Blackwell optimality in risk-sensitive stochastic control*

Marcin Pitera[1] and Łukasz Stettner[2]

Abstract—In this paper, we consider a discrete-time Markov Decision Process (MDP) on a finite state-action space with a long-run risk-sensitive criterion used as the objective function. We discuss the concept of Blackwell optimality and comment on intricacies which arise when the risk-neutral expectation is replaced by the risk-sensitive entropy. Also, we show the relation between the Blackwell optimality and ultimate stationarity and provide an illustrative example that helps to better understand the structural difference between these two concepts.

I. INTRODUCTION

The *risk-sensitive criterion* in its *averaged* and *discounted* form is a popular discrete-time MDP objective function choice if one wants to incorporate risk into the decision-making process and consider long-run time horizon; see [1], [2], [3] and references therein. Essentially, this criterion is a non-linear extension of the traditional risk-neutral mean criterion, in which the risk sensitivity parameter $\gamma \neq 0$ is encoded in the entropic-utility function $\frac{1}{\gamma} \ln \mathbb{E}\left[e^{\gamma S}\right]$ that is applied to the cumulative reward S. Determining whether to use the averaged approach, where risk is considered per unit of time before eventually applying a limit, or the discounted criterion, which involves discounting future returns to ensure finite outcomes, is not straightforward and often relies on the underlying problem specification.

In this paper, we focus on the *Blackwell optimality*, a specific strategy property that allows transfer from the discounted to the averaged setup, see [4], [5]. On the one hand, this property holds significance in risk-neutral finite state-action framework, particularly within theoretical algorithms relying on the vanishing-discount approach and numerous Reinforced Learning techniques that use approximate discounting mechanisms, see [6], [7], [8], [9], [10] and references therein. On the other hand, it has been only recently shown in [11] that the Blackwell property can be efficiently extended from the risk-neutral setup to the risk-sensitive setup. The transfer is unexpectedly challenging due to the structural difference of the risk-neutral and risk-sensitive setups, even in the finite framework. In this paper, we investigate the sources of this challenge, recall recent results obtained in [11], show how they are connected to the classical theory, and commenKalt how Blackwell optimality is linked to *ultimate stationarity* (also called *eventual-stationarity*), a concept introduced in [12] that links optimal policies in the risk-sensitive setup with optimal policies in the risk-neutral setup. For completeness, we also introduce an example in which the difference between

those two concepts is illustrated. We hope that improved comprehension of Blackwell optimality within risk-sensitive framework may foster advancements in developing efficient risk-sensitive reinforcement learning algorithms, an area under active and intensive exploration, see [13], [14], [15], and references therein.

II. LONG-RUN RISK-SENSITIVE STOCHASTIC CONTROL FOR MDPs

Consider a finite state-action space (E, U) and a time-homogeneous Markov Decision Process (MDP) given via the family of transition probabilities $(\mathbb{P}^a)_{a \in U}$. As usual, we use $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n \in \mathbb{N}})$ to denote the discrete-time filtered canonical probabability space and use Π to denote the set of all policies, that is, sequences of random variables $\pi = (a_n)_{n \in \mathbb{N}}$ such that $a_n \colon \Omega \to U$ is $\mathscr{F}_n$-measurable. Furthermore, for $\pi \in \Pi$ and $x \in E$ we use $\mathbb{P}_x^\pi$ to denote the controlled probability for the canonical process $(X_n)_{n \in \mathbb{N}}$ with $X_0 = x$. We also use Π' and Π'' to denote the sets of all stationary Markov policies and Markov policies, respectively. With a slight abuse of notation we link elements of Π' with functions $u \colon E \to U$ and sometimes write $u \in \Pi'$, having in mind that u constitutes policy $\pi = (a_n)_{n \in \mathbb{N}} \in \Pi$ such that $a_n = u(X_n)$; we also write $\mathbb{P}^u$ instead of $\mathbb{P}^\pi$. Similarly, we associate elements of Π'' with sequences of functions $(u_n)_{n \in \mathbb{N}}$. We refer to [2] and [11] for MDP construction details and formal definitions.

