

$$\|x_{t+1} - x_t\| = \alpha\|g_t\| \leq \alpha G, \tag{101}$$

so

$$|R_{t+1} - R_t| = |\|x_{t+1} - x^*\| - \|x_t - x^*\|| \leq \|x_{t+1} - x_t\| \leq \alpha G. \tag{102}$$

Then, $R_\sigma \leq R_{\sigma-1} + \alpha G \leq r_{\text{in}} + \alpha G$. Also, by choice of $\alpha$, $\alpha G \leq r_{\text{out}}/4$. Then,

$$R_\sigma \leq r_{\text{in}} + \alpha G \leq \frac{1}{2}r_{\text{out}} + \frac{1}{4}r_{\text{out}} = \frac{3}{4}r_{\text{out}} < r_{\text{out}}. \tag{103}$$

This means that all sample paths leaving the $r_{\text{out}}$-neighborhood of $x^*$ must first visit the annulus, or in other words, $\sigma < \tau$ in all sample paths.

Now, we will move on to property **(ii)**. We have

$$\begin{aligned} R_{t+1}^2 = \|x_{t+1} - x^*\|^2 = \|d_t - \alpha g_t\|^2 &= \|d_t\|^2 - 2\alpha\langle g_t, d_t\rangle + \alpha^2\|g_t\|^2 \\ &= R_t^2 - 2\alpha\langle v_t + b_t, d_t\rangle + \alpha^2\|g_t\|^2. \end{aligned} \tag{104}$$

Taking the expectation of both sides, conditioned on $x_t$,

$$\mathbb{E}[R_{t+1}^2 \mid x_t] = R_t^2 - 2\alpha\langle v_t + b_t, d_t\rangle + \alpha^2\mathbb{E}[\|g_t\|^2 \mid x_t]. \tag{105}$$

For any $R_t \leq r_{\text{out}}$, by (97),

$$\langle v_t, d_t\rangle \geq \frac{\kappa_{\text{sh}}}{2}R_t. \tag{106}$$

Also, $-2\alpha\langle b_t, d_t\rangle \leq 2\alpha\beta R_t$. Then, for $R_t \leq r_{\text{out}}$,

$$\begin{aligned} \mathbb{E}[R_{t+1}^2 \mid x_t] &= R_t^2 - 2\alpha\frac{\kappa_{\text{sh}}}{2}R_t + 2\alpha\beta R_t + \alpha^2 G^2 \\ &= R_t^2 - \alpha(\kappa_{\text{sh}} - 2\beta)R_t + \alpha^2 G^2. \end{aligned} \tag{107}$$

Let $\tilde{\kappa}_{\text{sh}} := \kappa_{\text{sh}} - 2\beta$. By assumption, $\beta < \kappa_{\text{sh}}/2$, so $\tilde{\kappa}_{\text{sh}} > 0$. Hence,

$$\mathbb{E}[R_{t+1}^2 - R_t^2 \mid x_t] \leq -\alpha\tilde{\kappa}_{\text{sh}}R_t + \alpha^2 G^2, \tag{108}$$

whenever $t < \tau$.

Specifically, within the annulus (i.e., $\sigma \leq t < \tau$), we have the uniform upper bound

$$\mathbb{E}[R_{t+1}^2 - R_t^2 \mid x_t] \leq -\alpha\tilde{\kappa}_{\text{sh}}\, r_{\text{in}} + \alpha^2 G^2 = -\alpha(\tilde{\kappa}_{\text{sh}}\, r_{\text{in}} - \alpha G^2). \tag{109}$$

Once again, by choice of $\alpha$, we have

$$\alpha \leq \frac{(\kappa_{\text{sh}} - 2\beta)\kappa_{\text{sh}}}{4G^2\rho} = \frac{\tilde{\kappa}_{\text{sh}}\, r_{\text{in}}}{2G^2}, \tag{110}$$

which ensures $\tilde{\kappa}_{\text{sh}}\, r_{\text{in}} - \alpha G^2 = \tilde{\kappa}_{\text{sh}}\, r_{\text{in}}/2 > 0$. Then, letting $\zeta := \tilde{\kappa}_{\text{sh}}\, r_{\text{in}}/2$, we have

$$\mathbb{E}[R_{t+1}^2 - R_t^2 \mid x_t] \leq -\zeta\alpha < 0. \tag{111}$$

We have now shown properties **(i)** and **(ii)**. The rest of the proof will leverage these properties to invoke the Azuma-Hoeffding concentration inequality [Chung and Lu, 2006] on a judiciously-defined stochastic process. Let us first define the stochastic process of the *stopped radius* $\tilde{R}_t := R_{\min\{t,\tau\}}$, and the stochastic process

$$S_j := \tilde{R}_{\sigma+j}^2 + \zeta\alpha \cdot \min\{j, \tau - \sigma\}. \tag{112}$$

Note that this process is adapted to $x_{\sigma+j}$. We will now show that $S_j$ is a supermartingale. We note that if $j \geq \tau - \sigma$, the exit out of the neighborhood has occurred, $\tilde{R}_{\sigma+j} = R_\tau$, so $S_{j+1} = S_j$. Otherwise, if $\sigma + j < \tau$, let $t^* := \sigma + j$. Because $\sigma \leq t^* < \tau$,

$$\mathbb{E}[\tilde{R}^2_{t^*+1} - \tilde{R}^2_{t^*} \mid x_{t^*}] = \mathbb{E}[R^2_{t^*+1} - R^2_{t^*} \mid x_{t^*}] \leq -\zeta\alpha, \tag{113}$$

hence,

$$\mathbb{E}[S_{j+1} - S_j \mid x_{\sigma+j}] \leq -\zeta\alpha + \zeta\alpha = 0. \tag{114}$$

Let us now express the event $\{\tau \leq T\}$ of exiting the neighborhood within $T$ steps after initialization in terms of the supermartingale we have just defined. Consider the event

$$\left\{ \max_{0 \leq j \leq T} (S_j - S_0) \geq \Delta \right\}, \tag{115}$$

where $\Delta := \frac{7}{16} r^2_{\text{out}} = \frac{7\mu^2}{16\rho^2}$. Assume $\tau \leq T$. Let $j^* := \tau - \sigma$. We have

$$S_{j^*} = \tilde{R}^2_\tau + \zeta\alpha(\tau - \sigma) = R^2_\tau + \zeta\alpha(\tau - \sigma) \geq r^2_{\text{out}}. \tag{116}$$

Furthermore, $S_0 = \tilde{R}^2_\sigma = R^2_\sigma$. Therefore, on $\{\tau \leq T\}$,

$$\max_{0 \leq j \leq T} (S_j - S_0) \geq S_{j^*} - S_0 \geq r^2_{\text{out}} - R^2_\sigma. \tag{117}$$

But, as established in (103), $R_\sigma \leq \frac{3}{4} r_{\text{out}}$, meaning

$$\max_{0 \leq j \leq T} (S_j - S_0) \geq r^2_{\text{out}} - \left(\frac{3}{4} r_{\text{out}}\right)^2 = \Delta. \tag{118}$$

Then,

$$\{\tau \leq T\} \subseteq \left\{ \max_{0 \leq j \leq T} (S_j - S_0) \geq \Delta \right\}. \tag{119}$$

Now, the only remaining step is to get a bound on the increment $|S_{j+1} - S_j|$. Note that

$$|S_{j+1} - S_j| \leq |\tilde{R}^2_{\sigma+j+1} - \tilde{R}^2_{\sigma+j}| + \zeta\alpha, \tag{120}$$

so it will be useful to get a bound on the increment $|\tilde{R}^2_{t+1} - \tilde{R}^2_t| = |\tilde{R}_{t+1} - \tilde{R}_t|(\tilde{R}_{t+1} - \tilde{R}_t)$. We consider three cases based on the time index $t$:

When $t \geq \tau$, $\tilde{R}_t = \tilde{R}_{t+1}$, so $|\tilde{R}^2_{t+1} - \tilde{R}^2_t| = 0$.

When $t, t+1 < \tau$, $\tilde{R}_t = R_t < r_{\text{out}}$ and $\tilde{R}_{t+1} = R_{t+1} < r_{\text{out}}$, hence

$$|\tilde{R}^2_{t+1} - \tilde{R}^2_t| \leq (\alpha G) \cdot 2r_{\text{out}}. \tag{121}$$

When $t < \tau$, $t + 1 = \tau$, $\tilde{R}_t = R_t < r_{\text{out}}$ and $\tilde{R}_{t+1} = R_{t+1} = R_\tau \leq r_{\text{out}} + \alpha G$. Then,

$$|\tilde{R}^2_{t+1} - \tilde{R}^2_t| \leq \alpha G \cdot (2r_{\text{out}} + \alpha G). \tag{122}$$

In all three cases, $|\tilde{R}^2_{t+1} - \tilde{R}^2_t| \leq \alpha G \cdot (2r_{\text{out}} + \alpha G)$ is a valid upper bound. Then, for any $j$,

$$\begin{aligned} |S_{j+1} - S_j| \leq |\tilde{R}^2_{t+1} - \tilde{R}^2_t| + \zeta\alpha &\leq \alpha G \cdot (2r_{\text{out}} + \alpha G) + \zeta\alpha \\ &=: B(\alpha). \end{aligned} \tag{123}$$

Finally, invoking the Azuma-Hoeffding inequality, we conclude

$$\mathbb{P}(\tau \leq T) \leq \mathbb{P}\left( \max_{0 \leq j \leq T} (S_j - S_0) \geq \Delta \right) \leq \exp\left( -\frac{\Delta^2}{2TB(\alpha)^2} \right), \tag{124}$$

or equivalently,

$$\mathbb{P}\left( \max_{0 \leq t \leq T} \|x_t - x^*\| \leq \frac{\mu}{\rho} \right) \geq 1 - \exp\left( -\frac{\Delta^2}{2TB(\alpha)^2} \right). \tag{125}$$

□

**Theorem D.11** (Convergence of ORDO). *Fix $\epsilon > 0$ and integer $k \geq 2$ and consider Algorithm 1 (ORDO). Let $G_{\max}$ be as in Proposition D.4, let $\rho$ be as in Corollary D.7, and let $\beta_N$ be the bias level from Proposition D.3. Let $\kappa_{\mathrm{sh}}$ be the sharpness constant from Assumption D.9 and suppose $\beta_N < \kappa_{\mathrm{sh}}/2$. Run ORDO with a constant inner-loop step size $\eta_{\bar{t}} \equiv \eta$, chosen as*

$$\eta := \min\left\{ \frac{(\kappa_{\mathrm{sh}} - 2\beta_N)\epsilon}{4G_{\max}^3}, \frac{(\kappa_{\mathrm{sh}} - 2\beta_N)\kappa_{\mathrm{sh}}}{4G_{\max}^2 \rho} \right\}. \tag{126}$$

*Suppose the inner loop at outer iteration $t$ is initialized in the sharpness tube, meaning $\mathrm{dist}(\theta_{t,0}, \Theta_t^*) \leq \kappa_{\mathrm{sh}}/(2\rho)$. Then, Lemma D.10 implies that, with high probability, the inner iterates remain inside the tube where the sharpness geometry is effective.*

*On the intersection of these high-probability tube events over all executed outer iterations, ORDO terminates after $O(\epsilon^{-2})$ total inner-loop updates and returns an iterate that is $\epsilon$-almost non-dominated in expectation.*

*Proof.* Let $t \in \{0, \ldots, T-1\}$. Let $E_{t,\bar{t}} := \|\theta_{t,\bar{t}} - \theta_t^*\|^2$, where $\|\theta_{t,0} - \theta_t^*\| \leq \kappa_{\mathrm{sh}}/2\rho$ by assumption. We would like to find a recursion on $\mathbb{E}[E_{t,\bar{t}+1} \mid \theta_{t,\bar{t}}]$. Using the update rule,

$$\begin{aligned} E_{t,\bar{t}+1} = \|\theta_{t,\bar{t}+1} - \theta_t^*\|^2 &= \|\theta_{t,\bar{t}} - \eta g_{t,\bar{t}} - \theta_t^*\|^2 \\ &= \|\theta_{t,\bar{t}} - \theta_t^*\|^2 - 2\eta\langle g_{t,\bar{t}}, \theta_{t,\bar{t}} - \theta_t^*\rangle + \eta^2\|g_{t,\bar{t}}\|^2. \end{aligned} \tag{127}$$

Taking expectation conditioned on $\theta_{t,\bar{t}}$,

$$\begin{aligned} \mathbb{E}[E_{t,\bar{t}+1} \mid \theta_{t,\bar{t}}] &= E_{t,\bar{t}} - 2\eta\langle \mathbb{E}[g_{t,\bar{t}} \mid \theta_{t,\bar{t}}], \theta_{t,\bar{t}} - \theta_t^*\rangle + \eta^2\mathbb{E}[\|g_{t,\bar{t}}\|^2 \mid \theta_{t,\bar{t}}] \\ &\leq E_{t,\bar{t}} - 2\eta\langle v_{t,\bar{t}} + b_{t,\bar{t}}, \theta_{t,\bar{t}} - \theta_t^*\rangle + \eta^2 G_{\max}^2. \end{aligned} \tag{128}$$

Using the same line of reasoning as in the proof of D.10, specifically, (97), we have

$$\langle v_{t,\bar{t}}, \theta_{t,\bar{t}} - \theta_t^*\rangle \geq \frac{\kappa_{\mathrm{sh}}}{2}\|\theta_{t,\bar{t}} - \theta_t^*\|, \tag{129}$$

as long as $\|\theta_{t,\bar{t}} - \theta_t^*\| \leq \kappa_{\mathrm{sh}}/\rho$, which is guaranteed for all $(t, \bar{t})$ by the conditioning on the good event of Lemma D.10.

