On the Gambler's Problem

April 2, 2019

1 The Gambler's Problem with Continuous State Space

The Gambler's problem is introduced in Sutton and Barto's book [SB18], Chapter 4, to model a gambler's decision. Assume that the gambler visits a slot machine with a initial amount of money. At each round, the gambler in possession of s amount of money has the option to bet any money between 0 and s. If the gambler wins that the round, the bet is doubled up and returned to the gambler; if they loses the bet, they no longer possesses the bet. Hence, when the gambler bets a amount of money, they will hold either $s-a$ or $s+a$ at the end of the round. Naturally, we assume that the win probability p of each round is a constant less than 0.5, and the outcomes of the bets are independent between each of the rounds. The gambler's objective is to win a certain amount of money before leaving the casino. Otherwise if they loses all their money, the gambler fails to achieve the goal. The gambler wants a strategy to decide the amount to bet at each round, to maximize the probability to achieve the goal. We also desire a function to value the utility of the gambler's money, which is similar to the Independent Chip Model [HR07].

We formulate the Gambler's problem described above as an MDP. Without loss of generality let the goal of the gambler be 1 and hence the gambler bets no more than $1-s$ money. Define $\mathcal{S}=[0,1]$ and $\mathcal{A}(s)=[0,\min(s,1-s)]$ to be the state space and the action space, respectively. At each step s represents the amount of money the gambler currently possesses, and the agent takes $a\in\mathcal{A}(s)$ which denotes the amount of bet. The consecutive state s' stochastically transits to $s-a$ and $s+a$ with probability $p>0.5$ and $1-p$, respectively. The process terminates if $s\in\{0,1\}$ and the agent receives an episodic reward $r=s$ at the terminate state. We discuss a general setting under the discount factor $0\leq\gamma\leq 1$. By definition we have the boundary conditions $v(0)=0$ and $v(1)=1$ which correspond to the Bellman property on the terminal states. An additional boundary condition is $v(s)<1$, which is by the fact that $v(s)$ represents a probability. We solve the optimal value function $v(s)$ and discuss its property.

Theorem 1. *$v(s)=\sum_{i=1}^{\infty}(1-p)\gamma^i b_i\prod_{j=1}^{i-1}((1-p)+(2p-1)b_j)$ is the optimal state-value function for any $0\leq\gamma\leq 1$ and $p>0.5$, where $s=0.b_1b_2\ldots b_l\ldots_{(2)}$ is the binary representation of the state $0\leq s<1$.*

It is obvious that the series converge for any $0\leq s<1$. The proof of $v(s)$ being the optimal state-value function is two-fold. We first verify that the Bellman property $v(s)=\max_{0<a\leq\min(s,1-s)}(1-p)\gamma v(s+a)+p\gamma v(s-a)$ for $0<s<1$ is satisfied by the value function $v(s)$ in the theorem. We then show that the solution of the Bellman equation subject to the boundary conditions is unique. Let $g_l=\{k2^{-l}|k\in\{1,\ldots,2^l-1\}\}$ such that g_l is the set of numbers that can be represented by l binary bits. It is easy to verify that $v(s)=\max_{a\in g_1}(1-p)\gamma v(s+a)+p\gamma v(s-a)$ for any $s\in g_1\cup g_2$. Assume that $v(s)=\max_{a\in g_l}(1-p)\gamma v(s+a)+p\gamma v(s-a)$ is satisfied for any $s\in g_l$, we show that property holds for $s\in g_{l+1}$, thus complete the induction. Note that without ambiguity we reuse the notation $\{b_i\}$ and $\{c_i\}$ as the binary representation of numbers.

Though the description of the Gambler's problem seems natural and straightforward, Theorem 1 shows that its simpleness is deceptive. The optimal value function presents its self-similar, fractal and non-rectifiable form, which cannot be described by any simple analytic formula. At any level of zooming-in, the value function keeps showing the same texture as itself. The "spur"s on the curve happen at 1/2, 1/4, 3/4, 1/8, and so forth any point on the dyadic rationals, shown in Figure 1. With the fractal nature, the value function does not possess many of the desired properties for analysis. Namely, the function is not continuous; not differentiable or weak-differentiable. Any point on the dyadic rational has a left-derivative of zero and a right derivative of infinity; no local linear or Taylor expansion; cannot be approximated by a neural network to arbitrary precision, albeit the universal approximation theorem [Csá01]. It is observed that those properties are not expected by the recent line of reinforcement learning studies, who commonly use a neural network approximation and minimizes the Bellman error. Instead,

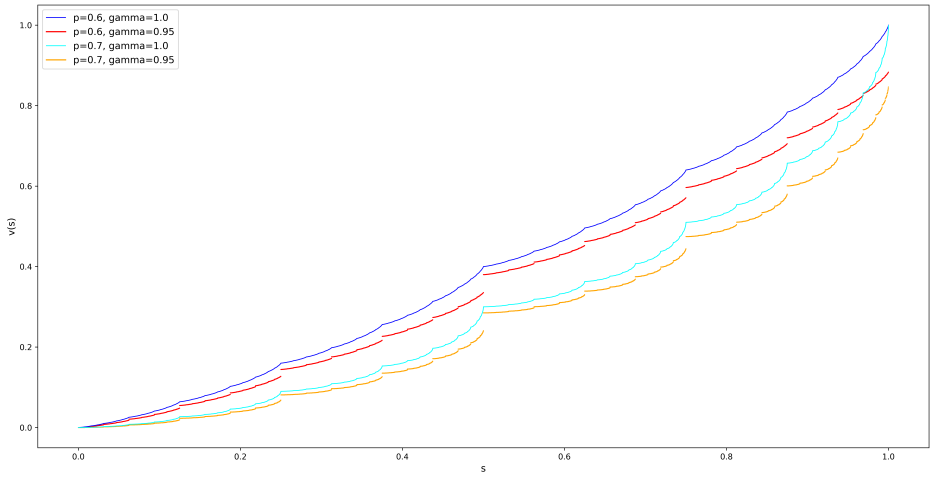

Figure 1: The optimal value function of the Gambler's problem.

the assumptions imposed in the recent studies are likely to be compromised, even for problems as "simple" as the Gambler's problem.

