

TRENDS IN MOBILEN UND VERTEILTEN SYSTEMEN

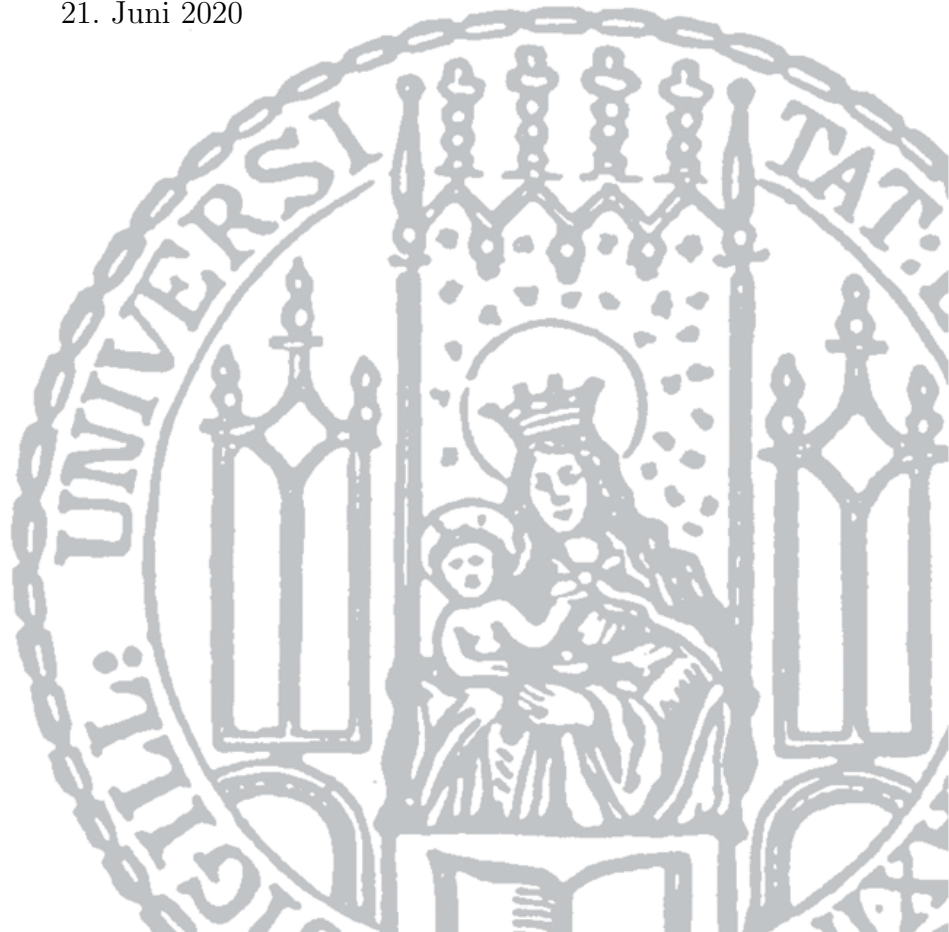
# Curiosity und Diversity im Kontext von Reinforcement Learning - Ein Vergleich

BEARBEITER: David Jonathan Müller  
Lukas Johannes Rieger

BETREUER: M. Sc. Thomas Gabor

AUFGABENSTELLER: Prof. Dr. Claudia Linnhoff-Popien

Dokument erstellt  
21. Juni 2020





# Curiosity und Diversity im Kontext von Reinforcement Learning - Ein Vergleich

## **Zusammenfassung**

Dieses Dokument ist die Vorlage für Seminararbeiten am Lehrstuhl für  
Mobile und Verteilte Systeme.

## **Inhaltsverzeichnis**

<b>1</b>	<b>Einleitung</b>	<b>3</b>
<b>2</b>	<b>Grundlagen</b>	<b>4</b>
2.1	Reinforcement Learning im Allgemeinen . . . . .	4
2.2	Markov Entscheidungsprozesse . . . . .	4
2.3	Bayes'sche Inferenz . . . . .	4
<b>3</b>	<b>Das Curiosity Modell nach Schmidhuber</b>	<b>5</b>
<b>4</b>	<b>Diversity im Reinforcement Learning</b>	<b>6</b>
4.1	Klärung wichtiger Begriffe aus der Informationstheorie . . . . .	6
4.2	Funktionsweise . . . . .	7
4.3	Experimente und Beispiele . . . . .	8
4.4	Verwendung der gelernten Fähigkeiten . . . . .	9
<b>5</b>	<b>Curiosity und Diversity im Vergleich</b>	<b>12</b>
<b>6</b>	<b>Verwandte Arbeiten</b>	<b>13</b>
<b>7</b>	<b>Schluss</b>	<b>14</b>
	<b>Literatur</b>	<b>15</b>