Furthermore, using Proposition D.3

$$\langle b_{t,\bar{t}}, \theta_{t,\bar{t}} - \theta_t^*\rangle \geq -\beta_N\|\theta_{t,\bar{t}} - \theta_t^*\|. \tag{130}$$

Combining (129) and (130), we have

$$\langle v_{t,\bar{t}} + b_{t,\bar{t}}, \theta_{t,\bar{t}} - \theta_t^*\rangle \geq \left(\frac{\kappa_{\mathrm{sh}}}{2} - \beta_N\right)\|\theta_{t,\bar{t}} - \theta_t^*\|. \tag{131}$$

Let us define $K := (\kappa_{\mathrm{sh}}/2 - \beta_N)$ for convenience. Note that $K > 0$ by assumption. Using this in (128),

$$\mathbb{E}[E_{t,\bar{t}+1} \mid \theta_{t,\bar{t}}] \leq E_{t,\bar{t}} - 2\eta K\|\theta_{t,\bar{t}} - \theta_t^*\| + \eta^2 G_{\max}^2. \tag{132}$$

Let $u_t := \|\theta_{t,\bar{t}} - \theta_t^*\| = \sqrt{E_t}$. Then, we can rewrite the previous as

$$\mathbb{E}[u_{t,\bar{t}+1}^2 \mid \theta_{t,\bar{t}}] \leq u_{t,\bar{t}}^2 - 2\eta K u_{t,\bar{t}} + \eta^2 G_{\max}^2. \tag{133}$$

Taking full expectation,

$$\mathbb{E}[u_{t,\bar{t}+1}^2] \leq \mathbb{E}[u_{t,\bar{t}}^2] - 2\eta K\mathbb{E}[u_{t,\bar{t}}] + \eta^2 G_{\max}^2. \tag{134}$$

Rearranging and summing over $\bar{t} = 0, \ldots \bar{T} - 1$,

$$2\eta K\sum_{\bar{t}=0}^{\bar{T}-1}\mathbb{E}[u_{t,\bar{t}}] \leq \mathbb{E}[u_{t,0}^2] - \mathbb{E}[u_{t,\bar{T}}^2] + \bar{T}\eta^2 G_{\max}^2. \tag{135}$$

Rearranging once more, and dividing each side by $\bar{T}$,

$$\frac{1}{\bar{T}}\sum_{\bar{t}=0}^{\bar{T}-1}\mathbb{E}[u_{t,\bar{t}}] \leq \frac{u_{t,0}^2}{2\eta K\bar{T}} + \frac{\eta G_{\max}^2}{2K}. \tag{136}$$

Then, there exists some iteration $\bar{t}^* < \bar{T}$ such that

$$\mathbb{E}[u_{t,\bar{t}^*}] \leq \frac{u_{t,0}^2}{2\eta K \bar{T}} + \frac{\eta G_{\max}^2}{2K} \leq \frac{\kappa_{\text{sh}}^2}{8\eta K \bar{T} \rho^2} + \frac{\eta G_{\max}^2}{2K} \tag{137}$$

Now, note that the gradient of $\phi_{k,t}$ is uniformly bounded by

$$\|\nabla \phi_{k,t}(\theta)\| \leq G_{\max}, \tag{138}$$

through a similar argument as in Proposition D.4. Then,

$$\phi_{k,t}(\theta_{t,\bar{t}}) - \phi_{k,t}(\theta_t^*) \leq G_{\max} \|\theta_{t,\bar{t}} - \theta_t^*\| = G_{\max} u_{t,\bar{t}}. \tag{139}$$

Taking expectation, for $\bar{t}^*$,

$$\mathbb{E}\left[\phi_{k,t}(\theta_{t,\bar{t}^*}) - \phi_{k,t}(\theta_t^*)\right] \leq G_{\max}\left(\frac{\kappa_{\text{sh}}^2}{8\eta K \bar{T} \rho^2} + \frac{\eta G_{\max}^2}{2K}\right). \tag{140}$$

Now, our current aim will be to show that if a progress of $\epsilon$ can be made on the subproblem at outer iteration $t$, then our algorithm also makes progress. That is, if $\theta_t^*$ is such that $\phi_{k,t}(\theta_t^*) < -\epsilon$ (i.e., $\theta_t^*$ stochastically dominates the previous outer iteration's solution $\theta_t$ by at least $\epsilon$,) then our algorithm is guaranteed to return a solution that stochastically dominates $\theta_t$ by at least $\epsilon/2$ on expectation.

To show this, suppose $\phi_{k,t}^* := \phi_{k,t}(\theta_t^*) < -\epsilon$. Furthermore,

$$\mathbb{E}[\phi_{k,t}(\theta_{t,\bar{t}^*})] \leq \phi_{k,t}^* + G_{\max}\left(\frac{\kappa_{\text{sh}}^2}{8\eta K \bar{T} \rho^2} + \frac{\eta G_{\max}^2}{2K}\right). \tag{141}$$

By choice of $\eta$, and further choosing

$$\bar{T} = \left\lceil \frac{G_{\max} \kappa_{\text{sh}}^2}{2\eta K \rho^2 \epsilon} \right\rceil, \tag{142}$$

there exists $\bar{t}^*$ such that

$$\mathbb{E}[\phi_{k,t}(\theta_{t,\bar{t}^*})] \leq -\frac{\epsilon}{2}. \tag{143}$$

If, on the other hand, $\phi_{k,t}^* \geq -\epsilon$, then for all $\theta \in \mathbb{R}^w$, $\phi_{k,t}(\theta) \geq -\epsilon$, meaning that there is no solution that dominates the previous solution $\theta_t$ by at least $\epsilon$, i.e., that $\theta_t$ is $\epsilon$-non-dominated.

The final step is to relate progress over different subproblems $\phi_{k,t}$ (over different outer iterations $t$) to each other. For this, let us view $\phi_{k,t}$ as a function of *two* parameters now, both of the decision parameter and also of the reference parameter resulting from the solution of the previous outer iteration, $\theta_t$. Namely, we define

$$V_k(\theta, \theta_t) := \phi_{k,t}(\theta). \tag{144}$$

Let us first show that $V_k$ admits a property analogous to the triangle inequality. To see this, let $\theta_1, \theta_2, \theta_3 \in \mathbb{R}^w$, and $X_i \sim p_{\theta_i}$ over compact support $M$. We have

$$\begin{aligned} V_k(X_1, X_3) &= \max_{m \in M} \left(F_{X_1}^k(m) - F_{X_3}^k(m)\right) \\ &= \max_{m \in M} \left(F_{X_1}^k(m) - F_{X_2}^k(m) + F_{X_2}^k(m) - F_{X_3}^k(m)\right) \\ &\leq \max_{m \in M} \left(F_{X_1}^k(m) - F_{X_2}^k(m)\right) + \max_{m \in M} \left(F_{X_2}^k(m) - F_{X_3}^k(m)\right) \\ &= V_k(X_1, X_2) + V_k(X_2, X_3). \end{aligned} \tag{145}$$

The results of the proof up to now have shown that each outer loop up to termination satisfies

$$\mathbb{E}[V_k(\theta_{t+1}, \theta_t)] \leq -\epsilon/2. \tag{146}$$

Furthermore, by the same argument as in B.4, $V_k$ is lower-bounded by $-2c_k D^{(k-1)d}$. Assume now that the outer loop runs for $T$ iterations, soon to be determined. By the triangle inequality in $V_k$ just proved, the progress certificate of $-\epsilon/2$ in each iteration, and the lower boundedness of $V_k$, we have

$$-2 \leq \mathbb{E}[V_k(\theta_0, \theta_T)] \leq \sum_{t=0}^{T-1} \mathbb{E}[V_k(\theta_{t+1}, \theta_t)] \leq -T \cdot \frac{\epsilon}{2}, \tag{147}$$

which guarantees

$$T \leq \frac{4}{\epsilon}. \tag{148}$$

Then, choosing

$$T = \left\lceil \frac{4}{\epsilon} \right\rceil, \tag{149}$$

we are guaranteed that the algorithm is guaranteed to terminate and return a solution via line 17 in at most $T$ outer iterations. Finally, to establish the total number of iterations needed, we examine the product

$$T \cdot \bar{T} = \left\lceil \frac{4}{\epsilon} \right\rceil \cdot \left\lceil \frac{G_{\max}\kappa_{\text{sh}}^2}{2\eta K \rho^2 \epsilon} \right\rceil, \tag{150}$$

we conclude that the total number of iterations needed is

$$O\left( \frac{4G_{\max}\kappa_{\text{sh}}^2}{2\eta K \rho^2 \epsilon^2} \right) = O(\epsilon^{-2}). \tag{151}$$

□

The previous result establishes the convergence of our algorithm to an $\epsilon$-non-dominated solution in $O(\epsilon^{-2})$ iteration complexity. Sharpness imposes a strong regularity condition, not unlike that of the concavity assumption present in Cen et al. [2026]. Furthermore, we require that the bias in the stochastic gradients due to the inexact inner solver for $\hat{m}$ be sufficiently small with respect to the sharpness parameter $\kappa_{\text{sh}}$, and that the solution returned by the algorithm for each subproblem $\phi_{k,t}$ be sufficiently close to an optimal solution for the next subproblem. We note, however, that such strong conditions are absolutely needed in order to prove a result as strong and global as $\epsilon$-non-dominatedness, especially in our more involved multivariate setting.

Acknowledging that some of these requirements may not be met in practical scenarios, in the next section, we analyze the proposed Algorithm 3 (Soft-ORDO), which relies on using a smooth surrogate for the inner maximization problem of our objective, lifting the dependence on many of the required criteria, and also resolving the problem of biased stochastic gradients.

**Algorithm 3** Soft-ORDO

**Require:** Number of maximum iterations $T$, progress threshold $\epsilon > 0$, initialization $\theta_0$, $\alpha$-net $M_J = \{m_1, \dots, m_J\}$, softmax temperature $\kappa$

**for** $t = 0, \dots, T-1$ **do**
  Sample $\{\tau_{t,i}\}_{i=1}^N, \{\tau_{t,i}^+\}_{i=1}^N, \{\tau_{t,i}^q\}_{i=1}^N \sim p_{\theta_t}(\tau)$ by rolling out MDP with policy $\pi_{\theta_t}$
  $\{R_{t,i}\}_{i=1}^N \leftarrow \{R(\tau_{t,i})\}_{i=1}^N$
  $\{R_{t,i}^+\}_{i=1}^N \leftarrow \{R(\tau_{t,i}^+)\}_{i=1}^N$
  $\{R_{t,i}^q\}_{i=1}^N \leftarrow \{R(\tau_{t,i}^q)\}_{i=1}^N$
  $q_j \leftarrow \mathrm{softmax}_\kappa(\hat{H}_k(\theta, m_j))$, where

$$\hat{H}_k(\theta, m_j) = \frac{c_k}{N} \sum_{i=1}^N \left( h_k(R_{t,i}^q, m_j) - h_k(R_{t,i}, m_j) \right)$$

$$g_t \leftarrow \frac{c_k}{N} \sum_{j=1}^J q_j \sum_{i=1}^N h_k(R_{t,i}^+, m_j) \nabla_\theta \log p_{\theta_t}(\tau_{t,i}^+)$$

  $\theta_t^+ \leftarrow \theta_t - \eta_t g_t$
  **if** $\|g_t\| > \epsilon$ **then**
    $\theta_{t+1} \leftarrow \theta_t^+$
    **break**
  **end if**
  **if** $\theta_t$ has not been updated **then**
    **return** $\theta_t$
  **end if**
**end for**

# E SOFT-ORDO

Section 4 sheds light on how the inner maximization problem of (9) gives rise to numerous obstacles in the name of optimization. Firstly, because we are not able to exactly solve this nonconvex inner problem to find a maximizer as Danskin's theorem (Theorem B.9) instructs, our stochastic gradient estimators end up being biased. Furthermore, $\phi_{k,t}$ having this *max-function* form for all $t$ extinguishes hopes of having a smooth, geometrically favorable optimization objective, leading us to settle for the loose and unfriendly setting of weak convexity (Corollary D.7).

A straightforward approach would then be to swap this *hard* maximization with the smooth and accessible *soft* maximization, over a sufficiently large and representative subset of $M$.

Let us begin by stating the notion of optimality that we would like to achieve within this section.

