東京電機大学

博　士　論　文

Softsatisficing: 確率論的満足化方策

Softsatisficing: Probabilistic satisficing policy

2022 年 3 月

東京電機大学大学院　先端科学研究科　情報学専攻

神谷匠

学籍番号 17UDJ91

目 次

図の目次

表の目次

第 1 章

序論

エージェントが環境の中で試行錯誤しながら適切な行動を探索する学習・制御の枠組みである強化学習は，より複雑な環境での最適方策の獲得のため，急速な発展を経て複雑化している．その発展は，深層強化学習でめざましく，Deep Q-Network(DQN) [Mnih 15] 以降，様々な側面に対して加えられた．工夫の統合による Rainbow [Hessel 18] に始まり，時系列を考慮した R2D2 [Kapturowski 19]，探索促進のために様々な要素を詰め込んだ NGU [Badia 20b]，そして，メタの視点を導入した Agent57 [Badia 20a] ではついに人レベルを超える性能を発揮した．それに伴い，エージェントが分散化し環境のコピーを必要とするなど，学習のためのリソースは，最適化アルゴリズムにとって深刻な障害となりつつある．

Simon は，そのようなリソース制約下でのエージェントの行動原理として，限定合理性（*bounded rationality*）という概念を提案した [Simon 57]．限定合理性により駆動するエージェントは一見非合理的な振る舞いをするように見えるかもしれないが，限界と制約を考慮するとエージェントの行動は合理的であると理解できる．近年，限定合理性は注目されており，脳科学（脳），認知科学（心），人工知能（機械）の 3 つの分野を統合すると主張されてきた計算合理性は，限定合理性を更新したものである [Gershman 15]．また，人間の柔軟で効率的な認知を可能にすると考えられてきた抽象化や階層化がその限界から生まれたものであり，限定合理性であることも議論されている [Genewein 15]．

限定合理性理論における代表的な意思決定方法は，満足化（*satisficing*） [Simon 55, Simon 56] である．満足化エージェントは，最適な行動を探し続けるのではなく，ある一定以上の質（基準）を目標とすることで，その制約下での最適化を行う．強化学習において満足化は，一部の研究 [片山 98, 岡田 01, Goodrich 04, Bendor 09, Reverdy 17] を除いてあまり注目されていなかった．満足化の実装として，著者の一人が Risk-sensitive Satisficing(RS) と呼ばれる単純な満足化価値関数を用いたアルゴリズムを提案し [Takahashi 16]，いくつかの強化学習タスクにおいて，RS の有効性を実証的に

検証してきた．[Oyo 17, Wakabayashi 21] また，通常のバンディット問題のアルゴリズムにおいて，学習時の後悔（期待損失) は，最適な場合にも選択回数に対する対数オーダーで発散する [Lai 85]．対して，*RS* では，後悔が有限でバウンドされることが示された [Tamatsukuri 19]．*RS* を初めとした満足化のアイディアの実装は，基準を持つことにより，様々な動機付けも可能とする．それは内発的な動機付けに基づく効率的な探索だけではなく，他者の成功体験や達成可能性といった外部情報を用いたエミュレーションをも可能とした [Wakabayashi 21, Shinriki 20]．

極限での振る舞いの分析と数値シミュレーションにより性能を示してきた *RS* だが，効率的な情報収集のための，探索あるいは学習過程についての研究は，未だ十分でない．情報収集に際して *RS* は，期待損失を均衡させ，価値の過大評価あるいは過小評価の抑制を実現している．後に示すように，その「損失均衡」において，*RS* は行動の信頼性を自律的に制御し，満足な行動の発見を可能としている．本論文では，学習過程の信頼性を近似する Softsatisficing を提案し，*RS* の信頼性制御における性質を分析する．そして，それにより実現される満足行動の探索性，つまり，満足化による意思決定について議論する．

1.1　本論文の構成

本論文の構成は以下の通りである．次章では，最適化の限界と最適化から満足化への緩和の意味について述べる．第 3 章では，多腕バンディット問題とその代表的な最適化アルゴリズムについて説明する．第 4 章では，*RS* アルゴリズムを概観し，その基本的な性質を提示する．第 5 章では，*RS* の分析に基づいた Softsatisficing を提案し，満足化により意思決定について議論する．そして第 6 章を結論とする．

第 2 章

最適化と満足化

強化学習エージェントの学習は，ある問題の最適解（最適な行動系列）の獲得を目標とする．しかし，エージェントは，その環境における最適性を自己判断できないため，十分な観測により行動の価値を推定することで収益の最大化を目指すことになる．

単一の状態しか持たない，行動とそれに対応する報酬のみを考慮する単純な強化学習タスクのクラスとして K 本腕バンディットタスクが存在する．プレイするスロットマシンの選択をモデル化したバンディットタスクでは，K 台のスロットマシンからもっとも報酬を獲得できるものを実際のプレイを通して予測する．報酬はスロットマシン毎に設定されたある確率に従い獲得でき，その最大化のためにエージェントは報酬確率が最大であるものをプレイする必要がある．

観測による行動選択の最適化を行う場合，たとえバンディット問題であっても，エージェントにとって最適性の判断は容易ではない．学習対象が複雑になるほど，探索空間は広大となり，最適性の判断に必要なリソースも増加する．

他方，人間や動物は，環境内の情報を適切に取捨選択し，それなりに適切な行動を行っているように見える．では，進化の結果であれ，経験からの学習であれ，人間や動物はどのようにそれを成しているのであろうか？その答えの 1 つとして，限定合理性による意思決定がある [Simon 57]．

合理性を背景とする最適化は，十分な観測による価値を通して環境の最適性の判断を必要とした．一方で限定合理性では，エージェントが完全に合理的に行動できることを前提としない．限定合理性エージェントは，自らの認知的限界の範囲で情報収集や意思決定を行う．

限定合理性における代表的な意思決定方法に，Simon により提唱された満足化（*Satisficing*）がある [Simon 55, Simon 56]．満足化とは，「満足（satisfy）」と「十分（suffice）」による造語で，ある環境下において，許容できる質（基準）を満たす行動を見つけるまで探索を続けるような意思決定原理である．最適化エージェントは，最適性の判断のために

俯瞰的に環境の全容を把握する必要があった．対して満足化エージェントは，基準によりその制約を緩め，基準を満たすことで緩和された最適性を判断できる．例えば，学業やアスリートのトレーニングにおいて，可能行動の空間全体から最適な行動を探索するのでなく，自己ベストの更新や世界記録を目標とするような意思決定は，よく行われるだろう．

本論文における満足化は，従来の満足化の概念に比べて，環境の探索の仕方の調整を含むより詳細なものである．従来の満足化概念においては，満足でないなら探索を続け，満足ならば探索を打ち切る，という 2 値的（binary） な意思決定である．それに対し，目標達成の程度に応じて探索の仕方を変えることは，無作為に探索するよりは妥当な意思決定であろう．アスリートの例であれば，自己ベストの更新を目標とするなら，自己ベストを出した際のフォームを微調整するだろう．あるいは世界記録を目指すとき，自己ベストが世界記録に遠く及ばないのであれば，フォームの大幅な変更や肉体改造も視野に入りうる．この事例のように，目標達成が不確実な場合において，行動の微修正などで満足を試みる場合もあれば，他方で，現状が満足からほど遠いなら，手順や方法の大きな変更を試みる場合もある．あるいは，その両者を比較することもあるだろう．この探索へのアプローチは，本質的に，環境の不確実性を扱うための方法である．また，最適な行動への到達可能性を考慮にいれる方法でもある．この側面は，従来の最適化アルゴリズムに対する意味のある代替案として満足化を特徴付けるものかもしれない．

