CVaR Deep Optimal Stopping

Yechan Jeong
Seoul National University
yechan@snu.ac.kr

Youngjoon Hong
Seoul National University
hongyj@snu.ac.kr

Abstract

Risk-sensitive optimal stopping is useful when poor stopping outcomes carry substantial costs, as in option exercise, sequential medical treatment, and safety-critical control. A natural objective is Conditional Value at Risk (CVaR) of the stopped reward. We develop a theoretical framework and algorithm for CVaR deep optimal stopping. The framework and algorithm address four difficulties in applying risk-sensitive reinforcement learning to optimal stopping: state augmentation, score-function pathology, blindness to success, and dependence on auxiliary networks. We isolate three structural properties of optimal stopping: single-point reward, action-independent dynamics, and sum-zero gradient. We exploit these properties to establish Markov optimality without state augmentation and derive a pathwise gradient estimator. Under bounded policy derivatives, the pathwise estimator has bounded second moments near deterministic policies, avoiding the divergence that can occur with score-function estimators. Building on this framework, we propose DIOS (Distribution-Informed Optimal Stopping) to address the remaining two difficulties. A per-time-step soft penalty combined with tail-level annealing mitigates blindness to success. Estimating the reward quantile directly from each batch removes the need for an auxiliary critic. For a fixed tail level, we establish a finite-time stationarity guarantee under regularity assumptions along the training iterates. Across the 24 main Bermudan max-call configurations, DIOS achieves the highest mean CVaR and the shortest recorded training time among the tested CVaR baselines.

1 Introduction

Risk-sensitive deep optimal stopping. Risk-sensitive optimal stopping is useful when poor stopping outcomes carry substantial costs, as in option exercise, sequential medical treatment, and safety-critical control. Maximizing expected reward alone may give insufficient weight to adverse stopping outcomes. We study the Conditional Value at Risk (CVaR) of the stopped reward, which measures the average reward over a chosen fraction of the worst outcomes [21, 22].

Coherent risk measure and CVaR. A coherent risk measure satisfies monotonicity, translation invariance, positive homogeneity, and subadditivity [1]. The negative of lower-tail CVaR is a coherent risk measure for rewards [22]. We choose CVaR to optimize the average reward over a prescribed fraction of the worst stopping outcomes. The Rockafellar–Uryasev representation decomposes CVaR optimal stopping into an outer optimization over a scalar variable and an inner risk-neutral stopping problem. Standard stopping theory and deep optimal stopping methods can therefore be applied to the inner stopping problem.

Deep optimal stopping. The Snell envelope characterizes risk-neutral optimal stopping [24], and the regression-based approach of Longstaff and Schwartz [16] provides a method for American option pricing. Deep learning is used to address the curse of dimensionality in optimal stopping.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Becker et al. [3] introduced *deep optimal stopping*, a backward-induction method that parameterizes each per-step stopping probability with a neural network and trains the networks in backward order. Neural-network regression methods approximate continuation values [11, 15]. Other methods jointly optimize stopping probabilities across time [4, 8], with shared policy parameters in Damera Venkata and Bhattacharyya [8]. Language models have also been applied to stopping decisions in sales conversations [19]. We study CVaR deep optimal stopping, replacing expected-reward maximization with lower-tail CVaR maximization.

Risk-sensitive reinforcement learning. For CVaR objectives and constraints, Bäuerle and Ott [2], Chow and Ghavamzadeh [6], and Chow et al. [7] obtain Bellman recursions by augmenting the state with risk or cumulative-reward variables. Tamar et al. [25] derive a score-function CVaR policy-gradient (CVaR-PG) estimator. Methods for improving sample efficiency in CVaR policy optimization modify sampling, policy parameterization, or gradient estimation. Greenberg et al. [10] pair a soft-risk schedule with a cross-entropy environment sampler. Luo et al. [18] combine a risk-averse policy with a risk-neutral policy trained via Implicit Q-Learning [14]. Kim and Min [13] weight individual costs by predicted tail probabilities from a separate network. Mead et al. [20] compute policy gradients from trajectory returns capped at a value updated during training. Luo and Delage [17] mix the CVaR-PG gradient with a quantile-gradient term computed by an auxiliary quantile critic. These CVaR methods can be applied to optimal stopping, a finite-horizon MDP with a binary stop/continue action and action-independent state process. Applying risk-sensitive reinforcement learning (RL) to optimal stopping raises four difficulties.

Four structural problems of risk-sensitive RL.

(P1) State augmentation. Static CVaR objectives can be time-inconsistent [23]. Existing CVaR-MDP methods obtain Bellman recursions by adding variables for accumulated costs to the state [2, 6, 7].

(P2) Score-function pathology. Near deterministic policies, the second moment of the trajectory-level score-function estimator can diverge or vanish, depending on the policy parameterization. Several methods retain a score-function CVaR-PG component [10, 17, 18].

(P3) Blindness to success. The CVaR-PG gradient estimator assigns zero weight to trajectories with rewards above the estimated quantile. The CVaR-PG gradient estimate can also vanish when the sampled lower-tail rewards coincide [10, 17].

(P4) Dependence on auxiliary networks. Several methods use learned tail probabilities, values, or quantiles in policy updates [13, 17, 18]. The auxiliary networks are trained on data generated by the policy.

Optimal stopping addresses (P1) and permits a pathwise treatment of (P2). DIOS addresses (P3) and removes the auxiliary critic in (P4).

Framework: addressing (P1) and (P2). The single-point reward (S1) and the envelope reduction give Markov optimality without state augmentation. Action-independent dynamics (S2) permit differentiation of the conditional stopping probabilities without a trajectory score term. The sum-zero gradient (S3) makes the pathwise gradient invariant to a common shift of the transformed rewards. Under the policy-derivative bounds, pathwise gradient moments remain bounded near deterministic policies.

DIOS and convergence: addressing (P3) and (P4). We propose *Distribution-Informed Optimal Stopping* (DIOS), a pathwise algorithm for CVaR optimal stopping. A per-time-step soft penalty allows rewards at other stopping times to contribute even when the sampled stopping reward exceeds the estimated quantile. Training starts with a tail level of 1, then gradually reduces the tail level to its target. Direct quantile estimation from each batch removes the auxiliary critic. Theorem 4.2 gives a finite-time best-iterate stationarity guarantee when the analyzed iterates remain in a domain satisfying the regularity assumptions.

Contributions.

- We exploit (S1)–(S3) to develop a CVaR optimal stopping framework addressing (P1) and (P2). We establish the envelope reduction (Proposition 3.1) and prove Markov optimality without state augmentation (Theorem 3.3).
- We derive a pathwise gradient estimator (Lemma 3.6) and bound its second moments under the policy-derivative and reward-moment assumptions (Lemma 3.9). We compare pathwise and score-function gradients near deterministic policies (Theorem 3.10).
- We propose *Distribution-Informed Optimal Stopping* (DIOS, Algorithm 1). DIOS combines a per-time-step soft penalty with tail-level annealing to mitigate (P3). Direct quantile estimation removes the auxiliary critic in (P4). At a fixed tail level, we establish a finite-time best-iterate stationarity guarantee under the regularity and parameter-domain conditions of Theorem 4.2.
- We demonstrate across 24 Bermudan max-call configurations that DIOS achieves the highest mean CVaR and the shortest recorded training time among the tested CVaR baselines [13, 17, 18, 20, 25].

2 Deep optimal stopping and CVaR

2.1 Deep optimal stopping

Setup. We work in the finite-horizon, discrete-time, model-free setting. Fix a probability space $(\Omega, \mathcal{F}, P)$ and a finite horizon $T \in \mathbb{N}$. Let $(S_t)_{t=0}^T$ be an $\mathbb{R}^d$-valued Markov process and $(\mathcal{F}_t)_{t=0}^T$ its natural filtration, $\mathcal{F}_t := \sigma(S_0, \dots, S_t)$. Define $R_t := g_t(S_t)$, where $g_t : \mathbb{R}^d \to \mathbb{R}_+$ is Borel-measurable. The joint distribution of $(S_0, \dots, S_T)$ is unknown and accessed only through samples. Appendix A discusses the assumptions for the Bermudan max-call experiments in Section 5.

Assumption 2.1 (Integrability)**.** $R_t \in L^1(P)$ for every $t \in \{0, \dots, T\}$.

The optimal stopping problem. Let $\mathcal{T}$ be the set of random variables $\tau : \Omega \to \{0, \dots, T\}$ satisfying $\{\tau = t\} \in \mathcal{F}_t$ for every t. The optimal stopping problem is to find $\tau^* \in \mathcal{T}$ attaining

$$V^* \;:=\; \sup_{\tau \in \mathcal{T}} \mathbb{E}[R_\tau]\,. \tag{1}$$

Existence of τ^* under Assumption 2.1 is classical [24]. Define the Snell envelope $(V_t)_{t=0}^T$ by backward induction:

$$V_T := R_T, \quad V_t := \max(R_t,\ \mathbb{E}[V_{t+1} \mid \mathcal{F}_t]) \quad \text{for } t < T. \tag{2}$$

Assumption 2.1 ensures that V_t is integrable for every $t = 0, \dots, T$. For $t < T$, an optimal rule stops when the immediate reward R_t satisfies $R_t \geq \mathbb{E}[V_{t+1} \mid \mathcal{F}_t]$ and continues otherwise, with a forced stop at T. The supremum in (1) equals $\mathbb{E}[V_0]$, attained at $\tau^* := \inf\{t \geq 0 : V_t = R_t\}$.

Stochastic stopping policies [8]. Let $\phi = (\phi_0, \dots, \phi_T)$ consist of Borel functions $\phi_t : \mathbb{R}^d \to [0, 1]$ with $\phi_T \equiv 1$. At time t, conditional on not having stopped earlier, the policy stops with probability $\phi_t(S_t)$. Let $U_0, \dots, U_T$ be independent uniform random variables on $[0, 1]$, independent of the state process. On the enlarged probability space, define

$$\tau \;:=\; \inf\{t \,:\, U_t \leq \phi_t(S_t)\}\,. \tag{3}$$

Conditioning on $\mathcal{F}_T$ fixes $S_0, \dots, S_T$, leaving only the randomness in $U_0, \dots, U_T$. Define $\psi_t := P(\tau = t \mid \mathcal{F}_T)$ and $\Psi_t := \prod_{s<t}(1 - \phi_s(S_s))$, with $\Psi_0 := 1$. Then

$$\psi_t \;=\; \phi_t(S_t) \cdot \prod_{s<t}(1 - \phi_s(S_s)) \;=\; \Psi_t\, \phi_t(S_t), \qquad t = 0, \dots, T. \tag{4}$$

The factor $\Psi_t = P(\tau \geq t \mid \mathcal{F}_T)$ gives the conditional probability of not stopping before time t. Since $\phi_T \equiv 1$, the probabilities ψ_t satisfy

$$\sum_{t=0}^{T} \psi_t \;=\; 1 \qquad \text{pointwise on } \Omega. \tag{5}$$

By the tower property,

$$\mathbb{E}[R_\tau] \;=\; \mathbb{E}[\mathbb{E}[R_\tau \mid \mathcal{F}_T]] \;=\; \mathbb{E}\left[\sum_{t=0}^{T} P(\tau = t \mid \mathcal{F}_T)\, R_t\right] \;=\; \mathbb{E}\left[\sum_{t=0}^{T} \psi_t\, R_t\right]. \tag{6}$$

Three structural properties of optimal stopping. We exploit the following three properties of the stopping formulation.

(S1) *Single-point reward.* The reward $R_\tau = g_\tau(S_\tau)$ is realized at the single time index τ, in contrast to cumulative-return MDPs where rewards $\sum_t r_t$ are summed over the trajectory.

(S2) *Action-independent dynamics.* The state process $(S_t)_{t=0}^T$ evolves under a transition kernel that does not depend on the stopping action.

(S3) *Sum-zero gradient.* Parameterize the policy by $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^p$ is open, and assume each $\theta \mapsto \psi_t^\theta$ is differentiable P-a.s. Since (5) holds for every θ, differentiating in θ yields

$$\sum_{t=0}^{T} \nabla_\theta \psi_t^\theta \;=\; 0 \qquad P\text{-a.s.}$$

2.2 CVaR optimal stopping

Fix $\beta \in (0,1)$. For $X \in L^1(P)$, define the lower β-quantile by $\mathrm{VaR}_\beta(X) := \inf\{x : P(X \le x) \ge \beta\}$. Define the *Conditional Value at Risk* by

$$\mathrm{CVaR}_\beta(X) \;:=\; \tfrac{1}{\beta}\int_0^\beta \mathrm{VaR}_u(X)\, du. \tag{7}$$

Define $\mathrm{CVaR}_1(X) := \mathbb{E}[X]$. Smaller β focuses the objective on a smaller fraction of the lowest outcomes. The Rockafellar-Uryasev variational form expresses CVaR as a one-dimensional supremum [22],

$$\mathrm{CVaR}_\beta(X) \;=\; \sup_{\nu\in\mathbb{R}}\Big\{\nu \;-\; \tfrac{1}{\beta}\,\mathbb{E}[(\nu - X)^+]\Big\}, \tag{8}$$

attained at $\nu^* = \mathrm{VaR}_\beta(X)$ for $0 < \beta < 1$. The quantity $(\nu - X)^+ = \max\{\nu - X, 0\}$ measures the shortfall below ν. Replacing the expectation in (1) with CVaR yields the ***CVaR optimal stopping value***

$$V_\beta^* \;:=\; \sup_{\tau\in\mathcal{T}} \mathrm{CVaR}_\beta(R_\tau). \tag{9}$$

3 CVaR within the deep optimal stopping framework

Building on the Rockafellar–Uryasev envelope reduction, we establish Markov optimality without state augmentation and derive a pathwise gradient formula.

3.1 Envelope reduction

Proofs are given in Appendix B.

Proposition 3.1 (CVaR envelope reduction)**.** *Under Assumption 2.1 and $\beta \in (0,1)$,*

$$V_\beta^* \;=\; \sup_{\nu\in\mathbb{R}}\Big\{\nu \;+\; \tfrac{1}{\beta}\,\sup_{\tau\in\mathcal{T}}\mathbb{E}[-(\nu - R_\tau)^+]\Big\}, \tag{10}$$

and there exist $\nu^* \in \mathbb{R}$ and $\tau^* \in \mathcal{T}$ with $\nu^* = \mathrm{VaR}_\beta(R_{\tau^*})$ jointly realizing the supremum.

Proposition 3.1 separates an outer maximization over the scalar ν from an inner risk-neutral stopping problem at fixed ν. Define the per-step *soft penalty* $\tilde{R}_t^\nu := -\frac{1}{\beta}(\nu - R_t)^+$. For the parameterized stopping policies ϕ^θ, define the objective for the inner maximization over θ by

$$J^{(\nu)}(\theta) \;:=\; \mathbb{E}\left[\sum_{t=0}^{T} \psi_t^\theta \cdot \tilde{R}_t^\nu\right]. \tag{11}$$

For fixed ν, the Snell-envelope recursion and stopping-policy representation in Section 2.1 apply with R_t replaced by $\tilde{R}_t^\nu$. Appendix I extends the outer–inner decomposition to an optimized certainty equivalent objective.

3.2 Markov optimality without augmentation

Applying CVaR recursively at each time generally changes the static CVaR objective [23]. CVaR-MDP methods instead obtain Bellman recursions by adding variables for accumulated costs to the state [2, 6, 7].

Definition 3.2 (Markov stopping time)**.** Define a stopping time $\tau \in \mathcal{T}$ to be *Markov* if there exist Borel sets $D_0, \dots, D_T \subseteq \mathbb{R}^d$ with $D_T = \mathbb{R}^d$ such that $\tau = \inf\{t \geq 0 : S_t \in D_t\}$ P-a.s.

Theorem 3.3 (Markov optimality without augmentation)**.** *Under Assumption 2.1 and $\beta \in (0,1)$, there exist $\nu^* \in \mathbb{R}$ and a Markov stopping time $\tau^* \in \mathcal{T}$ such that*

$$V_\beta^* \;=\; \nu^* \;+\; \tfrac{1}{\beta}\,\mathbb{E}\big[-(\nu^* - R_{\tau^*})^+\big]\,.$$

Remark 3.4. For fixed ν^*, (S1) makes the transformed reward $-(\nu^* - g_t(S_t))^+$ a function of the current state and time. By the Markov property, each conditional expectation in the Snell-envelope recursion is a function of t and S_t, with no additional dependence on $S_0, \dots, S_{t-1}$.

3.3 Pathwise and score-function gradients

For fixed ν, we show that pathwise gradient second moments remain uniformly bounded under bounded policy derivatives, whereas score-function second moments can diverge as stopping probabilities approach zero.

By (S2), the joint distribution of $(S_0, \dots, S_T)$ is independent of θ. At fixed ν, only ψ_t^θ in the integrand of (11) depends on θ.

Assumption 3.5 (Differentiable policy with bounded gradients)**.** Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^p$. On an event of probability one, each map $\theta \mapsto \phi_t^\theta(S_t)$, $0 \leq t \leq T$, is differentiable on the open domain Θ. There exist $G, C_\psi \geq 0$ such that, P-a.s.,

$$\sup_{\theta\in\Theta} \max_{0\leq t\leq T} \|\nabla_\theta \phi_t^\theta(S_t)\| \leq G, \qquad \sup_{\theta\in\Theta} \max_{0\leq t\leq T} \|\nabla_\theta \psi_t^\theta\| \leq C_\psi.$$

For every $\theta \in \Theta$ and $t < T$, $\phi_t^\theta(S_t) \in (0,1)$ P-a.s. The bound C_ψ may depend on T.

Assumptions 2.1 and 3.5 justify differentiation under the expectation in (11). Equation (21) in Appendix A shows that $C_\psi = G(T+1)$ is a valid choice.

Define $\mathrm{sg}[x]$ to have value x and zero derivative in automatic differentiation. Following Damera Venkata and Bhattacharyya [8], define the surrogate loss for the transformed rewards $\tilde{R}_t^\nu$ at fixed $\nu \in \mathbb{R}$ by

$$L(\theta) \;:=\; -\mathbb{E}\left[\sum_{t=0}^{T} \mathrm{sg}\Big[\tilde{R}_t^\nu\Big] \cdot \mathrm{sg}\big[\psi_t^\theta\big] \cdot \log \psi_t^\theta\right]. \tag{12}$$

Let $\nabla_\theta L$ denote the expected autograd gradient with $S_0, \dots, S_T$ and ν held fixed. The stop-gradient factors restrict differentiation to $\log \psi_t^\theta$, giving $\psi_t^\theta \nabla_\theta \log \psi_t^\theta = \nabla_\theta \psi_t^\theta$.