The most important result to supports the theorem is Claim 2. We introduce this claim first though it is depending the other claims.

Certain learning algorithms, such as q-learning, regard the solution of the Bellman equation as the value function. The solution of the Bellman equation is indeed the value function under discrete state space settings. However, it does not hold in general as the Bellman equation may have a continuum of finite solutions in an infinite state space. In fact, some necessary and sufficient conditions for a solution of the Bellman equation to be the value function has been discussed, for instance in [KLV15]. For our best knowledge, there has not been a general conclusion for continuous state spaces.

We show that, for the Gambler's problem, the solution toward the Bellman equation is indeed the value function.

Claim 2. *If the function $v(s)$ defined in Theorem 1 solves the system $v(s) = \max_{0<a\leq\min(s,1-s)}(1-p)\gamma v(s+a) + p\gamma v(s-a)$, $v(0) = 0$, $v(1) = 1$, $v(s) \leq 1$, it is the unique solution of the system.*

Proof. Assume that $v(s)$ is a solution of the system. We proof the claim by contradiction. Assume that $f(s)$ is also a solution of the system such that $f(s)$ is not identical with $v(s)$ at some s. Define $\delta = \sup_{0<s<1} f(s) - v(s)$. As $f(2^{-1}) \geq (1-p)\gamma f(1) + p\gamma f(0) = v(2^{-1})$, we have $\delta \geq 0$. We show that δ cannot be zero by contradiction. If δ is zero, as $v(s)$ and $f(s)$ are not identical, there exists an s such that $f(s) < v(s)$. Let $\bar{\delta} = \sup_{0<s<1} v(s) - f(s)$, for $\bar{\epsilon} = (1-p\gamma)\bar{\delta}$ we specify s_0 such that $v(s_0) - f(s_0) > \bar{\delta} - \bar{\epsilon}$. Let $a_0 = \min(s_0, 1-s_0)$, we have $v(s_0) = (1-p)\gamma v(s_0 - a_0) + p\gamma v(s_0 + a_0) \leq (1-p)\gamma f(s_0 - a_0) + p\gamma f(s_0 + a_0) + p\gamma\bar{\delta} \leq f(s_0) + p\gamma\bar{\delta}$. The factor-$p$ before $\gamma\bar{\delta}$ is due to either $v(s_0 - a_0) - f(s_0 - a_0)$ or $v(s_0 + a_0) - f(s_0 + a_0)$ being zero. It contradicts with $v(s_0) - f(s_0) > \bar{\delta} - \bar{\epsilon}$. Hence, δ cannot be zero. We discuss under $\delta > 0$ for the rest of the proof.

We first discuss under the existence of s_0 such that $f(s_0) - v(s_0) = \delta$. Let $\mathbb{S} = \{s | f(s) - v(s) = \delta\}$, for any $s_\delta \in \mathbb{S}$, for any $a_\delta \in \arg\max_{0<a\leq\min(s_\delta,1-s_\delta)}(1-p)\gamma f(s_\delta + a) + p\gamma f(s_\delta - a)$, we have

$$\begin{aligned} f(s_\delta) &= p\gamma f(s_\delta - a_\delta) + (1-p)\gamma f(s_\delta + a_\delta) \\ &\overset{(\heartsuit)}{\leq} p\gamma(v(s_\delta - a_\delta) + \delta) + (1-p)\gamma(v(s_\delta + a_\delta) + \delta) \\ &= p\gamma v(s_\delta - a_\delta) + (1-p)\gamma v(s_\delta + a_\delta) + \gamma\delta \\ &\overset{(\blacklozenge)}{\leq} v(s_\delta) + \gamma\delta. \end{aligned}$$

Thus, the equality of the inequation ($\heartsuit$) and ($\blacklozenge$) must hold. We specify $s_0 \in \mathbb{S}$, and by the equality of ($\heartsuit$) we have $f(s_0 - a_0) = v(s_0 - a_0) + \delta$, thus $s_0 - a_0 \in \mathbb{S}$. Let $s_1 = s_0 - a_0$, and we recursively specify $a_i \in \arg\max_{0<a\leq\min(s_i,1-s_i)}(1-p)\gamma f(s_i + a) + p\gamma f(s_i - a)$ and $s_{i+1} = s_i - a_i$, for $i = 0, 1, \ldots$ until $s_T = 0$

for some T, or indefinitely if such a T does not exist. For the first case where the sequence $\{s_t\}$ terminates at $s_T = 0$, we have $f(s_T) = v(s_T) + \delta$ by (♥), which contradicts with the boundary condition $f(s_T) = 0$. For the second case where the sequence $\{s_t\}$ is infinite, as it is strictly decreasing, it has a limit s^-. If $s_t \in g_l$ for some l, by similar arguments as Claim 4, we have $s_{t+1} \in g_l$, and inductively $s_{t'} \in g_l$ for any t'. As there is finite many elements in g_l, $\{s_t\}$ cannot be infinite. For $s_0 \notin \cup_{l=1}^{\infty} g_l$, as $\lim_{a_t \to 0}$, we have $(1-p)\gamma f(s_t + a_t) + p\gamma f(s_t - a_t)$ close to $\gamma f(s_t + a_t)$. It is obviously smaller than if $a_t = \min(s_t, 1 - s_t)$, which contradicts with the optimality of a_t.