第 3 章

K 本腕ベルヌーイバンディット問題

ここでは，RS の性能を示す際に用いる K 本腕バンディット問題について説明する．エージェントが知らない報酬確率 $\{p_1, p_2, \ldots, p_K\}$ に従い報酬 1 か 0 を得られる K 個の行動 $\{a_1, a_2, \ldots, a_K\}$ があるとする．エージェントが行動 a_i を選択した場合，確率 p_i で報酬 1 を獲得し，確率 $1 - p_i$ で 0 を獲得する．行動選択を繰り返す際の目標は，累積報酬の最大化である．最適行動選択率を表す accuracy は，エージェントの意思決定における最適性の指標である．期待損失を表す regret は，累積報酬のように報酬の確率分布 $\{p_i\}$ に依存しないアルゴリズムの性能指標である．i^* は，最大の報酬確率を持つ行動の添え字を示す（すなわち，$p_{i^*} = max_i p_i$）．t 番目のステップ（1 ステップは 1 回の行動選択を意味する）が終了したときの後悔は，以下のように定義される．

$$\mathrm{regret}(t) = \sum_{i=1}^{K} (p_{i^*} - p_i) E[n_i(t)] \tag{3.1}$$

この時，$E[\cdot]$ は期待値を表し，$n_i(t)$ は t ステップの終了までの行動 a_i の選択回数である（ステップ数が明示されていない場合は，単に n_i とする）．つまり後悔とは，「実際に選択された行動の累積期待報酬が，最初のステップから最適な行動を選択し続けた場合の累積期待報酬に比べてどれだけ劣るか」を意味する．後悔が小さければ小さいほどアルゴリズムの性能は高く，後悔がゼロであるときすべてのステップで最適な行動が選択されていることを示す．後悔は，ステップ数 t において，少なくとも $\mathcal{O}(\log t)$ で増加することが証明されている [Lai 85]．

エージェントの行動選択では，最も価値の高い行動をとることが基本的な方策となる（greedy 方策）．バンディットタスクでは一般的に，行動 a_i の価値 E_i は，各ステップ t

における報酬 r^t の平均として評価される．

$$E_i \leftarrow E_i + \frac{1}{n_i}(r^t - E_i) \tag{3.2}$$

しかし，平均報酬評価に対して greedy 方策によって行動選択すると，初期の試行で高い価値を持つ非最適行動 a_i $(i \neq i^*)$ があった場合，その a_i を選択し続ける危険性がある．最適行動を見つけるためには，他の行動を適当な回数だけ試して過小評価を抑制しなければならない．蓄積した知識に基づき単に価値の高いものを選ぶ（知識利用）だけでは不十分であり，様々な行動を選択（探索）する必要がある．これは探索と知識利用のトレードオフとしてよく知られ，K 本腕バンディット問題では両者のバランスをとる様々なアルゴリズムが提案されている．

RS の性能提示のため，当たり観測が密で簡単な環境 `Dense` と当たり観測が疎で困難な環境 `Sparse` の 2 種類の環境を用意した．`Dense` と `Sparse` ともに $K = 9$ であり，各行動 $\{a_1, a_2, \ldots, a_8, a_9\}$ に対して，`Dense` では報酬確率 $\{0.1, 0.2, \ldots, 0.8, 0.9\} \in P_{\mathbf{easy}}$，`Sparse` では報酬確率 $\{0.01, 0.02, \ldots, 0.08, 0.09\} \in P_{\mathbf{hard}}$ を持つ．また，K 本腕バンディット問題の後悔最小化（報酬最大化）モデルで理論的証明がありかつ単純な手法である Upper Confidence Bounds[Auer 02] と，Thompson Sampling[Thompson 33, Agrawal 12] をベースラインとする．

3.1 Upper Confidence Bounds (UCB) Algorithm

Upper Confidence Bounds(UCB) は，推定価値の信頼区間上限がもっとも高い行動を選択するアルゴリズムである．UCB では，各行動の価値に対しする信頼区間上限 R_{UCB_i} を Hoeffding's Inequality に基づき計算し，比較する．

$$\mathrm{UCB}_i = E_i + R_{\mathrm{UCB}_i} \tag{3.3}$$

$$R_{\mathrm{UCB}_i} = \sqrt{c \ln(t)/2n_i} \tag{3.4}$$

R_{UCB_i} の計算は，分母に試行数 n_i を用いるため，最初にすべての行動を一度ずつ選択する必要がある．理論的には $c = 1$ となるが，経験的に良いとされるしきい値として $c = 4$ を用いた．

3.2 Thompson Sampling (TS) Algorithm

Thompson Sampling(TS) は，行動選択により得られる報酬の分布をベイス推定によりモデル化し，各モデルからサンプリングした値がもっとも高い行動を選択する．これはベイズによる確率マッチング法の定式化であり，モデルからのサンプリングにより自ずと各

腕の期待値最大である確率程度に各々の行動が選択される．報酬が Bernoulli 分布に従う場合は，推定には Beta 分布を用いる．行動 a_i の Beta 分布のパラメータ α_i と β_i は次のように更新され，事後確率は更新された α_i と β_i による Beta 分布である．

$$\alpha_i \leftarrow \alpha_i + r^t \tag{3.5}$$
$$\beta_i \leftarrow \beta_i + (1 - r^t) \tag{3.6}$$

事後分布からサンプリングされた値は，UCB と同様に，行動 i の期待値 E_i を補正する次のような式で表せる．

$$\mathrm{TS}_i = E_i + R_{\mathrm{TS}_i} \tag{3.7}$$

このとき，Beta 分布における補正値 R_{TS_i} は，範囲 $[-E_i, 1-E_i]$ の値であり，$E[R_{\mathrm{TS}_i}] = 0$ である．

第 4 章

Risk-sensitive Satisficing アルゴリズムの概観

Risk-sensitive Satisficing（RS）は，greedy 方策の下で操作されたときに，満足化に基づく行動選択を実現する価値関数として提案された [Takahashi 16]．今までの RS に関する論文では満足化を組み込んだ価値関数を指して RS とすることが多いが，本論文ではその価値関数をコアとしたエージェントの意思決定の枠組みのことを RS と呼称する．

4.1 RS 価値関数と方策

RS の紹介のため，まず行動 a_i の価値 E_i と基準 $\aleph$ の差を損益 δ として定義する．

$$\delta_i = E_i - \aleph \tag{4.1}$$

損益 δ は，環境上で達成すべき目標を基準 $\aleph$ として持つことによる値である．価値 E_i は 0 を最低値としたものだが，損益 δ は，その最低値を $\aleph$ へと移動した形で，目標達成が可能な行動が存在するか否かの判断に利用される．そのような行動が存在しない場合は探索を行う．探索のもっとも単純な方法は，すべての行動からランダムに選択することだ．ランダムな探索であっても，制約下の最適性を判断できるため，行動空間全体を十分に探索する必要性を排除できる．