Lemma 3.6 (Pathwise gradient formula)**.** *Under Assumptions 2.1 and 3.5, for every fixed $\nu \in \mathbb{R}$,*

$$\nabla_\theta J^{(\nu)}(\theta) = \mathbb{E}\left[\sum_{t=0}^{T} \tilde{R}_t^\nu \nabla_\theta \psi_t^\theta\right] = -\nabla_\theta L(\theta). \tag{13}$$

Remark 3.7. Differentiating $\psi_t^\theta = \phi_t^\theta(S_t)\prod_{s<t}(1-\phi_s^\theta(S_s))$ gives

$$\nabla_\theta \psi_t^\theta = \Psi_t^\theta \nabla_\theta \phi_t^\theta(S_t) - \phi_t^\theta(S_t) \sum_{s<t} \left(\prod_{\substack{u<t\\u\neq s}} (1-\phi_u^\theta(S_u))\right) \nabla_\theta \phi_s^\theta(S_s).$$

No term divides by $\phi_s^\theta(S_s)$ or $1-\phi_s^\theta(S_s)$. Thus $\nabla_\theta \psi_t^\theta$ remains defined at probabilities 0 or 1 whenever $\nabla_\theta \phi_s^\theta(S_s)$ exists for every $s \leq t$.

Assumption 3.8 (L^2 payoff)**.** Assume $R_t \in L^2(P)$ for $0 \leq t \leq T$, and define $M := \max_{0\leq t\leq T} \|R_t\|_{L^2(P)}$.

Let $(S_t^{(i)})_{t=0}^T$, $i = 1, \dots, B$, be independent copies of $(S_t)_{t=0}^T$. Let the superscript (i) denote evaluation on trajectory i. Define the empirical gradient by

$$\hat{\nabla} L(\theta) := -\frac{1}{B} \sum_{i=1}^{B} \sum_{t=0}^{T} \tilde{R}_t^{\nu,(i)} \nabla_\theta \psi_t^{\theta,(i)}.$$

By (S3), subtracting the mean $\frac{1}{T+1} \sum_{s=0}^{T} \tilde{R}_s^{\nu,(i)}$ from each $\tilde{R}_t^{\nu,(i)}$ leaves $\hat{\nabla} L(\theta)$ unchanged. Let τ_θ be given by (3) for the policy ϕ^θ, and define $\nu^*(\theta) := \mathrm{VaR}_\beta(R_{\tau_\theta})$.

Lemma 3.9 (Second moment and variance)**.** *Under Assumptions 2.1, 3.5, and 3.8, fix $\theta \in \Theta$ and $\nu \in \mathbb{R}$. Define $\tilde{M} := \max_t \|\tilde{R}_t^\nu\|_{L^2(P)}$.*

(a) *For every $B \geq 1$,*

$$\mathbb{E}[\|\hat{\nabla} L(\theta)\|^2] \leq C_\psi^2 (T+1)^2 \tilde{M}^2.$$

(b) *For B independent state trajectories,*

$$\mathbb{E}[\|\hat{\nabla} L(\theta) - \nabla L(\theta)\|^2] \leq \frac{C_\psi^2 (T+1)^2 \tilde{M}^2}{B}.$$

(c) *If $M = 0$, define $c_\beta := 0$. If $M > 0$, define*

$$c_\beta := \sup_{\vartheta \in \Theta} \max_{0 \leq t \leq T} \frac{\mathbb{E}[((\nu^*(\vartheta) - R_t)^+)^2]}{M^2}.$$

Then $c_\beta \leq (T+1)/(1-\beta)$. At $\nu = \nu^*(\theta)$,

$$\mathbb{E}[\|\hat{\nabla} L(\theta) - \nabla L(\theta)\|^2] \leq \frac{4 C_\psi^2 c_\beta T^2 M^2}{\beta^2 B}. \tag{14}$$

In (14), the factor $C_\psi^2 T^2$ becomes $G^2 T^2 (T+1)^2$ for the choice $C_\psi = G(T+1)$. If C_ψ is independent of T, $C_\psi^2 T^2$ is proportional to T^2.

Let $\tau = \tau_\theta$. For $t < T$, let $a_t = \mathbf{1}\{U_t \leq \phi_t^\theta(S_t)\}$ and define $\pi_\theta(a_t \mid S_t) := (\phi_t^\theta(S_t))^{a_t}(1 - \phi_t^\theta(S_t))^{1-a_t}$. Define the likelihood of the sampled actions up to τ, conditional on $(S_0, \dots, S_T)$, by $\pi_\theta(\tau) := \prod_{t=0}^{\min\{\tau, T-1\}} \pi_\theta(a_t \mid S_t)$. For comparison with existing score-function CVaR policy-gradient methods [25], define the loss for automatic differentiation at fixed ν by

$$L_{\mathrm{score}}(\theta) \;:=\; -\mathbb{E}\Big[\mathrm{sg}\Big[\tilde{R}_\tau^\nu\Big] \cdot \log \pi_\theta(\tau)\Big].$$

With sampled states, actions, and τ held fixed, define

$$\hat{\nabla} L_{\mathrm{score}}(\theta) := -\tilde{R}_\tau^\nu \nabla_\theta \log \pi_\theta(\tau), \qquad \nabla_\theta L_{\mathrm{score}}(\theta) := \mathbb{E}[\hat{\nabla} L_{\mathrm{score}}(\theta)].$$

Under Assumptions 2.1 and 3.5, $-\nabla_\theta L_{\mathrm{score}}(\theta) = \nabla_\theta J^{(\nu)}(\theta)$. For the single-trajectory comparison $B = 1$, $\hat{\nabla} L(\theta)$ sums over all candidate stopping times, whereas $\hat{\nabla} L_{\mathrm{score}}(\theta)$ uses the reward at the sampled stopping time τ.

Theorem 3.10 (Bounded pathwise second moments and score-function divergence)**.** *Under Assumptions 2.1, 3.5, and 3.8, fix $\nu \in \mathbb{R}$ and a sequence $(\theta_n) \subset \Theta$. Let $B = 1$. Suppose there exist $t^* \in \{0, \dots, T-1\}$, $\epsilon_0 > 0$, and an event $A \in \sigma(S_{t^*})$ with $P(A) > 0$ such that $\phi_{t^*}^{\theta_n}(S_{t^*}) \to 0$ on A. Assume*

(a) $\inf_n \inf_{\omega \in A} \|\nabla_\theta \phi_{t^*}^{\theta_n}(S_{t^*}(\omega))\| > 0,$

(b) $\sup_n \sup_{\omega \in A} \phi_s^{\theta_n}(S_s(\omega)) \leq 1 - \epsilon_0$ *for every $s < t^*$,*

(c) $P(\{R_{t^*} < \nu\} \cap A) > 0.$

Then

$$\mathbb{E}[\|\hat{\nabla} L_{\mathrm{score}}(\theta_n)\|^2] \to \infty, \qquad \sup_n \mathbb{E}[\|\hat{\nabla} L(\theta_n)\|^2] < \infty.$$

Remark 3.11 (Sigmoid policies)**.** We implement the proposed algorithm, DIOS, using sigmoid policies. Fix $\nu \in \mathbb{R}$, let $B = 1$, and let $(\theta_n) \subset \Theta$. Assume that the logit gradients are uniformly bounded in n, t, and the state. Then condition (a) of Theorem 3.10 fails if $\phi_{t^*}^{\theta_n}(S_{t^*}) \to 0$. Under Assumptions 2.1, 3.5, and 3.8, if $\phi_t^{\theta_n}(S_t)$ converges almost surely to 0 or 1 for every $t < T$, then $\hat{\nabla} L_{\mathrm{score}}(\theta_n) \xrightarrow{L^2} 0$ and $\hat{\nabla} L(\theta_n) \xrightarrow{L^2} 0$, as proved in Appendix B.

4 Distribution-Informed Optimal Stopping

DIOS (*Distribution-Informed Optimal Stopping*) combines a per-time-step soft penalty with tail-level annealing. DIOS estimates the reward quantile directly from each batch without an auxiliary critic. Section 4.3 gives a finite-time stationarity guarantee for the fixed-tail-level update.

4.1 Blindness to success

CVaR-PG [25] estimates the gradient of $\mathrm{CVaR}_\beta(R_{\tau_\theta})$ using a trajectory score. On each of B independent state trajectories i, let $\tau_\theta^{(i)}$ be the stopping time generated by ϕ^θ, and let $\pi_\theta^{(i)}$ denote the action likelihood defined in Section 3.3, evaluated on trajectory i. Let $\widehat{\nu}_\beta$ be the empirical β-quantile of $R^{(1)}_{\tau_\theta^{(1)}}, \ldots, R^{(B)}_{\tau_\theta^{(B)}}$. The CVaR-PG gradient estimator is

$$\widehat{g}_{\mathrm{PG}}(\theta) = \frac{1}{\beta B} \sum_{i=1}^{B} \mathbf{1}\left\{ R^{(i)}_{\tau_\theta^{(i)}} \le \widehat{\nu}_\beta \right\} \left(R^{(i)}_{\tau_\theta^{(i)}} - \widehat{\nu}_\beta \right) \nabla_\theta \log \pi_\theta^{(i)}(\tau_\theta^{(i)}). \tag{15}$$

The CVaR-PG estimator $\widehat{g}_{\mathrm{PG}}(\theta)$ in (15) has *blindness to success*: trajectories with sampled stopping rewards above $\widehat{\nu}_\beta$ receive zero weight.

Methods for improving sample efficiency in CVaR policy optimization modify the gradient estimator, the policy, or the sampling procedure. Greenberg et al. [10] pair a soft-risk schedule with a cross-entropy environment sampler that requires controllable environment dynamics. Luo et al. [18] combine a risk-averse policy with a risk-neutral policy trained via Implicit Q-Learning [14]. Kim and Min [13] weight individual costs using an auxiliary tail-probability predictor. Mead et al. [20] compute policy gradients from capped trajectory returns and update the cap during training, subject to an externally supplied minimum cap. Luo and Delage [17] augment CVaR-PG with a quantile-gradient term computed by an auxiliary quantile critic.

4.2 The DIOS algorithm

For a fixed policy ϕ^θ and $\beta \in (0,1)$, the CVaR envelope (8) is maximized at $\nu = \mathrm{VaR}_\beta(R_{\tau_\theta})$. DIOS, presented in Algorithm 1, samples one stopping time on each trajectory according to the current policy. The reward quantile is then estimated from the batch of rewards obtained at the sampled stopping times. Let ν_k and β_k denote the quantile estimate and tail level at iteration k. With ν_k held fixed, the policy update sums the terms $-\beta_k^{-1}(\nu_k - R_t)^+ \nabla_\theta \psi_t^\theta$ over all stopping times on each trajectory. Thus a reward $R_t < \nu_k$ can contribute even when the reward at the sampled stopping time exceeds ν_k. Computing ν_k directly from the batch removes the auxiliary critic.

Tail-level annealing gradually shifts training toward the target lower tail. The fractions ρ_w and ρ_a specify the warmup and annealing durations, respectively. During warmup, $\beta_k = 1$ and ν_k is the largest of the B sampled stopping rewards. Since $\sum_{t=0}^{T} \nabla_\theta \psi_t^\theta = 0$ and $-(\nu_k - R_t)^+ = \min\{R_t, \nu_k\} - \nu_k$, the warmup gradient equals the pathwise gradient with rewards $\min\{R_t, \nu_k\}$. Warmup therefore trains on rewards truncated at the batch maximum of sampled stopping rewards. After warmup, β_k decreases linearly to β_{tgt} and remains at β_{tgt} for the rest of training.

4.3 Finite-time convergence

We establish a finite-time stationarity guarantee for the DIOS update at a fixed tail level and bound the reward-quantile estimation error.

Fix $\beta \in (0,1)$. Define

$$V_\beta(\theta) := \mathrm{CVaR}_\beta(R_{\tau_\theta}), \qquad F_\theta(x) := \mathbb{P}(R_{\tau_\theta} \le x).$$

Algorithm 1 DIOS for CVaR optimal stopping

Require: policy network $\phi^\theta = (\phi_0^\theta, \dots, \phi_{T-1}^\theta)$ with implicit $\phi_T^\theta \equiv 1$ (forced terminal stop); target tail level $\beta_{\mathrm{tgt}} \in (0,1)$; warmup and annealing fractions $\rho_w, \rho_a \geq 0$ with $\rho_w + \rho_a \leq 1$; positive integers K, B; learning rate $\eta > 0$

for $k = 0, 1, \dots, K-1$ **do**

If $\rho_a > 0$, set $\beta_k \leftarrow 1 + (\beta_{\mathrm{tgt}} - 1) \cdot \min\left\{1, \max\left\{0, \dfrac{k - \rho_w K}{\rho_a K}\right\}\right\}$. If $\rho_a = 0$, set $\beta_k = 1$ for $k < \rho_w K$ and $\beta_k = \beta_{\mathrm{tgt}}$ thereafter.

Sample B independent trajectories $\{(S_t^{(i)})_{t=0}^T\}_{i=1}^B$ from the environment

Compute $\psi_t^{\theta,(i)} \leftarrow \phi_t^\theta(S_t^{(i)}) \prod_{s<t}\left(1 - \phi_s^\theta(S_s^{(i)})\right)$ for all i, t

Independently for each i, sample $\tau^{(i)} \sim \mathrm{Categorical}\left((\psi_t^{\theta,(i)})_{t=0}^T\right)$, compute $G^{(i)} := R_{\tau^{(i)}}^{(i)}$

$m_k \leftarrow \max\{1, \lfloor \beta_k B \rfloor\}$, $\quad \nu_k \leftarrow$ the m_k-th smallest value in $\{G^{(i)}\}_{i=1}^B$

$\tilde{R}_t^{(i)} \leftarrow -\frac{1}{\beta_k}\left(\nu_k - R_t^{(i)}\right)^+$ for all i, t

$\widehat{L}(\theta) \leftarrow -\frac{1}{B}\sum_{i=1}^B \sum_{t=0}^T \mathrm{sg}\left[\tilde{R}_t^{(i)}\right] \cdot \mathrm{sg}\left[\psi_t^{\theta,(i)}\right] \cdot \log \psi_t^{\theta,(i)}$

$\theta \leftarrow \theta - \eta \nabla_\theta \widehat{L}(\theta)$

end for

Assumption 4.1 (Convergence regime)**.** There exist a nonempty open convex domain $U \subset \Theta$ and numbers $\nu_- > 0$, $f_- > 0$, $\delta_0 > 0$, and $L_V > 0$. Define

$$\Theta_{\nu_-} := \{\theta \in U : \nu^*(\theta) \geq \nu_-\}.$$

(a) For every $\theta \in \Theta_{\nu_-}$ and $u \in [0, \delta_0]$,

$$F_\theta\big(\nu^*(\theta) + u\big) \geq \beta + f_- u, \qquad F_\theta\big(\nu^*(\theta) - u\big) \leq \beta - f_- u.$$

(b) The function V_β is differentiable on U, and ∇V_β is L_V-Lipschitz on U.

For each integer batch size $B \geq 1$, fix an integer m_B satisfying

$$1 \leq m_B \leq B, \qquad |m_B - \beta B| \leq 1. \tag{16}$$

Let $\mathcal{H}_k$ be the sigma-field generated by θ_0 and all state trajectories and stopping draws before iteration k. Conditional on $\mathcal{H}_k$, sample B independent copies of $(S_t)_{t=0}^T$. For each trajectory i, draw $\tau_{\theta_k}^{(i)}$ with probabilities $(\psi_t^{\theta_k,(i)})_{t=0}^T$, independently across i conditional on the state trajectories. Let ν_k be the m_B-th smallest reward among $R_{\tau_{\theta_k}^{(i)}}^{(i)}$, $i = 1, \dots, B$. For a constant step size $\eta > 0$, define

$$\widehat{g}_k(\nu) := -\frac{1}{\beta B}\sum_{i=1}^B \sum_{t=0}^T (\nu - R_t^{(i)})^+ \, \nabla_\theta \psi_t^{\theta,(i)}\Big|_{\theta=\theta_k}, \qquad \theta_{k+1} = \theta_k + \eta \widehat{g}_k(\nu_k). \tag{17}$$

Assumption 4.1(b) gives [9]

$$V_\beta(\theta') \geq V_\beta(\theta) + \langle \nabla V_\beta(\theta), \theta' - \theta \rangle - \tfrac{L_V}{2}\|\theta' - \theta\|^2, \qquad \theta, \theta' \in U.$$

Theorem 4.2 (Finite-time convergence)**.** *Under Assumptions 3.5, 3.8, and 4.1, there exist an integer $B_0 \geq 1$ and constants $C_\nu > 0$ and $C_g \geq 0$, depending only on β, T, M, C_ψ, f_-, and δ_0. Let K, B be integers with $K \geq 1$ and $B \geq B_0$, and let $0 < \eta \leq 1/L_V$. Suppose $(\theta_k)_{k=0}^K$ satisfies* (17) *and $\theta_k \in \Theta_{\nu_-}$ almost surely for $0 \leq k \leq K$. Define*

$$D_0 := \sup_{\vartheta \in \Theta} V_\beta(\vartheta) - \mathbb{E}[V_\beta(\theta_0)].$$

Then the following bounds hold.

(a) *The reward-quantile estimates ν_k satisfy*

$$\max_{0 \leq k < K} \mathbb{E}\big[(\nu_k - \nu^*(\theta_k))^2\big] \leq \frac{C_\nu}{B}. \tag{18}$$

(b) *The average expected squared gradient satisfies*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla V_\beta(\theta_k)\|^2\big] \le \frac{2D_0}{\eta K} + \frac{C_g}{B}. \tag{19}$$

Consequently,

$$\min_{0\le k<K}\mathbb{E}\big[\|\nabla V_\beta(\theta_k)\|^2 + (\nu_k - \nu^*(\theta_k))^2\big] \le \frac{2D_0}{\eta K} + \frac{C_g + C_\nu}{B}. \tag{20}$$

For every $\epsilon > 0$, integers $K \ge 1$ and $B \ge B_0$ can be chosen so that

$$\min_{0\le k<K}\mathbb{E}\big[\|\nabla V_\beta(\theta_k)\|^2 + (\nu_k - \nu^*(\theta_k))^2\big] \le \epsilon,$$

provided the corresponding iterates satisfy $\theta_k \in \Theta_{\nu_-}$ almost surely. For fixed β and η and sufficiently small $\epsilon > 0$, the choices satisfy $K = O(\epsilon^{-1})$ and $B = O(\epsilon^{-1})$, giving $KB = O(\epsilon^{-2})$ sampled trajectories.