We then discuss under $\gamma < 1$. Let $\epsilon = (1-\gamma)\delta$, by the definition of δ we specify s_δ such that $f(s_\delta) > v(s_\delta) + \delta - \epsilon$. Similarly, let $a_\delta \in \arg\max_{0<a\leq \min(s_\delta, 1-s_\delta)} (1-p)\gamma f(s_\delta + a) + p\gamma f(s_\delta - a)$, we have

$$\begin{aligned} f(s_\delta) &= p\gamma f(s_\delta - a_\delta) + (1-p)\gamma f(s_\delta + a_\delta) \\ &\leq p\gamma(v(s_\delta - a_\delta) + \delta - \epsilon) + (1-p)\gamma(v(s_\delta + a_\delta) + \delta - \epsilon) \\ &= p\gamma v(s_\delta - a_\delta) + (1-p)\gamma v(s_\delta + a_\delta) + \gamma(\delta - \epsilon) \\ &\leq v(s_\delta) + \gamma(\delta - \epsilon). \end{aligned}$$

With $\gamma < 1$, $f(s_\delta) > v(s_\delta) + \delta - \epsilon$ contradicts with $f(s_\delta) \leq v(s_\delta) + \gamma(\delta - \epsilon)$. Hence, the claim follows under the case $\gamma < 1$.

Finally, we show that under $\gamma = 1$, there exists $\max_{0<s<1} f(s) - v(s)$, which contradicts with our arguments that it must not exist. In fact, a continuous function on a closed interval must attain its maximum on that interval. As $v(0) = f(0) = 0$, $v(1) = f(1) = 1$ and $f(s) \geq v(s)$ for any s, the maximum over $0 < s < 1$ exists if and only if the maximum over $0 \leq s \leq 1$ exists. It suffices to show that $f(s) - v(s)$ is continuous. With the similar arguments in Claim 6, $v(s)$ is continuous over $0 \leq s \leq 1$, including $s \in g_l$ when $\gamma = 1$. To show the continuity of $f(s)$ we observe that $f(s)$ must be monotonically strictly increasing, under $\gamma = 1$. Otherwise per Claim 3 we specify $s_1 < s_2$, $v(s_1) > v(s_2)$ and $s_3 = 1$ and yield contradiction. Armed with the monotonicity of $f(s)$, the continuity is immediate. In fact, if at any point s it is not continuous, there must exist a ϵ-neighborhood set of s such that $v(s - \epsilon/4)$ does not attain $pv(s - 3\epsilon/4) + (1-p)v(s + \epsilon/4)$. Thus, the claim follows under $\gamma = 1$ as well. □

Claim 3. *If the function $f(s)$ solves the system $v(s) = \max_{0<a\leq \min(s,1-s)} (1-p)v(s+a) + pv(s-a)$, $v(0) = 0$, $v(1) = 1$, then for any $s_1 < s_2 < s_3$, either $f(s_2) > f(s_1)$ or $f(s_2) > f(s_3)$ or $f(s_1) = f(s_2) = f(s_3)$ is satisfied.*

Proof. We show it by contradiction. Without loss of generality if there exists $s_1 < s_2 < s_3$ such that $f(s_1) \geq f(s_2)$ and $f(s_3) > f(s_2)$, we define g_l' as the dyadic ratio spanned by s_1 and s_3. Formally, g_l' is the set of s where $(s - s_1)/(s_3 - s_1)$ can be represented by at most l bits in binary. We choose an integer $l \geq \log((s_2 - s_1)/4(s_3 - s_1)) \log(1-p)$ and $s_2^- \in g_l'$ such that s_3 is the largest element in g_l' subject to $s_3 < s_2$. For any k, we have

$$\begin{aligned} f(s_1 + (s_3 - s_1)2^{-k}) &\geq pf(s_1) + (1-p)f(s_1 + (s_3 - s_1)2^{-(k-1)}) \\ &\geq pf(s_1) + p(1-p)f(s_1) + (1-p)^2 f(s_1 + (s_3 - s_1)2^{-(k-2)}) \\ &\geq pf(s_1) + p(1-p)f(s_1) + \cdots + (1-p)^k f(s_1 + (s_3 - s_1)2^{-(k-k)}) \\ &= f(s_1) + (1-p)^k (f(s_3) - f(s_1)). \end{aligned}$$

As s_2^- is on the dyadic rational, we have

$$\begin{aligned} f(s_2^-) &\geq f(s_1 + (s_3 - s_1)2^{\lfloor \log((s_2 - s_1)/2(s_3 - s_1)) \rfloor}) \\ &\geq f(s_1) + (1-p)^{\log((s_2 - s_1)/4(s_3 - s_1))} (f(s_3) - f(s_1)). \end{aligned}$$

Meanwhile, we have

$$\begin{aligned} f(s_2^- + (k+1)(s_2 - s_2^-)) - f(s_2^- + k(s_2 - s_2^-)) &\leq (p/(1-p))(f(s_2^- + k(s_2 - s_2^-)) - f(s_2^- + (k-1)(s_2 - s_2^-))) \\ &\leq (p/(1-p))^k (f(s_2) - f(s_2^-)) \\ &\leq f(s_2) - f(s_2^-). \end{aligned}$$

Thus, as $s_2 + 2^l(1 - s_2)(s_2 - s_2^-) < 1$, we have

$$\begin{aligned} f(s_2 + 2^l(1 - s_2)(s_2 - s_2^-)) - f(s_2^-) &\leq 2^l(1 - s_2)(f(s_2) - f(s_2^-)) \\ &\leq 2^l(1 - s_2)(f(s_2) - f(s_1) - (1-p)^{\log((s_2 - s_1)/4(s_3 - s_1))} (f(s_3) - f(s_1))) \end{aligned}$$

$$\leq -2^l(1-s_2)(1-p)^{\log((s_2-s_1)/4(s_3-s_1))}(f(s_3)-f(s_1))).$$

As l approaches infinity, $f(s_2+2^l(1-s_2)(s_2-s_2^-))$ approaches negative infinity. Per $v(s)=\max_{0<a\leq\min(s,1-s)}(1-p)v(s+a)+pv(s-a)$, $v(s)$ is constantly negative infinity on $(0,1)$, which contradicts with $v(s)=\max_{0<a\leq\min(s,1-s)}(1-p)v(s+a)+pv(s-a)$ if the a takes $\min(s,1-s)$. The claim follows. □