Given a fixed running reward function $c \colon E \times U \to \mathbb{R}$, we consider the *long-run risk-sensitive averaged criterion*, that is, the objective function given by

$$J_\gamma(x, \pi) := \begin{cases} \liminf_{n \to \infty} \frac{1}{n} \frac{1}{\gamma} \ln \mathbb{E}_x^\pi \left[e^{\gamma \sum_{i=0}^{n-1} c(X_i, a_i)}\right], & \text{if } \gamma \neq 0, \\ \liminf_{n \to \infty} \frac{1}{n} \mathbb{E}_x^\pi \left[\sum_{i=0}^{n-1} c(X_i, a_i)\right], & \text{if } \gamma = 0, \end{cases} \tag{1}$$

where $\mathbb{E}_x^\pi$ denotes the expectation for $\mathbb{P}_x^\pi$, $x \in E$, $\pi \in \Pi$, and $\gamma \in \mathbb{R}$ is the risk-aversion parameter; for $\gamma = 0$ we recover the classical risk-neutral criterion, while for $\gamma \neq 0$, we are in the risk-sensitive setup (see [1]). Our goal is to find optimal policy π^* for which we have $J_\gamma(x, \pi^*) = \sup_{\pi \in \Pi} J_\gamma(x, \pi)$.

For simplicity, in this paper we decided to impose a *strong uniform mixing* assumption on the underlying MDP, that is, assume condition

$$\inf_{a \in U} \inf_{x, y \in E} \mathbb{P}^a(x, y) > 0. \tag{C}$$

Assumption (C) implies the existence of a unique full-support invariant measure for any stationary Markov control and also ensures one-step transition equivalence for all controlled transition kernels. Although the results presented in this paper are true under much weaker assumptions linked to classical *mixing* and *multi-step communication*, we decided to present

[1]Marcin Pitera is with Institute of Mathematics, Jagiellonian University, Krakow, Poland `marcin.pitera@uj.edu.pl`

[2]Łukasz Stettner is with Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland `l.stettner@impan.pl`

them in the simplified setup; see [11] details. Notably, under (C), the optimal value of the objective function is independent of the starting point; see [11] for details.

The value of the objective criterion defined in (1) as well as the respective optimal strategies are often approximated using the discounted setup and so called vanishing discount approach, see [16]. For a given discount factor $\beta \in (0,1)$, the discounted analog of criterion (1) is given by

$$J_\gamma(x,\pi;\beta) := \begin{cases} \frac{1}{\gamma} \ln \mathbb{E}_x^\pi \left[e^{\gamma \sum_{i=0}^{n-1} \beta^i c(X_i,a_i)} \right], & \text{if } \gamma \neq 0, \\ \mathbb{E}_x^\pi \left[\sum_{i=0}^{n-1} \beta^i c(X_i,a_i) \right], & \text{if } \gamma = 0. \end{cases} \quad (2)$$

For consistency, we also set $J_\gamma(x,\pi;1) = J_\gamma(x,\pi)$. In contrast to (1), the optimal strategy for (2) with $\gamma \neq 0$ does not need to be Markov stationary even under assumption (C). In this paper, we study how the optimal strategies induced by (2) are linked to the optimal strategies induced by (1) and differentiate the concept of *Blackwell optimality* and *ultimate stationarity* that were introduced in [4] and [17], respectively.

For brevity, we say that a policy is *average optimal* (or optimal for averaged problem), if it is optimal for (1) and *β-discount optimal* (or optimal for β-discount problem), if it is optimal for (2).

III. BLACKWELL OPTIMALITY

In a nutshell, a policy π is said to be *Blackwell optimal*, if there exists $\beta_0 < 1$ such that π is β-discount optimal for any $\beta \in (\beta_0, 1)$; we also say that the policy π satisfies the *Blackwell property*, if it is Blackwell optimal. In this section we recall the result from [4] in which existence of Blackwell optimal policies was shown, and later show how to extend this concept to risk-sensitive setup based on the recent results presented in [11].