Remark 4.3. Theorem 4.2 applies when quantile estimation and the policy update use the same batch. Since ν_k depends on the sampled trajectories, $\widehat{g}_k(\nu_k)$ need not be conditionally unbiased for $\nabla V_\beta(\theta_k)$ given $\mathcal{H}_k$. The proof in Appendix B establishes

$$\mathbb{E}\big[\|\widehat{g}_k(\nu_k) - \nabla V_\beta(\theta_k)\|^2 \mid \mathcal{H}_k\big] \le C_g/B$$

by bounding the sampling error at $\nu^*(\theta_k)$ and the error from replacing $\nu^*(\theta_k)$ with ν_k.

5 Empirical validation

We validate DIOS on Bermudan max-call options against four CVaR baselines: CVaR-PG [25], MIX [18], RET-CAP [20], and CVaR-VaR [17]. Appendix D.1 reports additional results for PCVaR-PG [13]. The setup uses volatility $\sigma = 0.20$, $d \in \{20, 50, 100, 200\}$ underlying assets, initial price $s_0 \in \{90, 100, 110\}$, 10 candidate stopping times $t = 0, \dots, T$ with $T = 9$, and CVaR level $\beta = 0.2$, yielding 12 configurations. An asymmetric-volatility variant reported in Appendix E.1 contributes 12 more, for 24 in total. We estimate $\text{CVaR}_{0.2}$ by averaging the lowest 20% of evaluation rewards. Let $\bar{\tau} \in [0, T]$ denote the mean stopping time on the evaluation set. For each configuration, the top row of Table 1 reports the mean and standard deviation of $\text{CVaR}_{0.2}$ over ten training seeds. The bottom row reports training time in seconds, including periodic evaluation, and the fraction of seeds with $\bar{\tau} < 0.5$, which we classify as collapsed. Bold entries mark the highest mean CVaR and shortest training time. Implementation details are in Appendix C.

DIOS achieves the highest mean CVaR and the shortest recorded training time in all 24 symmetric and asymmetric configurations. CVaR-PG and MIX have large standard deviations and collapsed seeds, with collapse rates of $3/10$ and $5/10$, respectively, at symmetric $(d, s_0) = (20, 110)$. CVaR-VaR takes roughly 4–5 times as long as DIOS, while MIX and RET-CAP take roughly twice as long. Appendix D.5 compares CVaR-VaR with a variant that replaces the quantile critic by a batch mean and changes the tradeoff-weight schedule. For RET-CAP, we compute the minimum cap q^* by averaging empirical β-quantiles of DIOS evaluation rewards over ten seeds. Appendix D.2 specifies the calculation and Appendix E.1 reports the asymmetric results.

The mean advantage over CVaR-VaR is small in some configurations. At symmetric $(d, s_0) = (200, 110)$, mean CVaR is 89.378 for DIOS and 89.375 for CVaR-VaR. Two-sided paired t-tests over ten training seeds give $p < 0.05$ against CVaR-VaR in 5 of the 24 configurations. The p-values are not adjusted for testing multiple configurations. Appendix C gives the paired mean differences and p-values.

Additional experiments. Appendix F compares raw batch quantile estimates with exponential moving average (EMA) estimates within DIOS at $\beta = 0.1$ and 0.05. DIOS with EMA has the highest mean CVaR among the tested methods in all 12 configurations at each tail level. Paired t-tests give $p < 0.05$ against CVaR-VaR in 6/12 configurations at $\beta = 0.1$ and 12/12 at $\beta = 0.05$, with no adjustment for multiple comparisons. Appendix G reports house-selling experiments, where DIOS has the highest mean CVaR in 23/24 configurations. Appendix H compares CVaR, parameter counts, and training times for DIOS encoder widths 32–256.

Table 1: Symmetric Bermudan max-call ($\sigma = 0.20$). RET-CAP CVaR values use reruns with recomputed minimum caps and retuned hyperparameters. RET-CAP times use the earlier minimum caps and hyperparameters.

d	s_0		DIOS (ours)	CVaR-VaR	CVaR-PG	MIX	RET-CAP
20	90	CVaR ↑	**18.171** ± 0.285	18.068 ± 0.385	14.452 ± 7.232	14.331 ± 7.176	18.003 ± 0.290
		wall (coll) ↓	**12.3** (0/10)	54.7 (0/10)	14.8 (2/10)	22.8 (2/10)	23.5 (0/10)
	100	CVaR ↑	**29.963** ± 0.329	29.850 ± 0.441	29.872 ± 0.535	28.004 ± 5.385	29.677 ± 0.295
		wall (coll) ↓	**12.7** (0/10)	51.0 (0/10)	15.2 (0/10)	24.8 (0/10)	25.8 (0/10)
	110	CVaR ↑	**41.966** ± 0.339	41.893 ± 0.587	32.222 ± 14.557	25.787 ± 15.793	41.616 ± 0.366
		wall (coll) ↓	**13.0** (0/10)	51.6 (0/10)	14.3 (3/10)	23.3 (5/10)	26.0 (0/10)
50	90	CVaR ↑	**33.220** ± 0.126	33.213 ± 0.175	33.121 ± 0.144	33.146 ± 0.173	32.810 ± 0.225
		wall (coll) ↓	**11.2** (0/10)	48.8 (0/10)	15.3 (0/10)	23.9 (0/10)	26.0 (0/10)
	100	CVaR ↑	**46.802** ± 0.153	46.715 ± 0.187	46.614 ± 0.272	46.588 ± 0.289	45.913 ± 0.278
		wall (coll) ↓	**11.4** (0/10)	58.2 (0/10)	15.0 (0/10)	22.0 (0/10)	25.1 (0/10)
	110	CVaR ↑	**60.392** ± 0.171	60.241 ± 0.311	60.157 ± 0.275	60.257 ± 0.234	59.115 ± 0.379
		wall (coll) ↓	**11.6** (0/10)	59.6 (0/10)	15.2 (0/10)	25.0 (0/10)	25.8 (0/10)
100	90	CVaR ↑	**45.001** ± 0.167	44.859 ± 0.195	40.406 ± 13.470	40.408 ± 13.471	44.323 ± 0.359
		wall (coll) ↓	**11.9** (0/10)	52.6 (0/10)	15.1 (1/10)	22.5 (1/10)	25.4 (0/10)
	100	CVaR ↑	**59.838** ± 0.160	59.805 ± 0.148	47.853 ± 23.927	47.771 ± 23.886	58.956 ± 0.390
		wall (coll) ↓	**11.9** (0/10)	49.9 (0/10)	14.9 (2/10)	22.9 (2/10)	25.4 (0/10)
	110	CVaR ↑	**74.668** ± 0.161	74.543 ± 0.202	61.470 ± 25.738	61.580 ± 25.791	73.084 ± 0.600
		wall (coll) ↓	**12.0** (0/10)	59.5 (0/10)	15.6 (2/10)	22.5 (2/10)	25.3 (0/10)
200	90	CVaR ↑	**57.142** ± 0.108	57.110 ± 0.117	51.275 ± 17.093	51.302 ± 17.102	56.105 ± 0.423
		wall (coll) ↓	**11.8** (0/10)	59.0 (0/10)	14.3 (1/10)	25.0 (1/10)	26.8 (0/10)
	100	CVaR ↑	**73.259** ± 0.074	73.233 ± 0.126	65.788 ± 21.930	58.529 ± 29.264	71.396 ± 0.410
		wall (coll) ↓	**11.8** (0/10)	58.6 (0/10)	15.7 (1/10)	22.2 (2/10)	25.4 (0/10)
	110	CVaR ↑	**89.378** ± 0.100	89.375 ± 0.099	81.387 ± 23.796	81.287 ± 23.764	86.622 ± 0.795
		wall (coll) ↓	**11.7** (0/10)	57.0 (0/10)	15.8 (1/10)	23.3 (1/10)	25.3 (0/10)

6 Conclusion

We developed a framework and algorithm for CVaR deep optimal stopping. The properties (S1)–(S3) yield Markov optimality without state augmentation and a pathwise gradient formula. Theorem 3.10 contrasts bounded pathwise second moments with score-function divergence under the stated conditions. DIOS combines a per-time-step soft penalty, tail-level annealing, and direct batch quantile estimation. We establish a finite-time stationarity guarantee at a fixed tail level. DIOS achieves the highest mean CVaR and the shortest recorded training time across the 24 max-call configurations.

Limitations and future work. Theorem 4.2 bounds expected squared gradients and squared quantile errors at fixed β under Assumptions 3.5, 3.8, and 4.1, with $\theta_k \in \Theta_{\nu_-}$ throughout the analyzed updates. The theorem does not establish parameter convergence or global optimality. The score-function and pathwise estimators can each vanish under the sigmoid conditions in Appendix B. Appendix A explains why ReLU, unbounded inputs, training-mode BatchNorm, and reward atoms prevent automatic verification of the assumptions. Extending the convergence analysis to Adam, clipping, training-mode BatchNorm, annealing, and EMA remains open. At $\beta = 1$, unbounded rewards can preclude a finite maximizing ν in (8).

The annealing ablation also changes the quantile rank, and the critic ablation also changes the tradeoff-weight schedule. Recorded training times come from different sessions, sometimes on shared GPUs, and exclude training and evaluation of policies used to set RET-CAP's minimum cap. RET-CAP CVaR values and times use different settings, as specified in Table 1. The experiments cover max-call options, house selling, and encoder widths 32–256. Further work includes larger architectures, action-dependent dynamics, and convergence for other optimized certainty equivalent objectives. Medical and control applications require validation of tail estimates to limit human and financial costs.

Acknowledgments

This work was supported by the Korea government (MSIT) through the NRF (RS-2023-00219980 & RS-2026-25491270); and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2021-II211343, Artificial Intelligence Graduate School Program, Seoul National University).

References

[1] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999) Coherent measures of risk. *Mathematical Finance* **9**(3):203–228.

[2] Bäuerle, N. & Ott, J. (2011) Markov decision processes with Average-Value-at-Risk criteria. *Mathematical Methods of Operations Research* **74**(3):361–379.

[3] Becker, S., Cheridito, P. & Jentzen, A. (2019) Deep optimal stopping. *Journal of Machine Learning Research* **20**(74):1–25.

[4] Becker, S., Cheridito, P., Jentzen, A. & Welti, T. (2021) Solving high-dimensional optimal stopping problems using deep learning. *European Journal of Applied Mathematics* **32**(3):470–514.

[5] Ben-Tal, A. & Teboulle, M. (2007) An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance* **17**(3):449–476.

[6] Chow, Y. & Ghavamzadeh, M. (2014) Algorithms for CVaR optimization in MDPs. *Advances in Neural Information Processing Systems* **27**:3509–3517.

[7] Chow, Y., Ghavamzadeh, M., Janson, L. & Pavone, M. (2018) Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research* **18**(167):1–51.

[8] Damera Venkata, N. & Bhattacharyya, C. (2023) Deep recurrent optimal stopping. *Advances in Neural Information Processing Systems* **36**:12222–12244.

[9] Ghadimi, S. & Lan, G. (2013) Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* **23**(4):2341–2368.

[10] Greenberg, I., Chow, Y., Ghavamzadeh, M. & Mannor, S. (2022) Efficient risk-averse reinforcement learning. *Advances in Neural Information Processing Systems* **35**:32639–32652.

[11] Herrera, C., Krach, F., Ruyssen, P. & Teichmann, J. (2024) Optimal stopping via randomized neural networks. *Frontiers of Mathematical Finance* **3**(1):31–77.

[12] Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**(301):13–30.

[13] Kim, J.-H. & Min, S. (2024) Risk-sensitive policy optimization via predictive CVaR policy gradient. In *Proceedings of the International Conference on Machine Learning*, *Proceedings of Machine Learning Research* **235**:24354–24369.

[14] Kostrikov, I., Nair, A. & Levine, S. (2022) Offline reinforcement learning with implicit Q-learning. *International Conference on Learning Representations*.

[15] Lapeyre, B. & Lelong, J. (2021) Neural network regression for Bermudan option pricing. *Monte Carlo Methods and Applications* **27**(3):227–247.

[16] Longstaff, F. A. & Schwartz, E. S. (2001) Valuing American options by simulation: A simple least-squares approach. *The Review of Financial Studies* **14**(1):113–147.

[17] Luo, Y. & Delage, E. (2026) Boosting CVaR policy optimization with quantile gradients. In *Proceedings of the International Conference on Machine Learning*, *Proceedings of Machine Learning Research* **306**:82196–82219.

[18] Luo, Y., Pan, Y., Wang, H., Torr, P. & Poupart, P. (2024) A simple mixture policy parameterization for improving sample efficiency of CVaR optimization. *Reinforcement Learning Journal* **2**:573–592.

[19] Manzoor, E., Ascarza, E. & Netzer, O. (2025) Learning when to quit in sales conversations. *arXiv preprint* arXiv:2511.01181.

[20] Mead, H., Costen, C., Lacerda, B. & Hawes, N. (2025) Return capping: Sample efficient CVaR policy gradient optimisation. In *Proceedings of the International Conference on Machine Learning*, *Proceedings of Machine Learning Research* **267**:43503–43518.

[21] Rockafellar, R. T. & Uryasev, S. (2000) Optimization of conditional value-at-risk. *Journal of Risk* **2**(3):21–41.

[22] Rockafellar, R. T. & Uryasev, S. (2002) Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance* **26**(7):1443–1471.

[23] Shapiro, A. (2009) On a time consistency concept in risk averse multistage stochastic programming. *Operations Research Letters* **37**(3):143–147.

[24] Shiryaev, A. N. (2008) *Optimal Stopping Rules*. Stochastic Modelling and Applied Probability, Vol. 8. Springer.

[25] Tamar, A., Glassner, Y. & Mannor, S. (2015) Optimizing the CVaR via sampling. *Proceedings of the AAAI Conference on Artificial Intelligence* **29**(1):2993–2999.

A Assumptions in the experimental domains

We verify the payoff-moment assumptions and examine the additional policy and convergence conditions in the experimental domains.

A.1 Model and payoff moments

Max-call rewards. Let S_t^j denote the price of asset $j \in \{1, \dots, d\}$, with $S_0^j = s_0$. Asset prices follow independent geometric Brownian motions under the risk-neutral measure, with drift $r - \delta$ and volatility σ_j. Prices are sampled at $t = 0, \dots, T = 9$ with time step $\Delta t = 1/3$ year. Let $K_{\text{strike}} = 100$ and define

$$R_t := e^{-rt\Delta t} \left(\max_{1 \le j \le d} S_t^j - K_{\text{strike}} \right)^+ .$$

The experiments use $r = 0.05$, $\delta = 0.10$, $d \in \{20, 50, 100, 200\}$, and $s_0 \in \{90, 100, 110\}$. The symmetric domain has $\sigma_j = 0.20$. The asymmetric domain has $\sigma_j = 0.10 + j/(2d)$. Every asset price has finite moments of all orders. Since

$$0 \le R_t \le \sum_{j=1}^{d} S_t^j, \qquad R_t^2 \le \sum_{j=1}^{d} (S_t^j)^2,$$

Assumptions 2.1 and 3.8 hold in both domains, with $M = \max_{0 \le t \le T} \|R_t\|_{L^2(P)} < \infty$.

House-selling rewards. Let $X_j > 0$ be the offer at decision $j \in \{1, \dots, 9\}$ and let $c > 0$ be the cost per decision. Accepting at j gives $X_j - cj$, which can be negative. Adding $9c$ to the reward at every decision makes the rewards nonnegative. For each fixed policy, CVaR and VaR increase by $9c$. Replacing R_t by $R_t + 9c$ and ν by $\nu + 9c$ leaves $(\nu - R_t)^+$ and the pathwise gradient unchanged. The tables retain the original reward units. Lognormal offers and the tested Pareto shapes $a > 2$ have finite second moments.

A.2 Policy regularity

Sigmoid output. For a finite logit $z_t^\theta(S_t)$, the probability $\phi_t^\theta(S_t) = (1 + e^{-z_t^\theta(S_t)})^{-1}$ lies in $(0, 1)$ for $t < T$. Forced terminal stopping gives $\phi_T^\theta(S_T) = 1$ and $\nabla_\theta \phi_T^\theta(S_T) = 0$. A sigmoid output does not by itself imply the uniform derivative bound in Assumption 3.5. For example, $\phi^\theta(s) = (1 + e^{-\theta s})^{-1}$ has derivative $s/4$ at $\theta = 0$, which is unbounded as $s \to \infty$.

Derivatives of stopping probabilities. Suppose $\|\nabla_\theta \phi_t^\theta(S_t)\| \le G$ uniformly in t and θ, as required by Assumption 3.5. The derivative formula in Remark 3.7 contains probability factors in $[0,1]$. The triangle inequality therefore gives

$$\|\nabla_\theta \psi_t^\theta\| \le G(1+t) \le G(T+1). \tag{21}$$

Thus $C_\psi = G(T+1)$ is a valid choice in Assumption 3.5. For a bound independent of T, suppose additionally that $\phi_s^\theta(S_s) \ge \phi_-$ for every $s < T$ and $\theta \in \Theta$ on an event of probability one, where $\phi_- \in (0,1)$. For $1 \le t \le T$,

$$\|\nabla_\theta \psi_t^\theta\| \le G(1-\phi_-)^t + tG(1-\phi_-)^{t-1}.$$

At $t = 0$, $\|\nabla_\theta \psi_0^\theta\| \le G$. Consequently, a valid choice is

$$C_\psi = G \max\left\{1, \sup_{m \in \{1,2,\dots\}} \left[(1-\phi_-)^m + m(1-\phi_-)^{m-1}\right]\right\} < \infty.$$

Finiteness follows from $m(1-\phi_-)^{m-1} \to 0$. For fixed G and ϕ_-, the bound C_ψ is independent of T.

Training implementation. The network in Appendix C uses ReLU and BatchNorm. ReLU is not differentiable at zero, and the lognormal inputs are unbounded. During training, BatchNorm statistics use all $B(T+1)$ state-time pairs, so a probability at time t can depend on future states and other trajectories. The training computation therefore differs from the pointwise policy $\phi_t^\theta(S_t)$ assumed in the analysis. Evaluation uses fixed running statistics. The architecture alone does not establish Assumption 3.5.