Claim 4. *For any $s\in g_l$, $\max_{a\in g_{l+1}}(1-p)\gamma v(s+a)+p\gamma v(s-a)=\max_{a\in g_l}(1-p)\gamma v(s+a)+p\gamma v(s-a)$.*

Proof. It suffices to show that for any $s,a\in g_l$ and $2^{-l}\leq a\leq\max(s,1-s)$, either of $(1-p)v(s+a)+pv(s-a)\geq(1-p)v(s+a-2^{-(l+1)})+pv(s-a+2^{-(l+1)})$ or $(1-p)v(s+a-2^{-l})+pv(s-a+2^{-l})\geq(1-p)v(s+a-2^{-(l+1)})+pv(s-a+2^{-(l+1)})$ is satisfied. We discuss the first condition. According to the definition of $v(s)$,

$$\begin{aligned}v(s-a+2^{-(l+1)})-v(s-a)&=\sum_{i=1}^{l+1}(1-p)\gamma^i b_i\prod_{j=1}^{i-1}((1-p)+(2p-1)b_j)-\sum_{i=1}^{l}(1-p)\gamma^i b_i\prod_{j=1}^{i-1}((1-p)+(2p-1)b_j)\\&=(1-p)\gamma^{l+1}b_{l+1}\prod_{j=1}^{l}((1-p)+(2p-1)b_j)\\&=(1-p)\gamma^{l+1}\prod_{j=1}^{l}((1-p)+(2p-1)b_j),\end{aligned}\tag{1}$$

where $s-a=0.b_1b_2\dots b_{l(2)}$, and

$$\begin{aligned}&v(s+a)-v(s+a-2^{-(l+1)})\\&=(1-p)c_k\gamma^k\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j)\\&\quad-\sum_{i=k+1}^{l+1}(1-p)^2\gamma^i\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j)\prod_{j=k+1}^{i-1}((1-p)+(2p-1)c_j)\\&=(1-p)\gamma^k\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j)-\sum_{i=k+1}^{l+1}(1-p)^2p^{i-k-1}\gamma^i\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j)\\&=(1-\sum_{i=k+1}^{l+1}(1-p)p^{i-k-1}\gamma^{i-k})(1-p)\gamma^k\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j)\\&=(1-\sum_{i=0}^{l-k}(1-p)p^i\gamma^{i+1})(1-p)\gamma^k\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j)\\&=(1-(1-p)\gamma\frac{1-(p\gamma)^{l-k+1}}{1-p\gamma})(1-p)\gamma^k\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j)\\&\geq p^{l-k+1}(1-p)\gamma^{l+1}\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j),\end{aligned}\tag{2}\tag{3}$$

where $s+a=0.c_1c_2\dots c_{l(2)}$ and c_k is the last 1 bit of $s+a$. When $p^{l-k+1}\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j)\geq(p/(1-p))\prod_{j=1}^{l}((1-p)+(2p-1)b_j)$, we have $v(s+a)-v(s+a-2^{-(l+1)})\geq v(s-a+2^{-(l+1)})-v(s-a)$, which is one of the desired results. On the other hand, consider

$$v(s+a-2^{-(l+1)})-v(s+a-2^{-l})=(1-p)\gamma^{l+1}\prod_{j=1}^{l}((1-p)+(2p-1)c'_j)$$

and

$$v(s-a+2^{-l})-v(s-a+2^{-(l+1)})\geq p^{l-k'+1}(1-p)\gamma^{l+1}\prod_{j=1}^{k'-1}((1-p)+(2p-1)b'_j),$$

where $s + a - 2^{-l} = 0.c'_1 c'_2 \dots c'_{l(2)}$, $s - a + 2^{-l} = 0.b'_1 b'_2 \dots b'_{l(2)}$, and b'_k is the last 1 bit of $s - a + 2^{-l}$. We notice that by the definition of k, $c'_k = 0$ and $c'_{k+1} = c'_{k+2} = \dots = c'_l = 1$ (if $k = l$ there is no c'_{k+1} then). Hence,

$$\begin{aligned}(1-p)\gamma^{l+1}\prod_{j=1}^{l}((1-p)+(2p-1)c'_j) &= p^{l-k}(1-p)^2\gamma^{l+1}\prod_{j=1}^{k-1}((1-p)+(2p-1)c'_j)\\ &= p^{l-k}(1-p)^2\gamma^{l+1}\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j).\end{aligned}$$

Similarly, $b_{k'} = 0$ and $b_{k'+1} = b_{k'+2} = \dots = b_l = 1$, we have

$$(1-p)\gamma^{l+1}\prod_{j=1}^{l}((1-p)+(2p-1)b'_j) = p^{l-k'}(1-p)^2\gamma^{l+1}\prod_{j=1}^{k'-1}((1-p)+(2p-1)b_j).$$

Hence, the second sufficient condition $v(s-a+2^{-l}) - v(s-a+2^{-(l+1)}) \geq ((1-p)/p)(v(s+a-2^{-(l+1)}) - v(s+a-2^{-l}))$ holds whenever

$$\prod_{j=1}^{k'-1}((1-p)+(2p-1)b_j) \geq ((1-p)/p)\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j). \tag{4}$$