A. Risk-neutral setup

In this subsection we fix $\gamma = 0$ and consider risk-neutral setup. First, we recall the classical Blackwell property theorem which ensures existence of Blackwell optimal strategy in the finite setup

Theorem III.1. *Assume* (C) *and fix* $\gamma = 0$. *Then, there exists a stationary Markov policy* $\pi \in \Pi'$ *that satisfies the Blackwell property. Furthermore, policy* π *is optimal for the risk-neutral averaged problem.*

The first proof of Theorem III.1 was presented in [4], see also Section 10.1.2 in [5] for an extensive discussion about Blackwell optimality and related proof concepts. The traditional approach relies on Laurent series expansions along with Cramer's rule to demonstrate that every element of the inverse matrix $(I - \beta \mathbb{P}^u)^{-1}$ is a rational function of β, where $\mathbb{P}^u$ represents the transition matrix under decision rule u. This approach is then coupled with the insight that the optimal policy for any given β can be identified among a fixed set of stationary policies, thereby reducing the problem to a finite set of policies. This also shows that the interval $[0,1)$ for the discount factor can be partitioned into a finite number of

segments, each associated with a decision rule that specifies the optimal stationary policy for those discount factors.

The concept of Blackwell optimality is now considered as a canonical result in the risk-neutral criterion MDP literature, and was extensively studied in numerous papers. We refer to [18], [19], [20], [5] for details on early developments in this area and to [9], [21] for a more recent contributions, which include comprehensive discussion of the Blackwell property as well as its various modifications. Also, as noted in [10], the concept of Blackwell optimality is an important topic in Reinforced Learning for average reward optimality, as it allows a transition from discounter to average setup. In fact, better understanding of the Blackwell optimality criterion is stated as *one of the pressing questions* in the list of open research problems formulated in [8].

B. Risk-sensitive setup

Unfortunately, it is not possible to directly transfer Theorem III.1 and the previously discussed proof methods to the risk-sensitive setup. In contrast to the optimal discounted policies for risk-neutral setup, the risk-sensitive discounted optimal policies usually lack stationarity, resulting in a non-linear challenge that inhibits direct use of techniques based on Laurent series expansion. However, a modified Blackwell property can still be retrieved within the non-stationary framework through proofs relying on a blend of vanishing discount and span-contraction methods.

Consider a constant risk-sensitivity parameter $\gamma \neq 0$ with the optimal β-discounted non-stationary Markov policy represented as $(u_0^\beta, u_1^\beta, u_2^\beta, \ldots)$. Then, for each $n \in \mathbb{N}$, there exists $\beta_n < 1$ such that for any β within $(\beta_n, 1)$ the stationary Markov policy $(u_n^\beta, u_n^\beta, \ldots)$ is optimal for the risk-sensitive average problem.

To make this statement more formal, let us consider the optimal policy induced by the discounted Bellman equation for problem (2). Namely, for any $\beta \in (0,1)$ and $\gamma \neq 0$, let $\hat{\pi}^\beta := (\hat{u}_0^\beta, \hat{u}_1^\beta, \ldots)$, constitute the optimal β-discount policy given by

$$\hat{u}_n^\beta(x) := \arg\max_{a \in U} \left[c(x,a) + \frac{1}{\gamma_n} \ln \left(\sum_{y \in E} e^{w^\beta(y,\gamma_{n+1})} \mathbb{P}^a(x,y) \right) \right],$$
$$(3)$$

where $\gamma_n := \gamma \beta^n$ and $w^\beta(\cdot, \cdot)$ is the solution to the discounted Bellman equation

$$w^\beta(x,\gamma) = \max_{a \in U} \left[c(x,a) + \frac{1}{\gamma} \ln \sum_{y \in E} e^{w^\beta(y,\beta\gamma)} \mathbb{P}^a(x,y) \right]. \quad (4)$$

The policy $\hat{\pi}^\beta$ might be non-stationary as we might have $w(\cdot, \gamma_{n+1}) \not\equiv w(\cdot, \gamma_{m+1})$ and consequently $\hat{u}_n^\beta \not\equiv \hat{u}_m^\beta$, for $n \neq m$; note that the Bellman equation corresponding to (4) is in fact recursive and requires us to compute w^β on parameter grid $(\gamma_n)_{n \in \mathbb{N}}$. Also, one can show that $w^\beta(x,\gamma) = \sup_{\pi \in \Pi} J_\gamma(x,\pi;\beta)$. We are now ready to present the analogue of Theorem III.1 transferred to the risk-sensitive framework.