A.3 Convergence assumptions and training diagnostics

Terminal-stop benchmark. For nonnegative R_T, $\mathrm{CVaR}_\beta(R_T) > 0$ if and only if $P(R_T = 0) < \beta$. Let Φ denote the standard normal CDF. In the symmetric domain, $T\Delta t = 3$ and $r - \delta - \sigma^2/2 = -0.07$, giving

$$P(R_T = 0) = \Phi\left(\frac{\log(K_{\text{strike}}/s_0) + 0.21}{0.2\sqrt{3}}\right)^d.$$

The largest probability on the tested grid occurs at $(d, s_0) = (20, 90)$ and is below 0.0183. In the asymmetric domain,

$$P(R_T = 0) = \prod_{j=1}^{d} \Phi\left(\frac{\log(K_{\text{strike}}/s_0) - 3(r - \delta - \sigma_j^2/2)}{\sigma_j\sqrt{3}}\right).$$

The largest probability again occurs at $(d, s_0) = (20, 90)$ and is below 0.00861. Hence stopping at T gives positive CVaR at every tested target tail level in both domains.

Conditions along the iterates. The terminal-stop calculation establishes a positive-CVaR benchmark. Theorem 4.2 additionally requires the CDF bounds in Assumption 4.1(a), a differentiable V_β with an L_V-Lipschitz gradient on U, and $\nu^*(\theta_k) \ge \nu_-$ throughout the analyzed updates. Max-call rewards have an atom at zero, so the CDF bounds cannot be inferred from the continuous asset-price distributions. During the fixed-tail phase, the recorded batch estimate ν_k is zero in 6 of 180,000 symmetric-domain updates and 2 of 180,000 asymmetric-domain updates. Counts of zero ν_k do not verify $\nu^*(\theta_k) \ge \nu_-$.

B Proofs

Figure 1 shows which preceding results each proof uses.

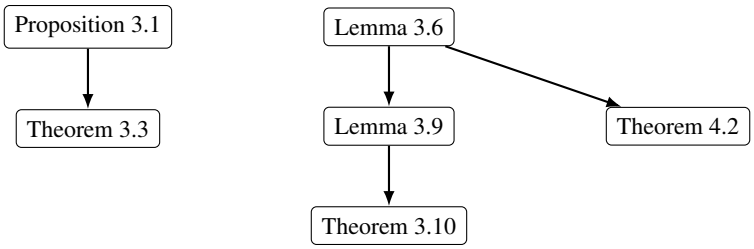

Figure 1: Proof dependencies. An arrow points from a result to a proof that uses it.

Proof of Proposition 3.1

Proof. For every $X \in L^1(P)$, the representation (8) gives

$$\mathrm{CVaR}_\beta(X) \;=\; \sup_{\nu\in\mathbb{R}} F(\nu, X) \quad \text{with} \quad F(\nu, X) \;:=\; \nu + \tfrac{1}{\beta}\,\mathbb{E}\big[-(\nu - X)^+\big]\,,$$

where the supremum is attained at $\nu = \mathrm{VaR}_\beta(X)$. Assumption 2.1 gives $R_\tau \in L^1(P)$ for every $\tau \in \mathcal{T}$, so we may set $X = R_\tau$.

By the commutativity of iterated suprema,

$$V_\beta^* \;=\; \sup_{\tau\in\mathcal{T}} \sup_{\nu\in\mathbb{R}} F(\nu, R_\tau) \;=\; \sup_{\nu\in\mathbb{R}} \Phi(\nu),$$

where

$$\Phi(\nu) \;:=\; \sup_{\tau\in\mathcal{T}} f_\tau(\nu), \qquad f_\tau(\nu) \;:=\; \nu + \tfrac{1}{\beta}\,\mathbb{E}[-(\nu - R_\tau)^+].$$

Equation (10) follows.

For each $x \in \mathbb{R}$, $\nu \mapsto -(\nu - x)^+$ is 1-Lipschitz:

$$\big|-(\nu_1 - x)^+ + (\nu_2 - x)^+\big| \;\le\; |\nu_1 - \nu_2|.$$

Hence $\nu \mapsto \mathbb{E}[-(\nu - R_\tau)^+]$ is 1-Lipschitz for every $\tau \in \mathcal{T}$, and f_τ is $(1 + 1/\beta)$-Lipschitz. Since $-(\nu - x)^+ \le 0$, $f_\tau(\nu) \le \nu$, so $\Phi(\nu) \le \nu < \infty$ for every $\nu \in \mathbb{R}$. A pointwise-finite supremum of L-Lipschitz functions is again L-Lipschitz, so Φ is $(1 + 1/\beta)$-Lipschitz.

For $\nu \le 0$, $R_\tau \ge 0$ (since each g_t takes values in $\mathbb{R}_+$) gives $(\nu - R_\tau)^+ = 0$, hence $\mathbb{E}[-(\nu - R_\tau)^+] = 0$ for every $\tau \in \mathcal{T}$ and

$$\Phi(\nu) \;=\; \nu \qquad (\nu \le 0).$$

In particular $\Phi(0) = 0$, and $\Phi(\nu) \to -\infty$ as $\nu \to -\infty$.

The bound $-(\nu - R_\tau)^+ \le -(\nu - R_\tau)$ is pointwise, so

$$\Phi(\nu) \;\le\; \nu + \tfrac{1}{\beta}\left(-\nu + \sup_{\tau\in\mathcal{T}} \mathbb{E}[R_\tau]\right) \;=\; (1 - 1/\beta)\,\nu \;+\; \tfrac{1}{\beta}\,\sup_{\tau\in\mathcal{T}} \mathbb{E}[R_\tau].$$

Since $R_t \ge 0$, $R_\tau = \sum_{t=0}^T R_t \mathbf{1}\{\tau = t\} \le \sum_{t=0}^T R_t$, hence

$$\sup_{\tau\in\mathcal{T}} \mathbb{E}[R_\tau] \;\le\; \sum_{t=0}^{T} \mathbb{E}[R_t] \;=:\; C \;<\; \infty$$

by Assumption 2.1. Hence $\Phi(\nu) \le (1 - 1/\beta)\,\nu + C/\beta$, and since $1 - 1/\beta < 0$, $\Phi(\nu) \to -\infty$ as $\nu \to +\infty$.

Φ is continuous and $\Phi(\nu) \to -\infty$ as $|\nu| \to \infty$, so there is $K > 0$ with $\Phi(\nu) < 0$ for $|\nu| > K$. Since $\Phi(0) = 0$, $\sup_\nu \Phi = \sup_{[-K,K]} \Phi$, attained by Weierstrass. Fix $\bar\nu \in [-K, K]$ with $\Phi(\bar\nu) = V_\beta^*$.

Fix $\nu \in \mathbb{R}$. The process $h_t^\nu := -(\nu - R_t)^+$ is $(\mathcal{F}_t)$-adapted with $|h_t^\nu| \le |\nu| + R_t \in L^1(P)$. Define the Snell envelope

$$V_T^\nu \;:=\; h_T^\nu, \qquad V_t^\nu \;:=\; \max\big(h_t^\nu,\; \mathbb{E}[V_{t+1}^\nu \mid \mathcal{F}_t]\big) \quad (t < T).$$

By Shiryaev [24],

$$\sup_{\tau\in\mathcal{T}} \mathbb{E}[h_\tau^\nu] \;=\; \mathbb{E}[V_0^\nu],$$

where the supremum is attained at $\tau^\nu := \inf\{t \geq 0 : V_t^\nu = h_t^\nu\}$. The boundary condition $V_T^\nu = h_T^\nu$ forces $\tau^\nu \leq T$, so $\tau^\nu \in \mathcal{T}$. Since $f_\tau(\nu) = \nu + \frac{1}{\beta}\,\mathbb{E}[h_\tau^\nu]$, $\Phi(\nu) = f_{\tau^\nu}(\nu)$.

Set $\tau^* := \tau^{\bar{\nu}}$. Then

$$V_\beta^* \;=\; \Phi(\bar{\nu}) \;=\; f_{\tau^*}(\bar{\nu}) \;\leq\; \sup_\nu f_{\tau^*}(\nu) \;=\; \mathrm{CVaR}_\beta(R_{\tau^*}) \;\leq\; V_\beta^*.$$

By (8), $\sup_\nu f_{\tau^*}(\nu)$ is attained at $\nu = \mathrm{VaR}_\beta(R_{\tau^*})$, so

$$f_{\tau^*}\big(\mathrm{VaR}_\beta(R_{\tau^*})\big) \;=\; V_\beta^*.$$

Hence $\Phi(\mathrm{VaR}_\beta(R_{\tau^*})) \geq V_\beta^*$. The definition $V_\beta^* = \sup_{\nu\in\mathbb{R}} \Phi(\nu)$ gives the reverse inequality. Setting $\nu^* := \mathrm{VaR}_\beta(R_{\tau^*})$, the pair (ν^*, τ^*) jointly realizes the supremum in (10). □

Proof of Theorem 3.3

Proof. By Proposition 3.1, there exists $\nu^* \in \mathbb{R}$ attaining the outer supremum in (10). Define $h_t^{\nu^*} := -(\nu^* - R_t)^+$ and the Snell envelope

$$V_T^{\nu^*} \;:=\; h_T^{\nu^*}, \qquad V_t^{\nu^*} \;:=\; \max\big(h_t^{\nu^*},\, \mathbb{E}[V_{t+1}^{\nu^*} \mid \mathcal{F}_t]\big) \quad (t < T).$$

By Shiryaev [24], $\tau^{\nu^*} := \inf\{t \geq 0 : V_t^{\nu^*} = h_t^{\nu^*}\}$ attains $\sup_{\tau\in\mathcal{T}} \mathbb{E}[h_\tau^{\nu^*}]$.

It suffices to find Borel sets $D_0, \ldots, D_T \subseteq \mathbb{R}^d$ with $D_T = \mathbb{R}^d$ and

$$\tau^{\nu^*} \;=\; \inf\{t \geq 0 : S_t \in D_t\} \quad P\text{-a.s.}$$

Define the Borel function $\tilde{h}_t^{\nu^*} : \mathbb{R}^d \to \mathbb{R}$ by $\tilde{h}_t^{\nu^*}(s) := -(\nu^* - g_t(s))^+$. By (S1), $h_t^{\nu^*} = \tilde{h}_t^{\nu^*}(S_t)$.

Backward induction gives Borel functions $v_t^{\nu^*} : \mathbb{R}^d \to \mathbb{R}$ with $v_t^{\nu^*}(S_t) \in L^1(P)$ satisfying

$$V_t^{\nu^*} \;=\; v_t^{\nu^*}(S_t) \quad P\text{-a.s.}$$

At $t = T$, define $v_T^{\nu^*} := \tilde{h}_T^{\nu^*}$. Then $V_T^{\nu^*} = v_T^{\nu^*}(S_T)$ and $v_T^{\nu^*}(S_T) \in L^1(P)$ by Assumption 2.1.

Suppose $V_{t+1}^{\nu^*} = v_{t+1}^{\nu^*}(S_{t+1})$ P-a.s. for some Borel $v_{t+1}^{\nu^*}$ with $v_{t+1}^{\nu^*}(S_{t+1}) \in L^1(P)$. Since $(S_t)_{t=0}^T$ is Markov,

$$\mathbb{E}[V_{t+1}^{\nu^*} \mid \mathcal{F}_t] \;=\; \mathbb{E}[v_{t+1}^{\nu^*}(S_{t+1}) \mid S_t].$$

By Doob-Dynkin, there is a Borel $q_t^{\nu^*} : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}[v_{t+1}^{\nu^*}(S_{t+1}) \mid S_t] = q_t^{\nu^*}(S_t)$ P-a.s. Define

$$v_t^{\nu^*} \;:=\; \max\big(\tilde{h}_t^{\nu^*},\, q_t^{\nu^*}\big),$$

which is Borel as the max of two Borel functions. Then $V_t^{\nu^*} = v_t^{\nu^*}(S_t)$ P-a.s. Integrability follows from

$$|v_t^{\nu^*}(S_t)| \;\leq\; |\tilde{h}_t^{\nu^*}(S_t)| + |q_t^{\nu^*}(S_t)|$$

combined with Assumption 2.1 and the conditional Jensen inequality.

Define

$$D_t \;:=\; \{s \in \mathbb{R}^d : v_t^{\nu^*}(s) = \tilde{h}_t^{\nu^*}(s)\},$$

which is Borel as the preimage of $\{0\}$ under the Borel map $v_t^{\nu^*} - \tilde{h}_t^{\nu^*}$. Then $V_t^{\nu^*} = h_t^{\nu^*} \iff S_t \in D_t$ P-a.s., so

$$\tau^{\nu^*} \;=\; \inf\{t \geq 0 : S_t \in D_t\} \quad P\text{-a.s.}$$

Since $v_T^{\nu^*} = \tilde{h}_T^{\nu^*}$, $D_T = \mathbb{R}^d$. Define $\tau^* := \tau^{\nu^*}$. The stopping time τ^* is Markov by Definition 3.2. Since ν^* attains the outer supremum in (10), $V_\beta^* = \nu^* + \frac{1}{\beta}\sup_\tau \mathbb{E}[h_\tau^{\nu^*}]$. Since $\tau^* = \tau^{\nu^*}$ maximizes $\mathbb{E}[h_\tau^{\nu^*}]$ over $\tau \in \mathcal{T}$, $V_\beta^* = \nu^* + \frac{1}{\beta}\mathbb{E}[-(\nu^* - R_{\tau^*})^+]$. □

Proof of Lemma 3.6

Proof. Since $\psi_t^\theta \nabla_\theta \log \psi_t^\theta = \nabla_\theta \psi_t^\theta$,

$$-\nabla_\theta L(\theta) \;=\; \mathbb{E}\left[\sum_{t=0}^{T} \tilde{R}_t^\nu \, \nabla_\theta \psi_t^\theta\right].$$

By (S2), the joint distribution of $(S_0, \ldots, S_T)$ does not depend on θ. At fixed ν, the factors $\tilde{R}_t^\nu$ also do not depend on θ. Since $\|\nabla_\theta \psi_t^\theta \cdot \tilde{R}_t^\nu\| \le \frac{C_\psi}{\beta}(|\nu| + R_t) \in L^1(P)$ by Assumptions 2.1 and 3.5, differentiation under the expectation gives

$$\nabla_\theta J^{(\nu)}(\theta) \;=\; \mathbb{E}\left[\sum_{t=0}^{T} \tilde{R}_t^\nu \, \nabla_\theta \psi_t^\theta\right] \;=\; -\nabla_\theta L(\theta).$$

□

Proof of Lemma 3.9

Proof. Fix $\theta \in \Theta$ and $\nu \in \mathbb{R}$. For each sampled trajectory $i = 1, \ldots, B$, define

$$\hat{g}^{(i)}(\theta) \;:=\; -\sum_{t=0}^{T} \tilde{R}_t^{\nu,(i)} \cdot \nabla_\theta \psi_t^{\theta,(i)}.$$

Define $\bar{a}^{(i)} := \frac{1}{T+1} \sum_{t=0}^{T} \tilde{R}_t^{\nu,(i)}$. By (S3), $\sum_{t=0}^{T} \nabla_\theta \psi_t^{\theta,(i)} = 0$ P-a.s., so

$$\hat{g}^{(i)}(\theta) \;=\; -\sum_{t=0}^{T} \big(\tilde{R}_t^{\nu,(i)} - \bar{a}^{(i)}\big) \cdot \nabla_\theta \psi_t^{\theta,(i)}.$$

By Cauchy–Schwarz,

$$\|\hat{g}^{(i)}\|^2 \;\le\; \Big(\sum_{t=0}^{T} (\tilde{R}_t^{\nu,(i)} - \bar{a}^{(i)})^2\Big) \cdot \Big(\sum_{t=0}^{T} \|\nabla_\theta \psi_t^{\theta,(i)}\|^2\Big).$$

By Assumption 3.5, $\|\nabla_\theta \psi_t^{\theta,(i)}\| \le C_\psi$ P-a.s., hence

$$\sum_{t=0}^{T} \|\nabla_\theta \psi_t^{\theta,(i)}\|^2 \;\le\; (T+1) C_\psi^2.$$

Since $\bar{a}^{(i)}$ minimizes $\sum_{t=0}^{T} (\tilde{R}_t^{\nu,(i)} - a)^2$ over $a \in \mathbb{R}$, comparison with $a = 0$ gives

$$\mathbb{E}\left[\sum_{t=0}^{T} (\tilde{R}_t^{\nu,(i)} - \bar{a}^{(i)})^2\right] \le \sum_{t=0}^{T} \mathbb{E}[(\tilde{R}_t^{\nu,(i)})^2] \le (T+1)\tilde{M}^2.$$

Thus $\mathbb{E}[\|\hat{g}^{(i)}\|^2] \le C_\psi^2 (T+1)^2 \tilde{M}^2$. By convexity, $\mathbb{E}[\|\hat{\nabla} L(\theta)\|^2] \le B^{-1} \sum_{i=1}^{B} \mathbb{E}[\|\hat{g}^{(i)}(\theta)\|^2] \le C_\psi^2 (T+1)^2 \tilde{M}^2$, proving part (a).

$\hat{\nabla} L(\theta) = (1/B) \sum_{i=1}^{B} \hat{g}^{(i)}$ is the mean of B i.i.d. samples with $\mathbb{E}[\hat{g}^{(i)}] = \nabla L(\theta)$ by Lemma 3.6, so

$$\mathbb{E}\big[\|\hat{\nabla} L(\theta) - \nabla L(\theta)\|^2\big] \;=\; \tfrac{1}{B}\, \mathbb{E}[\|\hat{g}^{(1)} - \mathbb{E}[\hat{g}^{(1)}]\|^2] \;\le\; \tfrac{1}{B}\, \mathbb{E}\big[\|\hat{g}^{(1)}\|^2\big] \;\le\; \frac{C_\psi^2 (T+1)^2 \tilde{M}^2}{B}.$$

The variance bound proves part (b). For part (c), if $M = 0$, then $R_t = 0$ P-a.s. for every t. Thus $\nu^*(\theta) = 0$, $\hat{\nabla} L(\theta) = \nabla L(\theta) = 0$ at $\nu = \nu^*(\theta)$, and (14) holds with $c_\beta = 0$. Suppose $M > 0$. At $\nu = \nu^*(\theta)$, $\tilde{M}^2 \le c_\beta M^2/\beta^2$ and $(T+1)^2 \le 4T^2$ for $T \ge 1$, which gives (14). For $T = 0$, $\psi_0 = 1$ and $\hat{\nabla} L(\theta) = \nabla L(\theta) = 0$. For $M > 0$, the lower-quantile property gives

$$(1-\beta)\nu^*(\theta)^2 \le \mathbb{E}[R_{\tau_\theta}^2] \le (T+1) M^2.$$

Since $R_t \ge 0$, $((\nu^*(\theta) - R_t)^+)^2 \le \nu^*(\theta)^2$, which gives $c_\beta \le (T+1)/(1-\beta)$. □

Proof of Theorem 3.10

Proof. Define $\tilde{M} := \max_t \|\tilde{R}_t^\nu\|_{L^2(P)}$.