Rewrite $\prod_{j=1}^{l}((1-p)+(2p-1)b_j) = p^{l-k}(1-p)\prod_{j=1}^{k'-1}((1-p)+(2p-1)b_j)$, the first sufficient condition $p^{l-k+1}\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j) \geq (p/(1-p))\prod_{j=1}^{l}((1-p)+(2p-1)b_j)$ is equivalent to

$$\prod_{j=1}^{k-1}((1-p)+(2p-1)c_j) \geq \prod_{j=1}^{k'-1}((1-p)+(2p-1)b_j). \tag{5}$$

Let N_b and N_c indicates the number of 1s in $b_1, \dots b_{k'-1}$ and $c_1, \dots c_{k-1}$, respectively. If $N_b \leq N_c$, equation (5) holds which indicates $(1-p)v(s+a)+pv(s-a) \geq (1-p)v(s+a-2^{-(l+1)})+pv(s-a+2^{-(l+1)})$. Otherwise $N_b > N_c$ then $N_b \geq N_c + 1$, thus equation (4) holds which indicates $(1-p)v(s+a-2^{-l})+pv(s-a+2^{-l}) \geq (1-p)v(s+a-2^{-(l+1)})+pv(s-a+2^{-(l+1)})$. We conclude that $\arg\max_{a\in g_{l+1}}(1-p)\gamma v(s+a)+p\gamma v(s-a) \notin g_{l+1}$ as desired. □

Claim 5. *For any $s \in g_{l+1}$, $\min(s, 1-s)$ is an element of $\arg\max_{a\in g_{l+1}}(1-p)\gamma v(s+a)+p\gamma v(s-a)$.*

Proof. We proof by induction by assuming for any $s \in g_l$, $\min(s, 1-s) \in \arg\max_{a\in g_l}(1-p)\gamma v(s+a)+p\gamma v(s-a)$. We note that the base case $l = 1$ is trivial since $a \in g_1$ has only one element and $l = 2$ can also be verified by exhausting $a \in \{2^{-1}, 2^{-2}\}$ for $s = 2^{-1}$. As shown in Claim 4, for $s \in g_l$, $\arg\max_{a\in g_{l+1}}(1-p)\gamma v(s+a)+p\gamma v(s-a) \in g_l$. By the induction assumption $\min(s, 1-s) \in \arg\max_{a\in g_l}(1-p)\gamma v(s+a)+p\gamma v(s-a) \in \arg\max_{a\in g_{l+1}}(1-p)\gamma v(s+a)+p\gamma v(s-a)$. Hence, the claim holds for $s \in g_{l+1}-g_l$ and we assume $s \in g_{l+1}-g_l$ for the rest of the proof. We first discuss under $s \geq 2^{-1}+2^{-2}$. As $a \leq 1-s$, both $s-a \geq 2^{-1}$ and $s+a \geq 2^{-1}$ are satisfied. Hence, the first bit after the decimal of s, $s-a$, and $s+a$ is 1. For any $s = 0.c_1c_2\dots c_{l(2)}$ with $c_1 = 1$,

$$\begin{aligned}v(s) &= \sum_{i=1}^{l}(1-p)\gamma^i c_i\prod_{j=1}^{i-1}((1-p)+(2p-1)c_j)\\ &= (1-p)\gamma + \sum_{i=2}^{l}(1-p)\gamma^i c_i\prod_{j=1}^{i-1}((1-p)+(2p-1)c_j)\\ &= (1-p)\gamma + \sum_{i=1}^{l-1}(1-p)\gamma^{i+1}c_{i+1}((1-p)+(2p-1)c_1)\prod_{j=1}^{i-1}((1-p)+(2p-1)c_{j+1})\\ &= (1-p)\gamma + p\gamma v(2s-1).\end{aligned}$$

Hence,

$$(1-p)v(s+a)+pv(s-a) = (1-p)\gamma + p\gamma((1-p)v(2s+2a-1)+pv(2s-2a-1))$$

$$= (1-p)\gamma + p\gamma((1-p)v((2s-1)+2a) + pv((2s-1)-2a)).$$

We have both $2s-1 \in g_l$ and $2a \in g_l$ hence according to the induction assumption the maximum of $(1-p)v((2s-1)+2a) + pv((2s-1)-2a)$ is taken at $2a = \min(2s-1, 1-(2s-1)) = 2-2s$. As $a = 1-s$ is a feasible point of $a \leq \min(s, 1-s), a \in g_{l+1}$, we have $1-s \in \arg\max_{a \leq \min(s,1-s), a \in g_{l+1}} (1-p)v(s+a) + pv(s-a)$ as desired. We then discuss under the case $2^{-1} \leq s \leq 112^{-1} + 2^{-2}$. Similarly, for $s - a \leq 2^{-1}$, we have

$$\begin{aligned} v(s) &= \sum_{i=1}^{l-1} (1-p)^2 \gamma^{i+1} c_{i+1} \prod_{j=1}^{i-1} ((1-p) + (2p-1)c_{j+1}) \\ &= (1-p)\gamma v(2s). \end{aligned}$$

Thus,

$$\begin{aligned} (1-p)v(s+a) + pv(s-a) &= (1-p)^2\gamma + p(1-p)\gamma v(2s+2a-1) + p(1-p)\gamma v(2s-2a) \\ &= (1-p)\gamma(pv((2s-2^{-1}) - (2a-2^{-1})) + (1-p)v((2s-2^{-1}) + (2a-2^{-1}))) \\ &\quad + (1-p)\gamma(2p-1)v((2s-2^{-1}) + (2a-2^{-1})) + (1-p)^2\gamma. \end{aligned}$$