# 1 Einleitung

Das wohl herausstechendste Merkmal des Menschen ist seine Fähigkeit, komplexe Probleme in unterschiedlichsten Umgebungen und oftmals ohne vorheriges Wissen zu lösen. Diese Fähigkeit ist nicht nur für organische Lebewesen, sondern auch für künstliche Aktoren in einer Vielzahl an Szenarien äußerst nützlich. Ein häufig verwendeter Ansatz im Bereich der künstlichen Intelligenz stellt das “Reinforcement Learning” dar. Dieser Ansatz erlaubt es Agenten, diverse Aktionen im Kontext ihrer jeweiligen Umgebung zu erlernen und zu verbessern, ohne dabei notwendigerweise durch eine dritte Instanz überwacht werden zu müssen.[5] Ein zentraler Problembereich stellt hierbei das Finden und Konfigurieren von “Belohnungen”, die dem Agenten, ähnlich wie beim Menschen, Rückmeldung über die von ihm ausgeführten Aktionen geben soll.

Da in den meisten realen Anwendungsfällen externe Belohnungen nur spärlich oder überhaupt nicht vorhanden sind, benötigt man zusätzliche Methoden, um einen Lernfortschritt zu erzielen.[6] Intuitiv lässt sich ein Lernanreiz entweder als extrinsisch, oder als intrinsisch definieren. Dient der Lernvorgang dem Erreichen eines externen, vorgegebenen Ziels, lässt sich dieser als extrinsisch kategorisieren. In diesem Falle fällt es ebenfalls leicht, die Belohnung eines Aktors in Abhängigkeit seines Zielfortschritts zu definieren. Da in realistischen Szenarien eine solche direkte Bewertung meist nur schwer vorzunehmen ist, muss eine etwaige Belohnung ebenfalls intrinsische Anreize in Betracht ziehen. Von solchen intrinsischen Anreizen spricht man dann, wenn die zugrundeliegende Aktion *inhärent* sinnvoll erscheint, also nicht von externen Gegebenheiten abhängt.

So scheint es etwa für den Menschen intuitiv sinnvoll, in einer unbekannten Umgebung möglichst viele unterschiedliche Vorgehensweisen auszuprobieren und sich dabei auch Strategien zu merken, welche vielleicht erst zu einem späteren Zeitpunkt Anwendung finden. Ein solches Vorgehen wäre auch bei einer künstlichen Intelligenz wünschenswert. Zu diesem Zweck existieren unterschiedliche Ansätze, welche sich allerdings in gewissen Punkten durchaus ähnlich sind.

In dieser Arbeit betrachten und vergleichen wir zwei Methodiken, welche das klassische Reinforcement Learning in dieser Hinsicht erweitern. Zu diesem Zweck wird zunächst der Begriff Reinforcement Learning erklärt. Die anschließenden Kapitel fokussieren sich auf zwei Ansätze zur Bewertung des Lernfortschritts. In diesem Rahmen stellen wir zuerst den Ansatz der “Curiosity” nach [7] vor. Wir betrachten außerdem das Konzept der “Diversity” und dessen Anwendung im Reinforcement Learning. Diese stellen wir daraufhin vergleichend gegenüber und gehen auf deren Gemeinsamkeiten und Kernunterschiede ein.

## **2 Grundlagen**

### **2.1 Reinforcement Learning im Allgemeinen**

### **2.2 Markov Entscheidungsprozesse**

### **2.3 Bayes'sche Inferenz**

### **3 Das Curiosity Modell nach Schmidhuber**

## 4 Diversity im Reinforcement Learning

In diesem Kapitel beschäftigen wir uns damit, wie das Konzept von Diversity im Bereich des Reinforcement Learning (RL) Anwendung finden kann. Ziel ist, dass ein RL Agent in einer unüberwachten Phase zunächst Fähigkeiten erlernt, welche das Bewältigen von Aufgaben in der darauf folgenden, überwachten Phase erleichtern sollen.