**Definition E.1** ($k$th-order $\epsilon$-Pareto stationarity)**.** Let $X_\theta$ and $X_{\theta'}$ be RVs with compact support $M \subseteq \mathbb{R}^d$. $X_\theta$ is $k$th-order $\epsilon$-Pareto stationary with respect to $X_{\theta'}$ if

$$\mathrm{dist}\left(0, \partial_\theta \max_m \mathcal{L}_k(X_\theta, X_{\theta'}, m)\right) \leq \epsilon. \tag{152}$$

That is, there is no infinitesimal change in $\theta$ that can decrease the worst-case $k$th-order dominance violation against the reference $X_{\theta'}$ at a rate larger than $\epsilon$. Additionally, we say $X_\theta$ is $k$th-order $\epsilon$-Pareto stationary if there exists $X_{\theta'}$ such that $X_\theta$ is $k$th-order $\epsilon$-Pareto stationary with respect to $X_{\theta'}$.

Let $M_J = \{m_1, \dots, m_J\}$ be an $\alpha$-net covering $M$. That is, for any $m \in M$, there exists $j \in [J]$ such that $\|m_j - m\| \leq \alpha$. We will define our smooth surrogate function to be the log-sum-exp function

$$\tilde{\phi}_k(\theta) := \kappa \log \left( \sum_{j=1}^J \exp\left( \frac{H_k(\theta, m_j)}{\kappa} \right) \right). \tag{153}$$

The gradient of the surrogate is, correspondingly,

$$\nabla\tilde{\phi}_k(\theta) = \sum_{j=1}^{J} q_j \nabla_\theta H_k(\theta, m_j), \tag{154}$$

where

$$q_j = \frac{\exp(H_k(\theta, m_j)/\kappa)}{\sum_{j'=1}^{J} \exp(H_k(\theta, m_{j'})/\kappa)} =: \mathrm{softmax}_\kappa(H_k(\theta, m_j)). \tag{155}$$

Our first aim will be to show how, with careful choices of $J$ and $\kappa$, the surrogate gradient $\nabla\tilde{\phi}_k$ will estimate a true subgradient of the original objective, $\phi_k$. That is, we want to show that $\mathrm{dist}(\nabla\tilde{\phi}_k(\theta), \partial\phi_k(\theta))$ can be made arbitrarily small.

Firstly, by Theorem B.9, we have

$$\partial\phi_k(\theta) = \mathrm{conv}\left\{\nabla_\theta H_k(\theta, m^*) : m^* \in M^*(\theta)\right\}. \tag{156}$$

By the $\alpha$-net argument, we have that for each $m^* \in M^*(\theta)$, there exists a $j^* \in [J]$ such that $\|m^* - m_{j^*}\| \le \alpha$.

Let $J^*(\theta) := \{j : \exists\, m^* \in M^*(\theta), \|m_j - m^*\| \le \alpha\}$, the set of the $\alpha$-net indices which are nearby a maximizer. Conversely, let $\bar{J}(\theta) := [J] \setminus J^*(\theta)$.

We now define

$$\tilde{\partial}\phi_k(\theta) := \mathrm{conv}\left\{\nabla_\theta H_k(\theta, m_j) : j \in J^*(\theta)\right\}, \tag{157}$$

an estimate of the subgradient $\partial\phi_k(\theta)$, using the elements of the $\alpha$-net instead of the true maximizer.

We begin by showing

$$\mathrm{dist}(\partial\phi_k(\theta), \tilde{\partial}\phi_k(\theta)) \le L_m\alpha, \tag{158}$$

where $L_m$ is the Lipschitz continuity constant of $\nabla_\theta H_k$ in $m$, as defined in Proposition D.2. We will do this through the following two lemmas.

**Lemma E.2.**

$$\sup_{v \in \partial\phi_k(\theta)} \mathrm{dist}\left(v, \tilde{\partial}\phi_k(\theta)\right) \le L_m\alpha. \tag{159}$$

*Proof.* Let $v \in \partial\phi_k(\theta)$. Then, there exist $m_1^*, \ldots, m_r^* \in M^*(\theta)$ and coefficients $\lambda_1, \ldots, \lambda_r \ge 0$, $\sum_{i=1}^r \lambda_i = 1$ such that

$$v = \sum_{i=1}^{r} \lambda_i \nabla_\theta H_k(\theta, m_i^*). \tag{160}$$

For each $i \in [r]$, we can pick $j_i \in J^*(\theta)$ from the $\alpha$-net such that $\|m_{j_i} - m_i^*\| \le \alpha$.

Let $w := \sum_{i=1}^r \lambda_i \nabla_\theta H_k(\theta, m_{j_i})$. Clearly, $w \in \tilde{\partial}\phi_k(\theta)$. Also,

$$\begin{aligned}
\|v - w\| &= \left\| \sum_{i=1}^{r} \lambda_i \left(\nabla_\theta H_k(\theta, m_i^*) - \nabla_\theta H_k(\theta, m_{j_i})\right) \right\| \\
&\le \sum_{i=1}^{r} \lambda_i \left\|\nabla_\theta H_k(\theta, m_i^*) - \nabla_\theta H_k(\theta, m_{j_i})\right\| \\
&\le \sum_{i=1}^{r} \lambda_i L_m \|m_i^* - m_{j_i}\| \\
&\le \sum_{i=1}^{r} \lambda_i L_m \alpha = L_m\alpha \sum_{i=1}^{r} \lambda_i = L_m\alpha.
\end{aligned} \tag{161}$$

Then,

$$\mathrm{dist}\left(v, \tilde{\partial}\phi_k(\theta)\right) \le L_m\alpha, \tag{162}$$

for all $v \in \partial\phi_k(\theta)$. Hence, we conclude

$$\sup_{v \in \partial\phi_k(\theta)} \operatorname{dist}\left(v, \tilde{\partial}\phi_k(\theta)\right) \leq L_m \alpha. \tag{163}$$

□

This result shows that any element of the subgradient $\partial\phi_k(\theta)$ is $L_m\alpha$-close to the subgradient set approximated by the $\alpha$-net elements, $\tilde{\partial}\phi_k(\theta)$.

The following lemma presents the complementary result, that any element of the approximated subgradient set is $L_m\alpha$-close to the subgradient set.

**Lemma E.3.**

$$\sup_{\tilde{v} \in \tilde{\partial}\phi_k(\theta)} \operatorname{dist}\left(\tilde{v}, \partial\phi_k(\theta)\right) \leq L_m \alpha \tag{164}$$

*Proof.* A symmetric argument to the proof of Lemma E.2 furnishes the result. □

From the two preceding lemmas, we get the immediate corollary:

**Corollary E.4** (Distance of subgradient and the $\alpha$-net approximated subgradient)**.**

$$\operatorname{dist}\left(\partial\phi_k(\theta), \tilde{\partial}\phi_k(\theta)\right) \leq L_m \alpha. \tag{165}$$

*Proof.* The proof immediately follows from the equivalency

$$\operatorname{dist}\left(\partial\phi_k(\theta), \tilde{\partial}\phi_k(\theta)\right) = \max\left\{\sup_{v \in \partial\phi_k(\theta)} \operatorname{dist}\left(v, \tilde{\partial}\phi_k(\theta)\right), \sup_{\tilde{v} \in \tilde{\partial}\phi_k(\theta)} \operatorname{dist}\left(\tilde{v}, \partial\phi_k(\theta)\right)\right\}, \tag{166}$$

and the results of Lemmas E.2 and E.3. □

Next, we will show that the softmax function of (155) will concentrate all weight on the near-maximal elements of the $\alpha$-net, $J^*(\theta)$.

Let us make explicit the gap between the maximal value of $H_k$ for a given $\theta$ and the submaximal values with a function $\Delta$:

$$\phi_k(\theta) - H_k(\theta, m) \geq \Delta(\theta) > 0, \tag{167}$$

for all $m \notin M^*(\theta)$. Then, for any $j \in \bar{J}(\theta)$,

$$H_k(\theta, m_j) \leq \phi_k(\theta) - \Delta(\theta). \tag{168}$$

But, for any $j^* \in J^*(\theta)$, there exists, by definition, $m^* \in M^*(\theta)$ such that $\|m_{j^*} - m^*\| \leq \alpha$. Then,

$$|H_k(\theta, m_{j^*}) - H_k(\theta, m^*)| \leq L\alpha, \tag{169}$$

or equivalently,

$$H_k(\theta, m_{j^*}) \geq \phi_k(\theta) - L\alpha, \tag{170}$$

where $L$ is the Lipschitz continuity constant of $H_k$ in $m$, as per Proposition B.5.

Combining (168) and (170), for any $j^* \in J^*(\theta)$ and $j \in \bar{J}(\theta)$,

$$H_k(\theta, m_{j^*}) - H_k(\theta, m_j) \geq \Delta(\theta) - L\alpha. \tag{171}$$

Suppose we choose $J$ large enough such that $\alpha$ is small enough to satisfy

$$\Delta(\theta) - L\alpha \geq \delta_\Delta > 0. \tag{172}$$

We would then have

$$\frac{q_j}{q_{j^*}} \leq \exp\left(-\frac{\delta_\Delta}{\kappa}\right), \tag{173}$$

implying

$$q_j \leq \exp\left(-\frac{\delta_\Delta}{\kappa}\right), \tag{174}$$

since $q_{j^*} \leq 1$. Then,

$$\sum_{j \in \bar{J}(\theta)} q_j \leq \sum_{j \in \bar{J}(\theta)} \exp\left(-\frac{\delta_\Delta}{\kappa}\right) \leq J \exp\left(-\frac{\delta_\Delta}{\kappa}\right). \tag{175}$$

This shows that choosing $\tau$ "small enough", the weight assigned to the "bad" $\alpha$-net points belonging to $\bar{J}(\theta)$ can be made arbitrarily small, meaning that almost all of the weight can be made to sit on the "good" points belonging to $J^*(\theta)$. (We will discuss how the choices for $J$ and $\tau$ can be made to ensure this more rigorously in Remark E.6.)

As the final step, we will show that the gradient $\nabla\tilde{\phi}_k(\theta)$ of our smooth surrogate is close to the subgradient of the true function $\partial\phi_k(\theta)$. To do this, we will show that it is close to the approximated subgradient $\tilde{\partial}\phi_k(\theta)$ and then use the result of Corollary E.4.

**Theorem E.5.**

$$\operatorname{dist}\left(\nabla\tilde{\phi}_k(\theta), \partial\phi_k(\theta)\right) \leq L_m\alpha + G_{\max} J \exp\left(-\frac{\delta_\Delta}{\kappa}\right). \tag{176}$$

*Proof.* We have

$$\nabla\tilde{\phi}_k(\theta) = \sum_{j \in J^*(\theta)} q_j \cdot \nabla_\theta H_k(\theta, m_j) + \sum_{j \in \bar{J}(\theta)} q_j \cdot \nabla_\theta H_k(\theta, m_j) =: v^*(\theta) + \bar{v}(\theta). \tag{177}$$

Note that $v^*(\theta) \in \tilde{\partial}\phi_k(\theta)$. Also,

$$\|\bar{v}(\theta)\| \leq \sum_{j \in \bar{J}(\theta)} q_j \cdot \|\nabla_\theta H_k(\theta, m_j)\| \leq G_{\max} \sum_{j \in \bar{J}(\theta)} q_j \leq G_{\max} J \exp\left(-\frac{\delta_\Delta}{\kappa}\right). \tag{178}$$

Then,

$$\operatorname{dist}\left(\nabla\tilde{\phi}_k(\theta), \tilde{\partial}\phi_k(\theta)\right) \leq G_{\max} J \exp\left(-\frac{\delta_\Delta}{\kappa}\right). \tag{179}$$

Finally, using Corollary E.4 and triangle inequality, we have

$$\operatorname{dist}\left(\nabla\tilde{\phi}_k(\theta), \partial\phi_k(\theta)\right) \leq L_m\alpha + G_{\max} J \exp\left(-\frac{\delta_\Delta}{\kappa}\right). \tag{180}$$

□

*Remark* E.6 (Choosing $J$ and $\kappa$). If we would like to have an $\alpha$-net $M_J$ covering $M \subseteq \mathbb{R}^d$ with the property that for any $m \in M$, there exists $j \in [J]$ such that $\|m_j - m\| \leq \alpha$, a natural choice is a regular grid over the $d$ dimensions, resulting in a need for

$$J \leq \left(1 + \frac{D}{\alpha}\right)^d = O(\alpha^{-d}) \tag{181}$$

grid points in total.