We have both $2s - 2^{-1} \in g_l$ and $2a - 2^{-1} \in g_l$ whenever $l \geq 2$. Thus, according to the induction assumption $pv((2s-2^{-1}) - (2a-2^{-1})) + (1-p)v((2s-2^{-1}) + (2a-2^{-1}))$ takes its maximum at $2a - 2^{-1} = 1 - (2s-2^{-1})$, which is equivalent to $a = 1-s$. We verify that $a = 1-s$ is a feasible point of $a \leq \min(s, 1-s), a \in g_{l+1}$. Meanwhile, according to the arguments in equation (1) and equation (3), we observe that the function is monotonically increasing on g_l for any l. Hence, $v((2s-2^{-1}) + (2a-2^{-1}))$ takes the maximum at the maximum possible a, which is $a = 1-s$. Since both parts takes their respective maximum at $a = 1-s$, we conclude that $1-s \in \arg\max_{a \leq \min(s,1-s), a \in g_{l+1}} (1-p)v(s+a) + pv(s-a)$ as desired. In similar arguments we show that $a = s$ is a maxima when $s \leq 2^{-1} - 2^{-2}$ and when $2^{-1} - 2^{-2} \leq s2^{-1}$, respectively. The claim follows. □

Claim 6. *Both $v(s)$ and $\max_a (1-p)\gamma v(s+a) + p\gamma v(s-a)$ are continuous at s if there does not exist an l such that $s \in g_l$.*

Proof. We first proof the continuity of $v(s)$. For $s = b_1 b_2 \dots b_l \dots_{(2)}$, $s \notin g_l$ indicates that for any integer N there exists an $n_1 \geq N$ such that $b_{n_1} = 1$ and an $n_0 \geq N$ such that $b_{n_0} = 0$. For any $s - 2^{-n_1} \leq s' \leq s + 2^{-n_0}$, by the monotonicity of $v(s)$, we have

$$\begin{aligned} v(s) - v(s') &\leq v(s) - v(s - 2^{-n_1}) \\ &= (1-p)\gamma^{n_1} \prod_{j=1}^{n_1-1} ((1-p) + (2p-1)b_j) \cdot (1 + \sum_{i=n_1+1}^{\infty} \gamma^{i-n_1} b_i p \prod_{j=n_1+1}^{i-1} ((1-p) + (2p-1)b_j)) \\ &\quad - (1-p)\gamma^{n_1} \prod_{j=1}^{n_1-1} ((1-p) + (2p-1)b_j) \sum_{i=n_1+1}^{\infty} \gamma^{i-n_1} b_i (1-p) \prod_{j=n_1+1}^{i-1} ((1-p) + (2p-1)b_j) \\ &= (1-p)\gamma^{n_1} \prod_{j=1}^{n_1-1} ((1-p) + (2p-1)b_j) \cdot (1 + \sum_{i=n_1+1}^{\infty} \gamma^{i-n_1} b_i (2p-1) \prod_{j=n_1+1}^{i-1} ((1-p) + (2p-1)b_j)) \\ &\leq (1-p)\gamma^{n_1} p^{n_1-1} \cdot (1 + \sum_{i=n_1+1}^{\infty} \gamma^{i-n_1} (2p-1) p^{n_1-i-1}) \\ &\leq 2(1-p)\gamma^N p^{N-1}, \end{aligned}$$

and similarly

$$\begin{aligned} v(s) - v(s') &\geq v(s) - v(s + 2^{-n_0}) \\ &\geq -(1-p)\gamma^{n_0} p^{n_0-1} \cdot (1 + \sum_{i=n_0+1}^{\infty} \gamma^{i-n_0} (2p-1) p^{n_0-i-1}) \\ &\geq -2(1-p)\gamma^N p^{N-1}. \end{aligned}$$

Hence, $|v(s) - v(s')|$ is bounded by $-2(1-p)\gamma^N p^{N-1}$ for $s - 2^{-n_1} \leq s' \leq s + 2^{-n_0}$. As $-2(1-p)\gamma^N p^{N-1}$ converges to zero as N approaches infinity, $v(s)$ is continuous as desired.

We then show the continuity of $v'(s) = \max_a (1-p)\gamma v(s+a) + p\gamma v(s-a)$. We first show that $v'(s)$ is monotonically increasing. In fact, for $s' \geq s$ and $0 \leq a \leq \min(s, 1-s)$, either $0 \leq a \leq \min(s', 1-s')$ or $0 \leq a+s-s' \leq \min(s', 1-s')$ must be satisfied. Let a' be a or $a+s-s'$ whoever is feasible, we have both $s'+a' \geq s+a$ and $s'-a' \geq s-a$. Specify a such that $v'(s) = (1-p)\gamma v(s+a) + p\gamma v(s-a)$, we have $v'(s') \geq (1-p)\gamma v(s'+a') + p\gamma v(s'-a') \geq v'(s)$. The monotonicity follows. Similarly we let $s = b_1 b_2 \dots b_l \dots_{(2)}$ and $n_1 \geq N$ such that $b_{n_1} = 1$ and an $n_0 \geq N+2$ such that $b_{n_0} = 0$. Also let $s_0 = b_1 b_2 \dots b_{N(2)}$. Then for the neighbourhood set $s_0 - 2^{-(N+1)} \leq s' \leq s_0 + 2^{-(N+1)}$, $v'(s) = v(s)$ for both the ends $s_0 - 2^{-(N+1)} \in g_{N+1}$ and $s_0 + 2^{-(N+1)} \in g_{N+1}$. $|v'(s) - v'(s')|$ is then bounded by $|v(s_0 - 2^{-(N+1)}) - v(s_0 + 2^{-(N+1)})|$. As shown in equation (1) and equation (3), the bound converges to zero as N approaches infinity. The continuity of $v'(s)$ follows. □

Proof of Theorem 1. Let $v'(s) = \max_a (1-p)\gamma v(s+a) + p\gamma v(s-a)$, per Claim 5 we have $v(s) = v'(s)$ on the dyadic rationals $\cup_{l=1}^{\infty} g_l$. Since $\cup_{l=1}^{\infty} g_l$ is a dense set, $v(s) = v'(s)$ whenever both $v(s)$ and $v'(s)$ are continuous at s. Thus, for any s if there does not exist an l such that $s \in g_l$, $v(s)$ and $v'(s)$ are continuous per Claim 6, which indicates $v(s) = v'(s)$. Otherwise if there exists an l such that $s \in g_l$, per Claim 5 we have $v(s) = v'(s)$. Hence, the Bellman equation and the boundary conditions are verified for $v(s)$. Per Claim 2, $v(s)$ is the unique solution to the Bellman equation and the boundary conditions. Since the optimal value function must satisfy the Bellman equation and the boundary conditions, $v(s)$ is the optimal value function, as desired. □

The optimal value function induces one of the optimal deterministic policies immediately.