Theorem III.2. *Assume* (C) *and fix* $\gamma \neq 0$. *Then, for any* $n \in \mathbb{N}$ *there is* $\beta_n \in (0,1)$ *such that for any* $\beta \in (\beta_n, 1)$ *the stationary Markov policy* $\hat{u}_n^\beta \in \Pi'$ *defined in* (3) *is optimal for the risk-sensitive averaged problem.*

The proof of Theorem III.2 can be found in [11]. The proof techniques used therein are essentially different from those used to prove Theorem III.1 and rely on a combination of span-contraction methods with vanishing discount approach. One of the key results (also presented and proved in [11]) which contributes to the proof of Theorem III.2 is the vanishing discount approximation result, which states that for appropriately chosen sequences we can recover the solution to the averaged Bellman equation from the discounted value functions. Namely, let us fix $z \in E$ and consider centered function $\bar{w}^\beta(x, \gamma) := w^\beta(x, \gamma) - w^\beta(z, \gamma)$. Furthermore, for $n \in \mathbb{N}$, let

$$\lambda_n^\beta(\gamma) := w^\beta(z, \gamma\beta^n) - w^\beta(z, \gamma\beta^{n+1}),$$
$$\bar{w}_n^\beta(x, \gamma) := \bar{w}^\beta(x, \gamma\beta^n).$$

Then, we can formulate the following theorem.

Theorem III.3. *Assume* (C); *fix* $\gamma \neq 0$. *Then, for* $x \in E$ *and* $\gamma \in \mathbb{R} \setminus \{0\}$, *there exists limits*

$$w(x, \gamma) := \lim_{\beta \uparrow 1} \bar{w}_n^\beta(x, \gamma) \text{ and } \lambda(\gamma) := \lim_{\beta \uparrow 1} \lambda_n^\beta(\gamma).$$

Furthermore, the function $w(\cdot, \gamma)$ *and the constant* $\lambda(\gamma)$ *constitute the optimal stationary policy to the averaged risk-sensitive problem, that is, Markov stationary policy*

$$u(x) := \arg\max_{a \in U}\left[c(x,a) - \lambda(\gamma) + \frac{1}{\gamma} \ln \sum_{y \in E} e^{\gamma w(y, \gamma)} \mathbb{P}^a(x, y) \right],$$

is optimal for the averaged risk-sensitive criterion (1).

Theorem III.3 proof follows directly from the proof of Theorem 4.4 presented in [11]. The proof is based on recursive selection schemes and use finite space induced uniform bounds for value functions and optimal constants. We refer to Proposition 4.2, Proposition 4.3, and proof of Theorem 4.4 in [11] for more details.

As in the risk-neutral setup, the Bellman property could be used to provide a link between optimal discounted and optimal averaged policy for the risk-sensitive criterion. This property is important in the context of theoretical algorithms based on the vanishing-discount approach as well as many Machine Learning methods based on approximating discounting schemes, see [22], [6], [7], [8] and references therein. Consequently, we believe that better understanding of Blackwell optimality in the risk-sensitive setup could help to boost development of efficient risk-sensitive reinforced learning algorithms which is an area of on-going and intensive research, see [13], [14], [15] and references therein.

IV. STRUCTURAL DIFFERENCE BETWEEN RISK-NEUTRAL AND RISK-SENSITIVE SETUP

In this section, we explain in details the structural difference between Blackwell optimality in the risk-neutral and risk-sensitive setup. They key difference is linked to the fact

that while optimal discount policies in the risk-neutral setup induced by Bellman equation are inherently stationary, this is not the case for the risk-sensitive setup. Consequently, it is structurally impossible to directly transfer Theorem III.1 to the risk-sensitive framework, and one should adjust for non-stationary policies, as done in Theorem III.2.