Suppose now that we would like to make the distance in Theorem E.5 smaller than some desired threshold $\epsilon$, i.e.,

$$\operatorname{dist}\left(\nabla\tilde{\phi}_k(\theta), \partial\phi_k(\theta)\right) \leq L_m\alpha + G_{\max} J \exp\left(-\frac{\delta_\Delta}{\kappa}\right) \leq \epsilon. \tag{182}$$

One way to achieve this is to ensure that both of the terms of the right-hand side of the inequality are smaller than $\epsilon/2$, that is,

$$L_m\alpha \leq \frac{\epsilon}{2}, \quad G_{\max} J \exp\left(-\frac{\delta_\Delta}{\kappa}\right) \leq \frac{\epsilon}{2}. \tag{183}$$

Then, we would like

$$\alpha \leq \frac{\epsilon}{2L_m}. \tag{184}$$

But, recalling that we also need

$$\Delta(\theta) - L\alpha \geq \delta_\Delta > 0 \tag{185}$$

for all $\theta \in \mathbb{R}^w$, as per (172). Let $\Delta_{\min} := \inf_{\theta \in \mathbb{R}^w} \Delta(\theta)$. Let us ensure that

$$\Delta_{\min} - L\alpha \geq \frac{\Delta_{\min}}{2} =: \delta_\Delta. \tag{186}$$

Rearranging, this requires

$$\alpha \leq \frac{\Delta_{\min}}{2L}. \tag{187}$$

Then, combining the two requirements, we need

$$\alpha = \min\left\{\frac{\epsilon}{2L}, \frac{\Delta_{\min}}{2L}\right\}, \tag{188}$$

and $J = O(\alpha^{-d})$, as argued above.

For the choice of $\tau$, we turn to the second term of the right-hand side of (183). We need

$$G_{\max} J \exp\left(-\frac{\delta_\Delta}{\kappa}\right) \leq \frac{\epsilon}{2}, \tag{189}$$

which points to choosing

$$\kappa \leq \frac{\Delta_{\min}}{2\log\left(\frac{2GK}{\epsilon}\right)} = O\left(\frac{1}{\log(1/\epsilon)}\right). \tag{190}$$

Before we begin with the analysis of the convergence of Algorithm 3 (Soft-ORDO), we will present a final result on the nature of $\tilde{\phi}_k$. It is well known that the log-sum-exp function is smooth, and that is in fact precisely one of the reasons why we would like to handle it rather than the inhospitable hard-maximization. In the following proposition, we quantify this smoothness.

**Proposition E.7** ($\tilde{L}$-smoothness of $\tilde{\phi}_k$)**.** *$\tilde{\phi}_k$ is $\tilde{L}$-smooth, with*

$$\tilde{L} := \rho + \frac{G_{\max}^2}{\kappa}. \tag{191}$$

*Proof.* Calculating the Hessian of $\tilde{\phi}_k$,

$$\nabla^2 \tilde{\phi}_k(\theta) = \sum_{j=1}^{J} q_j \nabla_\theta^2 H_k(\theta, m_j) + \frac{1}{\kappa}\mathrm{cov}(\nabla_\theta H_k(\theta, m_j)). \tag{192}$$

Focusing on the first term on the right-hand side,

$$\left\| \sum_{j=1}^{J} q_j \nabla_\theta^2 H_k(\theta, m_j) \right\| \leq \sum_{j=1}^{J} q_j \left\| \nabla_\theta^2 H_k(\theta, m_j) \right\| \leq \sum_{j=1}^{J} q_j \cdot \rho = \rho, \tag{193}$$

by Proposition D.5.

For the covariance term, note that for any unit vector $u$,

$$u^\top \mathrm{cov}(\nabla_\theta H_k(\theta, m_j))u = \mathrm{var}(u^\top \nabla_\theta H_k(\theta, m_j)) \leq \mathbb{E}[(u^\top \nabla_\theta H_k(\theta, m_j))^2] \leq G_{\max}^2. \tag{194}$$

Hence,

$$\left\| \nabla^2 \tilde{\phi}_k(\theta) \right\| \leq \rho + \frac{G_{\max}^2}{\kappa}. \tag{195}$$

□

Finally, we qualify, in terms of bias and variance, the stochastic gradient estimator $g_t$ we will be using in the algorithm in the following result.

**Proposition E.8** ($O(1/N)$ bias and variance of the stochastic gradient)**.** *The stochastic gradient $g_t$, as computed in line 7 of Algorithm 3 (Soft-ORDO) satisfies*

1. *Conditional unbiasedness:*
$$\mathbb{E}[g_t \mid \theta_t, q] = \sum_{j=1}^{J} q_j \nabla_\theta H_k(\theta_t, m_j). \tag{196}$$
2. *Conditional variance bound:*
$$\mathbb{E}\left[\|g_t - \mathbb{E}[g_t \mid \theta, q]\|^2 \mid \theta, q\right] \leq \frac{4G_{\max}^2}{N}. \tag{197}$$
3. *Bias bounded in $O(1/N)$:*
$$\left\|\mathbb{E}[g_t \mid \theta_t] - \nabla\tilde{\phi}_k(\theta)\right\|^2 \leq \frac{\beta^2}{N}. \tag{198}$$
4. *Variance bounded in $O(1/N)$:*
$$\mathbb{E}\left[\|g_t - \mu_t\|^2 \mid \theta_t\right] \leq \frac{\sigma^2}{N}, \tag{199}$$
*where $\mu_t := \mathbb{E}[g_t \mid \theta_t]$.*

*Proof.* 1. We have
$$\nabla\tilde{\phi}_k(\theta_t) = \sum_{j=1}^{J} q_j^*(\theta_t) \nabla_\theta H_k(\theta_t, m_j), \tag{200}$$
where $q_j^* := \mathrm{softmax}_\kappa(H_k(\theta_t, m_j))$. Let
$$Z_i := c_k \sum_{j=1}^{J} q_j h_k(R_{t,i}^+, m_j) \nabla_\theta \log p_{\theta_t}(\tau_{t,i}^+). \tag{201}$$
Then, $g_t := \frac{1}{N}\sum_{i=1}^{N} Z_i$. We write
$$\mathbb{E}[g_t \mid \theta_t, q] = \mathbb{E}[Z_1 \mid \theta_t, q] = c_k \sum_{j=1}^{J} q_j \mathbb{E}_{\tau\sim p_{\theta_t}}[h_k(R(\tau), m_j) \nabla_\theta \log p_{\theta_t}(\tau)]. \tag{202}$$
But, by Proposition B.7,
$$\mathbb{E}[g_t \mid \theta_t, q] = \sum_{j=1}^{J} q_j \nabla_\theta H_k(\theta_t, m_j). \tag{203}$$

2. Let $\mu_t := \mathbb{E}[Z_1 \mid \theta_t, q]$. We have
$$0 \leq \sum_{j=1}^{J} q_j h_k(R_{t,i}^+, m_j) \leq D^{(k-1)d}. \tag{204}$$
Also, $\|\nabla_\theta \log p_\theta(\tau)\| \leq B_p$ for all $\theta$, as per Remark B.3. Then,
$$\|Z_1\| \leq B_p = G_{\max}, \tag{205}$$
almost surely. Now,
$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}(Z_i - \mu)\right\|^2\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\|Z_i - \mu\|^2\right]. \tag{206}$$
But, $\|Z_i - \mu\| \leq \|Z_i\| + \|\mu\| \leq 2G_{\max}$, hence,
$$\mathbb{E}\left[\|Z_i - \mu\|^2\right] \leq (2G_{\max})^2 = 4G_{\max}^2. \tag{207}$$
So,
$$\mathbb{E}\left[\|g_t - \mu_t|^2 \mid \theta_t, q\right] \leq \frac{1}{N^2} \cdot N \cdot 4G_{\max}^2 = \frac{4G_{\max}^2}{N} = O(1/N). \tag{208}$$

3. Taking expectation over $q$ in 1.,

$$\mathbb{E}[g_t \mid \theta_t] = \mathbb{E}\left[\sum_{j=1}^{J} q_j \nabla_\theta H_k(\theta_t, m_j) \mid \theta_t\right] = \sum_{j=1}^{J} \mathbb{E}[q_j \mid \theta_t] \nabla_\theta H_k(\theta_t, m_j). \tag{209}$$

Then,

$$\mathbb{E}[g_t \mid \theta_t] - \nabla \tilde{\phi}_k(\theta_t) = \sum_{j=1}^{J} (\mathbb{E}[q_j \mid \theta_t] - q_j^*) \nabla_\theta H_k(\theta_t, m_j). \tag{210}$$

This means

$$\begin{aligned}
\left\| \mathbb{E}[g_t \mid \theta_t] - \nabla \tilde{\phi}_k(\theta_t) \right\|^2 &\leq \left( \sum_{j=1}^{J} \mathbb{E}[|q_j - q_j^*| \mid \theta_t] \|\nabla_\theta H_k(\theta_t, m_j)\| \right)^2 \\
&\leq G_{\max}^2 \mathbb{E}\left[ \sum_{j=1}^{J} |q_j - q_j^*| \mid \theta_t \right]^2 \\
&\leq G_{\max}^2 \mathbb{E}\left[ \|q - q^*\|_1^2 \mid \theta_t \right] \\
&\leq G_{\max}^2 \frac{4}{\kappa^2} \mathbb{E}\left[ \left\| \hat{H}_k - H_k \right\|_\infty^2 \mid \theta_t \right] \\
&\leq G_{\max}^2 \frac{4}{\kappa^2} \sum_{j=1}^{J} \mathbb{E}\left[ (\hat{H}_k(\theta_t, m_j) - H_k(\theta_t, m_j))^2 \mid \theta_t \right] \\
&\leq G_{\max}^2 \frac{4}{\kappa^2} \sum_{j=1}^{J} \frac{1}{N} \leq \frac{4 G_{\max}^2 J}{\kappa^2 N} =: \frac{\beta^2}{N},
\end{aligned} \tag{211}$$

where $\hat{H}_k$ and $H_k$ are vectors in $\mathbb{R}^J$ whose $j^{\text{th}}$ coordinates are $\hat{H}_k(\theta_t, m_j)$ and $H_k(\theta_t, m_j)$, respectively.

4. A similar argument to that in Proposition D.8 yields the result.

□

We have now fully specified the optimization problem handled by Algorithm 3 (Soft-ORDO). At each iteration $t$, we minimize an $\tilde{L}$-smooth function $\tilde{\phi}_{k,t}$ through a stochastic subgradient method, taking a single update step, where we have access to stochastic gradients with bias and variance scaling in $O(1/N)$. [2] Unlike the setting of Section 4, we do not run into the problem of an irreducible bias term, as in the $\delta_c$ of Proposition D.3.

With all of these settled, we now state the main convergence theorem for Algorithm 3 (Soft-ORDO).

**Theorem E.9** (Convergence of Soft-ORDO)**.** *Fix $\epsilon > 0$ and integer $k \geq 2$ and consider Algorithm 3 (Soft-ORDO). Let $\tilde{L}$ be the smoothness constant from Proposition E.7. Let $\sigma^2$ and $\beta^2$ be the variance and bias constants from Proposition E.8. Choose batch size and constant step size*

$$N \geq \frac{\sigma^2 + 10\beta^2}{\epsilon^2}, \qquad \eta_t \equiv \eta := \frac{1}{\tilde{L}}. \tag{212}$$

*Then, Soft-ORDO returns an iterate $\theta_T$ within $O(\epsilon^{-2})$ iterations such that $\|\nabla \tilde{\phi}_{k,t}(\theta_T)\| \leq \epsilon$.*

*Moreover, due to Theorem E.5, the same iterate is also $(\epsilon + \epsilon')$-Pareto stationary for the original nonsmooth objective $\phi_{k,t}$, where $\epsilon' := L_m \alpha + G_{\max} J \exp(-\delta_\Delta / \kappa)$.*

*Proof.* We begin by stating the well-known approximation error between log-sum-exp and the maximum [Boyd and Vandenberghe, 2004]. Namely,

$$0 \leq \max_{j \in [J]} H_{k,t}(\theta, m_j) - \tilde{\phi}_{k,t}(\theta) \leq \kappa \log J. \tag{213}$$

[2] At this point, we once again readopt the $t$ subscript.