Corollary 7. *The policy $\pi(s) = \min(s, 1-s)$ is (Blackwell) optimal under any γ in the aforementioned MDP.*

Some properties of the optimal value function are described below.

Corollary 8. *The curve of the value function on the interval $[k2^{-l}, (k+1)2^{-l}]$ is similar (in geometry) to the entire value function curve defined on $[0, 1]$, for any integer $l \geq 0$ and $0 \leq k \leq 2^l - 1$.*

Corollary 9. *The expectation $\int_0^1 v(s)ds = (1-p)\gamma$.*

Corollary 10. *The curve of $v(s)$ has a fractal dimension of 1.64.*

Corollary 11. $\arg\min v(s) - s = \frac{2}{3}$.

2 Q-Learning for the Gambler's Problem

We have proved the value function we proposed in Theorem 1, by showing the uniqueness of the solution of the system $v(s) = \max_{0<a\leq\min(s,1-s)}(1-p)\gamma v(s+a) + p\gamma v(s-a)$, $v(0) = 0$, $v(1) = 1$, $v(s) \leq 1$, in Claim 2. We will inspect the system from the Q-learning algorithmic perspective.

The Q-learning algorithm's objective is to satisfy the Bellman equation, by minimizing the Bellman error. The system $v(s) = \max_{0<a\leq\min(s,1-s)}(1-p)\gamma v(s+a) + p\gamma v(s-a)$, $v(0) = 0$, $v(1) = 1$ is exactly the Bellman equation of the Gambler's problem on both the non-terminal states and the terminal states. However, the condition $v(s) \leq 1$ is not originated from the Bellman equations. Instead, it is by the definition of the problem that the value function is a probability of some event. We will study the system that removes this additional condition to learn the characteristics of the Q-learning algorithm.

The value function $v(s)$ is obviously still a solution of the system without the $v(s) \leq 1$ condition. The question we will be discussing is if such a Bellman equation system is sufficient. when $\gamma < 1$, $v(s) \leq 1$ is not used in the proof of Theorem 1. Hence, a function $f(s)$ solving the Bellman equation system is sufficient to $f(s)$ being the value function. However, we show that this does not hold when $\gamma = 1$. To show this, we give all possible solutions to the Bellman equation system.

Corollary 12. *The function $f(s)$ solves the system $v(s) = \max_{0<a\leq\min(s,1-s)}(1-p)v(s+a) + pv(s-a)$, $v(0) = 0$, $v(1) = 1$, if and only if either*

- *$f(s)$ is exactly $v(s)$ defined in Theorem 1, or*
- *$f(s) = C$ for all $s \in (0, 1)$, for some constant $C > 1$.*

Proof. It is obvious that both $f(s)$ defined above are the solutions of the system. We only have to show that they are all the solutions. If $f(s) \leq 1$ for any s, the case has been already been discussed in the proof of Claim 2, where $v(s)$ defined Theorem 1 is the unique solution. For the rest of the proof, we show that $f(s) = C$ for some $C > 1$ is the unique solution if there exists an s such that $f(s) > 1$.

There exists some $s_l < s_m < s_h$, where $f(s_m) > f(s_l)$ and $f(s_m) > f(s_h)$, as $s_l = 0$, $s_h = 1$, and s_m be the point where $f(s_m) > 1$ is an instance. For any $s_l < s_m < s_h$ where $f(s_m) > f(s_l)$ and $f(s_m) > f(s_h)$, denote the following cases. H_1: there exists $s_m < s'_h < s_h$ such that $f(s'_h) > f(s_m)$; H_2: $f(s'_h) \leq f(s_m)$ for $s_m < s'_h < s_h$ and there exists $s_m < s'_h < s_h$ such that $f(s'_h) = f(s_m)$; H_3: $f(s'_h) < f(s_m)$ for all $s_m < s'_h < s_h$. Respectively, L_1, L_2 and L_3 for the interval $s'_l \in [s_l, s_m]$.

Case $\mathbf{L_1H_1, L_2H_1, L_1H_2}$ We have $s'_l < s_m < s'_h$ and $s'_l > s_m$, $s'_h > s_m$. This contradicts immediately with Claim 3.

Case $\mathbf{L_3H_3, L_2H_3, L_3H_2}$ If there exists an $s \in [0, s_l) \cup (s_h, 1]$ such that $f(s) > f(s_m)$, either (s, s_l, s_m) or (s_m, s_h, s) will contradict with Claim 3. If there does not exists such an s, we have $f(s_m) > pf(s_m - a) + (1 - p)f(s_m + a)$ for any $a > 0$, which contradicts with $f(s_m) = \max_{0<a\leq\min(s_m, 1-s_m)}(1-p)f(s_m+a)+pf(s_m-a)$.

Case $\mathbf{L_2H_2}$ We first show that in this case, $\max_{0\leq s'\leq 1} f(s') = f(s_m)$. If fact, we have $f(s'_l) = f(s_m) = f(s'_h)$ for some s'_l and s'_h. If $f(s) > f(s_m)$ for some s, either one of (s, s'_l, s_m), (s'_l, s, s_m), (s_m, s, s'_h), (s_m, s'_h, s) will contradict with Claim 3, where have exhausted all the possible interval s is in.