To better understand this, let us recall that the value function in the risk-sensitive discounted Bellman equation given in (4), that is, the value of $w^\beta(x, \gamma)$ in equation

$$w^\beta(x, \gamma) = \max_{a \in U}\left[c(x,a) + \frac{1}{\gamma} \ln \sum_{y \in E} e^{w^\beta(y, \beta\gamma)} \mathbb{P}^a(x, y) \right] \quad (5)$$

might strongly depend on the underlying risk sensitivity parameter γ and consequently we might get $\hat{u}_n^\beta \not\equiv \hat{u}_m^\beta$, for $n \neq m$, where $\hat{u}_n^\beta$ is defined in (3). On the other hand, when $\gamma = 0$, the Bellman equation (5) simplifies to equation

$$w^\beta(x, 0) = \max_{a \in U}\left[c(x,a) + \beta \sum_{y \in E} w^\beta(y, 0) \mathbb{P}^a(x, y) \right],$$

so that we only need to consider the value $w^\beta(\cdot, 0)$ and the respective stationary Markov policy $\hat{u}^\beta \in \Pi'$ given by

$$\hat{u}^\beta(x) := \arg\max_{a \in U}\left[c(x,a) + \beta \left(\sum_{y \in E} w^\beta(y, 0) \mathbb{P}^a(x, y) \right) \right]. \quad (6)$$

Essentially, this is due to the fact that the expectation operator embedded in the sum in (5) is positively homogeneous, while the entropy operator encoded in the logarithm of the expectation of the exponent in 5 is a proper convex (or concave) function which cannot be linearly scaled. Furthermore, it should be noted that while the set of stationary Markov policies Π' is finite, the set of Markov policies Π'' is not, which inhibits the usage of methods described in Section III-A. This is essentially why the equivalent of Theorem III.1 cannot be true in the risk-sensitive setup and why the related proof techniques do not work in this setup. Namely, we cannot restrict our analysis to the finite number of stationary Markov policies in the risk-sensitive cane. On the other hand, this raises a natural question under which additional conditions there exists a stationary solution to the risk-sensitive discounted problem for any discount factor. In particular, one expects this to be the case for values of γ which are sufficiently close to zero. In this case, the entropy operator becomes almost linear (as, in the limit, it converges to the expectation operator). This simple idea leads to the concept of *ultimate stationarity* which is formulated in the next subsection.

V. INTERACTION BETWEEN BLACKWELL OPTIMALITY AND ULTIMATE STATIONARITY IN THE RISK-SENSITIVE SETUP

The concept of *ultimate stationarity* was initially developed in [12], where a specific form of risk-neutral optimal strategy, also optimal in the risk-aversion neighborhood of zero, was studied. We say that $\pi \in \Pi''$ is an *ultimately stationary policy*, if it can be represented as

$$\pi = (u_0, u_1, \ldots, u_{N-1}, u, u, \ldots), \quad (7)$$

for some $N \in \mathbb{N}$ and one-step policies $u_1, \ldots, u_{N-1}, u \in \Pi'$. In other words, the ultimate stationary policy π becomes a stationary Markov policy after a fixed number of steps. With a slight abuse of notation, we also refer to policy u in (7) as the ultimate stationary policy. The first (essential) step in showing the existence of an ultimate stationarity policy is formulated in the following theorem.

Theorem V.1. *Assume* (C) *and fix* $\beta \in (0,1)$. *Then, there is* $\gamma_0 < 0$ *(resp.* $\gamma_0 > 0$*) and* $u \in \Pi'$ *such that* u *is optimal for the discounted problem* (2) *for any* $\gamma \in [\gamma_0, 0]$ *(resp.* $\gamma \in [0, \gamma_0]$*).*

The proof of Theorem V.1 can be found in [17]. Knowing this result, one can deduce that, for fixed $\beta \in (0,1)$ and $\gamma \in \mathbb{R} \setminus \{0\}$, the nth one-step policy in (3) could be seen as the 1st one-step policy for the discounted problem with initial risk-aversion equal to $\beta^n \gamma$, which is closer to zero than the initial parameter γ. Note that ultimately stationary policy u in representation (7) corresponds to policy from Theorem V.1. This observation can be used to prove the existence of an ultimate stationary policy for any value of $\beta \in (0,1)$; of course, one expects that the closer the value of β to 1 is, the more steps one needs to make to ensure the stationarity property in the optimal policy. For details, we refer to [12] where this result is formulated and proved.

Upon initial inspection, one would hope that we can use Theorem V.1 to somehow recover the statement of Theorem III.1 or Theorem III.2. Unfortunately, this is not the case, as those theorems are of different nature. Namely

- *The Blackwell policies* are linked to the situation in which risk aversion γ is fixed and we let $\beta \to 1$;
- *The ultimate stationary policies* are linked to the situation in which β is fixed and we let $\gamma \to 0$.