RS では，価値関数に信頼性を導入することによって，信頼性が低い行動に楽観的なバイアスをかけ，その過小評価と過大評価を抑制する．一方で，信頼性の高い行動に対するバイアスは抑制し，本来の価値を見据えて知識利用を行う．RS 価値関数 は，損益 δ と信頼性 τ によって，次のように定義される [Takahashi 16]．

$$RS_i = \tau_i \delta_i = \tau_i (E_i - \aleph) \tag{4.2}$$

$$\tau_i = \frac{n_i}{t} \tag{4.3}$$

ここで, n_i は行動 a_i の試行回数で, t は総試行回数である．RS による行動選択は, greedy 方策によりこの価値 RS_i を最大化するように行われる．

$$a_i \leftarrow \arg\max_i RS_i \tag{4.4}$$

RS への理解の促進と，UCB と TS との比較のため，式 4.2 を価値とバイアスの加算として変形した $RS\beta$ 価値関数についても紹介する．

$$\begin{aligned} RS\beta_i &= \tau_i(E_i - \aleph) + \aleph & (4.5)\\ &= E_i - (1-\tau_i)E_i + \aleph(1-\tau_i) \\ &= E_i + (\aleph - E_i)(1-\tau_i) & (4.6)\end{aligned}$$

このとき，$R_{\mathbf{RS}_i} = (\aleph - E_i)(1-\tau_i)$ とすると，次のように表記できる．

$$\begin{aligned} RS\beta_i &= E_i + R_{\mathbf{RS}i} & (4.7)\\ R_{\mathbf{RS}_i} &= (\aleph - E_i)(1-\tau_i) & (4.8)\end{aligned}$$

RS では価値の最低値を，0 から $\aleph$ へずらしていた．$RS\beta$ は，再度最低値を 0 に戻す操作に相当し（式 4.5），この時，式 4.7 のように，価値にバイアスを加算する形となる．バイアス $R_{\mathbf{RS}_i}$ は，基準に満たない量 $\aleph - E_i$（$= -\delta$）に依存する．よって $RS\beta$ においては，価値が信頼できるなら価値 E そのものとしてみなし，信頼できない（不確かである）のであれば価値 E を基準 $\aleph$ に近づけるような楽観的なバイアスを加えることになる．

4.2 性能の概観

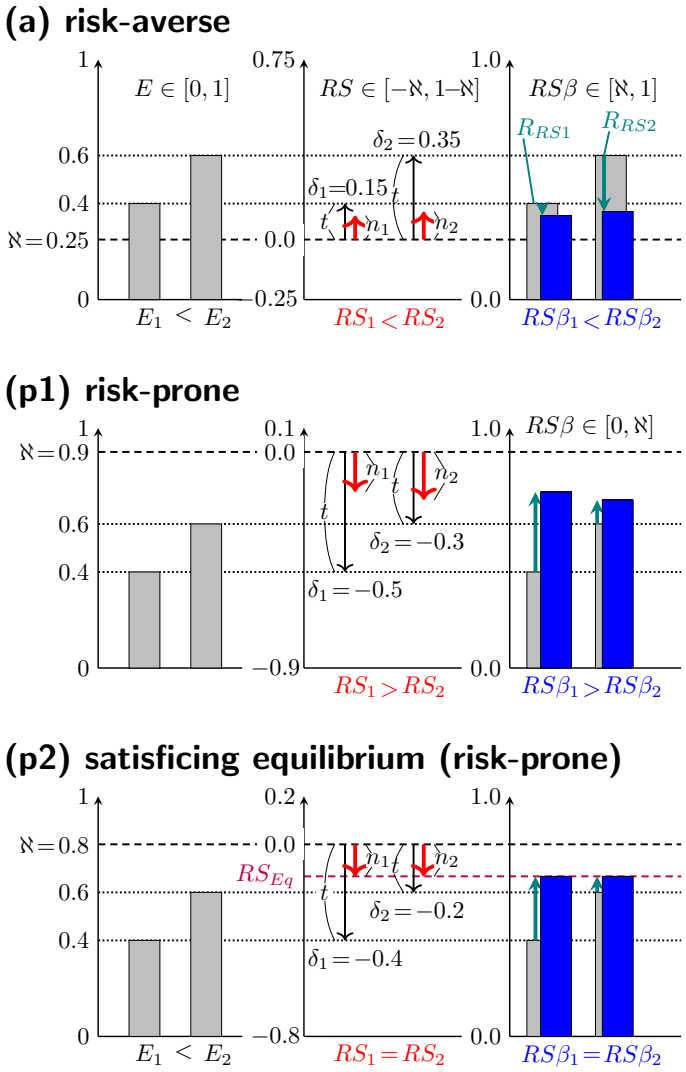
意思決定に満足化を用いる効果を示すため，3 章にて定義した `Dense` と `Sparse` を用いて，バンディットタスクにおける RS の性能を提示する．バンディットタスクの指標 accuracy と regret により，様々な基準に応じた RS の振る舞いを示し，greedy 方策により意思決定を行う RS がどのような探索を行っているのかを解説する．

まず，より RS の理解を促すため，環境に応じた複数の基準を定義する．以下に提示する基準は，環境と基準の関係性による RS の振る舞いを示すためのものであり，報酬の最大化を目的としたものだけでないことに注意してほしい．

環境上で満足可能な行動が 1 つに絞られ最適化が行えるような基準値として，真の報酬確率の上位 2 つの間に基準を置いたものを最適化基準と呼ぶ．そのもっとも単純なものを次のように定義する．

$$\aleph_{\mathrm{opt}} = (p_{1\mathrm{st}} + p_{2\mathrm{nd}})/2 \tag{4.9}$$

最適化基準を用いたとき RS は後悔を定数に抑えられることが示されている [Tamatsukuri 19]．この他に，達成可能な価値水準を大幅に超える過剰基準，



action	a_1		a_2
E	4/10=0.4	$<$	3/5=0.6
τ	10/15		5/15
δ	0.15		0.35
RS	-1/6	$>$	-1/5
R_{RS}	1/3		1/10
$RS\beta$	11/15	$>$	7/10

action	a_1		a_2
E	2/5=0.4	$<$	6/10=0.6
τ	5/15		10/15
δ	-0.5		-0.3
RS	-1/6	$>$	-1/5
R_{RS}	1/3		1/10
$RS\beta$	11/15	$>$	7/10

action	a_1		a_2
E	2/5=0.4	$<$	6/10=0.6
τ	5/15		10/15
δ	-0.4		-0.2
RS	-2/15	$=$	-2/15
R_{RS}	4/15		1/15
$RS\beta$	2/3	$=$	2/3

図 4.1: 行動価値（E_i），RS 価値（RS_i），$RS\beta$ 価値（$RS\beta_i$）の比較

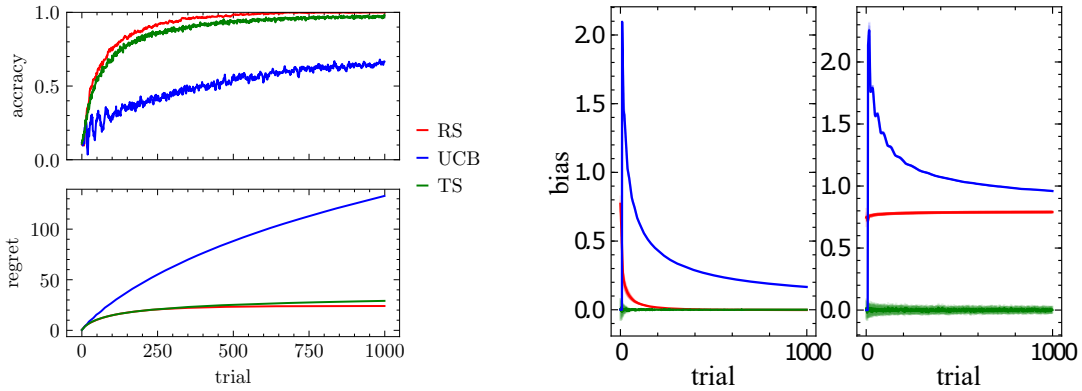
報酬の値域上限に近いが満足不可能な基準，どのような収益でも満足する全許容の基準を用いる（表 4.1）．