Since $\{s|f(s) = f(s_m)\}$ is non-empty, let $s^- = \inf\{0 < s < s_m|f(s) = f(s_m)\}$. If $f(s^-) = f(s_m)$, then $f(s) < f(s^-)$ for all $s < s^-$, consequently for all $a > 0$ we have $f(s^-) > pf(s^- - a) + (1 - p)f(s^- + a)$. That will contradict with $f(s') = \max_{0<a\leq\min(s',1-s')}(1 - p)f(s' + a) + pf(s' - a)$. If $s^- > 0$, we have $f(s)$ monotonically strictly increasing on $(0, s^-)$. Otherwise, (s_1^-, s_2^-, s^-) contradicts with Claim 3 for some $s_1^- < s_2^- < s^-$ and $f(s_1^-) > f(s_2^-)$. Thus, $pf(s' - a) + (1 - p)f(s' + a)$ is monotonically increasing as a decreases. The value $f(s^-) = \max_{0<a\leq\min(s^-,1-s^-)}(1 - p)f(s^- + a) + pf(s^- - a)$ will not exist.

We have shown that s^- cannot be greater than zero and $f(s^-) < f(s_m)$. Similarly, we have $s^+ = \sup\{s_m < s < 1|f(s) = f(s_m)\} = 1$. Thus, for arbitrary small ϵ, there exists $f(s) = f(s_m)$ in both $s \in (0, \epsilon)$ and $s \in (1 - \epsilon, 1)$. If $f(s) < f(s_m)$ for some s, it will contradict with Claim 3 as we specify $s_\epsilon^- \in (0, s)$ and $s_\epsilon^+ \in (s, 1)$, where $f(s_\epsilon^-) = f(s_\epsilon^+) = f(s_m)$. Hence $f(s) \geq f(s_m)$, which, combined with $f(s) \leq f(s_m)$ we have shown before, yields $f(s) = C$ for some $C = f(s_m) > 1$.

Case $\mathbf{L_1H_3}$, $\mathbf{L_3H_1}$ In this case, in exactly one of the intervals (s_l, s_m) and (s_m, s_h) there exists an s such that $f(s) > f(s_m)$. Let $\epsilon = |s_l - s_m|/4$, we inspect the value of $f(s_m - \epsilon)$ and $f(s + \epsilon)$. **(A)** If both the values are smaller than $f(s_m)$, we let $s_l \leftarrow s_m - \epsilon$ and $s_h \leftarrow s + \epsilon$ and repeat this case. **(B)** Consider otherwise exactly one of them is be greater than $f(s_m)$. If $f(s_m - \epsilon) > f(s_m)$, let $s_m \leftarrow s_m - \epsilon$ and $s_h \leftarrow s_m$ and repeat the case. Respectively, if $f(s_m + \epsilon) > f(s_m)$, let $s_m \leftarrow s_m + \epsilon$ and $s_l \leftarrow s_m$ and repeat the case. In all conditions the new interval will have halved length. **(C)** Consider otherwise exactly one of them equals to $f(s_m)$. If $f(s_m - \epsilon) = f(s_m)$, we have either $f(s_m - \epsilon) = f(s_m - \epsilon/2) = f(s_m)$ or $f(s_m - \epsilon/2) > f(s_m)$. The former case falls under Case L_2H_2. For the latter case we let $s_l \leftarrow s_m - \epsilon$, $s_m \leftarrow s_m - \epsilon/2$, and $s_h \leftarrow s_m$ and repeat the process. Respectively, if $f(s_m - \epsilon) = f(s_m)$, we have either $f(s_m + \epsilon) = f(s_m + \epsilon/2) = f(s_m)$ or $f(s_m + \epsilon/2) > f(s_m)$. The former falls into Case L_2H_2 and the later falls into the repeat of this process. **(D)** Otherwise exactly one of them equals to $f(s_m)$, which falls into Case L_2H_2.

We repeat the process in Case L_1H_3 and L_3H_1. There are two possibilities: either the process continues indefinitely, or it stops at some iteration where it falls into Case L_2H_2. Case L_2H_2 will yield the $f(s) = C$ solution. We need to show that it is not possible that the process continues indefinitely. We observe that for each iteration of the process, the length of the interval $|s_l - s_h| = 0$ becomes at most half of the length at the previous iteration. By contradiction, if it does continues indefinitely, we have $\lim|s_l - s_h| = 0$ as the number of iterations approaches infinity. By the way s_l, s_m, and s_h are generated, $f(s_m) > f(s)$ for any $s \in [0, s_l] \cup [s_h, 1]$. Let $s_m^* = \lim s_m$ be the limit point of the s_m sequence. For any a, there will be always an $a' < a$ such that $pf(s_m^* - a') + (1 - p)f(s_m^* + a') > pf(s_m^* - a) + (1 - p)f(s_m^* + a)$. Therefore $f(s_m^*) = \max_{0<a\leq\min(s_m^*,1-s_m^*)}(1-p)f(s_m^*+a)+pf(s_m^*-a)$ will not exist. We conclude that Case L_2H_2 will eventually happen at some iteration of the Case L_1H_3 and L_3H_1 series we have constructed. The claim follows by the constant $f(s) = C$ solution under Case L_2H_2. □

References

[Csá01] Balázs Csanád Csáji. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24:48, 2001.

[HR07] Dan Harrington and Bill Robertie. *Harrington On Modern Tournament Poker*. Two Plus Two Publishing LLC, 2007. ISBN: 978-1880685563.

[KLV15] Takashi Kamihigashi and Cuong Le Van. Necessary and sufficient conditions for a solution of the bellman equation to be the value function: A general principle. *Documents de travail du Centre dÉconomie de la Sorbonne 2015.07*, 2015. ISSN: 1955-611X.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.