On $\{\tau = u\}$, the sampled action likelihood equals ψ_u^θ. For $u < T$,

$$\nabla_\theta \log \psi_u^\theta \;=\; \frac{\nabla \phi_u^\theta(S_u)}{\phi_u^\theta(S_u)} \;-\; \sum_{s<u} \frac{\nabla \phi_s^\theta(S_s)}{1-\phi_s^\theta(S_s)}.$$

At $u = T$,

$$\nabla_\theta \log \psi_T^\theta = -\sum_{s<T} \frac{\nabla \phi_s^\theta(S_s)}{1-\phi_s^\theta(S_s)}.$$

For each u, $\nabla_\theta \log \psi_u^\theta$ is $\mathcal{F}_u$-measurable.

Since $\mathbb{E}[\mathbf{1}\{\tau = u\} \mid \mathcal{F}_T] = \phi_u^\theta(S_u) \prod_{s<u}(1-\phi_s^\theta(S_s)) = \psi_u^\theta$,

$$\mathbb{E}\big[\|\hat{\nabla} L_{\text{score}}(\theta_n)\|^2\big] \;=\; \sum_{u=0}^{T} \mathbb{E}\Big[(\tilde{R}_u^\nu)^2 \,\|\nabla_\theta \log \psi_u^{\theta_n}\|^2 \,\psi_u^{\theta_n}\Big].$$

By condition (c) and continuity of probability under increasing unions, choose $\eta > 0$ such that $P(\{R_{t^*} \le \nu - \eta\} \cap A) > 0$. Define $A' := \{R_{t^*} \le \nu - \eta\} \cap A$ and $\delta := \eta^2/\beta^2$. Then $(\tilde{R}_{t^*}^\nu)^2 \ge \delta > 0$ on A'.

Nonnegativity of $(\tilde{R}_u^\nu)^2 \|\nabla_\theta \log \psi_u^{\theta_n}\|^2 \psi_u^{\theta_n}$ gives

$$\mathbb{E}\big[\|\hat{\nabla} L_{\text{score}}(\theta_n)\|^2\big] \;\ge\; \mathbb{E}\Big[\mathbf{1}_{A'}\,(\tilde{R}_{t^*}^\nu)^2 \,\|\nabla_\theta \log \psi_{t^*}^{\theta_n}\|^2 \,\psi_{t^*}^{\theta_n}\Big].$$

Define

$$V_n := \sum_{s<t^*} \frac{\nabla_\theta \phi_s^{\theta_n}(S_s)}{1-\phi_s^{\theta_n}(S_s)}, \qquad c := \inf_n \inf_{\omega \in A} \|\nabla_\theta \phi_{t^*}^{\theta_n}(S_{t^*}(\omega))\|^2 > 0.$$

On A', condition (b) and Assumption 3.5 give

$$\prod_{s<t^*}(1-\phi_s^{\theta_n}(S_s)) \ge \epsilon_0^{t^*}, \qquad \|V_n\|^2 \le t^* \sum_{s<t^*} \frac{G^2}{\epsilon_0^2} = \frac{(t^*)^2 G^2}{\epsilon_0^2}.$$

Apply $\|u-v\|^2 \ge \frac{1}{2}\|u\|^2 - \|v\|^2$ to $\nabla_\theta \log \psi_{t^*}^{\theta_n} = \nabla_\theta \phi_{t^*}^{\theta_n}(S_{t^*})/\phi_{t^*}^{\theta_n}(S_{t^*}) - V_n$. Using $\psi_{t^*}^{\theta_n}/\phi_{t^*}^{\theta_n}(S_{t^*}) = \prod_{s<t^*}(1-\phi_s^{\theta_n}(S_s))$, $\psi_{t^*}^{\theta_n} \le 1$, and $\mathbb{E}[(\tilde{R}_{t^*}^\nu)^2] \le \tilde{M}^2$ gives

$$\mathbb{E}[\|\hat{\nabla} L_{\text{score}}(\theta_n)\|^2] \ge \frac{\delta c \epsilon_0^{t^*}}{2} \mathbb{E}\left[\frac{\mathbf{1}_{A'}}{\phi_{t^*}^{\theta_n}(S_{t^*})}\right] - \frac{\tilde{M}^2 (t^*)^2 G^2}{\epsilon_0^2}.$$

Since $\phi_{t^*}^{\theta_n}(S_{t^*}) \to 0$ on A' and $P(A') > 0$, Fatou's lemma gives $\mathbb{E}[\mathbf{1}_{A'}/\phi_{t^*}^{\theta_n}(S_{t^*})] \to \infty$. Therefore $\mathbb{E}[\|\hat{\nabla} L_{\text{score}}(\theta_n)\|^2] \to \infty$.

For fixed ν, nonnegativity of R_t gives $|\tilde{R}_t^\nu| \le \nu^+/\beta$. Lemma 3.9(a) therefore gives

$$\mathbb{E}[\|\hat{\nabla} L(\theta)\|^2] \le (T+1)^2 C_\psi^2 (\nu^+)^2/\beta^2,$$

for every $\theta \in \Theta$.

□

Sigmoid policies

Under Assumptions 2.1, 3.5, and 3.8, fix $\nu \in \mathbb{R}$ and a sequence $(\theta_n) \subset \Theta$. Let $B = 1$. Define $\varsigma(x) := (1+e^{-x})^{-1}$. Let $\phi_t^\theta(S_t) = \varsigma(z_t^\theta(S_t))$ with $z_t^\theta(S_t) \in \mathbb{R}$ for $t < T$. Assume

$$\sup_n \sup_{\omega \in \Omega} \max_{t<T} \|\nabla_\theta z_t^{\theta_n}(S_t(\omega))\| < \infty,$$

$$\phi_t^{\theta_n}(S_t) \to \phi_t^{\lim} \in \{0,1\} \quad P\text{-a.s. for every } t < T.$$

Then

$$\mathbb{E}[\|\hat{\nabla} L_{\text{score}}(\theta_n)\|^2] \to 0, \qquad \mathbb{E}[\|\hat{\nabla} L(\theta_n)\|^2] \to 0.$$

Proof. Define $\tilde{M} := \max_{0\le t\le T} \|\tilde{R}_t^\nu\|_{L^2(P)}$. Define $\tau_n := \tau_{\theta_n}$ and $a_t^{(n)} := \mathbf{1}\{U_t \le \phi_t^{\theta_n}(S_t)\}$ for $t < T$, with $U_0, \dots, U_T$ as in (3). For $t < T$, the sigmoid parameterization gives $\nabla_\theta \log \pi_\theta(a_t \mid S_t) = (a_t - \phi_t^\theta(S_t))\nabla_\theta z_t^\theta(S_t)$. Therefore

$$\nabla_\theta \log \pi_\theta(\tau) = \sum_{t=0}^{\min(\tau, T-1)} (a_t - \phi_t^\theta(S_t))\nabla_\theta z_t^\theta(S_t).$$

Define $C_z := \sup_n \sup_{\omega\in\Omega} \max_{t<T} \|\nabla_\theta z_t^{\theta_n}(S_t(\omega))\| < \infty$. Since $|a_t^{(n)} - \phi_t^{\theta_n}(S_t)| \le 1$, the triangle inequality gives $\|\nabla_\theta \log \pi_{\theta_n}(\tau_n)\| \le (T+1)C_z$.

Since $(\tilde{R}_{\tau_n}^\nu)^2 \le \sum_{s=0}^T (\tilde{R}_s^\nu)^2$ and $\|\nabla_\theta \log \pi_{\theta_n}(\tau_n)\| \le (T+1)C_z$,

$$\|\hat{\nabla} L_{\text{score}}(\theta_n)\|^2 \;\le\; (1+T)^2 C_z^2 \sum_{s=0}^T (\tilde{R}_s^\nu)^2 \quad \text{pointwise,}$$

with $\mathbb{E}[\sum_{s=0}^T (\tilde{R}_s^\nu)^2] \le (T+1)\tilde{M}^2 < \infty$, independent of n.

For each $t < T$, $\phi_t^{\theta_n}(S_t) \to \phi_t^{\lim} \in \{0, 1\}$ and $U_t \ne \phi_t^{\lim}$ almost surely. Hence $a_t^{(n)} = \phi_t^{\lim}$ for all sufficiently large n, so $a_t^{(n)} - \phi_t^{\theta_n}(S_t) \to 0$ almost surely. Since T is finite and $\|\nabla_\theta z_t^{\theta_n}(S_t)\| \le C_z$, the finite score sum satisfies $\nabla_\theta \log \pi_{\theta_n}(\tau_n) \to 0$ almost surely.

Hence $\|\hat{\nabla} L_{\text{score}}(\theta_n)\|^2 \to 0$ almost surely. Dominated convergence with the integrable bound $(T+1)^2 C_z^2 \sum_{s=0}^T (\tilde{R}_s^\nu)^2$ gives $\mathbb{E}[\|\hat{\nabla} L_{\text{score}}(\theta_n)\|^2] \to 0$.

For $t < T$, $\nabla_\theta \phi_t^{\theta_n}(S_t) = \phi_t^{\theta_n}(S_t)(1 - \phi_t^{\theta_n}(S_t))\nabla_\theta z_t^{\theta_n}(S_t) \to 0$. The derivative formula in Remark 3.7 gives $\nabla_\theta \psi_t^{\theta_n} \to 0$ for $0 \le t \le T$. Thus $\hat{\nabla} L(\theta_n) \to 0$ almost surely. Since $|\tilde{R}_t^\nu| \le \nu^+/\beta$, the pathwise formula gives $\|\hat{\nabla} L(\theta_n)\| \le (T+1)C_\psi \nu^+/\beta$. Dominated convergence yields $\mathbb{E}[\|\hat{\nabla} L(\theta_n)\|^2] \to 0$.

For each $t < T$, the sigmoid derivative also gives

$$\|\nabla_\theta \phi_t^{\theta_n}(S_t)\| \le C_z \phi_t^{\theta_n}(S_t)(1 - \phi_t^{\theta_n}(S_t)).$$

Thus $\nabla_\theta \phi_t^{\theta_n}(S_t) \to 0$ wherever $\phi_t^{\theta_n}(S_t) \to 0$, which excludes condition (a) of Theorem 3.10 when the logit gradients are uniformly bounded. □

Proof of Theorem 4.2

Proof. Let $\theta \in \Theta_{\nu_-}$ and define $q := \nu^*(\theta)$. Let $X_1, \dots, X_B$ be independent copies of R_{τ_θ}. Let $X_{(1)} \le \cdots \le X_{(B)}$ denote the sorted sample and define $\widehat{q} := X_{(m_B)}$. Define

$$H := \max_{0\le t\le T} R_t, \qquad M_H := \sqrt{T+1}\, M, \qquad Q := \frac{M_H}{\sqrt{1-\beta}}.$$

Then $\mathbb{E}[H^2] \le M_H^2$. Assumption 4.1(a) at $u = 0$ gives $F_\theta(q) = \beta$. Since $0 \le R_{\tau_\theta} \le H$, the quantile q satisfies $(1-\beta)q^2 \le \mathbb{E}[(R_{\tau_\theta})^2] \le M_H^2$, and hence $0 \le q \le Q$. Define

$$B_0 := \left\lceil \max\left\{3, \frac{2}{f_-\delta_0}, 1 + \frac{2\beta}{f_-\delta_0}\right\}\right\rceil. \tag{22}$$

Let $B \ge B_0$. For $x \in \mathbb{R}$, define $N_x := \sum_{i=1}^B \mathbf{1}_{\{X_i \le x\}}$. The order statistic $\widehat{q}$ satisfies

$$\{\widehat{q} > q + u\} \subseteq \{N_{q+u} \le m_B - 1\}, \qquad \{\widehat{q} < q - u\} \subseteq \{N_{q-u} \ge m_B\}.$$

For $2/(f_-B) \le u \le \delta_0$, Assumption 4.1(a) and the rank condition (16) give

$$\mathbb{E}[N_{q+u}] - (m_B - 1) \ge Bf_-u, \qquad m_B - \mathbb{E}[N_{q-u}] \ge Bf_-u - 1 \ge Bf_-u/2.$$

Hoeffding's inequality [12, Theorem 2] for the binomial counts N_{q+u} and N_{q-u} therefore yields

$$\mathbb{P}(|\widehat{q} - q| > u) \le 2\exp\left(-\frac{Bf_-^2u^2}{2}\right). \tag{23}$$

Integrating (23) over $2/(f_-B) \le u \le \delta_0$ and using $\mathbb{P}(|\widehat{q}-q|>u) \le 1$ for $0 \le u < 2/(f_-B)$ gives

$$\begin{aligned}\mathbb{E}\big[\min\{(\widehat{q}-q)^2, \delta_0^2\}\big] &= \int_0^{\delta_0} 2u\mathbb{P}(|\widehat{q}-q|>u)\,du \\ &\le \frac{4}{f_-^2B^2} + \int_0^{\infty} 4u\exp\left(-\frac{Bf_-^2u^2}{2}\right)du \le \frac{8}{f_-^2B}. \end{aligned} \tag{24}$$

Define $E_+ := \{\widehat{q} > q+\delta_0\}$ and $E_- := \{\widehat{q} < q-\delta_0\}$. For each $i = 1, \dots, B$, define

$$N_{q+\delta_0}^{(-i)} := \sum_{j\neq i} \mathbf{1}_{\{X_j \le q+\delta_0\}}.$$

The count $N_{q+\delta_0}^{(-i)}$ is independent of X_i, has distribution $\mathrm{Bin}(B-1, F_\theta(q+\delta_0))$, and satisfies $E_+ \subseteq \{N_{q+\delta_0}^{(-i)} \le m_B - 1\}$. By (22), (16), and Assumption 4.1(a),

$$\frac{m_B-1}{B-1} \le \beta + \frac{\beta}{B-1} \le \beta + \frac{f_-\delta_0}{2}, \qquad F_\theta(q+\delta_0) \ge \beta + f_-\delta_0.$$

Independence of X_i and $N_{q+\delta_0}^{(-i)}$ and Hoeffding's inequality give

$$\mathbb{E}[X_i^2\mathbf{1}_{E_+}] \le \mathbb{E}[X_i^2]\mathbb{P}(N_{q+\delta_0}^{(-i)} \le m_B - 1) \le M_H^2 \exp\left(-\frac{(B-1)f_-^2\delta_0^2}{2}\right).$$

Since $\widehat{q}^2 \le \sum_{i=1}^B X_i^2$,

$$\mathbb{E}[\widehat{q}^2\mathbf{1}_{E_+}] \le BM_H^2 \exp\left(-\frac{(B-1)f_-^2\delta_0^2}{2}\right). \tag{25}$$

Hoeffding's inequality applied to $N_{q+\delta_0}$ and $N_{q-\delta_0}$ gives

$$\mathbb{P}(E_+) \le e^{-Bf_-^2\delta_0^2/2}, \qquad \mathbb{P}(E_-) \le e^{-Bf_-^2\delta_0^2/2}. \tag{26}$$

On E_+, $(\widehat{q}-q)^2 \le 2\widehat{q}^2 + 2Q^2$. On E_-, nonnegativity of $X_1, \dots, X_B$ gives $(\widehat{q}-q)^2 \le Q^2$. Combining (25) and (26) gives

$$\mathbb{E}[(\widehat{q}-q)^2\mathbf{1}_{E_+\cup E_-}] \le (2BM_H^2 + 3Q^2)\exp\left(-\frac{(B-1)f_-^2\delta_0^2}{2}\right). \tag{27}$$

Since $B(2BM_H^2+3Q^2)e^{-(B-1)f_-^2\delta_0^2/2}$ is bounded for $B \ge B_0$, (24) and (27) imply that there exists $C_\nu > 0$, depending only on β, T, M, f_-, and δ_0, such that

$$\mathbb{E}[(\widehat{q}-q)^2] \le C_\nu/B \quad \text{for every } \theta \in \Theta_{\nu_-} \text{ and } B \ge B_0. \tag{28}$$

Conditional on $\mathcal{H}_k$, the rewards $R^{(i)}_{\tau^{(i)}_{\theta_k}}$, $i = 1, \dots, B$, are independent with CDF F_{θ_k}. Applying (28) with $\theta = \theta_k$ therefore gives

$$\mathbb{E}[(\nu_k - \nu^*(\theta_k))^2 \mid \mathcal{H}_k] \le C_\nu/B. \tag{29}$$

Taking expectation in (29) proves (18).