Furthermore, by the $\alpha$-net construction and Proposition B.5,

$$\left|\phi_{k,t}(\theta) - \max_{j\in[J]} H_{k,t}(\theta, m_j)\right| \le L\alpha. \tag{214}$$

Then, by triangle inequality

$$\left|\phi_{k,t}(\theta) - \tilde{\phi}_{k,t}(\theta)\right| \le L\alpha + \kappa \log J =: \delta_{\text{val}}. \tag{215}$$

By $\tilde{L}$-smoothness, we have

$$\begin{aligned}\tilde{\phi}_{k,t}(\theta_t^+) &\le \tilde{\phi}_k(\theta_t) + \langle \nabla\tilde{\phi}_{k,t}(\theta_t), \theta_t^+ - \theta_t\rangle + \frac{\tilde{L}}{2}\|\theta_t^+ - \theta_t\|^2 \\ &\le \tilde{\phi}_{k,t}(\theta_t) - \eta\langle \nabla\tilde{\phi}_{k,t}(\theta_t), g_t\rangle + \frac{\tilde{L}}{2}\eta^2\|g_t\|^2.\end{aligned} \tag{216}$$

Taking conditional expectation with respect to $\theta_t$,

$$\mathbb{E}[\tilde{\phi}_{k,t}(\theta_t^+) \mid \theta_t] \le \tilde{\phi}_{k,t}(\theta_t) - \eta\langle \nabla\tilde{\phi}_{k,t}(\theta_t), \mathbb{E}[g_t \mid \theta_t]\rangle + \frac{\tilde{L}}{2}\eta^2\mathbb{E}[\|g_t\|^2 \mid \theta_t]. \tag{217}$$

Let us treat the inner product term first. We have

$$\begin{aligned}\langle \nabla\tilde{\phi}_{k,t}(\theta_t), \mathbb{E}[g_t \mid \theta_t]\rangle = \langle \nabla\tilde{\phi}_{k,t}(\theta_t), \nabla\tilde{\phi}_{k,t}(\theta_t) + b_t\rangle &= \|\nabla\tilde{\phi}_{k,t}(\theta_t)\|^2 + \langle \nabla\tilde{\phi}_{k,t}(\theta_t), b_t\rangle \\ &\ge \|\nabla\tilde{\phi}_{k,t}(\theta_t)\|^2 - \|\nabla\tilde{\phi}_{k,t}(\theta_t)\| \cdot \|b_t\|.\end{aligned} \tag{218}$$

Now, focusing on the $\mathbb{E}[\|g_t\|^2 \mid \theta_t]$ term, we get

$$\mathbb{E}[\|g_t\|^2 \mid \theta_t] = \|\nabla\tilde{\phi}_{k,t}(\theta_t) + b_t\|^2 + \mathbb{E}[\|g_t - \mathbb{E}[g_t \mid \theta_t]\|^2 \mid \theta_t] \le \left(\|\nabla\tilde{\phi}_{k,t}(\theta_t) + \|b_t\|\right)^2 + \frac{\sigma^2}{N}. \tag{219}$$

Then,

$$\begin{aligned}\mathbb{E}[\tilde{\phi}_{k,t}(\theta_t^+) \mid \theta_t] &\le \tilde{\phi}_{k,t}(\theta_t) - \eta\|\nabla\tilde{\phi}_{k,t}(\theta_t)\|^2 + \eta\|\nabla\tilde{\phi}_{k,t}(\theta_t)\| \cdot \|b_t\| + \frac{\tilde{L}}{2}\eta^2\left((\|\nabla\tilde{\phi}_{k,t}(\theta_t) + \|b_t\|)^2 + \frac{\sigma^2}{N}\right) \\ &\le \tilde{\phi}_{k,t}(\theta_t) - \frac{\eta}{2}\|\nabla\tilde{\phi}_{k,t}(\theta_t)\|^2 + 2\eta\|\nabla\tilde{\phi}_{k,t}(\theta_t)\| \cdot \|b_t\| + \frac{\tilde{L}}{2}\eta^2\left(\|b_t\|^2 + \frac{\sigma^2}{N}\right).\end{aligned} \tag{220}$$

Using Young's inequality,

$$2\eta\|\nabla\tilde{\phi}_{k,t}(\theta_t)\| \cdot \|b_t\| \le \frac{\eta}{4}\|\nabla\tilde{\phi}_{k,t}(\theta_t)\|^2 + 4\eta\|b_t\|^2. \tag{221}$$

Then,

$$\mathbb{E}[\tilde{\phi}_{k,t}(\theta_t^+) \mid \theta_t] \le \tilde{\phi}_{k,t}(\theta_t) - \frac{\eta}{4}\|\nabla\tilde{\phi}_{k,t}(\theta_t)\|^2 + \frac{\tilde{L}}{2}\eta^2\frac{\sigma^2}{N} + 4\eta\|b_t\|^2 + \frac{\tilde{L}}{2}\eta^2\|b_t\|^2. \tag{222}$$

Since $\eta \le 1/\tilde{L}$ and $\|b_t\|^2 \le \beta^2/N$, we get the main inequality

$$\mathbb{E}[\tilde{\phi}_{k,t}(\theta_t^+) \mid \theta_t] \le \tilde{\phi}_{k,t}(\theta_t) - \frac{\eta}{4}\|\nabla\tilde{\phi}_{k,t}(\theta_t)\|^2 + \frac{\tilde{L}}{2}\eta^2\frac{\sigma^2}{N} + 5\eta\frac{\beta^2}{N}. \tag{223}$$

Now, note that $\tilde{\phi}_{k,t}(\theta_t) = 0$. Then,

$$\mathbb{E}[\tilde{\phi}_{k,t}(\theta_t^+) \mid \theta_t] \le -\frac{\eta}{4}\|\nabla\tilde{\phi}_{k,t}(\theta_t)\|^2 + \frac{\tilde{L}}{2}\eta^2\frac{\sigma^2}{N} + 5\eta\frac{\beta^2}{N}. \tag{224}$$

Assume $\|\tilde{\phi}_{k,t}(\theta_t)\|^2 > \epsilon^2$, so the line 9 of Algorithm 3 is valid. This means, by the previous inequality,

$$\mathbb{E}[\tilde{\phi}_{k,t}(\theta_t^+) \mid \theta_t] \le -\frac{\eta}{4}\epsilon^2 + \frac{\tilde{L}}{2}\eta^2\frac{\sigma^2}{N} + 5\eta\frac{\beta^2}{N} = -\eta\left(\frac{1}{4}\epsilon^2 - \frac{\tilde{L}}{2}\eta\frac{\sigma^2}{N} - 5\frac{\beta^2}{N}\right) \le -\eta\frac{\epsilon^2}{2} = -\frac{\epsilon^2}{2\tilde{L}}. \tag{225}$$

This means that whenever line 9 of Algorithm 3 is triggered, we are guaranteed to make at least $\epsilon^2/2\tilde{L}$ progress. Note that otherwise, we have $\|\tilde{\phi}_{k,t}(\theta_t)\|^2 \leq \epsilon^2$, which guarantees that we have achieved $\epsilon$-first-order stationarity.

We will now show that progress and stationarity in this smooth surrogate $\tilde{\phi}_{k,t}$ translates to progress and stationarity in the true objective $\phi_{k,t}$. By (215),

$$\mathbb{E}[\phi_{k,t}(\theta_t^+) \mid \theta_t] \leq \mathbb{E}[\tilde{\phi}_{k,t}(\theta_t^+) \mid \theta_t] + \delta_{\text{val}} \leq -\frac{\epsilon^2}{2\tilde{L}} + \delta_{\text{val}} =: -\epsilon_\phi. \tag{226}$$

By Remark E.6, we have shown that one can choose $J$ and $\kappa$ accordingly, so that $\delta_{\text{val}} < -\epsilon^2/2\tilde{L}$, ensuring that whenever there is progress in $\tilde{\phi}_{k,t}$, there is also true progress of at least $\epsilon_\phi$ in $\phi_{k,t}$.

Similarly, using Theorem E.5, and triangle inequality, we have that

$$\operatorname{dist}(0, \partial\phi_{k,t}(\theta_t)) \leq \|\nabla\tilde{\phi}_{k,t}(\theta_t)\| + \operatorname{dist}\left(\nabla\tilde{\phi}_{k,t}(\theta_t), \partial\phi_{k,t}(\theta_t)\right), \tag{227}$$

hence,

$$\mathbb{E}\left[\operatorname{dist}(0, \partial\phi_{k,t}(\theta_t))\right] \leq \epsilon + \epsilon', \tag{228}$$

where

$$\epsilon' := L_m\alpha + G_{\max}J\exp\left(-\frac{\delta_\Delta}{\kappa}\right). \tag{229}$$

Once again, $\alpha$, $J$, and $\kappa$ can be judiciously tuned such that whenever we have $\epsilon$-first-order stationarity with respect to $\tilde{\phi}_{k,t}$, we also have $(\epsilon + \epsilon')$-first-order stationarity with respect to the true objective $\phi_{k,t}$, for a desired value of $\epsilon'$.

To finalize, let us choose $T$ such that the algorithm is guaranteed to return a solution through line 14 within $T$ iterations, and establish an upper bound on $T$. Suppose that the loop runs for $T$ iterations, yet to be determined. We employ the same argument to the proof of Theorem 4.1: a triangle inequality in $V_k(\theta, \theta_t) := \phi_{k,t}(\theta)$, and lower-boundedness. We have,

$$-2 \leq \mathbb{E}[V_k(\theta_0, \theta_t)] \leq \sum_{t=0}^{T-1}\mathbb{E}[V_k(\theta_{t+1}, \theta_t)] \leq -T\epsilon_\phi. \tag{230}$$

Then, we are guaranteed that the algorithm will terminate and output a solution with

$$T \leq \frac{2}{\epsilon_\phi}. \tag{231}$$

Pick

$$T = \left\lceil \frac{2}{\frac{\epsilon^2}{2\tilde{L}} + L\alpha + \kappa\log J} \right\rceil = O\left(\frac{1}{\epsilon^2}\right). \tag{232}$$

Also, by construction, the final iterate $\theta_T$ satisfies

$$\|\nabla\tilde{\phi}_{k,t}(\theta_T)\|^2 \leq \epsilon^2, \tag{233}$$

□

yielding an $\epsilon$-Pareto stationary point.

One final detail that may be ironed out is the fact that the object of analysis of the previous theorem is the norm of the true gradient $\|\nabla\tilde{\phi}_{k,t}\|$, whereas we are limited to observing its stochastic estimates $g_t$ in practice. The following remark bridges this final gap.

*Remark* E.10. It is straightforward to relate the norm of the true gradient $\|\nabla\tilde{\phi}_{k,t}\|$ and of the stochastic gradient $g_t$ through triangle inequality and the well-known Markov's inequality. Suppose, at some time $t$, $\|g_t\| \leq \epsilon$. We have

$$\|\nabla\tilde{\phi}_{k,t}(\theta_t)\| \leq \|g_t\| + \|g_t - \nabla\tilde{\phi}_{k,t}(\theta_t)\|. \tag{234}$$

Because

$$\mathbb{E}\left[\left\|g_t - \nabla\tilde{\phi}_k(\theta_t)\right\|^2 \mid \theta_t\right] \leq \frac{\sigma^2 + \beta^2}{N}, \tag{235}$$

we can use Markov's inequality to get

$$\mathbb{P}\left(\left\|g_t - \nabla\tilde{\phi}_k(\theta_t)\right\| \geq \delta\right) \leq \frac{(\sigma^2 + \beta^2)}{N\delta^2} =: p_\delta, \tag{236}$$

which means that

$$\|\nabla\tilde{\phi}_{k,t}(\theta_t)\| \leq \epsilon + \delta, \tag{237}$$

with probability $1 - p_\delta$.

# F ANALYSIS OF SMOOTH-RELU

We note that smooth-ReLU is basically the Unit Normal Loss Integral (UNLI), a function used in decision analysis and risk modeling to calculate various metrics related to the expected value of information Wilson [2015].

Furthermore, smooth-ReLU is related to Gaussian Error Linear Unit (GELU) Hendrycks [2016], namely $x\Phi_1(x)$. But, while GELU is a smooth nonconvex under-approximation of ReLU, smooth-ReLU is a smooth convex over-approximation of ReLU. As the ensuing analysis shows, convexity and over-approximation properties are essential for our derivations.

*Remark* F.1 (Isomorphism to logistic smoothing and softplus). The ensuing derivations are invariant under the choice of the smoothing kernel, provided the kernel admits full support and finite moments. Specifically, the results extend directly to the softplus operator Dugas et al. [2000], $\zeta_s^{\text{sp}}(x) := s\log(1+e^{x/s})$, which is identified as the convolution of the ReLU with a logistic distribution $\mathcal{L}(0,s)$, and crucially is also a smooth convex over-approximation of ReLU. The validity of the ordering relations (Proposition F.3) relies exclusively on strict Jensen's inequality applied to the convex monomial $x \mapsto x^{k-1}$ $(k>2)$ over a non-degenerate random variable; this argument holds independently of the specific noise distribution. Similarly, the approximation error bounds (Proposition F.8) rely on the Lipschitz continuity properties of the smoothed function. By substituting the Gaussian measure $d\Phi_\mu$ with the logistic measure $dP_s$, all derived guarantees hold as long as the Gaussian approximation constant $\frac{\mu}{\sqrt{2\pi}}$ is replaced by the logistic constant $s\ln 2$.

The next result compares the two approximations to the cdf that leverage Gaussian smoothing for different SD orders $k$.

**Proposition F.2** (Comparison of smoothing approximations)**.** *For any $m \in M$ and $\mu > 0$, the relationship between the approximation and the true convolution depends on the order $k$:*

1. *For $k = 2$, the approximation is exact: $\tilde{F}_2^{approx}(m) = \tilde{F}_2^{true}(m)$.*
2. *For integer $k > 2$, the approximation strictly lower bounds the true convolution: $\tilde{F}_k^{approx}(m) < \tilde{F}_k^{true}(m)$.*
3. *For $k = 1$, using the convention $0^0 = 0$ and $x^0 = 1$ for $x > 0$, the approximation degenerates to a constant upper bound: $1 \equiv \tilde{F}_1^{approx}(m) > \tilde{F}_1^{true}(m)$.*

*Proof.* We analyze the relationship term-wise. Since the expectation operator $\mathbb{E}_R$ and the product over independent dimensions $i \in [d]$ are linear and monotonicity-preserving for non-negative terms, it suffices to compare the inner expectations for a fixed realization of $R_i$.