以下に提示する図では，すべて行動選択 1 回を 1 試行とした 1000 シミュレーションの平均である．まず，`Dense` の最適基準を持つ RS と主要な手法について accuracy と regret を，図 4.2 に示す．UCB は，step 数 t を水準に，各行動が最適である可能性を考慮する．よって，最適でない行動についても，少しでも可能性がある限り選択し続ける．対して TS は，報酬分布をモデルを用いて推定することで，早々に不適な行動を選択肢から除外する．RS は両者の中間に位置するような挙動である．$\aleph - E$ をバイアスの最大値として，満足できそうな行動の探索を続けつつ，より満足に近い行動を重点的に選択する．満足できる行動を発見後は，その行動を選択し続ける．

次に，`Dense` と `Sparse` の基準ごとの RS の accuracy と regret を図 4.3 に示す．

表 4.1: `Dense` と `Sparse` において用いる基準

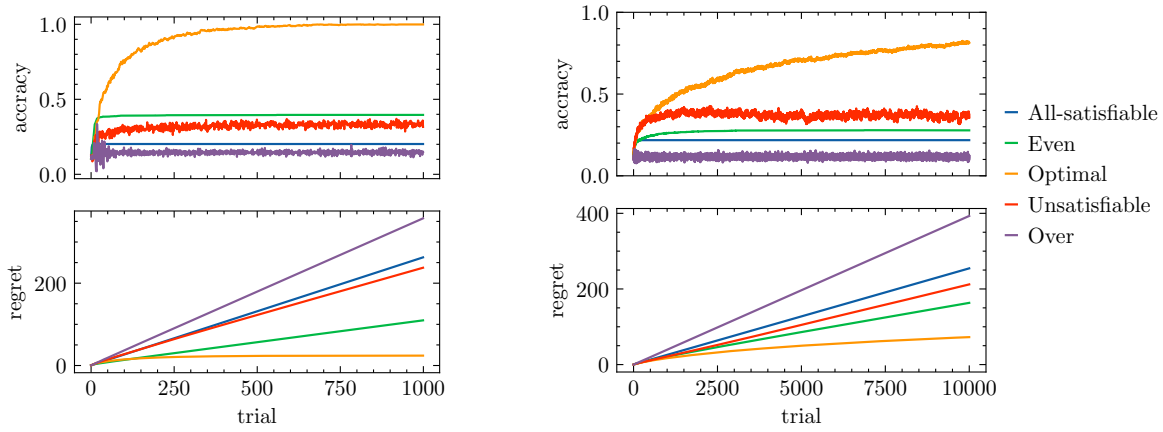
基準	`Dense`	`Sparse`
最適	0.85	0.085
過剰	2.0	1.0
満足不可能	1.0	0.1
成否等価	0.5	0.05
全許容	0.0	0.0



(a) accuracy と regret (b) バイアス ($p = 0.9$, $p = 0.1$)

図 4.2: `Dense` における最適基準 RS と主要な手法の比較

`Dense` の最適化基準の RS は accuracy の向上が速く，400 step 時点でほぼ最適行動のみを選択しており，regret も収束していることがわかる．一方で最適基準でない RS は，満足不可能基準 $\aleph = 1.0$ と成否等価基準 $\aleph = 0.5$ の accuracy が同程度へ収束している．前者が振動しているのは，満足できないため探索を続けつつ損失のもっとも少ない最適行動の選択率を増やしているためだ．後者は，当たり確率 0.5 以上の行動のうちどれかに満足した上で，より当たり確率の大きい最適行動を満足な行動として選択しているためである．両基準の違いは regret にも現れており，満足不可能基準が 1000 試行時点で成否等価基準の 2 倍程度と，大幅な差が開いている．過剰基準 $\aleph = 2.0$ と全許容基準 $\aleph = 0.0$ についても，おおむね同等の理由で accuracy が振動，収束している．また，RS の accuracy はどの基準においても，早期に収束していることが確認できる．`Sparse` についても，`Dense` 同様に最適基準 $\aleph = 0.085$ はほぼ最適行動のみの選択し，他基準についても，accuracy と regret の両者が同様の振る舞いをしていることが確認できる．



(a) `Dense` (b) `Sparse`

図 4.3: 様々な基準の RS の Accuracy と regret

4.2.1 非満足時損失均衡

すべての行動の価値が基準より小さい，非満足状況にある RS は，信頼性が低く不確実な行動に楽観的なバイアスをかけることで過小評価を避ける．信頼性が高く損失の少ない行動でも，満足できない可能性を避けるように損失を均衡させることで過大評価を避け，満足可能な行動を探索する．この均衡のことを，「非満足時損失均衡」と呼ぶ．

図 4.4 は `Sparse` における探索時の価値 RS である．報酬の観測が疎であるため，価値

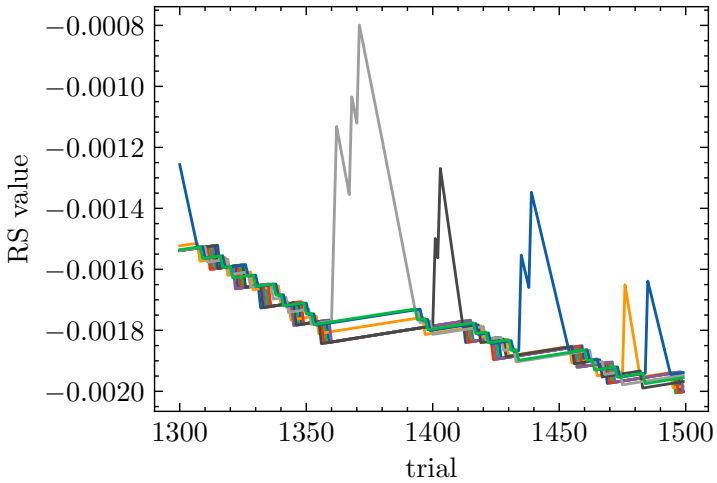


図 4.4: `Sparse` において探索中の RS 価値．報酬の獲得がない場合，選択された行動は信頼度を増加させ，RS 価値を減少させる．選択されなかった行動は信頼度を低下させ，RS 価値を増加させる．その結果，RS 価値はバランスされ，エージェントは探索を行う．

RS を均衡させるようにすべての行動を同程度に探索していることがわかる.