For $\theta \in \Theta_{\nu_-}$ and $q = \nu^*(\theta)$, define

$$W_q(\vartheta) := q - \frac{1}{\beta}\mathbb{E}\left[\sum_{t=0}^{T} \psi_t^\vartheta (q - R_t)^+\right], \qquad \vartheta \in U. \tag{30}$$

The representation (8) gives $V_\beta(\vartheta) \ge W_q(\vartheta)$ for every $\vartheta \in U$ and $V_\beta(\theta) = W_q(\theta)$. The bounds $0 \le (q-R_t)^+ \le q$ and $\|\nabla_\vartheta \psi_t^\vartheta\| \le C_\psi$ justify differentiating W_q under the expectation. By Assumption 4.1(b), the function $V_\beta - W_q$ is differentiable on U and has a minimum at $\theta \in U$. Hence

$$\nabla V_\beta(\theta) = \nabla W_q(\theta) = -\frac{1}{\beta}\mathbb{E}\left[\sum_{t=0}^{T}(q-R_t)^+\nabla_\theta \psi_t^\theta\right]. \tag{31}$$

Define

$$\Lambda := \frac{(T+1)C_\psi}{\beta}, \qquad q_k := \nu^*(\theta_k), \qquad v_k := \nabla V_\beta(\theta_k).$$

For each sampled trajectory i, define

$$Y_k^{(i)} := -\frac{1}{\beta}\sum_{t=0}^{T}(q_k - R_t^{(i)})^+ \nabla_\theta \psi_t^{\theta_k,(i)}.$$

Then $\widehat{g}_k(q_k) = B^{-1}\sum_{i=1}^{B} Y_k^{(i)}$. Conditional on $\mathcal{H}_k$, the vectors $Y_k^{(1)}, \dots, Y_k^{(B)}$ are independent. Equation (31) gives $\mathbb{E}[Y_k^{(i)} \mid \mathcal{H}_k] = v_k$, and Assumption 3.5 with $q_k \le Q$ gives $\|Y_k^{(i)}\| \le \Lambda Q$. Define $\xi_k := \widehat{g}_k(q_k) - v_k$. Conditional independence of $Y_k^{(1)}, \dots, Y_k^{(B)}$ gives

$$\mathbb{E}[\xi_k \mid \mathcal{H}_k] = 0, \qquad \mathbb{E}[\|\xi_k\|^2 \mid \mathcal{H}_k] \le \frac{\Lambda^2 Q^2}{B}. \tag{32}$$

Define $\Delta_k := \widehat{g}_k(\nu_k) - \widehat{g}_k(q_k)$. The inequality $|(\nu_k - R_t^{(i)})^+ - (q_k - R_t^{(i)})^+| \le |\nu_k - q_k|$ and the derivative bound $\|\nabla_\theta \psi_t^{\theta_k,(i)}\| \le C_\psi$ in Assumption 3.5 give $\|\Delta_k\| \le \Lambda|\nu_k - q_k|$. Applying (29) yields

$$\mathbb{E}[\|\Delta_k\|^2 \mid \mathcal{H}_k] \le \frac{\Lambda^2 C_\nu}{B}. \tag{33}$$

Define

$$e_k := \widehat{g}_k(\nu_k) - v_k, \qquad C_g := 2\Lambda^2(Q^2 + C_\nu).$$

Since $e_k = \xi_k + \Delta_k$, the inequality $\|\xi_k + \Delta_k\|^2 \le 2\|\xi_k\|^2 + 2\|\Delta_k\|^2$, together with (32) and (33), gives

$$\mathbb{E}[\|e_k\|^2 \mid \mathcal{H}_k] \le C_g/B. \tag{34}$$

Since $\theta_k, \theta_{k+1} \in U$ and U is convex, Assumption 4.1(b) and $\theta_{k+1} = \theta_k + \eta(v_k + e_k)$ imply

$$\begin{aligned} V_\beta(\theta_{k+1}) - V_\beta(\theta_k) &\ge \eta\langle v_k, v_k + e_k\rangle - \frac{L_V\eta^2}{2}\|v_k + e_k\|^2 \\ &= \frac{\eta}{2}\big(\|v_k\|^2 - \|e_k\|^2\big) + \frac{\eta}{2}(1 - L_V\eta)\|v_k + e_k\|^2. \end{aligned}$$

Since $1 - L_V\eta \ge 0$, taking conditional expectation and applying (34) yields

$$\mathbb{E}[V_\beta(\theta_{k+1}) - V_\beta(\theta_k) \mid \mathcal{H}_k] \ge \frac{\eta}{2}\|v_k\|^2 - \frac{\eta C_g}{2B}. \tag{35}$$

For every $\vartheta \in \Theta$, the inequalities $0 \le V_\beta(\vartheta) \le \mathbb{E}[H] < \infty$ imply that D_0 is finite. Taking expectation in (35), summing over $k = 0, \dots, K-1$, and applying $\mathbb{E}[V_\beta(\theta_K) - V_\beta(\theta_0)] \le D_0$ gives

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|v_k\|^2] \le \frac{2D_0}{\eta K} + \frac{C_g}{B},$$

which proves (19).

The quantile bound (18) and the gradient bound (19) give

$$\begin{aligned} &\min_{0\le k<K}\mathbb{E}[\|\nabla V_\beta(\theta_k)\|^2 + (\nu_k - \nu^*(\theta_k))^2] \\ &\quad\le \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla V_\beta(\theta_k)\|^2 + (\nu_k - \nu^*(\theta_k))^2] \le \frac{2D_0}{\eta K} + \frac{C_g + C_\nu}{B}. \end{aligned}$$

Equation (20) follows. Let $\epsilon > 0$ and define

$$K = \max\left\{1, \left\lceil \frac{4D_0}{\eta\epsilon} \right\rceil\right\}, \qquad B = \max\left\{B_0, \left\lceil \frac{2(C_g + C_\nu)}{\epsilon} \right\rceil\right\}. \tag{36}$$

By (36), $2D_0/(\eta K) \le \epsilon/2$ and $(C_g + C_\nu)/B \le \epsilon/2$. For fixed β and η, $K = O(\epsilon^{-1})$, $B = O(\epsilon^{-1})$, and $KB = O(\epsilon^{-2})$ for sufficiently small $\epsilon > 0$. □

C Implementation details for the symmetric Bermudan max-call domain

We follow the LunarLander protocol of Luo and Delage [17] with the following changes. (i) The discount $\gamma = 1$ since the max-call payoff $e^{-rt\Delta t}(\max_j S_t^j - K_{\text{strike}})^+$ already discounts through the $e^{-rt\Delta t}$ factor (with $\Delta t = 1/3$ year). (ii) The feature vector is $(S_t/100,\ t/T,\ R_t)$, providing the time index and the immediate payoff suited to optimal stopping. (iii) BatchNorm is applied to the encoder. (iv) The policy logit at $t = T$ is set to $+20$. (v) Policy gradients are clipped to a maximum norm of 1.0. (vi) The IQL batch size of MIX scales linearly with horizon T from the LunarLander value in Luo et al. [18]: 2×10^5 transitions per update at $T = 500$ becomes $3{,}600$ at $T = 9$. The IQL replay buffer holds 10^6 transitions. Tables 2, 3, and 4 specify the domain components, training protocol, and per-method hyperparameters, respectively.

Policy and quantile implementation. The encoder has two Linear–BatchNorm–ReLU blocks of width 128, followed by a $128 \to 64 \to 1$ policy head with ReLU and sigmoid output. During training, BatchNorm computes statistics across the $B(T+1)$ sampled state-time pairs. Dropout is disabled. Rewards are divided by the standard deviation of the $B(T+1)$ payoffs in the first batch, and the normalization scale is held fixed throughout training. The implementation estimates the reward quantile using the order statistic of rank $\max\{1, \lfloor \beta_k B \rfloor\}$ from the sampled stopping rewards. Floating-point schedule arithmetic gives hold-phase ranks $3, 1, 1$ at target levels $0.2, 0.1, 0.05$ with $B = 20$. Two annealing updates at target 0.2 use ranks 7 and 5 instead of 8 and 6. With annealing disabled, the separate branch uses rank 4 at 0.2. The loss is the stop-gradient surrogate in Algorithm 1, with $\log \psi_t$ computed from log-sigmoid sums and ν_k held fixed during differentiation.

Compute. Experiments ran on a server with eight NVIDIA RTX A6000 GPUs, with each job assigned to one GPU. Some jobs shared a GPU. Training times include trajectory simulation, gradient updates, logging, and six periodic evaluations of $5{,}000$ paths. Model creation and the final $20{,}000$-path evaluation are excluded. RET-CAP times also exclude training and evaluation of the policies used to set the minimum cap. The RET-CAP CVaR values come from reruns with recomputed minimum caps and retuned hyperparameters, while the tables retain the training times measured with the earlier minimum caps and hyperparameters.

Evaluation. Each final policy is evaluated on $20{,}000$ paths with seed $10000 + s$ for training seed $s \in \{0, \ldots, 9\}$. Methods share evaluation paths for each seed. Evaluation uses BatchNorm running statistics and stops at the first t with $\phi_t^\theta(S_t) > 0.5$. At $\beta = 0.2$, CVaR is the mean of exactly the lowest $4{,}000$ rewards. Means and population standard deviations are over ten training seeds.

Table 2: Bermudan max-call domain parameters.

Parameter	Value
Strike K_{strike}	100
Risk-free rate r	0.05
Dividend yield δ	0.10
Volatility σ	0.20 (shared across all d underlyings)
Asset correlation ρ	0 (independent)
Maturity	3 years
Last time index T	9 (10 candidate times including $t = 0$)
Time step Δt	$1/3$ year
Payoff	$g_t(S_t) = e^{-rt\Delta t}(\max_j S_t^j - K_{\text{strike}})^+$
Policy input features	$(S_t/100,\ t/T,\ R_t)$
d-grid	$\{20, 50, 100, 200\}$
s_0-grid	$\{90, 100, 110\}$
Configurations	12

Paired comparisons. Table 5 reports two-sided paired t and exact Wilcoxon tests over training seeds, using shared evaluation paths per seed and no multiple-comparison correction. DIOS has a higher mean CVaR than CVaR-VaR in all 24 max-call configurations, with paired t-test $p < 0.05$ against CVaR-VaR in 2 symmetric and 3 asymmetric configurations.

Table 3: Shared training protocol.

Item	Value
Trajectories per update B	20
Outer iterations K	6,000
Risk level β	0.2
Discount γ	1
Seeds	10
Optimizer	Adam
Policy LR sweep set	$\{7\mathrm{e}{-4},\ 5\mathrm{e}{-4},\ 3\mathrm{e}{-4},\ 1\mathrm{e}{-4},\ 7\mathrm{e}{-5}\}$
Value LR multiplier sweep set	$\{1, 2, 5, 10\}$ times the policy learning rate
Advantage normalization sweep set	{yes, no} where applicable
Hidden width	128
Hidden layers	2
Activation	ReLU
Encoder normalization	BatchNorm
Terminal logit overwrite	$+20$ at $t = T$
Multi-step λ (CVaR-VaR, RET-CAP only)	0.95
Gradient clip (max-norm)	1.0
Final evaluation batch	20,000

Table 4: Per-method best hyperparameters from the learning-rate sweep on the representative configuration $d = 100$, $s_0 = 100$, applied uniformly to all twelve configurations.

Method	Policy LR	Critic LR mult.	Norm. adv.	Method-specific
DIOS (ours)	$7\mathrm{e}{-5}$	—	—	3-phase β_k-schedule with warmup fraction $\rho_w = 0.25$, anneal fraction $\rho_a = 0.50$, hold fraction 0.25: $\beta_k = 1$ for the first $\rho_w K$ iterations, linear $1 \to 0.2$ over the next $\rho_a K$, $\beta_k = 0.2$ thereafter
CVaR-VaR	$1\mathrm{e}{-4}$	$1\times$	no	$N_{\mathrm{q}} = 10$ atoms in the quantile critic; tradeoff weight $\omega = 0.5$ throughout training, selected from the three ω-schedules of Luo and Delage [17] (Maze: $\omega \equiv 0.5$; LunarLander: $\omega = 0.5$ for first 40% then linear decay to 0; Pendulum: $\omega = 0.5$ for first 25% then 0)
CVaR-PG	$3\mathrm{e}{-4}$	—	no	trajectory-level score-function estimator with no value critic [25]
RET-CAP (cap-anchor)	$1\mathrm{e}{-4}$	$5\times$	yes	minimum cap q^* computed from DIOS evaluation quantiles as in Appendix D.2; advantage-shaping weight 0.8
RET-CAP (RN-anchor)	$7\mathrm{e}{-5}$	$10\times$	yes	risk-neutral anchor variant of Appendix D.3; advantage-shaping weight 0.8
MIX	$7\mathrm{e}{-4}$	—	—	IQL component: LR $1\mathrm{e}{-4}$, update interval 50, expectile 0.8, temperature $\beta_{\mathrm{IQL}} = 1$, batch size 3,600 per update, replay buffer capacity 10^6 transitions

D Additional baseline experiments on the symmetric Bermudan max-call domain

D.1 PCVaR-PG

Table 6 reports PCVaR-PG [13] on all twelve symmetric configurations. At $s_0 = 110$, all ten seeds stop immediately and receive reward 10. At $s_0 \in \{90, 100\}$, the mean stopping time ranges from 1.0 to 3.4.

D.2 RET-CAP minimum cap and the choice of anchor

RET-CAP updates a return cap during training and constrains the cap to be at least an externally supplied value q^* [20]. In the max-call experiments, we compute the empirical β-quantile of the evaluation rewards for each trained DIOS policy. We average the quantiles over ten seeds and round to three decimal places to obtain q^*. The tables label the DIOS-derived minimum cap "cap-anchor".

Table 5: Seed-paired difference DIOS − CVaR-VaR in $\text{CVaR}_{0.2}$ (10 seeds, shared evaluation paths per seed). t: two-sided paired t; W: two-sided exact Wilcoxon signed-rank; no multiple-comparison correction.

d	s_0	$\bar{\Delta}$	95% CI	t p	W p	$p_t < 0.05$
Symmetric						
20	90	$+0.103$	[-0.064, +0.271]	0.197	0.625	no
20	100	$+0.113$	[-0.086, +0.311]	0.23	0.492	no
20	110	$+0.073$	[-0.204, +0.350]	0.565	1	no
50	90	$+0.007$	[-0.081, +0.095]	0.857	1	no
50	100	$+0.088$	[-0.010, +0.186]	0.0736	0.232	no
50	110	$+0.151$	[-0.078, +0.381]	0.169	0.232	no
100	90	$+0.142$	[+0.010, +0.274]	0.0374	0.0273	yes
100	100	$+0.032$	[-0.065, +0.129]	0.472	0.322	no
100	110	$+0.125$	[+0.017, +0.232]	0.0274	0.00977	yes
200	90	$+0.033$	[-0.071, +0.136]	0.494	0.232	no
200	100	$+0.025$	[-0.051, +0.102]	0.467	0.375	no
200	110	$+0.003$	[-0.090, +0.096]	0.94	0.77	no
Asymmetric						
20	90	$+0.812$	—	0.00523	0.00586	yes
20	100	$+0.427$	—	0.195	0.193	no
20	110	$+0.857$	—	0.0187	0.0273	yes
50	90	$+0.451$	—	0.0522	0.0371	no
50	100	$+0.568$	—	0.0828	0.0488	no
50	110	$+1.418$	—	0.018	0.00977	yes
100	90	$+0.745$	—	0.114	0.193	no
100	100	$+0.058$	—	0.817	0.922	no
100	110	$+0.654$	—	0.281	0.922	no
200	90	$+0.265$	—	0.184	0.232	no
200	100	$+0.526$	—	0.172	0.432	no
200	110	$+0.071$	—	0.844	0.77	no

Table 6: PCVaR-PG [13] on the Bermudan max-call domain across all 12 configurations, with selected policy learning rate 3e$-$4 from a sweep on $(d, s_0) = (100, 100)$. $\text{CVaR}_{0.2}$ mean $\pm$ std over 10 seeds, per-run wall time, and mean stopping time $\bar{\tau} \in [0, T]$ with $T = 9$.

d	s_0	$\text{CVaR}_{0.2}$	wall (s)	$\bar{\tau}$
20	90	2.741 ± 0.638	29.5	2.4
	100	11.407 ± 0.254	30.1	3.4
	110	10.000 ± 0.000	27.5	0.0
50	90	6.461 ± 0.149	33.6	3.2
	100	18.024 ± 0.050	32.5	3.1
	110	10.000 ± 0.000	26.4	0.0
100	90	8.265 ± 4.137	31.6	1.7
	100	22.267 ± 0.053	32.9	2.1
	110	10.000 ± 0.000	30.9	0.0
200	90	12.401 ± 4.134	31.5	1.3
	100	26.196 ± 0.042	26.5	1.0
	110	10.000 ± 0.000	30.1	0.0

Appendix D.3 compares the DIOS-derived minimum cap with a minimum cap computed from separately trained risk-neutral policies. The choice of an evaluation VaR as the minimum cap is part of our experimental protocol.

D.3 RET-CAP with risk-neutral anchor

We additionally set the RET-CAP minimum cap from the evaluation quantiles of separately trained risk-neutral (RN) policies on the symmetric domain. Table 7 compares the DIOS-derived and risk-

Table 7: RET-CAP with two anchor choices on the Bermudan max-call domain across all 12 configurations: the default cap-anchor (used in our main results) and the risk-neutral anchor variant. $\text{CVaR}_{0.2}$ mean $\pm$ std over 10 seeds; Δ is RN-anchor CVaR minus cap-anchor CVaR; $\bar{\tau}_{\text{RN}}$ is the mean stopping time under the RN-anchor variant ($T = 9$).

d	s_0	cap $\text{CVaR}_{0.2}$	RN $\text{CVaR}_{0.2}$	Δ	$\bar{\tau}_{\text{RN}}$
20	90	18.003 ± 0.290	16.447 ± 0.395	-1.556	3.2
	100	29.677 ± 0.295	27.107 ± 0.377	-2.570	2.8
	110	41.616 ± 0.366	37.921 ± 0.459	-3.696	2.5
50	90	32.810 ± 0.225	30.483 ± 0.378	-2.327	3.9
	100	45.913 ± 0.278	42.976 ± 0.705	-2.936	3.7
	110	59.115 ± 0.379	55.346 ± 0.580	-3.768	3.4
100	90	44.323 ± 0.359	42.148 ± 0.655	-2.175	4.6
	100	58.956 ± 0.390	56.053 ± 0.712	-2.904	4.4
	110	73.084 ± 0.600	69.170 ± 0.910	-3.914	4.1
200	90	56.105 ± 0.423	53.672 ± 0.402	-2.433	5.0
	100	71.396 ± 0.410	68.896 ± 0.620	-2.500	4.9
	110	86.622 ± 0.795	84.240 ± 0.589	-2.381	4.8

Table 8: Risk-neutral DIOS baseline vs DIOS on the symmetric domain. $\text{CVaR}_{0.2}$ at evaluation, mean over 10 seeds; $\Delta = \text{RN} - \text{DIOS}$. RN training drops the soft penalty and trains the same architecture under risk-neutral expectation; the resulting policy is then evaluated against the $\text{CVaR}_{0.2}$ target.

d	s_0	DIOS	RN	Δ
20	90	18.17	9.59	-8.58
	100	29.96	19.92	-10.05
	110	41.97	30.56	-11.41
50	90	33.22	25.79	-7.43
	100	46.80	38.25	-8.56
	110	60.39	50.70	-9.69
100	90	45.00	38.35	-6.65
	100	59.84	52.19	-7.65
	110	74.67	66.03	-8.64
200	90	57.14	50.88	-6.26
	100	73.26	66.12	-7.14
	110	89.38	81.36	-8.02

neutral minimum caps, with hyperparameters selected separately for each choice. The asymmetric counterpart is reported in Appendix E.3.