Let $y_i := m_i - R_i$ and define the random variable $V_i := (y_i - Z_i)_+$. Note that $V_i$ is non-degenerate (it has nonzero variance) due to the Gaussian noise $Z_i$. The comparison reduces to analyzing $A_i := (\mathbb{E}_{Z_i}[V_i])^{k-1}$ for the approximation versus $B_i := \mathbb{E}_{Z_i}[V_i^{k-1}]$ for the true convolution.

**Case $k = 2$:** The exponent is $k-1=1$. By the linearity of the expectation operator, $(\mathbb{E}[V_i])^1 = \mathbb{E}[V_i^1]$. Therefore, $A_i = B_i$, and the objectives are identical.

**Case $k > 2$:** The function $f(x) = x^{k-1}$ is strictly convex on $[0,\infty)$ for integers $k > 2$ (since $f''(x) > 0$ for $x > 0$). Because $V_i$ is a non-degenerate random variable, strict Jensen's inequality applies

$$(\mathbb{E}_{Z_i}[V_i])^{k-1} < \mathbb{E}_{Z_i}[V_i^{k-1}]. \tag{238}$$

Thus $A_i < B_i$. Since all terms are non-negative, the product over $d$ dimensions preserves this strict inequality, yielding $\tilde{F}_k^{\text{approx}}(m) < \tilde{F}_k^{\text{true}}(m)$.

**Case $k = 1$:** The exponent is 0. For the true convolution term $B_i$

$$B_i = \mathbb{E}_{Z_i}[(y_i - Z_i)_+^0] = \mathbb{E}_{Z_i}[\mathbb{I}_{\{y_i \geq Z_i\}}] = \Phi_\mu(y_i) < 1. \tag{239}$$

For the approximation term $A_i$, recall from Definition 5.1 that $\mathbb{E}_{Z_i}[V_i] = \zeta_\mu(y_i) > 0$. Therefore

$$A_i = (\mathbb{E}_{Z_i}[V_i])^0 = 1. \tag{240}$$

Consequently, $A_i > B_i$. The approximation $\tilde{F}_1^{\text{approx}}(m)$ evaluates to $c_1\mathbb{E}_R[\prod_{i=1}^d 1] = 1$, which strictly upper bounds the true smoothed probability mass. □

We now explore the relationships between the two approximations we explored above and the true cdf.

**Proposition F.3** (Ordering of exact and smoothed objectives)**.** *Let $F_R^k(m)$ be the exact $k$-th order cdf objective, and let $\tilde{F}_k^{true}(m)$ and $\tilde{F}_k^{approx}(m)$ be the smoothed variations defined in Definition 5.2. For any $m \in M$ and $\mu > 0$, the following ordering relations hold almost everywhere (with respect to the measure of $R$):*

1. ***For** $k = 2$ **(Overestimation):** Both approximations are strictly greater than the exact objective, and they are equal to each other:*

$$\tilde{F}_2^{true}(m) = \tilde{F}_2^{approx}(m) > F_R^2(m). \tag{241}$$

2. ***For integers** $k > 2$ **(Strict overestimation chain):** Both approximations strictly overestimate the exact objective, with the true convolution providing a looser bound than the approximation:*

$$\tilde{F}_k^{true}(m) > \tilde{F}_k^{approx}(m) > F_R^k(m). \tag{242}$$

3. ***For** $k = 1$ **(Indeterminate ordering):** The approximation $\tilde{F}_1^{approx}(m)$ is a trivial upper bound (constant 1). However, the true convolution $\tilde{F}_1^{true}(m)$ admits no global ordering with respect to $F_R^1(m)$; it overestimates $F_R^1(m)$ in regions of low probability density and underestimates it in regions of high probability density.*

*Proof.* Let $y_i := m_i - R_i$ for $i \in [d]$. Since the expectations over $R$ and the positive scaling constant $c_k$ preserve inequality order, it suffices to prove the inequalities for the kernels inside the expectation $\mathbb{E}_R$. We rely on the following lemma regarding convex convolutions.

**Lemma F.4** (Jensen's inequality for convolution)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a convex function and $Z \sim \mathcal{N}(0, \mu^2)$. Then for any $y \in \mathbb{R}$:*

$$\mathbb{E}_Z[g(y - Z)] \geq g(y). \tag{243}$$

*If $g$ is strictly convex on a set of nonzero measure in the support of $y - Z$, the inequality is strict.*

*Proof of lemma.* This is a direct application of Jensen's inequality.

$$\mathbb{E}_Z[g(y - Z)] \geq g(\mathbb{E}_Z[y - Z]) = g(y - \mathbb{E}[Z]) = g(y). \tag{244}$$

□

**Proof of Case 1 ($k = 2$):** Consider the kernel function $g(u) = (u)_+$. This function is convex. For the true convolution term $\tilde{F}_2^{\text{true}}$, the inner kernel is $\mathbb{E}_Z[(y_i - Z_i)_+]$. Applying Lemma F.4, we have $\mathbb{E}_Z[(y_i - Z_i)_+] \geq (y_i)_+$. Due to the infinite support of the Gaussian pdf, the convolution $\zeta_\mu(y_i) = \mathbb{E}_Z[(y_i - Z_i)_+]$ is strictly positive everywhere, whereas $(y_i)_+ = 0$ for $y_i \leq 0$. Thus, for any $y_i$, $\zeta_\mu(y_i) > (y_i)_+$ is possible, and specifically $\zeta_\mu(y_i) \geq (y_i)_+$ holds everywhere. Since $\tilde{F}_2^{\text{approx}}$ is defined by the product of these expectations $\zeta_\mu(y_i)$, and we established in Proposition F.2 that $\tilde{F}_2^{\text{true}} = \tilde{F}_2^{\text{approx}}$, it follows that

$$\prod_{i=1}^{d} \zeta_\mu(y_i) \geq \prod_{i=1}^{d} (y_i)_+. \tag{245}$$

Taking the expectation $\mathbb{E}_R$ yields $\tilde{F}_2^{\text{approx}}(m) \geq F_R^2(m)$. The inequality is strict if probability density exists where $(y_i)_+ = 0$ (which implies $y_i \leq 0$ where $\zeta_\mu > 0$).

**Proof of Case 2 ($k > 2$):** We establish the chain $\tilde{F}_k^{\text{true}} > \tilde{F}_k^{\text{approx}} > F_R^k$.

1. $\tilde{F}_k^{\textbf{true}} > \tilde{F}_k^{\textbf{approx}}$**:** This was proven in Proposition F.2 via strict Jensen's inequality on the convex function $x \mapsto x^{k-1}$.
2. $\tilde{F}_k^{\textbf{approx}} > F_R^k$**:** Recall the kernel for $\tilde{F}_k^{\text{approx}}$ is $\prod(\zeta_\mu(y_i))^{k-1}$ and for $F_R^k$ is $\prod((y_i)_+)^{k-1}$. From Case 1, we know $\zeta_\mu(y_i) \geq (y_i)_+ \geq 0$. Since the power function $x \mapsto x^{k-1}$ is strictly monotonically increasing for $x \geq 0$ (given $k > 2$), it preserves the order

$$(\zeta_\mu(y_i))^{k-1} \geq ((y_i)_+)^{k-1}. \tag{246}$$

Consequently, $\tilde{F}_k^{\text{approx}}(m) \geq F_R^k(m)$. The inequality is strictly enforced by the strict positivity of $\zeta_\mu$ versus the sparsity of ReLU.

**Proof of Case 3 ($k = 1$):**

1. $\tilde{F}_1^{\text{approx}} > F_R^1$**:** $\tilde{F}_1^{\text{approx}}(m) = c_1 \mathbb{E}_R[\prod 1] = 1$. Since $F_R^1(m)$ is a cumulative probability bounded by 1 (and strictly less than 1 if $m$ is not the maximal support), the approximation is a strict upper bound.
2. $\tilde{F}_1^{\text{true}}$ **vs** $F_R^1$ **(Indeterminate):** Consider the univariate setting for simplicity. $F_R^1(m) = \mathbb{I}_{\{m \geq R\}}$. $\tilde{F}_1^{\text{true}}(m) = \Phi_\mu(m - R)$.
   - Let $m \ll R$ (far left tail). Then $F_R^1(m) = 0$. But $\tilde{F}_1^{\text{true}}(m) = \Phi_\mu(\text{negative}) > 0$. Here $\tilde{F}_1^{\text{true}} > F_R^1$.
   - Let $m = R$. Then $F_R^1(m) = 1$. But $\tilde{F}_1^{\text{true}}(m) = \Phi_\mu(0) = 0.5$. Here $\tilde{F}_1^{\text{true}} < F_R^1$.

   Since the relationship depends on the specific value of $m$ relative to $R$, no global dominance order exists.

□

Next, we adopt some results on the approximation power of Gaussian smoothing (primarily from the optimization literature, see e.g. Nesterov and Spokoiny [2017]) to our setup. These results precisely characterize the error incurred when approximating the true cdf with the proposed smoothed approximations defined in Definition 5.2. We start with the approximation error of $\tilde{F}_k^{\text{true}}$.

**Proposition F.5** (Gaussian smoothing approximation error)**.** *Let $k \geq 2$. Let $\tilde{F}_k^{true}(m)$ be the Gaussian smoothing of $F_R^k(m)$ with parameter $\mu > 0$, where $Z \sim \mathcal{N}(0, \mu^2 I_d)$. The approximation error is uniformly bounded by:*

$$\left| F_R^k(m) - \tilde{F}_k^{true}(m) \right| \leq d(k-1) D^{-1} \mu. \tag{247}$$

*Proof.* Let $g(m, R) = \prod_{i=1}^d (m_i - R_i)_+^{k-1}$. By definition

$$\left| F_R^k(m) - \tilde{F}_k^{\text{true}}(m) \right| = \left| c_k \mathbb{E}_R \left[ g(m, R) \right] - c_k \mathbb{E}_R \left[ \mathbb{E}_Z \left[ g(m - Z, R) \right] \right] \right| \leq c_k \mathbb{E}_R \left[ \mathbb{E}_Z \left[ \left| g(m, R) - g(m - Z, R) \right| \right] \right]. \tag{248}$$

Using the Lipschitz continuity property from Proposition B.5,

$$|g(m, R) - g(m - Z, R)| \leq \left( \sqrt{d}(k-1) D^{d(k-1)-1} \right) \|Z\|_2. \tag{249}$$

Taking expectations over $Z$,

$$\mathbb{E}_Z[\|Z\|_2] = \mu \mathbb{E}_U[\|U\|_2], \quad \text{where } U \sim \mathcal{N}(0, I_d). \tag{250}$$

By Jensen's inequality,

$$\mathbb{E}_U[\|U\|_2] \leq \sqrt{\mathbb{E}[\|U\|_2^2]} = \sqrt{d}. \tag{251}$$

Substituting back,

$$\left| F_R^k(m) - \tilde{F}_k^{\text{true}}(m) \right| \leq c_k \left( \sqrt{d}(k-1) D^{d(k-1)-1} \right) (\mu \sqrt{d}) = d(k-1) D^{-1} \mu. \tag{252}$$

□

*Remark* F.6 (Smoothness-approximation trade-off). While the approximation error scales with $O(\mu)$, the **smoothness** of $\tilde{F}_k^{\text{true}}(m)$ (the Lipschitz constant of its gradient, $\nabla \tilde{F}$) scales with $O(1/\mu)$. Specifically, for any bounded function, convolution with a Gaussian of variance $\mu^2 I$ yields a function with $\frac{C}{\mu}$-Lipschitz continuous gradients Nesterov and Spokoiny [2017]. This represents the fundamental trade-off in Gaussian smoothing: smaller $\mu$ yields lower approximation error but poorer conditioning (higher curvature), while larger $\mu$ improves smoothness at the cost of accuracy.

Next, we quantify the approximation error of $\tilde{F}_k^{\text{approx}}$. We start with a simple lemma first.

**Lemma F.7** (Uniform approximation of smooth-ReLU)**.** *For any $x \in \mathbb{R}$ and $\mu > 0$, the smooth-ReLU function $\zeta_\mu(x)$ approximates the ReLU function $(x)_+$ with uniform error bounded by the scaled Gaussian peak:*

$$0 < \zeta_\mu(x) - (x)_+ \leq \frac{\mu}{\sqrt{2\pi}}. \tag{253}$$

*Proof.* Consider the difference function $\Delta(x) := \zeta_\mu(x) - (x)_+$. Using the explicit form $\zeta_\mu(x) = x\Phi(\frac{x}{\mu}) + \mu\phi(\frac{x}{\mu})$

- For $x \le 0$: $(x)_+ = 0$, so $\Delta(x) = x\Phi(\frac{x}{\mu}) + \mu\phi(\frac{x}{\mu})$. As $x \to -\infty$, $\Delta(x) \to 0$. The derivative $\Delta'(x) = \Phi(\frac{x}{\mu}) > 0$. Thus $\Delta(x)$ is increasing on $(-\infty, 0]$.
- For $x > 0$: $(x)_+ = x$, so $\Delta(x) = x(\Phi(\frac{x}{\mu}) - 1) + \mu\phi(\frac{x}{\mu})$. The derivative $\Delta'(x) = \Phi(\frac{x}{\mu}) - 1 < 0$. Thus $\Delta(x)$ is decreasing on $[0, \infty)$.