Those properties are structurally different and there is no direct connection between them. In particular, apart for some special situations, in which the initial risk-aversion parameter is very close to zero, the sets of Blackwell policies and ultimate stationary could be disjoint. This is presented in the illustrative example in Section VI.

Remark 1. *The existence of ultimate stationary policies (and Blackwell policies) is inherently linked to the finite space structure, in both risk-neutral and risk-sensitive cases. In fact, in the denumerable case, one can show direct existence counterexamples if no additional ergodic-mixing assumptions are imposed. We refer to [12] and [23] for details.*

VI. Illustrative example

In this section, we present an extended and refined analysis of the example introduced in [11] that was based on [12]. We use it to illustrate the structural difference between risk-neutral and risk-sensitive setups, and to visualise different nature of the Blackwell optimality and the ultimate stationarity property.

Let $(E, U) := (\{1, 2, 3\}, \{0, 1\})$ with transition matrices given by

$$P^0 = \begin{bmatrix} 2\varepsilon & 0.5 - \varepsilon & 0.5 - \varepsilon \\ 1 - 2\varepsilon & \varepsilon & \varepsilon \\ 1 - 2\varepsilon & \varepsilon & \varepsilon \end{bmatrix},$$

$$P^1 = \begin{bmatrix} 2\varepsilon & 0.9 - \varepsilon & 0.1 - \varepsilon \\ 1 - 2\varepsilon & \varepsilon & \varepsilon \\ 1 - 2\varepsilon & \varepsilon & \varepsilon \end{bmatrix},$$

for $\varepsilon \in (0, 0.1)$, and cost function $c : E \times U \to \mathbb{R}$ defined as

$$c(x, 0) = \begin{cases} 0 & \text{if } x = 1 \\ 0 & \text{if } x = 2 \\ 8 & \text{if } x = 3 \end{cases} \quad \text{and} \quad c(x, 1) = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{if } x = 2 \\ 8 & \text{if } x = 3 \end{cases}.$$

In this example, we have eight Markov stationary policies. Noting that $\mathbb{P}^1(i, \cdot) \equiv \mathbb{P}^2(i, \cdot)$ and $c(i, \cdot) \equiv \text{const}$ for $i \in \{2, 3\}$, we conclude that it does not matter which action is assigned to state 2 and 3. Consequently, without loss of generality, it is sufficient to consider Markov policies $\pi = (a_n)_{n \in \mathbb{N}}$, where $a_n = u_0(X_n)$ or $a_n = u_1(X_n)$ for every $n \in \mathbb{N}$, where $u_0(\cdot) \equiv 0$ and $u_1(\cdot) \equiv 1$. Furthermore, for ε close to zero, following similar reasoning to the one introduced in [11], one can conclude that for any $\beta \in (0, 1)$ and $\gamma \in \mathbb{R}$ the optimal Markov policy $\hat{\pi}(\beta, \gamma) \in \Pi''$ must be of the form

$$\pi(\beta, \gamma) = (u_0, \ldots, u_0, u_1, u_1, \ldots),$$

that is, in the fixed number of initial steps, say $n(\beta, \gamma)$, we use policy u_0, and then switch to policy u_1. As shown in [11], the policy u_0 is typically Blackwell optimal and u_1 is ultimately stationary. To illustrate this, we have computed the switch value points, i.e. values of $n(\beta, \gamma)$ on the $(0.5, 1) \times [-5, 5]$ dyadic grid of step-size $\delta := 0.01$. For simplicity we performed the calculation for the limit case $\varepsilon = 0$ as it allows decomposing $(X_n)_{n \in \mathbb{N}}$ into a series of independent (two-step) innovations which in turn allows decomposing entropic utility of the sum into entropic utilities of individual components. For brevity, we skip the technical calculation and focus on the output results. The obtained structure of optimal policies, represented by the value of $n(\beta, \gamma)$, is presented via a heat map in Figure 1.