D.4 DIOS ablations

We compare variants of two design choices in DIOS: the soft-penalty CVaR objective and the β-anneal schedule.

Risk-neutral baseline. We train the same DIOS architecture under the risk-neutral expectation $J_{\text{RN}}(\theta) = \mathbb{E}[\sum_t \psi_t^\theta R_t]$, replacing the soft-penalty objective with expected reward and omitting quantile estimation. We evaluate the risk-neutral policy and DIOS using $\text{CVaR}_{0.2}$ on the same evaluation paths for each seed. Table 8 reports $\text{CVaR}_{0.2}$ for the risk-neutral policy and DIOS, together with the difference, risk-neutral minus DIOS.

β-anneal ablation. We disable both the warmup and anneal phases, setting $\rho_w = \rho_a = 0$ so that $\beta_k = \beta_{\text{tgt}} = 0.2$ from iteration $k = 0$. Table 9 reports the comparison on the symmetric domain. Without the schedule, all ten seeds collapse at $(d, s_0) = (20, 110)$, and two of ten seeds collapse in each configuration with $d \in \{100, 200\}$. The full-schedule and no-schedule runs use hold-phase quantile ranks three and four, respectively.

Table 9: β-anneal ablation on the symmetric domain: DIOS with full schedule ($\rho_w = 0.25$, $\rho_a = 0.50$) vs no schedule ($\rho_w = \rho_a = 0$, $\beta = \beta_{\text{tgt}} = 0.2$ from iteration 0). $\text{CVaR}_{0.2}$ mean over 10 seeds; coll reports the number of collapsed seeds out of ten, with collapse defined by $\bar{\tau} < 0.5$.

d	s_0	DIOS	No schedule	coll
20	90	18.17	17.65	0/10
	100	29.96	29.57	0/10
	110	41.97	10.00	10/10
50	90	33.22	33.01	0/10
	100	46.80	46.51	0/10
	110	60.39	60.02	0/10
100	90	45.00	35.88	2/10
	100	59.84	47.68	2/10
	110	74.67	61.51	2/10
200	90	57.14	45.66	2/10
	100	73.26	58.55	2/10
	110	89.38	73.47	2/10

D.5 CVaR-VaR critic ablation

CVaR-VaR [17] uses a state-dependent quantile critic to track quantile levels and compute multi-step advantages for the VaR-gradient term. Our implementation uses a separate 128-unit MLP trained by multi-step quantile regression at each iteration, adding a critic forward pass and optimizer step to each policy update.

The no-critic variant replaces the critic prediction at every state with the preceding batch's mean stopping reward after normalization. At the first update, the baseline is the mean unnormalized terminal payoff in the batch. The no-critic variant also holds the tradeoff weight ω at 0.5 for the first 40% of training and then decreases ω linearly to zero. The main CVaR-VaR baseline uses $\omega = 0.5$ throughout.

In every configuration in Table 10, the no-critic variant has shorter recorded training time and lower mean CVaR than the main CVaR-VaR baseline.

Table 10: CVaR-VaR critic ablation on the symmetric Bermudan max-call domain. "Full" is our CVaR-VaR implementation with the state-dependent quantile critic; "no-critic" replaces the critic with a global mean baseline, with the changed tradeoff schedule described in the text. $\text{CVaR}_{0.2}$ mean $\pm$ std and wall time in seconds, both over 10 seeds. Δ_{CVaR} = no-critic mean minus full mean. Speedup = wall(full) / wall(no-critic).

d	s_0	full CVaR	wall	no-critic CVaR	wall	Δ_{CVaR}	speedup
20	90	18.068 ± 0.385	54.7	11.675 ± 0.682	29.1	-6.393	$1.88\times$
	100	29.850 ± 0.441	51.0	22.475 ± 0.508	28.4	-7.375	$1.80\times$
	110	41.893 ± 0.587	51.6	33.078 ± 0.561	27.5	-8.815	$1.88\times$
50	90	33.213 ± 0.175	48.8	27.879 ± 0.449	26.7	-5.334	$1.83\times$
	100	46.715 ± 0.187	58.2	40.830 ± 0.863	29.0	-5.885	$2.00\times$
	110	60.241 ± 0.311	59.6	53.875 ± 0.690	27.5	-6.366	$2.17\times$
100	90	44.859 ± 0.195	52.6	40.908 ± 0.698	28.5	-3.951	$1.85\times$
	100	59.805 ± 0.148	49.9	54.537 ± 0.630	27.5	-5.268	$1.81\times$
	110	74.543 ± 0.202	59.5	68.702 ± 0.699	29.7	-5.841	$2.01\times$
200	90	57.110 ± 0.117	59.0	52.955 ± 0.385	29.3	-4.155	$2.01\times$
	100	73.233 ± 0.126	58.6	68.589 ± 0.493	29.7	-4.644	$1.97\times$
	110	89.375 ± 0.099	57.0	84.030 ± 0.460	29.0	-5.345	$1.97\times$

E Experiments on the asymmetric Bermudan max-call domain

E.1 Setting and main results

We further evaluate the methods of §5 on the heterogeneous-volatility variant of the Bermudan max-call, which sets $\sigma_j = 0.10 + j/(2d)$ for $j = 1, \ldots, d$ in place of the shared $\sigma = 0.20$. The remaining domain parameters and the shared training protocol are given in Tables 2 and 3. Per-method best learning rates from a re-run sweep on $(d, s_0) = (100, 100)$, applied uniformly across all twelve configurations, are listed in Table 11. In the symmetric and asymmetric domains, DIOS uses $\rho_w = 0.25$, $\rho_a = 0.50$, and $\eta = 7\text{e}{-5}$, leaving the final 25% of training at the target tail level. DIOS attains the highest mean CVaR and shortest recorded training time in all twelve configurations.

Table 11: Per-method best learning rates on the asymmetric Bermudan max-call domain, from a re-run sweep on the representative configuration $d = 100$, $s_0 = 100$, applied uniformly to all twelve configurations without re-tuning. Method-specific items not shown are inherited from Table 4.

Method	Policy LR	Critic LR mult.	Norm. adv.	Notes
DIOS (ours)	$7\text{e}{-5}$	—	—	same as symmetric (Table 4)
CVaR-VaR	$3\text{e}{-4}$	$2\times$	no	Maze schedule ($\omega \equiv 0.5$), as in Table 4
CVaR-PG	$3\text{e}{-4}$	—	no	selected by our learning-rate sweep
MIX	$7\text{e}{-4}$	—	—	same as symmetric (Table 4)
RET-CAP (cap-anchor)	$1\text{e}{-4}$	$5\times$	yes	DIOS-derived anchor as in §5
RET-CAP (RN-anchor)	$7\text{e}{-4}$	$1\times$	yes	risk-neutral anchor variant of Appendix E.3

Table 12: Asymmetric Bermudan max-call ($\sigma_j = 0.10 + j/(2d)$). $\text{CVaR}_{0.2}$ ↑ on top, wall (s) with collapse rate ↓ on bottom; bold marks the per-row best. The RET-CAP timing caveat in Table 1 also applies here.

d	s_0		DIOS (ours)	CVaR-VaR	CVaR-PG	MIX	RET-CAP
20	90	CVaR ↑	**49.572** ± 0.587	48.761 ± 0.831	43.825 ± 14.656	38.857 ± 19.481	49.028 ± 0.572
		wall (coll) ↓	**12.4** (0/10)	55.8 (0/10)	15.7 (1/10)	21.7 (2/10)	25.6 (0/10)
	100	CVaR ↑	**64.962** ± 0.753	64.535 ± 1.296	62.877 ± 2.676	58.096 ± 19.378	64.277 ± 0.782
		wall (coll) ↓	**11.8** (0/10)	62.1 (0/10)	15.6 (0/10)	21.7 (1/10)	24.8 (0/10)
	110	CVaR ↑	**80.425** ± 0.822	79.568 ± 1.027	58.401 ± 31.707	50.893 ± 33.456	79.806 ± 0.696
		wall (coll) ↓	**12.1** (0/10)	52.2 (0/10)	15.9 (3/10)	23.1 (4/10)	24.3 (0/10)
50	90	CVaR ↑	**95.068** ± 0.532	94.617 ± 0.907	94.485 ± 0.771	94.037 ± 1.703	93.887 ± 1.039
		wall (coll) ↓	**11.7** (0/10)	63.5 (0/10)	15.2 (0/10)	21.7 (0/10)	25.3 (0/10)
	100	CVaR ↑	**115.542** ± 0.587	114.974 ± 1.131	114.553 ± 0.969	103.134 ± 34.408	114.633 ± 0.905
		wall (coll) ↓	**11.8** (0/10)	61.5 (0/10)	15.8 (0/10)	23.0 (1/10)	24.2 (0/10)
	110	CVaR ↑	**135.981** ± 0.629	134.563 ± 1.741	132.866 ± 3.655	122.375 ± 37.468	134.756 ± 0.827
		wall (coll) ↓	**11.8** (0/10)	54.7 (0/10)	14.8 (0/10)	22.3 (1/10)	26.9 (0/10)
100	90	CVaR ↑	**139.929** ± 0.828	139.184 ± 1.232	125.096 ± 41.709	125.357 ± 41.805	139.242 ± 0.935
		wall (coll) ↓	**11.8** (0/10)	53.5 (0/10)	15.9 (1/10)	22.6 (1/10)	26.7 (0/10)
	100	CVaR ↑	**165.289** ± 0.862	165.230 ± 1.085	148.721 ± 49.581	148.515 ± 49.525	164.846 ± 1.192
		wall (coll) ↓	**12.2** (0/10)	59.8 (0/10)	14.2 (1/10)	23.7 (1/10)	26.6 (0/10)
	110	CVaR ↑	**190.629** ± 1.000	189.975 ± 2.070	171.654 ± 53.899	172.614 ± 54.209	189.273 ± 1.732
		wall (coll) ↓	**12.9** (0/10)	55.4 (0/10)	15.2 (1/10)	23.7 (1/10)	24.4 (0/10)
200	90	CVaR ↑	**195.053** ± 1.137	194.788 ± 1.104	171.552 ± 58.019	175.488 ± 58.503	193.957 ± 1.301
		wall (coll) ↓	**13.4** (0/10)	62.3 (0/10)	15.7 (1/10)	22.9 (1/10)	25.9 (0/10)
	100	CVaR ↑	**226.624** ± 1.351	226.099 ± 1.399	198.697 ± 67.013	203.191 ± 67.779	224.970 ± 1.756
		wall (coll) ↓	**13.6** (0/10)	54.2 (0/10)	14.3 (1/10)	24.5 (1/10)	24.9 (0/10)
	110	CVaR ↑	**258.131** ± 1.444	258.060 ± 1.147	228.388 ± 73.747	232.825 ± 74.287	255.473 ± 2.795
		wall (coll) ↓	**12.5** (0/10)	51.0 (0/10)	15.8 (1/10)	23.2 (1/10)	27.1 (0/10)

Table 13: PCVaR-PG [13] on the asymmetric Bermudan max-call domain across all 12 configurations, with selected policy learning rate $3e{-}4$ from a sweep on $(d, s_0) = (100, 100)$. $\text{CVaR}_{0.2}$ mean $\pm$ std over 10 seeds, per-run wall time, and mean stopping time $\bar{\tau} \in [0, T]$ with $T = 9$.

d	s_0	$\text{CVaR}_{0.2}$	wall (s)	$\bar{\tau}$
20	90	7.205 ± 4.774	29.8	1.4
	100	16.519 ± 8.289	33.1	2.4
	110	10.000 ± 0.000	29.8	0.0
50	90	23.860 ± 0.295	34.4	3.1
	100	37.740 ± 1.456	33.4	4.5
	110	10.000 ± 0.000	30.2	0.0
100	90	28.266 ± 14.134	33.6	2.5
	100	50.220 ± 0.219	31.9	2.5
	110	10.000 ± 0.000	28.7	0.0
200	90	42.551 ± 14.185	30.3	0.9
	100	50.752 ± 25.377	28.8	0.9
	110	10.000 ± 0.000	29.1	0.0

Table 14: RET-CAP with two anchor choices on the asymmetric Bermudan max-call domain across all 12 configurations: cap-anchor (default, used in main results, Table 12) and risk-neutral anchor variant. $\text{CVaR}_{0.2}$ mean $\pm$ std over 10 seeds; Δ is RN-anchor CVaR minus cap-anchor CVaR; $\bar{\tau}_{\text{RN}}$ is the mean stopping time under the RN-anchor variant ($T = 9$).

d	s_0	cap $\text{CVaR}_{0.2}$	RN $\text{CVaR}_{0.2}$	Δ	$\bar{\tau}_{\text{RN}}$
20	90	49.028 ± 0.572	43.606 ± 2.596	-5.422	2.7
	100	64.277 ± 0.782	55.169 ± 4.124	-9.109	2.4
	110	79.806 ± 0.696	66.089 ± 3.564	-13.717	2.1
50	90	93.887 ± 1.039	84.134 ± 2.521	-9.753	3.4
	100	114.633 ± 0.905	102.160 ± 2.142	-12.473	3.3
	110	134.756 ± 0.827	119.518 ± 1.263	-15.237	3.1
100	90	139.242 ± 0.935	125.880 ± 2.753	-13.362	3.9
	100	164.846 ± 1.192	151.648 ± 3.372	-13.198	4.0
	110	189.273 ± 1.732	173.946 ± 2.998	-15.328	3.9
200	90	193.957 ± 1.301	179.846 ± 6.094	-14.111	4.6
	100	224.970 ± 1.756	212.469 ± 3.740	-12.500	4.6
	110	255.473 ± 2.795	239.484 ± 5.234	-15.989	4.5

E.2 PCVaR-PG

Table 13 reports PCVaR-PG on the asymmetric domain. At $s_0 = 110$, every seed stops at $t = 0$ and receives reward 10. At $s_0 \in \{90, 100\}$, mean stopping times are at most 4.5, compared with $T = 9$. The standard deviation of $\text{CVaR}_{0.2}$ reaches 14.185 at $(d, s_0) = (200, 90)$.

E.3 RET-CAP with risk-neutral anchor

Table 14 compares RET-CAP with DIOS-derived and risk-neutral minimum caps on the asymmetric domain. Hyperparameters are selected separately for each minimum cap.

E.4 DIOS ablations

We repeat the risk-neutral-objective and annealing comparisons from Appendix D.4 on the asymmetric domain.

Risk-neutral baseline. Table 15 reports mean $\text{CVaR}_{0.2}$ for the risk-neutral policy and DIOS, together with the difference, risk-neutral minus DIOS.

Table 15: Risk-neutral DIOS baseline vs DIOS on the asymmetric domain. $\text{CVaR}_{0.2}$ at evaluation, mean over 10 seeds; $\Delta = \text{RN} - \text{DIOS}$.

d	s_0	DIOS	RN	Δ
20	90	49.57	24.53	-25.05
	100	64.96	36.63	-28.33
	110	80.42	48.86	-31.56
50	90	95.07	66.16	-28.91
	100	115.54	83.08	-32.46
	110	135.98	100.00	-35.98
100	90	139.93	108.71	-31.22
	100	165.29	130.37	-34.92
	110	190.63	152.01	-38.62
200	90	195.05	160.66	-34.40
	100	226.62	188.10	-38.52
	110	258.13	215.50	-42.63

Table 16: β-anneal ablation on the asymmetric domain: DIOS with full schedule ($\rho_w = 0.25$, $\rho_a = 0.50$) vs no schedule ($\rho_w = \rho_a = 0$, $\beta = \beta_{\text{tgt}} = 0.2$ from iteration 0). $\text{CVaR}_{0.2}$ mean over 10 seeds; coll reports the number of collapsed seeds out of ten, with collapse defined by $\bar{\tau} < 0.5$.

d	s_0	DIOS	No schedule	coll
20	90	49.57	49.27	0/10
	100	64.96	64.53	0/10
	110	80.42	10.00	10/10
50	90	95.07	94.20	0/10
	100	115.54	114.58	0/10
	110	135.98	134.88	0/10
100	90	139.93	111.36	2/10
	100	165.29	131.55	2/10
	110	190.63	153.71	2/10
200	90	195.05	154.34	2/10
	100	226.62	179.28	2/10
	110	258.13	206.25	2/10

β-anneal ablation. Table 16 reports the comparison on the asymmetric domain. The no-schedule variant has more collapsed seeds, defined by $\bar{\tau} < 0.5$, than DIOS with the full schedule.

E.5 CVaR-VaR critic ablation

We repeat the CVaR-VaR critic ablation of Appendix D.5 on the asymmetric domain. The no-critic variant uses the batch-mean baseline defined in Appendix D.5 and decreases ω to zero after holding $\omega = 0.5$ for the first 40% of training. Table 17 compares the no-critic variant with CVaR-VaR using a critic and $\omega = 0.5$ throughout.