第 5 章

Softsatisficing

決定論的に振る舞う RS による満足化の探索過程の分析のため，その振る舞いを近似する確率論的なバリエーションを提案する．このバリエーションは，確率的方策としての定義であり，また，試行回数を必要しない．このモデルを，Softsatisficing と呼称する．Softsatisficing は RS の探索過程の選択分布を近似し，満足と非満足の 2 値でなく非満足時の満足化を緩やかに確率で表現する．以降では，Softsatisficing と，arg max を 2 値でなく確率的に行う Softmax との比較を通して，満足化の探索過程を定性的に分析する．

5.1 Softsatisficing Policy

非満足時の損失均衡を前提とした時，RS は Softmax （式 5.1）に類似した形式である Softsatisficing として表現できる．Softmax 関数は，arg max が価値を最大化する行動を決定論的に選択するのに対し，価値の差に応じて確率的に行動を選択するものである．同様に，Softsatisficing は，損失が最大の行動を決定論的に選択するのでなく，損失に応じて確率的に選択するものである．その行動選択確率は，RS の損失均衡時の信頼性 τ とほぼ等価となる

$$\mathrm{Softmax}(a_i) = \frac{e^{E_i \lambda}}{\Sigma_{j=1}^{K} e^{E_j \lambda}} \tag{5.1}$$

非満足均衡時の RS の信頼性 τ は，均衡値から計算できる．均衡値を $RS_{\mathbf{Eq}}$ とすると，信頼性 τ_i は次のように表現できる．

$$\tau_i = RS_{\mathbf{Eq}} \delta_i^{-1} \tag{5.2}$$

各行動の信頼性の総和は 1 である（式 5.3）．そのため，均衡値 $RS_{\mathbf{Eq}}$ は，式 5.4 のようになる．

$$\Sigma_{j=1}^{K} \tau_j = RS_{\mathbf{Eq}} \Sigma_{j=1}^{K} \delta_j^{-1} = 1 \tag{5.3}$$

$$RS_{\mathbf{Eq}} = \frac{1}{\Sigma_{j=1}^{K} \delta_j^{-1}} \tag{5.4}$$

よって式 5.2 と式 5.4 より，非満足均衡時の τ_i は式 5.5 のように記述できる．

$$\tau_i = \frac{-\delta_i^{-1}}{\Sigma_{j=1}^{K} - \delta_j^{-1}} \tag{5.5}$$

信頼性 τ_i はさらに，$-\delta^{-1} = e^{-\ln(-\delta)}$ となることから次のように変形できる．

$$\begin{aligned} \tau_i &= \frac{e^{-\ln(\aleph - E_i)}}{\Sigma_{j=1}^{K} e^{-\ln(\aleph - E_j)}} \\ &= \frac{e^{E_i \frac{-\ln(\aleph - E_i)}{E_i}}}{\Sigma_{j=1}^{K} e^{E_j \frac{-\ln(\aleph - E_j)}{E_j}}} \end{aligned} \tag{5.6}$$

このとき，$E_{\mathrm{RS_i}} = -\ln(\aleph - E_i)$ とすると，次のように Softsatisficing 関数を定義できる．

$$\mathrm{Softsatisficing}(a_i) = \frac{e^{E_{\mathrm{RS}_i}}}{\Sigma_{j=1}^{K} e^{E_{\mathrm{RS}_j}}} \tag{5.7}$$

$$E_{\mathrm{RS}_i} = -\ln(\aleph - E_i) \tag{5.8}$$

以降，$E_{\mathrm{RS_i}}$ を満足値と呼ぶ．また，逆温度関数 $\lambda_{\mathrm{RS}_i} = -\ln(\aleph - E_i)/E_i$ を定義すると，価値 E_i から満足値 E_{RS_i} への変調がより明確になる．

$$\mathrm{Softsatisficing}(a_i) = \frac{e^{E_i \lambda_{\mathrm{RS}_i}}}{\Sigma_{j=1}^{K} e^{E_j \lambda_{\mathrm{RS}_j}}} \tag{5.9}$$

$$\lambda_{\mathrm{RS}_i} = \frac{-\ln(\aleph - E_i)}{E_i} \tag{5.10}$$

$$\tag{5.11}$$

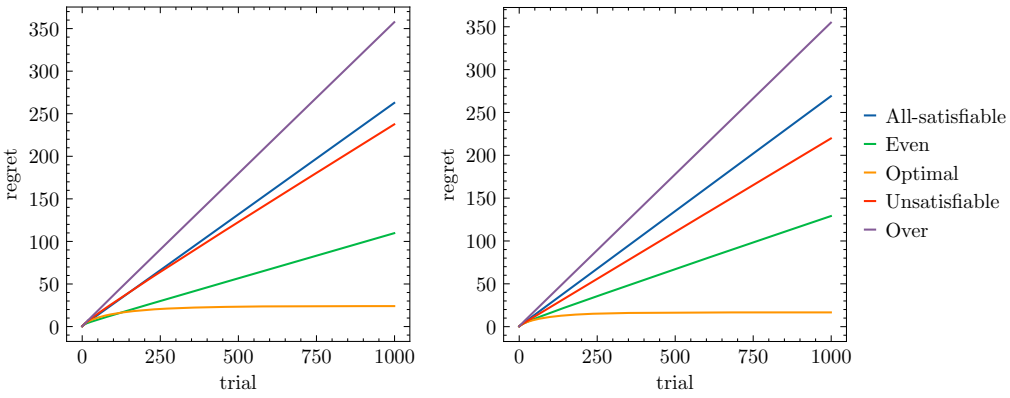
$$E_{\mathrm{RS}_i} = E_i \lambda_{\mathrm{RS}_i} \tag{5.12}$$

以上が，非満足均衡時における RS の意思決定を近似する Softsatisficing 関数である．動的な基準の変動を考慮しない場合，満足する行動が存在するならば単一である．そのため，方策としての Softsatisficing は，非満足では式 5.9 に従う．また，満足な行動があるなら，その行動を確率 1 で選択する．

図 5.1 は，`Dense` における RS と Softsatisficing の指標 regret による比較である．学習初期の振る舞いに多少の差異があるが，同等の性能を示していることがわかる．

5.2 Softmax と Softsatisficing の比較: 価値と満足値

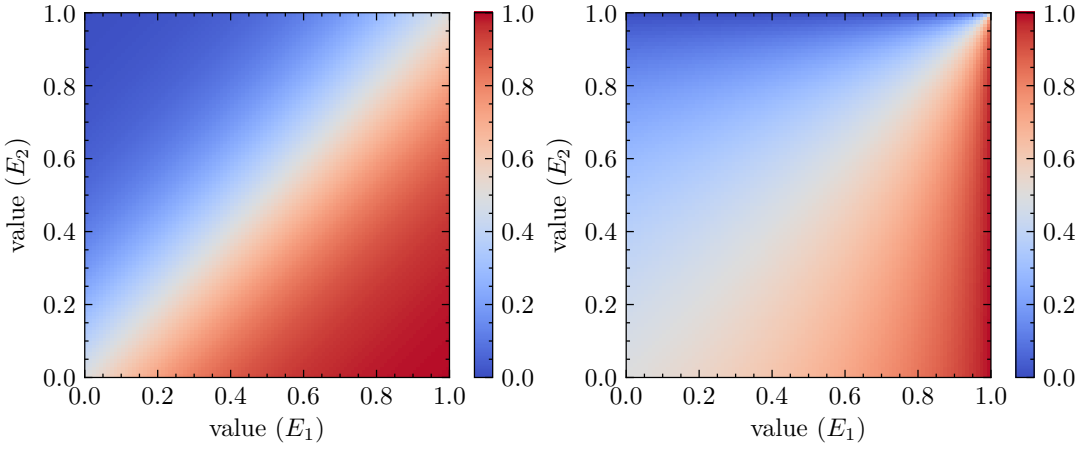
強化学習において Softmax は，価値の差にもとづく探索のための方策として用いられる [Sutton 18]．ただし，Softmax は，価値が収束するほどに十分な探索が前提である，