The plot nicely illustrates the structural difference between *ultimate stationarity* and *Blackwell property*:

- The *ultimate stationarity property* is represented by horizontal lines in Figure 1. Namely, for any fixed value of β, we ultimately switch to u_1 after $n(\beta, \gamma)$ steps. Furthermore, the closer we are to the risk-neutral case, the faster we switch; as expected, the policy u_1 becomes optimal stationary for values of γ sufficiently close to zero.
- The *Blackwell property* is represented by vertical lines in Figure 1. Namely, for any fixed value of γ the number of steps $n(\beta, \gamma)$ goes either to 0 or ∞ when $(1 - \beta) \to 0$. The limit is determined by the optimal policy for the risk-sensitive averaged case: if u_1 is optimal for a given γ, then
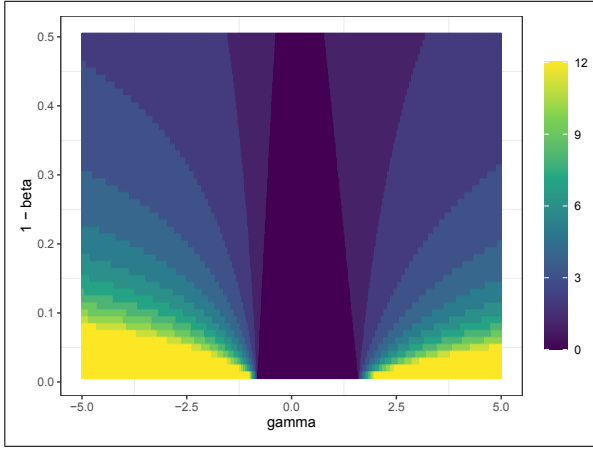
Fig. 1. The plot presents the optimal policy switching point $n(\beta, \gamma)$ from Markov policy u_0 to Markov policy u_1 for different values of $\gamma \in [-5, 5]$ and $\beta \in (0.5, 1)$.

we get $n(\beta, \gamma) \to 0$ as $\beta \to 1$, and if u_0 is optimal, then we get $n(\beta, \gamma) \to \infty$ as $\beta \to 1$.

We note that the Blackwell optimal policy might be different than the ultimate stationary policy. In our example, this is the case for values of γ sufficiently far away from zero. To better understand why this is the case, we can plot the difference between averaged risk-sensitive objective functions for u_1 and u_2, for different values of $\gamma \in \mathbb{R}$. The plot of the function $g(\gamma) := [J_\gamma(1, u_1; 1) - J_\gamma(1, u_0; 1)]$ is presented in Figure 2.
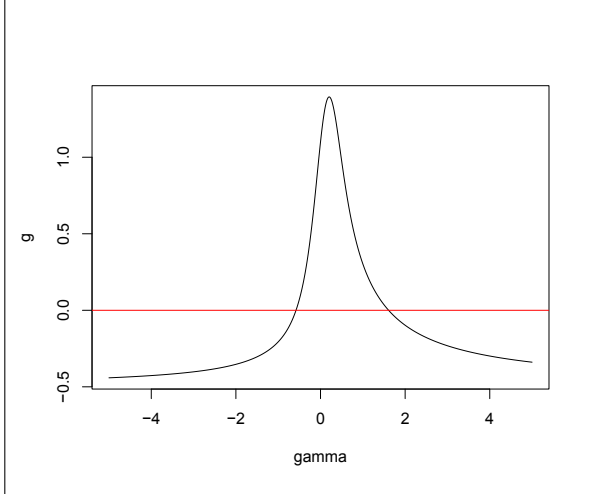

Fig. 2. The plot presents the difference between $J_\gamma(1, u_1; 1)$ and $J_\gamma(1, u_0; 1)$ for different values of $\gamma \in [-5, 5]$. Positive difference indicates optimality u_1, while negative difference indicates optimality of u_0.

In a nutshell, Blackwell optimal policy recovers the risk averaged optimal policy which depends on the value of $\gamma \in \mathbb{R}$. As expected, the switching points in Figure 1 are matching the zeroes of function g presented in Figure 2. This illustrates that while ultimate stationary policy is essentially recovering optimal behavior under almost risk-neutral risk regime, the Blackwell policy is recovering optimal behavior for the risk-sensitive case, when the discount is vanishing.