Table 17: CVaR-VaR critic ablation on the asymmetric Bermudan max-call domain. "Full" is our CVaR-VaR implementation with the state-dependent quantile critic; "no-critic" replaces the critic with a global mean baseline, with the changed tradeoff schedule described in the text. $\text{CVaR}_{0.2}$ mean $\pm$ std and wall time in seconds, both over 10 seeds. Δ_{CVaR} = no-critic mean minus full mean. Speedup = wall(full) / wall(no-critic).

d	s_0	full CVaR	wall	no-critic CVaR	wall	Δ_{CVaR}	speedup
20	90	48.761 ± 0.831	55.8	29.581 ± 0.949	28.2	-19.180	$1.98\times$
	100	64.535 ± 1.296	62.1	43.363 ± 3.072	27.7	-21.172	$2.24\times$
	110	79.568 ± 1.027	52.2	55.406 ± 1.149	26.2	-24.162	$1.99\times$
50	90	94.617 ± 0.907	63.5	71.847 ± 1.509	28.1	-22.770	$2.26\times$
	100	114.974 ± 1.131	61.5	91.802 ± 1.437	29.1	-23.172	$2.11\times$
	110	134.563 ± 1.741	54.7	106.933 ± 0.933	27.2	-27.630	$2.01\times$
100	90	139.184 ± 1.232	53.5	116.141 ± 2.076	28.1	-23.043	$1.91\times$
	100	165.230 ± 1.085	59.8	140.633 ± 3.651	29.4	-24.597	$2.03\times$
	110	189.975 ± 2.070	55.4	162.248 ± 2.673	28.7	-27.727	$1.93\times$
200	90	194.788 ± 1.104	62.3	170.293 ± 2.300	29.0	-24.495	$2.15\times$
	100	226.099 ± 1.399	54.2	197.651 ± 2.369	26.6	-28.448	$2.03\times$
	110	258.060 ± 1.147	51.0	228.504 ± 2.271	29.4	-29.556	$1.73\times$

F Smaller tail levels and quantile smoothing

Table 18 reports the symmetric max-call grid at $\beta = 0.1$ and 0.05. Raw DIOS uses the current batch order statistic as the quantile estimate. For DIOS with an exponential moving average (EMA), denoted ν-EMA, let $\widehat{\nu}_k$ be the batch order statistic and define

$$\nu_0 = \widehat{\nu}_0, \qquad \nu_k = (1-\alpha)\nu_{k-1} + \alpha\widehat{\nu}_k \quad (k \geq 1).$$

The smoothing is applied from the beginning of training, including warmup and annealing. The weight α multiplies the new batch statistic. We select $\alpha = 0.9$ at $\beta = 0.1$ on $(d, s_0) = (200, 100)$ and $\alpha = 0.1$ at $\beta = 0.05$ on $(d, s_0) = (20, 90)$, using seeds 0–2. The final comparison uses ten seeds. DIOS retains learning rate 7×10^{-5}. CVaR-VaR uses learning rate 10^{-4}, value multiplier one, and fixed tradeoff weight $\omega = 0.5$. CVaR-PG uses learning rate 3×10^{-4}.

DIOS with EMA exceeds CVaR-VaR and CVaR-PG in mean CVaR in all 12 configurations at each tail level. Two-sided paired t-tests over ten seeds give $p < 0.05$ against CVaR-VaR in 6/12 configurations at $\beta = 0.1$ and 12/12 at $\beta = 0.05$. The p-values are not adjusted for multiple comparisons. Raw DIOS has a higher mean CVaR than CVaR-VaR and CVaR-PG in 1 of 12 configurations at $\beta = 0.1$ and 8 of 12 configurations at $\beta = 0.05$. The improvement from smoothing is observed with batch size $B = 20$.

Table 18: Lower tail levels, symmetric max-call. CVaR_β mean $\pm$ SD over 10 seeds. DIOS raw: batch order statistic; DIOS ν-EMA: $\nu_k = (1-\alpha)\nu_{k-1} + \alpha\widehat{\nu}_k$. Bold: highest mean.

d	s_0	DIOS (raw)	DIOS (ν-EMA)	CVaR-VaR	CVaR-PG
$\beta = 0.1$ (ν-EMA $\alpha = 0.9$)					
20	90	14.117 ± 0.489	**14.759** ± 0.282	14.548 ± 0.392	3.974 ± 4.598
20	100	25.467 ± 0.576	**26.202** ± 0.359	25.823 ± 0.573	10.850 ± 11.412
20	110	37.086 ± 0.857	**37.854** ± 0.393	37.547 ± 0.571	12.678 ± 8.033
50	90	29.108 ± 0.462	**29.410** ± 0.214	29.266 ± 0.374	13.527 ± 11.258
50	100	42.299 ± 0.492	**42.676** ± 0.206	42.334 ± 0.532	24.711 ± 15.589
50	110	55.428 ± 0.559	**55.943** ± 0.233	55.531 ± 0.456	36.452 ± 16.694
100	90	40.863 ± 0.204	**41.125** ± 0.171	40.922 ± 0.374	14.152 ± 16.300
100	100	55.181 ± 0.229	**55.567** ± 0.212	55.226 ± 0.480	21.000 ± 22.708
100	110	69.559 ± 0.282	**70.014** ± 0.234	69.813 ± 0.306	32.822 ± 24.598
200	90	52.868 ± 0.161	**53.127** ± 0.136	52.813 ± 0.435	27.531 ± 23.773
200	100	68.620 ± 0.124	**68.808** ± 0.132	68.634 ± 0.289	36.662 ± 31.060
200	110	84.236 ± 0.158	**84.506** ± 0.162	84.395 ± 0.225	49.657 ± 33.571
$\beta = 0.05$ (ν-EMA $\alpha = 0.1$)					
20	90	7.472 ± 0.892	**12.123** ± 0.246	9.363 ± 2.778	1.536 ± 2.131
20	100	17.404 ± 1.203	**23.239** ± 0.295	21.396 ± 1.225	5.057 ± 6.297
20	110	27.787 ± 1.374	**34.698** ± 0.302	29.226 ± 2.444	15.150 ± 6.519
50	90	22.412 ± 0.707	**26.456** ± 0.244	24.768 ± 1.108	9.611 ± 7.502
50	100	34.889 ± 1.015	**39.333** ± 0.285	33.346 ± 1.786	19.106 ± 10.825
50	110	47.367 ± 1.224	**52.280** ± 0.316	44.453 ± 2.490	30.110 ± 11.272
100	90	34.293 ± 0.509	**38.086** ± 0.175	31.031 ± 1.267	10.986 ± 12.276
100	100	47.926 ± 0.461	**52.198** ± 0.203	41.660 ± 2.179	17.358 ± 18.225
100	110	61.674 ± 0.515	**66.321** ± 0.217	53.797 ± 1.910	28.809 ± 19.661
200	90	46.731 ± 0.635	**49.903** ± 0.190	37.470 ± 2.399	16.877 ± 17.550
200	100	61.650 ± 0.554	**65.238** ± 0.213	50.787 ± 2.648	23.700 ± 24.174
200	110	76.879 ± 0.710	**80.592** ± 0.239	61.297 ± 2.082	35.638 ± 26.123

G House selling

Offers $X_1, \ldots, X_9$ arrive independently with mean 50. The seller observes one offer at each decision point $j = 1, \ldots, 9$, can accept only the current offer, and must accept by $j = 9$. Accepting at j gives $X_j - cj$, so the first observed offer already incurs cost c. There is no stopping decision before the first offer. Relabeling the offer times as $t = 0, \ldots, 8$ gives the time convention of Section 2. The policy observes the normalized current offer, time, and immediate payoff.

The lognormal grid uses $\log X_j \sim N(\log 50 - \sigma^2/2, \sigma^2)$ with $\sigma \in \{0.3, 0.5, 0.7, 0.9\}$. The Pareto grid uses shape $a \in \{6, 4, 3, 2.5\}$ and scale $50(a-1)/a$. Each family is crossed with $c \in \{1, 2, 3\}$,

giving 24 cells. Finite second moments hold in both families. Negative rewards are possible in the lognormal grid, as discussed in Appendix A.

For every method, we select the policy learning rate from the sweep set in Appendix C. Selection uses $(\sigma, c) = (0.5, 2)$ for lognormal offers and $(a, c) = (3, 2)$ for Pareto offers. Selection and final experiments use seeds 0–9. Other settings follow the max-call protocol. RET-CAP uses the seed-averaged DIOS evaluation VaR, rounded to six decimals, as the minimum cap, and the selected RET-CAP policy learning rates are 10^{-4} for lognormal offers and 7×10^{-5} for Pareto offers. Tables 19 and 20 report the complete results.

RET-CAP has the largest mean only for $(\sigma, c) = (0.9, 2)$, exceeding DIOS by approximately 0.004, with paired t-test $p = 0.99$. An alternative anchor from a policy trained with $\beta_k \equiv 1$ gives RET-CAP 41.155 ± 0.649 at $(\sigma, c) = (0.9, 2)$, compared with DIOS 41.729 ± 0.974. The policy used to compute the house-selling alternative minimum cap is trained at $\beta_k = 1$ with rewards truncated at the batch maximum of sampled stopping rewards. The max-call risk-neutral policies instead maximize expected payoff directly.

Table 19: House selling: lognormal offers (9 offers, $\beta = 0.2$). $\text{CVaR}_{0.2}$ mean $\pm$ SD over 10 seeds (population SD); wall time (s) below. Bold: highest mean. RET-CAP uses the DIOS-derived anchor.

cell		DIOS	CVaR-VaR	CVaR-PG	RET-CAP	MIX
ln0.3_c1	CVaR	$\mathbf{49.877 \pm 0.225}$	49.651 ± 0.437	49.503 ± 0.523	48.653 ± 0.328	49.029 ± 0.756
	wall	12.3	49.3	15.4	23.5	23.3
ln0.3_c2	CVaR	$\mathbf{45.492 \pm 0.306}$	45.148 ± 0.382	45.123 ± 0.464	44.843 ± 0.394	45.067 ± 0.544
	wall	12.4	48.6	15.1	23.2	23.4
ln0.3_c3	CVaR	$\mathbf{41.661 \pm 0.395}$	40.771 ± 0.590	40.829 ± 0.600	41.392 ± 0.249	40.507 ± 0.651
	wall	12.0	55.1	15.2	23.0	23.3
ln0.5_c1	CVaR	$\mathbf{50.647 \pm 0.453}$	50.048 ± 0.943	49.940 ± 0.645	49.959 ± 0.412	49.797 ± 0.685
	wall	12.4	56.0	15.4	23.8	23.6
ln0.5_c2	CVaR	$\mathbf{46.141 \pm 0.253}$	45.446 ± 0.526	45.351 ± 0.611	45.649 ± 0.268	45.396 ± 0.562
	wall	11.9	57.9	15.5	23.6	23.9
ln0.5_c3	CVaR	$\mathbf{41.528 \pm 0.472}$	40.439 ± 0.827	40.616 ± 0.819	41.206 ± 0.502	40.432 ± 0.728
	wall	12.8	57.6	15.6	23.4	24.6
ln0.7_c1	CVaR	$\mathbf{49.790 \pm 0.387}$	48.318 ± 1.920	48.444 ± 1.578	49.220 ± 0.447	49.317 ± 0.620
	wall	12.3	56.8	15.3	23.6	24.2
ln0.7_c2	CVaR	$\mathbf{44.933 \pm 0.307}$	44.103 ± 0.962	43.890 ± 1.153	44.604 ± 0.449	43.768 ± 1.343
	wall	12.8	57.7	15.8	23.6	24.3
ln0.7_c3	CVaR	$\mathbf{40.137 \pm 0.461}$	38.671 ± 1.283	38.799 ± 1.184	40.068 ± 0.600	38.696 ± 1.124
	wall	11.9	53.3	15.0	24.2	23.6
ln0.9_c1	CVaR	$\mathbf{47.129 \pm 0.454}$	45.557 ± 1.703	45.994 ± 2.246	46.764 ± 0.620	46.524 ± 1.272
	wall	12.3	55.4	15.4	23.4	23.9
ln0.9_c2	CVaR	41.729 ± 0.974	40.881 ± 0.974	41.152 ± 1.269	$\mathbf{41.732 \pm 0.580}$	39.683 ± 1.740
	wall	12.5	54.2	15.4	23.8	23.9
ln0.9_c3	CVaR	$\mathbf{37.309 \pm 0.269}$	35.500 ± 2.221	35.377 ± 1.260	37.029 ± 0.498	35.201 ± 1.923
	wall	12.0	53.7	15.2	23.4	23.1

Table 20: House selling: Pareto offers (9 offers, $\beta = 0.2$). $\text{CVaR}_{0.2}$ mean $\pm$ SD over 10 seeds (population SD); wall time (s) below. Bold: highest mean. RET-CAP uses the DIOS-derived anchor.

cell		DIOS	CVaR-VaR	CVaR-PG	RET-CAP	MIX
pa6_c1	CVaR	$\mathbf{45.780 \pm 0.122}$	45.514 ± 0.520	45.322 ± 0.303	43.934 ± 0.205	45.387 ± 0.366
	wall	12.0	58.3	15.3	22.2	23.2
pa6_c2	CVaR	$\mathbf{42.202 \pm 0.140}$	41.958 ± 0.550	41.318 ± 0.608	40.674 ± 0.367	41.772 ± 0.342
	wall	12.4	54.7	15.0	22.8	23.3
pa6_c3	CVaR	$\mathbf{39.699 \pm 0.269}$	39.414 ± 0.016	38.714 ± 0.477	39.418 ± 0.010	39.123 ± 0.625
	wall	12.0	63.8	14.9	21.2	21.4
pa4_c1	CVaR	$\mathbf{45.279 \pm 0.171}$	45.088 ± 0.286	44.738 ± 0.696	43.688 ± 0.300	44.737 ± 0.452
	wall	12.1	57.2	15.2	22.7	22.9
pa4_c2	CVaR	$\mathbf{41.014 \pm 0.189}$	40.959 ± 0.383	40.590 ± 0.394	40.185 ± 0.175	40.454 ± 0.379
	wall	12.1	58.0	15.3	23.0	23.1
pa4_c3	CVaR	$\mathbf{37.748 \pm 0.241}$	37.081 ± 0.699	36.984 ± 0.476	37.405 ± 0.197	37.036 ± 0.704
	wall	12.6	56.1	14.9	22.7	23.1
pa3_c1	CVaR	$\mathbf{44.265 \pm 0.268}$	43.830 ± 0.734	43.909 ± 0.322	42.520 ± 0.443	44.092 ± 0.294
	wall	12.1	56.6	15.5	23.0	23.2
pa3_c2	CVaR	$\mathbf{39.655 \pm 0.200}$	39.423 ± 0.225	39.449 ± 0.232	38.974 ± 0.346	39.254 ± 0.419
	wall	12.5	54.0	15.4	22.8	23.2
pa3_c3	CVaR	$\mathbf{35.861 \pm 0.277}$	34.902 ± 0.826	35.036 ± 0.734	35.482 ± 0.208	35.229 ± 0.517
	wall	12.5	55.5	15.1	22.9	23.1
pa2.5_c1	CVaR	$\mathbf{42.991 \pm 0.206}$	42.556 ± 0.466	42.536 ± 0.686	41.532 ± 0.438	42.863 ± 0.372
	wall	12.3	51.8	15.5	23.4	23.4
pa2.5_c2	CVaR	$\mathbf{38.166 \pm 0.325}$	38.112 ± 0.268	37.812 ± 0.329	37.816 ± 0.407	36.430 ± 2.975
	wall	12.8	51.5	15.9	24.1	24.6
pa2.5_c3	CVaR	$\mathbf{34.007 \pm 0.476}$	33.188 ± 0.928	33.484 ± 0.977	33.831 ± 0.276	33.161 ± 1.193
	wall	12.4	51.4	15.4	24.5	23.6

H Network width

Table 21 compares CVaR, parameter counts, and training times for encoder widths 32, 64, 128, and 256 on symmetric $(d, s_0) = (100, 100)$ at $\beta = 0.2$. Mean CVaR ranges from 59.749 to 59.838. For the dense feedforward policy used here, a forward/backward pass costs $O(B(T+1)P)$ for P parameters. Sorting the sampled rewards costs $O(B \log B)$. DIOS adds no critic network.

Table 21: Network width, symmetric max-call $(d, s_0) = (100, 100)$, $\beta = 0.2$. $\text{CVaR}_{0.2}$ mean $\pm$ SD over 10 seeds.

hidden width	parameters	$\text{CVaR}_{0.2}$	wall (s)
32	5,025	59.785 ± 0.190	13.5
64	13,121	59.812 ± 0.201	13.0
128	38,529	59.838 ± 0.160	12.7
256	126,209	59.749 ± 0.110	13.1

I An optimized certainty equivalent example

For a nondecreasing concave utility u, define the optimized certainty equivalent [5] by $\mathrm{OCE}_u(X) := \sup_\nu\{\nu + \mathbb{E}[u(X-\nu)]\}$. If $u(R_t - \nu)$ is integrable for every t and every fixed ν, exchanging the two suprema as in Proposition 3.1 gives

$$\sup_{\tau\in\mathcal{T}} \mathrm{OCE}_u(R_\tau) = \sup_{\nu\in\mathbb{R}} \left\{ \nu + \sup_{\tau\in\mathcal{T}} \mathbb{E}[u(R_\tau - \nu)] \right\}.$$

For fixed ν, the inner optimization is an expected-reward stopping problem with reward $u(R_t - \nu)$. The outer optimization selects ν. For each fixed ν, the Snell envelope gives an optimizer of the inner problem. If the outer supremum is attained at ν^*, a stopping time that optimizes the inner problem at ν^* also optimizes the OCE stopping problem. Define the truncated quadratic utility by

$$u(x) = \begin{cases} x - x^2/2, & x \le 1, \\ 1/2, & x > 1, \end{cases} \qquad u'(x) = (1-x)^+.$$

For $X \in L^2$, any maximizing ν satisfies $\mathbb{E}[(1 - X + \nu)^+] = 1$. If $X \le 1 + \mathbb{E}[X]$ almost surely, then $\mathrm{OCE}_u(X) = \mathbb{E}[X] - \frac{1}{2}\mathrm{Var}(X)$. The risk measure $-\mathrm{OCE}_u$ is convex but generally noncoherent. For nonnegative R_t, the transformed reward satisfies $u(-\nu) \le u(R_t - \nu) \le 1/2$ at each fixed ν, so $u(R_t - \nu)$ gives bounded weights in the pathwise gradient.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction distinguish structural results, conditional stationarity guarantees, and empirical comparisons.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 discusses limitations. Appendix A checks payoff moments and explains the policy-derivative and convergence assumptions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Sections 3 and 4.3 state the assumptions. Appendix B contains the proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The domains, policy architecture, threshold ranks, optimizer settings, evaluation protocol, and per-method tuning are described in the appendices.

   Guidelines:

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

5. **Open access to data and code**

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

Answer: [No]

Justification: The manuscript does not provide a public code repository. Implementation and evaluation details are given in Appendix C.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Appendix C specifies the training and evaluation protocol. Appendices D–H give settings for the additional experiments.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Tables report ten-seed means and population standard deviations. Appendix C specifies paired tests and the absence of multiplicity correction.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The hardware and per-run times are reported in Appendix C. A complete project-wide compute total is not reported.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The introduction describes potential applications. Section 6 discusses the human and financial costs of inaccurate tail estimates in deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release data or models with a misuse risk.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [No]

    Justification: The baseline papers are cited, but licenses for all implementation components are not documented in this manuscript.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: The paper does not involve human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.