The global maximum occurs at $x = 0$, where $\Delta(0) = \mu\phi(0) = \frac{\mu}{\sqrt{2\pi}}$. Thus, the uniform error is bounded by $\frac{\mu}{\sqrt{2\pi}}$. □

**Proposition F.8** (Approximation error bound)**.** *Let $k \ge 2$ be an integer. Let $F_R^k(m)$ be the exact objective and $\tilde{F}_k^{approx}(m)$ be the approximation using the smooth-ReLU defined as:*

$$F_R^k(m) = c_k \mathbb{E}_R \left[ \prod_{i=1}^d (m_i - R_i)_+^{k-1} \right], \tag{254}$$

$$\tilde{F}_k^{approx}(m) = c_k \mathbb{E}_R \left[ \prod_{i=1}^d (\zeta_\mu(m_i - R_i))^{k-1} \right]. \tag{255}$$

*Given the compactness of domain discussed as a consequence of Assumption B.1, the error is bounded as*

$$\left| F_R^k(m) - \tilde{F}_k^{approx}(m) \right| \le \frac{d(k-1)\left(D + \frac{\mu}{\sqrt{2\pi}}\right)^{d(k-1)-1}}{(k-1)!^d \sqrt{2\pi}} \mu. \tag{256}$$

*Proof.* Let $y_i = m_i - R_i$. Define $u_i = (y_i)_+$ and $v_i = \zeta_\mu(y_i)$. By Lemma F.7, we have $|v_i - u_i| \le \frac{\mu}{\sqrt{2\pi}}$. We bound the difference of the product terms inside the expectation. Let $\mathcal{P}(u) = \prod_{i=1}^d u_i^{k-1}$ and $\mathcal{P}(v) = \prod_{i=1}^d v_i^{k-1}$. Using the telescoping sum decomposition for products

$$\left| \prod_{i=1}^d v_i^{k-1} - \prod_{i=1}^d u_i^{k-1} \right| \le \sum_{j=1}^d \left| v_j^{k-1} - u_j^{k-1} \right| \left( \prod_{l<j} v_l^{k-1} \prod_{l>j} u_l^{k-1} \right). \tag{257}$$

First, we bound the terms $u_l$ and $v_l$. By Assumption B.1, $u_l \le D$. By Lemma F.7, $v_l \le D + \frac{\mu}{\sqrt{2\pi}}$. Let $D_\mu := D + \frac{\mu}{\sqrt{2\pi}}$. Then $\max(u_l, v_l) \le D_\mu$. The product term for any $j$ contains $d-1$ factors, each bounded by $D_\mu^{k-1}$. Thus

$$\left( \prod_{l<j} v_l^{k-1} \prod_{l>j} u_l^{k-1} \right) \le (D_\mu^{k-1})^{d-1} = D_\mu^{(d-1)(k-1)}. \tag{258}$$

Next, we bound $|v_j^{k-1} - u_j^{k-1}|$. Consider the function $g(t) = t^{k-1}$. For $k \ge 2$, $g(t)$ is differentiable with $g'(t) = (k-1)t^{k-2}$. By the Mean Value Theorem on $[0, D_\mu]$

$$|v_j^{k-1} - u_j^{k-1}| \le \left( \max_{t \in [0, D_\mu]} |g'(t)| \right) |v_j - u_j| \le (k-1) D_\mu^{k-2} \cdot \frac{\mu}{\sqrt{2\pi}}. \tag{259}$$

Substituting these back into the summation

$$|\mathcal{P}(v) - \mathcal{P}(u)| \le \sum_{j=1}^d \left( (k-1) D_\mu^{k-2} \frac{\mu}{\sqrt{2\pi}} \right) D_\mu^{(d-1)(k-1)} \tag{260}$$

$$= d(k-1) D_\mu^{dk-d-k+1+k-2} \frac{\mu}{\sqrt{2\pi}} \tag{261}$$

$$= d(k-1) D_\mu^{d(k-1)-1} \frac{\mu}{\sqrt{2\pi}}. \tag{262}$$

Finally, applying the expectation operator $\mathbb{E}_R$ and the constant $c_k$

$$\left| F_R^k(m) - \tilde{F}_k^{\text{approx}}(m) \right| \le c_k \mathbb{E}_R[|\mathcal{P}(v) - \mathcal{P}(u)|] \le \frac{1}{(k-1)!^d} \cdot \frac{d(k-1)}{\sqrt{2\pi}} \left( D + \frac{\mu}{\sqrt{2\pi}} \right)^{d(k-1)-1} \mu. \tag{263}$$

□

*Remark* F.9 (Consistency and tightness of the smooth-ReLU approximation). The error bounds derived in Proposition F.5 and Proposition F.8, in conjunction with the ordering relations in Proposition F.3, present a coherent geometric picture of the smoothing mechanisms.

First, both approximations are consistent estimators of the non-smooth objective, converging linearly as $O(\mu)$. However, for higher-order dominance ($k > 2$), the ordering relation

$$\tilde{F}_k^{\text{true}}(m) > \tilde{F}_k^{\text{approx}}(m) > F_R^k(m) \tag{264}$$

indicates that the *approximate* formulation (smoothing the ReLU components individually) yields a **tighter** upper bound on the exact objective than the *true* convolution (smoothing the product of ReLUs).

This structural tightness is reflected in the derived error bounds. The bound for the true convolution scales with $\mu\sqrt{d}$ (arising from the expectation of the Euclidean norm of the $d$-dimensional noise vector), whereas the bound for the approximation scales principally with $\mu$ (arising from the component-wise deviation of the smooth-ReLU). Since $\sqrt{d} \geq 1$, the looser bound for $\tilde{F}_k^{\text{true}}$ allows for the larger overestimation gap observed in the ordering proposition.

Consequently, for optimization purposes where $k \geq 2$, the proposed approximation $\tilde{F}_k^{\text{approx}}$ is superior to the true convolution $\tilde{F}_k^{\text{true}}$: it is computationally tractable (avoiding high-dimensional integration), retains necessary smoothness properties, and lies strictly closer to the original non-smooth objective manifold.

The following theorem discusses the implication for solving the proposed task in (9).

**Theorem F.10** (Total bias of symmetric smooth-ReLU approximation)**.** *Let $k \geq 2$ be an integer. Consider the exact objective $\mathcal{L}_k(\theta, m)$ and the symmetric smoothed surrogate $\tilde{\mathcal{L}}_k^\mu(\theta, m)$ defined as:*

$$\mathcal{L}_k(\theta, m) := F_{R(\tau)}^k(m) - F_{R(\tau_t)}^k(m), \tag{265}$$

$$\tilde{\mathcal{L}}_k^\mu(\theta, m) := \tilde{F}_k^{approx}(m; \theta) - \tilde{F}_k^{approx}(m; \theta_t). \tag{266}$$

*Let $p_\theta$ and $p_{\theta_t}$ denote the probability density functions of the trajectories induced by $\theta$ and $\theta_t$, respectively. Under Assumption B.1, the total bias is bounded by:*

$$\left|\mathcal{L}_k(\theta, m) - \tilde{\mathcal{L}}_k^\mu(\theta, m)\right| \leq \left(\frac{2d(k-1)\left(D + \frac{\mu}{\sqrt{2\pi}}\right)^{d(k-1)-1}}{(k-1)!^d\sqrt{2\pi}}\right) \cdot \mu \cdot \|p_\theta - p_{\theta_t}\|_{\text{TV}}. \tag{267}$$

*Proof.* We begin by defining the kernels for the exact and approximate objectives. Let $y(\tau) := m - R(\tau) \in \mathbb{R}^d$. Define

$$g(y) := c_k \prod_{i=1}^d (y_i)_+^{k-1}, \tag{268}$$

$$\tilde{g}_\mu(y) := c_k \prod_{i=1}^d (\zeta_\mu(y_i))^{k-1}. \tag{269}$$

Let $\Delta_\mu(\tau) := \tilde{g}_\mu(y(\tau)) - g(y(\tau))$ represent the pointwise approximation error for a given trajectory $\tau$. The total bias can be expanded as

$$\text{Bias} := \left|\mathcal{L}_k(\theta, m) - \tilde{\mathcal{L}}_k^\mu(\theta, m)\right| = \left|(F_\theta^k - F_{\theta_t}^k) - (\tilde{F}_\theta^k - \tilde{F}_{\theta_t}^k)\right| = \left|(\tilde{F}_\theta^k - F_\theta^k) - (\tilde{F}_{\theta_t}^k - F_{\theta_t}^k)\right|. \tag{270}$$

Substituting the expectations over trajectory densities

$$\tilde{F}_\theta^k - F_\theta^k = \mathbb{E}_{\tau\sim p_\theta}[\Delta_\mu(\tau)] = \int \Delta_\mu(\tau) p_\theta(\tau) \mathrm{d}\tau. \tag{271}$$

Thus, the bias becomes

$$\text{Bias} = \left|\int \Delta_\mu(\tau) p_\theta(\tau)\mathrm{d}\tau - \int \Delta_\mu(\tau) p_{\theta_t}(\tau)\mathrm{d}\tau\right| \tag{272}$$

$$= \left|\int \Delta_\mu(\tau)(p_\theta(\tau) - p_{\theta_t}(\tau))\mathrm{d}\tau\right|. \tag{273}$$

By Hölder's inequality for integrals, this is bounded by the supremum of the error kernel and the $L_1$ norm of the density difference

$$\text{Bias} \leq \left( \sup_{\tau} |\Delta_\mu(\tau)| \right) \int |p_\theta(\tau) - p_{\theta_t}(\tau)| \mathrm{d}\tau. \tag{274}$$

Recalling the definition of total variation distance, $\int |p - q| = 2\|p - q\|_{\text{TV}}$, we have

$$\text{Bias} \leq 2 \left( \sup_{\tau} |\Delta_\mu(\tau)| \right) \|p_\theta - p_{\theta_t}\|_{\text{TV}}. \tag{275}$$

It remains to bound the supremum of the kernel error $\Delta_\mu$. From the proof of Proposition F.8, we established the pointwise bound

$$|\Delta_\mu(\tau)| \leq c_k \cdot d(k-1) \left( D + \frac{\mu}{\sqrt{2\pi}} \right)^{d(k-1)-1} \frac{\mu}{\sqrt{2\pi}}. \tag{276}$$

Substituting this supremum back into the bias inequality,

$$\begin{aligned} \text{Bias} &\leq 2 \left[ \frac{1}{(k-1)!^d} \cdot \frac{d(k-1)\mu}{\sqrt{2\pi}} \left( D + \frac{\mu}{\sqrt{2\pi}} \right)^{d(k-1)-1} \right] \\ &\quad \cdot \|p_\theta - p_{\theta_t}\|_{\text{TV}} \\ &= \left( \frac{2d(k-1) \left( D + \frac{\mu}{\sqrt{2\pi}} \right)^{d(k-1)-1}}{(k-1)!^d \sqrt{2\pi}} \right) \cdot \mu \\ &\quad \cdot \|p_\theta - p_{\theta_t}\|_{\text{TV}}. \end{aligned} \tag{277}$$

□

It is worth recalling (74) from the proof of Proposition D.5, namely,

$$\|p_\theta - p_{\theta_t}\|_{\text{TV}} \leq \frac{1}{2} B_p \|\theta - \theta_t\|. \tag{278}$$

Using Theorem F.10, we can establish the following sufficient condition to ensure a small approximation bias.