(a) *RS* (b) Softsatisficing

図 5.1: 様々な基準 ℵ による RS と Softsatisficing の後悔

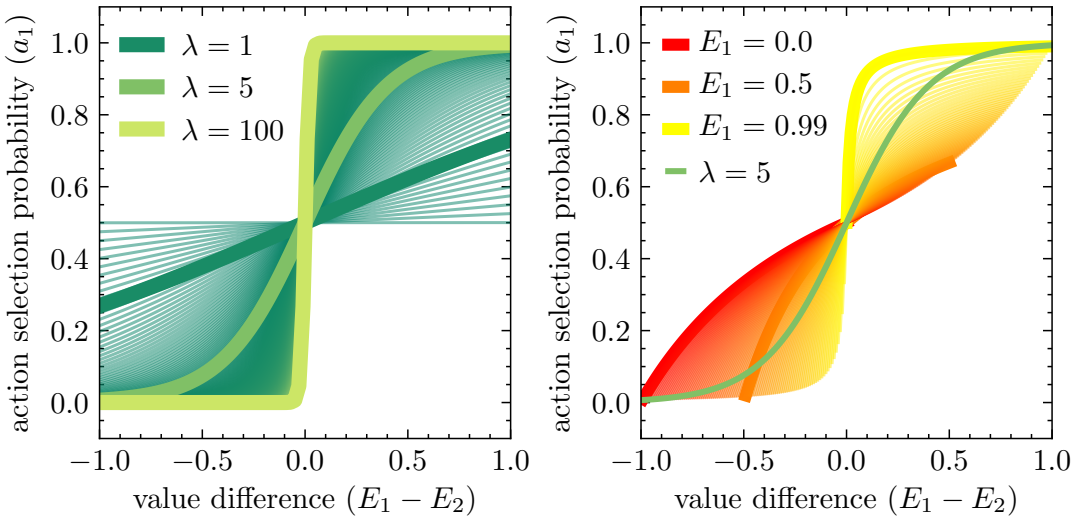
あるいは，学習のために適切な逆温度 λ を要するなど，効率的な探索が行えるとは言い難い．一方で，Softsatisficing は，同様の式を用いるにも関わらず，効率的な探索を可能とする．ここでは，Softmax と Softsatisficing の差異から，エージェントの効率的な探索に寄与する要因を示す．以降では簡単のため，行動を a_1 と a_2，それに対応して，範囲 $[0, 1]$ の価値 E_1 と E_2 を想定する．



(a) Softmax (b) Softsatisficing

図 5.2: (a) Softmax($\lambda = 5$) と (b) Softsatisficing の行動 a_1 の選択確率

図 5.2a と図 5.3a は，Softmax の各価値に対応する行動 a_1 の選択確率と，価値の差による累積確率分布である．Softmax は，価値の差 $E_1 - E_2$ に対して等価な確率を返す（図 5.2a）．また，選択確率 1 と 0 の間は緩やかに補完されるが，その程度は逆温度 λ に依存する（図 5.3a）．



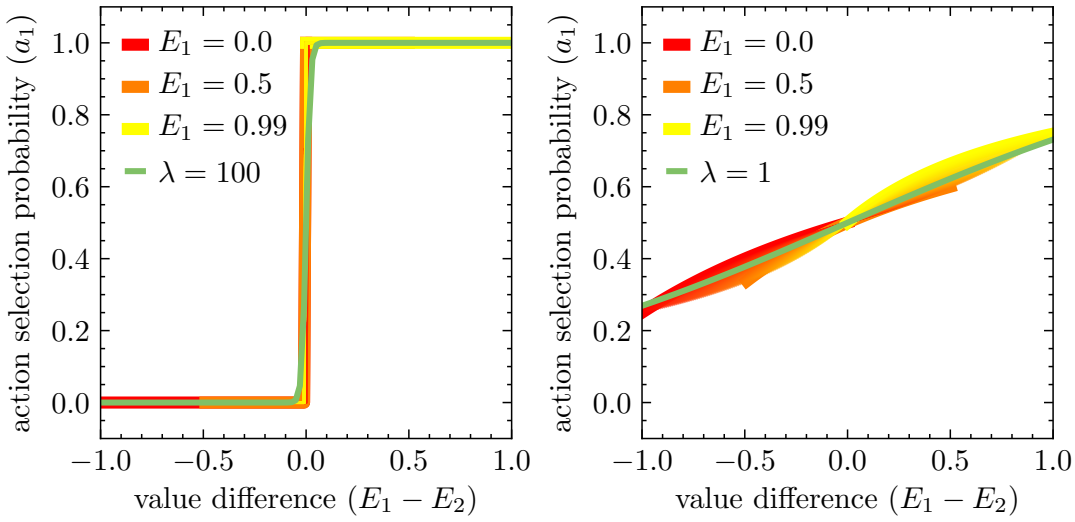
(a) Softmax (b) Softsatisficing

図 5.3: (a) Softmax と (b)$\aleph = 1.0$ の Softsatisficing の行動 a_1 の選択確率

図 5.2b と図 5.3b は，Softsatificing の各価値に対応する行動 a_1 の選択確率と，価値の差による累積確率分布を示す．Softsatisficing では，差が 0 の場合を除いて，価値の差に対応する行動選択率は単一ではない．行動選択確率は満足値 E_{RS} により決まるため，価値と基準の関係が，行動選択確率に影響を与える．

価値 E_1 をある数値に固定したとき，Softsatisficing は特徴的なカーブを描く Softmax が S 時カーブを描くのに対し，Softsatisficing は，常に上に凸のカーブとなる．$E_1 - E_2 > 0$ となる場合，価値の差 0 から離れるにつれ，価値 E_2 は減少する．これは，満足値関数は負の対数なので，Softmax の逆温度 λ を緩やかに 1 まで減少する操作に近い．逆に，$E_1 - E_2 < 0$ となる場合は，逆温度 λ を緩やかに ∞ へ近づけることに相当する．満足値を用いることで，Softsatisficing には，このような価値の差 0 を中心とした非対称性が存在する．これにより Softsatisficing は，価値が基準より遠いあるいは価値の差が小さい場合はランダム探索を行う．また，価値が基準に近い場合は，より価値を重視した指向性探索を行う．

Softsatisficing では，基準 $\aleph$ の操作も，逆温度 λ の操作に近い振る舞いを示す．行動選択確率 1 と 0 の間隔は，基準 $\aleph$ に制約される．価値に対し基準が低いのであれば即座に満足し，Softsatisficing による行動選択は満足と非満足の 2 値的な挙動となる（図 5.4a）．あらゆる価値より基準が十分に高いのであれば間隔は広がり，Softsatisficing による行動選択は一様分布に近似する（図 5.4b）．



(a) $\aleph = 0$ (b) $\aleph = 1.5$

図 5.4: (a) *aleph* = 0.0 の Softsatisficing と (b) *aleph* = 1.5 の Softsatisficing の行動 a_1 の選択確率