REFERENCES

[1] P. Whittle, *Risk-sensitive optimal control*. Wiley New York, 1990.
[2] N. Bäuerle and U. Rieder, *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011.
[3] N. Bäuerle and A. Jaśkiewicz, "Markov decision processes with risk-sensitive criteria: an overview," *Mathematical Methods of Operations Research*, vol. 99, no. 1, pp. 141–178, 2024.
[4] D. Blackwell, "Discrete dynamic programming," *The Annals of Mathematical Statistics*, pp. 719–726, 1962.
[5] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2009, vol. 414.
[6] G. Portillo-Ramírez, R. Cavazos-Cadena, and H. Cruz-Suárez, "Contractive approximations in average Markov decision chains driven by a risk-seeking controller," *Mathematical Methods of Operations Research*, vol. 98, no. 1, pp. 75–91, 2023.
[7] A. Biswas and V. S. Borkar, "Ergodic risk-sensitive control—a survey," *Annual Reviews in Control*, vol. 55, pp. 118–141, 2023.
[8] V. Dewanto, G. Dunn, A. Eshragh, M. Gallagher, and F. Roosta, "Average-reward model-free reinforcement learning: a systematic review and literature mapping," *arXiv preprint arXiv:2010.08920*, 2020.
[9] V. Dewanto and M. Gallagher, "Examining average and discounted reward optimality criteria in reinforcement learning," in *AI 2022: Advances in Artificial Intelligence*, H. Aziz, D. Corrêa, and T. French, Eds. Springer International Publishing, 2022, pp. 800–813.
[10] J. Grand-Clément and M. Petrik, "Reducing Blackwell and Average Optimality to Discounted MDPs via the Blackwell Discount Factor," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 52 628–52 647.
[11] N. Bäuerle, M. Pitera, and Ł. Stettner, "Blackwell optimality and policy stability for long-run risk-sensitive stochastic control," *SIAM Journal on Control and Optimization*, vol. 62, no. 6, pp. 3172–3194, 2024.
[12] S. C. Jaquette, "A utility criterion for Markov decision processes," *Management Science*, vol. 23, no. 1, pp. 43–49, 1976.
[13] Y. Fei, Z. Yang, and Z. Wang, "Risk-sensitive reinforcement learning with function approximation: A debiasing approach," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3198–3207.
[14] Y. Ding, M. Jin, and J. Lavaei, "Non-stationary risk-sensitive reinforcement learning: Near-optimal dynamic regret, adaptive detection, and separation design," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 7405–7413.
[15] E. Noorani and J. S. Baras, "Risk-sensitive reinforce: A Monte Carlo policy gradient algorithm for exponential performance criteria," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 1522–1527.
[16] R. Cavazos-Cadena and D. Hernández-Hernández, "Vanishing discount approximations in controlled Markov chains with risk-sensitive average criterion," *Advances in Applied Probability*, vol. 50, no. 1, pp. 204–230, 2017.
[17] S. C. Jaquette, "Markov decision processes with a new optimality criterion: Discrete time," *The Annals of Statistics*, vol. 1, no. 3, pp. 496–505, 1973.
[18] R. Dekker and A. Hordijk, "Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards," *Mathematics of Operations Research*, vol. 13, no. 3, pp. 395–420, 1988.
[19] A. Hordijk and A. A. Yushkevich, "Blackwell optimality," in *Handbook of Markov Decision Processes: Methods and Applications*. Springer, 2002, pp. 231–267.
[20] R. Cavazos-Cadena and J. B. Lasserre, "A direct approach to Blackwell optimality," *Morfismos*, vol. 3, no. 1, pp. 9–33, 1999.
[21] J. Grand-Clément, M. Petrik, and N. Vieille, "Beyond discounted returns: Robust Markov decision processes with average and Blackwell optimality," 2025. [Online]. Available: https://arxiv.org/abs/2312.03618
[22] R. Cavazos-Cadena and D. Hernández-Hernández, "Discounted approximations for risk-sensitive average criteria in Markov decision chains with finite state space," *Mathematics of Operations Research*, vol. 36, no. 1, pp. 133–146, 2011.
[23] A. Brau-Rojas, R. Cavazos-Cadena, and E. Fernández-Gaucherand, "Controlled Markov chains with risk-sensitive criteria: some (counter) examples," in *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No. 98CH36171)*, vol. 2. IEEE, 1998, pp. 1853–1858.