**Corollary F.11** (Selection of smoothing parameter $\mu$ for prescribed accuracy)**.** *Let $\epsilon > 0$ be a target error tolerance, and let $\Lambda := \|p_\theta - p_{\theta_t}\|_{\text{TV}} \leq \frac{1}{2} B_p \|\theta - \theta_t\|$ denote the total variation distance between the current and target trajectory densities. Assume $\Lambda > 0$ (otherwise the bias is zero). Define the dimension-dependent exponent $\alpha := d(k-1) - 1$ and the scaling constant:*

$$\mathcal{K} := \frac{(k-1)!^d \sqrt{2\pi}}{2d(k-1)(2D)^\alpha}. \tag{279}$$

*Then, the symmetric smoothed objective bias satisfies*

$$\left| \mathcal{L}_k(\theta, m) - \tilde{\mathcal{L}}_k^\mu(\theta, m) \right| \leq \epsilon \tag{280}$$

*provided that the smoothing parameter $\mu$ is chosen to satisfy*

$$\mu \leq \min \left( \sqrt{2\pi} D, \quad \mathcal{K} \cdot \frac{\epsilon}{\Lambda} \right). \tag{281}$$

*Proof.* The proof relies on establishing a uniform upper bound for the polynomial term in Theorem F.10. We impose a preliminary constraint that the smoothing scale does not exceed the domain scale: $\mu \leq \sqrt{2\pi} D$. Under this constraint, the term $\left( D + \frac{\mu}{\sqrt{2\pi}} \right)$ is bounded by

$$D + \frac{\mu}{\sqrt{2\pi}} \leq D + \frac{\sqrt{2\pi} D}{\sqrt{2\pi}} = 2D. \tag{282}$$

Since $k \geq 2$ and $d \geq 1$, the exponent $\alpha = d(k-1) - 1 \geq 0$. The power function $x \mapsto x^\alpha$ is non-decreasing for $x > 0$. Therefore,

$$\left(D + \frac{\mu}{\sqrt{2\pi}}\right)^\alpha \leq (2D)^\alpha. \tag{283}$$

Substituting this into the bias bound from Theorem F.10,

$$\text{Bias} \leq \left(\frac{2d(k-1)}{(k-1)!^d\sqrt{2\pi}}\right)(2D)^\alpha \cdot \mu \cdot \Lambda. \tag{284}$$

We require this quantity to be at most $\epsilon$, hence,

$$\left(\frac{2d(k-1)(2D)^\alpha}{(k-1)!^d\sqrt{2\pi}}\right)\Lambda \cdot \mu \leq \epsilon. \tag{285}$$

Rearranging for $\mu$, we obtain the condition

$$\mu \leq \left(\frac{(k-1)!^d\sqrt{2\pi}}{2d(k-1)(2D)^\alpha}\right)\frac{\epsilon}{\Lambda} = \mathcal{K}\frac{\epsilon}{\Lambda}. \tag{286}$$

Thus, taking $\mu = \min(\sqrt{2\pi}D, \mathcal{K}\frac{\epsilon}{\Lambda})$ satisfies both the polynomial bound assumption and the target accuracy requirement. □

**Proposition F.12** (Control of parameter divergence $\Delta_\theta$)**.** *Let $\theta_t$ be the parameter at the start of outer iteration $t$ of Algorithm 1. Let $\theta$ be the parameter vector obtained after $\bar{T}$ steps of the inner loop in Algorithm 1, utilizing a learning rate schedule $\{\eta_j\}_{j=0}^{\bar{T}-1}$. The parameter divergence $\Delta_\theta := \|\theta - \theta_t\|$ is bounded by:*

$$\Delta_\theta \leq G_{\max}\sum_{j=0}^{\bar{T}-1}\eta_j. \tag{287}$$

*Specifically, for a constant learning rate $\eta$ and a maximum of $\bar{T}$ inner iterations:*

$$\|\theta - \theta_t\| \leq \eta \cdot \bar{T} \cdot \left(\frac{D^{d(k-1)}B_p}{(k-1)!^d}\right). \tag{288}$$

*Proof.* The update rule in Algorithm 1 (line 11) is

$$\theta_{t,\bar{t}+1} \leftarrow \theta_{t,\bar{t}} - \eta_{\bar{t}}g_{t,\bar{t}}. \tag{289}$$

Let $\theta_{t,0} = \theta_t$. By recursive expansion,

$$\theta_{t,\bar{T}} - \theta_{t,0} = -\sum_{j=0}^{\bar{T}-1}\eta_j g_{t,j}. \tag{290}$$

Applying the Euclidean norm and the triangle inequality,

$$\|\theta_{t,\bar{T}} - \theta_t\| = \left\|\sum_{j=0}^{\bar{T}-1}\eta_j g_{t,j}\right\| \leq \sum_{j=0}^{\bar{T}-1}\eta_j\|g_{t,j}\|. \tag{291}$$

Using the uniform gradient bound $G_{\max}$ from Proposition D.4,

$$\|\theta_{t,\bar{T}} - \theta_t\| \leq G_{\max}\sum_{j=0}^{\bar{T}-1}\eta_j. \tag{292}$$

□

*Remark* F.13 (Choosing the smoothing parameter $\mu$). This result closes the loop with Corollary F.11. To guarantee that the symmetric smoothing bias is within $\epsilon$, one must choose $\mu$ based on $\Delta_\theta$. By controlling the learning rate $\eta$ and inner steps $\bar{T}$, we enforce an upper bound on $\Delta_\theta$, which in turn provides a strict lower bound on the required smoothing parameter $\mu$, preventing numerical instability associated with vanishing $\mu$.

# G EXPERIMENTAL DETAILS

This section provides a complete account of the experimental protocol used to produce the results in Section 6, including environment choices, training procedure, PPO and ORDO mixing, scheduling conventions, and evaluation settings.

We evaluate on five MORL benchmarks from MO-Gymnasium [Felten et al., 2023]: `deep-sea-treasure-v0`, `fruit-tree-v0`, `mo-mountaincarcontinuous-v0`, `mo-hopper-v4`, and `mo-halfcheetah-v4`. Following the protocol of prior work by Cai et al. [2023], we use the two-objective setting and retain only the first two reward coordinates when environments expose more than two reward dimensions. For each environment, we train a set of $N = 20$ policies and evaluate the resulting policy set with the same four metrics reported in prior work: expected utility (EU), hypervolume (HV), constraint satisfaction, and variance objective. These metrics are defined as follows:

**Expected utility (EU).** Expected Utility is similar to R-metrics used in multi-objective optimization. It measures the expected utility of the best policy in the set under randomly sampled linear preferences:

$$\mathrm{EU}(\Pi) = \mathbb{E}_{w\sim\mathcal{W}}\left[\max_{\pi\in\Pi} f_w(z)\right], \tag{293}$$

where $f_w$ denotes the user's utility function linearly parameterized by $w$ and $z$ denotes the (undiscounted) returns of one evaluation episode.

**Hypervolume (HV).** Hypervolume measures the Lebesgue volume in the objective space dominated by the policy set with respect to a reference point:

$$\mathrm{HV}(\Pi, r) = \mathrm{Leb}\big(\{y \in \mathbb{R}^K \mid \exists\pi \in \Pi : z \le y \le r\}\big), \tag{294}$$

where $r$ is the reference point and Leb denotes the Lebesgue measure (volume). Following Cai et al. [2023], the reference points in our experiments are computed as the minimal mean returns of each evaluated method.

**Distributional MORL metrics.** We additionally report *Constraint Satisfaction* and *Variance Objective*. For both metrics, we define $u(p, \pi)$ to be the satisfaction score of policy $\pi$ under preference $p$, generate $M = 100$ preferences $p_1, \ldots, p_M$, and score the policy set by

$$\frac{1}{M}\sum_{i=1}^{M} \max_{j\in\{1,\ldots,N\}} u(p_i, \pi_j). \tag{295}$$

**Constraint Satisfaction.** For Constraint Satisfaction, we define $u(p, \pi)$ to be the probability that a random episodic return $z \sim Z(\pi)$ satisfies a (randomly generated) constraint $p$. Each constraint $p$ is defined by $n$ linear inequalities $\{z \mid \forall i,\ w_i^\top z \ge c_i\}$, where each $w_i$ is sampled uniformly from $\{w \succeq 0,\ \mathbf{1}^\top w = 1\}$ and each $c_i$ is sampled from $\mathcal{U}(\min_{z\in\mathcal{Z}} w_i^\top z,\ \max_{z\in\mathcal{Z}} w_i^\top z)$, where $\mathcal{Z}$ is the set of all return samples of the final policies learned by our method.

**Variance Objective.** For Variance Objective, we define $u(p, \pi)$ to be a weighted sum of the expected return $\mathbb{E}_{z\sim\mu(\pi)}[z]$ and the standard deviation $\sqrt{\mathrm{Var}_{z\sim\mu(\pi)}[z]}$. We generate weights $w_i \in \mathbb{R}^{2K}$ with $w_i \succeq 0$ and $\mathbf{1}^\top w_i = 1$, write $w_i = [w_i^{(1)}, w_i^{(2)}]$ with $w_i^{(1)}, w_i^{(2)} \in \mathbb{R}^K$, and use $w_i^{(1)}$ and $-w_i^{(2)}$ as the weights for the expected return and the standard deviation, respectively.

## G.1 POLICY-SET CONSTRUCTION

Each policy is trained independently using a fixed scalarization weight vector $w \in \Delta_2$. For the two-objective setting, we use a predefined grid of preference vectors to cover the trade-off spectrum. Concretely, for each seed/policy index $i$, we assign a weight vector $w_i \in \Delta_2$ and optimize the scalarized utility $u_{w_i}(\tau) = w_i^\top R(\tau)$. This produces a portfolio of 20 policies intended to span qualitatively different trade-offs before distributional filtering and refinement through ORDO.

ORDO compares policies through return distributions and stochastic dominance violations. However, obtaining a diverse and competitive set of policies still benefits from initializing optimization around different scalarized preferences. The weight grid provides a simple, reproducible mechanism for this diversity.

### G.2 PPO-ORDO MIXED TRAINING OBJECTIVE

Our final training procedure uses a mixture of PPO and ORDO updates. At each update, we form

$$g_t = (1 - \alpha_t) g_t^{\mathrm{PPO}} + \alpha_t g_t^{\mathrm{ORDO}}, \tag{296}$$

and apply the optimizer update using this mixed gradient.

The term $g_t^{\mathrm{PPO}}$ is the gradient of a surrogate objective similar to PPO, clipped for the scalarized return $u_w$, together with the usual value-function and entropy-regularization terms. In particular, the standard component of the loss takes the form

$$\mathcal{L}_{\mathrm{PPO}} = \mathcal{L}_{\mathrm{clip}} + c_v \mathcal{L}_v - c_{\mathrm{ent}} \mathcal{H}(\pi_\theta), \tag{297}$$

where $\mathcal{L}_{\mathrm{clip}}$ is the clipped policy objective, $\mathcal{L}_v$ is the value loss, and $\mathcal{H}$ is the entropy. For continuous-action environments, we align the policy and value parameterization and optimization conventions with standard Stable Baselines3 [Raffin et al., 2021] settings for the PPO parameters to reduce implementation confounds in the comparison against PPO-based baselines.

The term $g_t^{\mathrm{ORDO}}$ is the policy-gradient estimator induced by the ORDO dominance-violation objective, computed using rollout samples together with an approximate solution of the inner maximization over orthant generators.

**PPO hyperparameters.** We use Adam with learning rate $3 \times 10^{-4}$, clip range $0.2$, GAE $\lambda = 0.95$, value coefficient $0.5$, max grad-norm $0.5$, 5 PPO epochs per update, and minibatch size 256. Each update collects 20 on-policy episodes.

### G.3 ORDO INNER MAXIMIZATION AND SMOOTHING

To obtain the ORDO signal, we compare the current policy against a periodically refreshed reference snapshot $\pi_{\mathrm{ref}}$ (refreshed every 10 update steps in our main runs). We compute the empirical dominance violation and approximately solve $\max_{m \in \mathcal{M}} \widehat{L}_k(R_{\mathrm{cand}}, R_{\mathrm{ref}}, m)$ by projected gradient ascent on $m$ (Adam, 50 steps, step size 0.1). The constraint set $M$ is instantiated as a box using a running min/max of observed episodic returns in each coordinate, expanded by a small margin. For stability we use a smooth hinge surrogate (softplus) inside the kernel with fixed smoothing parameter $\mu = 0.5$.

### G.4 MIXING SCHEDULES

We schedule $\alpha_t$ by environment transitions so that early training is dominated by PPO and ORDO refinement is applied later. Hopper, HalfCheetah, and DeepSeaTreasure use $\alpha_t = 0$ for the first 7M transitions followed by a linear ramp to $0.2$ over the final 3M transitions. MountainCar uses $\alpha_t = 0$ for the first 9M transitions followed by a linear ramp to $0.05$ over the final 1M transitions. FruitTree begins its linear ramp after 2M transitions, reaches $0.4$ after a further 5M transitions, and remains at $0.4$ thereafter. These schedules were selected independently for the different environment dynamics and reward scales.

### G.5 TRAINING BUDGETS, ROLLOUT COLLECTION AND SCHEDULING

Each policy is trained for a fixed budget of $10^7$ environment transitions, counting both candidate-policy and reference-policy rollouts collected by the mixed update. We use second-order dominance ($k = 2$) in the experiments. For practical stability, the integrated-cdf kernel uses the smoothed surrogate described in Section 5 (softplus/smooth-hinge), with fixed smoothing hyperparameters. The smoothing is used only to stabilize optimization and gradient estimation.

The scalarization and PPO-ORDO mixture operate directly on the benchmark reward coordinates for Hopper, HalfCheetah, MountainCar, and DeepSeaTreasure. FruitTree uses coordinatewise return normalization because its reward coordinates have substantially different scales. For continuous-action environments, we use a Gaussian policy with learned mean and log-standard deviation.

### G.6 EVALUATION PROTOCOL

Each trained policy is evaluated for 256 episodes, and the empirical return samples are aggregated to compute the four metrics for the final policy set. Evaluation metrics are computed from undiscounted episodic returns. We report the mean and sample standard deviation across three independently trained 20-policy portfolios. The same sampled utility weights and constraints are used across replicate portfolios so that variability reflects the learned portfolios rather than independently resampled evaluation preferences.