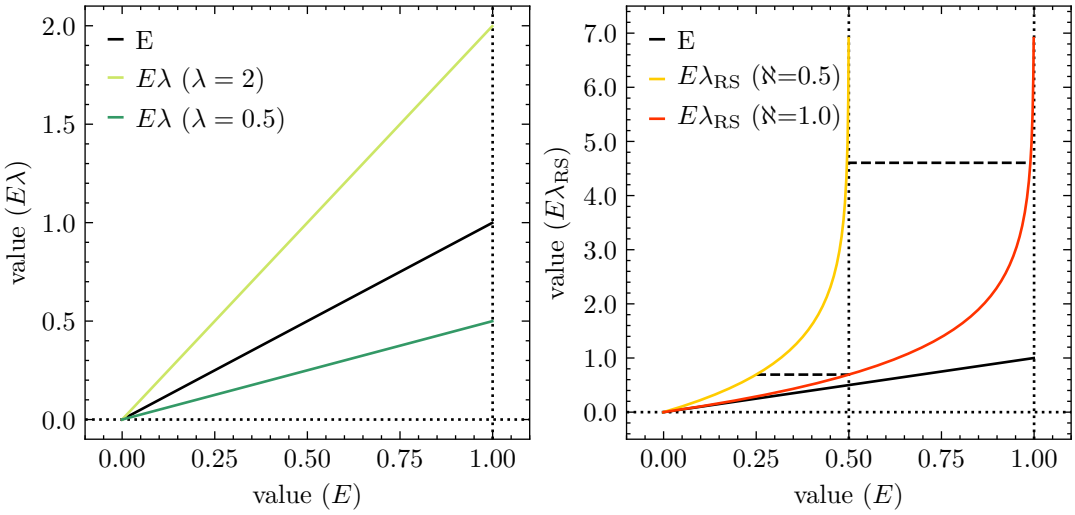
5.3 逆温度関数 λ_{RS} の役割

満足値 E_{RS} のみを考慮すればよいため，式 5.11 の逆温度関数 λ_{RS} の定義は，本来必要ない．しかし，より一般化したとき，満足化による基準を考慮した，適応的な逆温度関数とみなせるため，必要となる．例えば，*RS* の信頼性とは異なるが，$\lambda_{\mathrm{RS}_i} = -\ln(\aleph - E_i)$ など，より単純な定義も可能だ．この場合の逆温度関数 λ_{RS_i} は，より損失を強調する．ここでは，論旨からずれるため，現状の定義の逆温度 $\lambda_{\mathrm{RS}} = -\ln(\aleph - E)/E$ についてのみ触れる．

一般的に Softmax（式 5.1）で扱われる逆温度 λ は，すべての行動に対して一律に適用されるパラメータである．関数などにより動的に調整されるにしても，同様に，すべての行動に対して一律に適用される [Reverdy 14]．Softmax の逆温度 λ は，価値の差の増減を役割とする．$\lambda > 1$ であれば，価値 E の差異を大きく見積もり，Softmax による行動選択は arg max に近づく．$\lambda < 1$ であれば，差異を小さく見積もり，行動選択はランダム選択に近づく．

対して，Softsatisficing の逆温度関数 λ_{RS} は，価値を所与とするため，行動に対して個別のパラメータである．また，基準も所与とするため，逆温度は非満足度（つまり，価値下限から基準までの範囲に対する相対的な量）に応じて自律的に調整される．このとき，非満足度に対する満足値は，同じ値となる．簡単のため，以下では，満足値 $E\lambda_{\mathrm{RS}_i}$ を，

$E = 0$ のとき 0 になるように平行移動している．Softsatisficing では，Softmax と同様に，差が等しければ同様の値となるので，平行移動によって性質は変わらないことに注意して欲しい．このように，逆温度関数 λ_{RS} は，価値と基準の関係性に応じて，共通の満足



(a) 逆温度 λ による価値　　(b) 逆温度関数 λ_{RS} による価値

図 5.5: (a) Softmax の逆温度 λ による価値関数の傾きの制御と (b) Softsatisficing の基準を所与とした逆温度関数 λ_{RS} による価値の変調

値へと価値を変調する役割を担う．

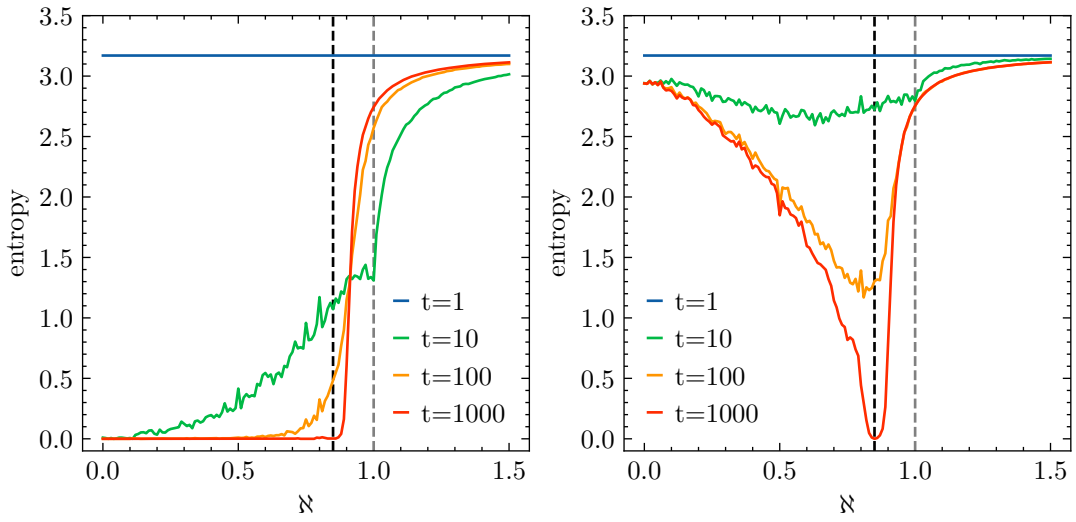
5.4 Satisficing による意思決定: 指向性探索とランダム探索

環境の不確実性を扱う時，人は指向性探索とランダムな探索の両方を行うと示唆されている [Gershman 18, Krueger 17]．先に示した通り，Softsatisficing は，環境の不確実性に対してその両探索を行う．

図 5.6a は，`Dense` における Softsatisficing の行動選択率の平均エントロピーである．基準が低く非満足度の減衰が早いほど，行動選択率のエントロピーは素早く減衰する．また，環境に対して適切な基準を持つ場合，徐々に非満足度が減衰するため，エントロピーはそれに伴い緩やかに減衰する．満足できない状況が続くのであれば，エントロピーは減衰せず，ランダム性を高く保つ．

図 5.6b は，`Dense` における Softsatisficing の平均行動選択率のエントロピーである．環境に対して最適な基準に近いほど，指向性が強く，エントロピーは素早く減衰する．一方で，最適な基準から遠いほど，行動選択はランダムに近づく．

Softmax は，人あるいは動物の意思決定においては，ランダム探索のモデルに相当す



(a) 行動選択率の平均エントロピー　　(b) 平均行動選択率のエントロピー

図 5.6: `Dense` における Softsatisficing による各試行時の (a) 行動選択率の平均エントロピーと (b) 平均行動選択率のエントロピー

る [Gershman 18, Tomov 20]．しかし，Softsatisficing は，ランダム探索のみでなく指向性探索も行えることを示した．むしろ，Softsatisficing に基づく分析によって，個人の行動データからその人の基準を推定することも可能だ．

第 6 章

結論

本論文では，Risk-sensitive Satisficing(RS) の分析に基づき，満足化の意思決定モデルである Softsatisficing を提案した．また，その分析により，満足化による行動の探索が，指向性探索とランダム探索を併せ持つことが示唆された．Softsatisficing は，RS の信頼性の分析モデルであり，2 値でない，緩やかな満足化の確率論的方策である．その行動選択確率は，価値と基準から非満足度を考慮することで満足値により計算され，比較された行動価値の順序によって非対称性を持つ．また，Softsatisficing に導入された逆温度関数は，満足化による意思決定を Softmax で扱う枠組みを提供する．

Softsatisficing の提案により，満足化による意思決定の工学的応用，実証の幅が広がった．人や動物の意思決定が Softsatisficing で記述できるか，あるいは，個人の基準の推定が可能かどうかの検討と，その実証を今後の課題とする．

謝辞

本研究を進めるにあたり，ご協力いただいた皆様に深く感謝いたします．特に，高橋達二教授には，長くご指導いただき，論文執筆にとどまらない多岐にわたるご助言をいただきました．心より感謝申し上げます．

参考文献

[Agrawal 12] Agrawal, S. and Goyal, N.: Analysis of Thompson Sampling for the Multi-armed Bandit Problem, in *Proceedings of the 25th Annual Conference on Learning Theory* (2012)

[Auer 02] Auer, P., Cesa-Bianchi, N., and Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem, *Machine Learning*, Vol. 47, No. 2, pp. 235–256 (2002)

[Badia 20a] Badia, A. P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., and Blundell, C.: Agent57: Outperforming the Atari Human Benchmark, in III, H. D. and Singh, A. eds., *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, pp. 507–517, PMLR (2020)

[Badia 20b] Badia, A. P., Sprechmann, P., Vitvitskyi, A., Guo, D., Piot, B., Kapturowski, S., Tieleman, O., Arjovsky, M., Pritzel, A., Bolt, A., and Blundell, C.: Never Give Up: Learning Directed Exploration Strategies, in *International Conference on Learning Representations* (2020)

[Bendor 09] Bendor, J. B., Kumar, S., and Siegel, D. A.: Satisficing: A 'pretty good' heuristic, *B. E. J. Theor. Econ.*, Vol. 9, No. 1 (2009)

[Genewein 15] Genewein, T., Leibfried, F., Grau-Moya, J., and Braun, D. A.: Bounded Rationality, Abstraction, and Hierarchical Decision-Making: An Information-Theoretic Optimality Principle, *Frontiers in Robotics and AI*, Vol. 2, p. 27 (2015)

[Gershman 15] Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B.: Computational rationality: A converging paradigm for intelligence in brains, minds, and machines, *Science*, Vol. 349, No. 6245, pp. 273–278 (2015)

[Gershman 18] Gershman, S. J.: Deconstructing the human algorithms for exploration, *Cognition*, Vol. 173, pp. 34–42 (2018)

[Goodrich 04] Goodrich, M. A. and Quigley, M.: Satisficing Q-learning: efficient

learning in problems with dichotomous attributes, in *2004 International Conference on Machine Learning and Applications, 2004. Proceedings.*, pp. 65–72 (2004)

[Hessel 18] Hessel, M., Modayil, J., Hasselt, H. V., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M. G., and Silver, D.: Rainbow: Combining Improvements in Deep Reinforcement Learning, in *AAAI* (2018)

[Kapturowski 19] Kapturowski, S., Ostrovski, G., Dabney, W., Quan, J., and Munos, R.: Recurrent Experience Replay in Distributed Reinforcement Learning, in *International Conference on Learning Representations* (2019)

[Krueger 17] Krueger, P. M., Wilson, R. C., and Cohen, J. D.: Strategies for exploration in the domain of losses, *Judgment and Decision Making*, Vol. 12, pp. 104–117 (2017)

[Lai 85] Lai, T. L. and Robbins, H.: Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*, Vol. 6, No. 1, pp. 4–22 (1985)

[Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015)

[Oyo 17] Oyo, K. and Takahashi, T.: Optimization through satisficing with prospects, in *AIP Conference Proceedings*, Vol. 1863, p. 360013 (2017)

[Reverdy 14] Reverdy, P. B., Srivastava, V., and Leonard, N. E.: Modeling Human Decision Making in Generalized Gaussian Multiarmed Bandits, *Proceedings of the IEEE*, Vol. 102, No. 4, pp. 544–571 (2014)

[Reverdy 17] Reverdy, P., Srivastava, V., and Leonard, N. E.: Satisficing in Multi-Armed Bandit Problems, *IEEE Trans. Automat. Contr.*, Vol. 62, No. 8, pp. 3788–3803 (2017)

[Shinriki 20] Shinriki, M., Wakabayashi, H., Kono, Y., and Takahashi, T.: Flexibility of emulation learning from pioneers in nonstationary environments, in *Advances in Intelligent Systems and Computing*, Advances in intelligent systems and computing, pp. 90–101, Springer International Publishing, Cham (2020)

[Simon 55] Simon, H. A.: A Behavioral Model of Rational Choice, *The Quarterly Journal of Economics*, Vol. 69, No. 1, pp. 99–118 (1955)

[Simon 56] Simon, H. A.: Rational choice and the structure of the environment., *Psychological Review*, Vol. 63, No. 2, pp. 129–138 (1956)

[Simon 57] Simon, H. A.: *Models of Man: Social and Rational*, John Wiley and Sons,

Inc., New York (1957)

[Sutton 18] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, A Bradford Book, Cambridge, MA, USA (2018)

[Takahashi 16] Takahashi, T., Kohno, Y., and Uragami, D.: Cognitive Satisficing: Bounded Rationality in Reinforcement Learning, *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 31, No. 6, pp. AI30–M_1–11 (in Japanese) (2016)

[Tamatsukuri 19] Tamatsukuri, A. and Takahashi, T.: Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function, *Biosystems.*, Vol. 180, pp. 46–53 (2019)

[Thompson 33] Thompson, W. R.: On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples, *Biometrika*, Vol. 25, No. 3/4, pp. 285–294 (1933)

[Tomov 20] Tomov, M. S., Truong, V. Q., Hundia, R. A., and Gershman, S. J.: Dissociable neural correlates of uncertainty underlie different exploration strategies, *Nat. Commun.*, Vol. 11, No. 1, p. 2371 (2020)

[Wakabayashi 21] Wakabayashi, H., Kamiya, T., and Takahashi, T.: Balancing Policy Improvement and Evaluation in Risk-Sensitive Satisficing Algorithm, *Advances in Intelligent Systems and Computing*, Vol. 1357, pp. 175–182 (2021)

[岡田 01] 岡田 浩之, 山川 宏, 大森 隆司：環境同定と報酬獲得のトレードオフを解消する報酬・嫌悪の二次元評価強化学習の提案, 日本ロボット学会誌, Vol. 19, No. 2, pp. 244–251 (2001)

[片山 98] 片山 晋, 武市 正人, 小林 重信：満足化原理に基づく強化学習のための確率的探査戦略, 人工知能, Vol. 13, No. 6, pp. 971–980 (1998)