Dieses Kapitel stützt sich zu einem Großteil auf die wissenschaftliche Arbeit *Diversity is all you need: Learning skills without a reward function*[3]. Falls nicht anders angegeben, wurden die Informationen hieraus entnommen.

### 4.1 Klärung wichtiger Begriffe aus der Informationstheorie

Im Folgenden werden Begriffe aus der Informationstheorie verwendet, welche es zunächst zu klären gilt. Wir betrachten *Entropie* und *Transinformation* nach [1] und [8].

Die *Entropie* beschreibt in der Informationstheorie den mittleren Informationsgehalt bzw. die Ungewissheit einer Quelle. Ist beispielsweise bei einer Ereignismenge jedes Ereignis gleich wahrscheinlich, so ist die Ungewissheit maximal [8].

Nach [1] besitzt eine diskrete Zufallsvariable  $X$  mit dem Zeichenvorrat  $\mathcal{X}$  und der Wahrscheinlichkeitsfunktion  $p(x)$  die *Entropie*

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log p(x)$$

Diese lässt sich nach [1] auch über den Erwartungswert  $E$  berechnen unter Verwendung von

$$H(X) = E_p \log \frac{1}{p(X)}$$

, woraus unter Anwendung der Rechenregeln für Logarithmus und Erwartungswert trivial

$$H(X) = -E_p \log p(X) \quad (1)$$

folgt.

Desweiteren ist die Formel für die *bedingte Entropie* nach [1] gegeben durch

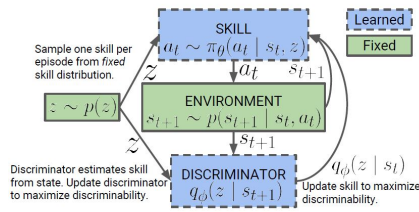
$$H(Y|X) = -E_{p(x,y)} \log p(Y|X) \quad (2)$$

Die *Transinformation* beschreibt die Menge an Information, die eine Zufallsvariable über eine andere enthält bzw. die Reduktion der Ungewissheit aufgrund des Wissens um die jeweils andere Zufallsvariable [1].

Mathematisch wichtig für uns ist lediglich der Zusammenhang zwischen *Transinformation* und *Entropie*, welcher nach [1] gegeben ist durch

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (3)$$





**Algorithm 1: DIAYN**

---

```

while not converged do
  Sample skill  $z \sim p(z)$  and initial state  $s_0 \sim p_0(s)$ 
  for  $t \leftarrow 1$  to  $steps\_per\_episode$  do
    Sample action  $a_t \sim \pi_\theta(a_t | s_t, z)$  from skill.
    Step environment:  $s_{t+1} \sim p(s_{t+1} | s_t, a_t)$ .
    Compute  $q_\phi(z | s_{t+1})$  with discriminator.
    Set skill reward  $r_t = \log q_\phi(z | s_{t+1}) - \log p(z)$ 
    Update policy ( $\theta$ ) to maximize  $r_t$  with SAC.
    Update discriminator ( $\phi$ ) with SGD.
  
```

---

Abbildung 1: “Diversity is all you need” Algorithmus

Quelle: [3]

## 4.2 Funktionsweise

Das selbstständige Erlernen von brauchbaren Fähigkeiten, so wie in [3] beschrieben, baut auf drei Grundideen auf.

(1) Unterschiedliche Fähigkeiten sollten andere Zustände besuchen, sodass ihre Unterscheidbarkeit gewährleistet ist.

(2) Um besagte Fähigkeiten zu unterscheiden, werden nicht die Aktionen, sondern die Zustände betrachtet. Das liegt daran, dass Aktionen, welche die Umgebung nicht beeinflussen, für einen Beobachter nicht erkennbar sind. Das Paper verdeutlicht dies mit einem treffenden Beispiel. Für einen außenstehenden Betrachter ist nicht ersichtlich, wie fest ein Roboterarm eine Tasse in der Hand hält, falls sich diese nicht bewegt.

(3) Schließlich soll erreicht werden, dass die Fähigkeiten so vielfältig (*diverse*) wie möglich sind. Die Idee ist, dass unterscheidbare Fähigkeiten mit einer hohen Entropie einen Zustandsraum abdecken, welcher sich weit entfernt von anderen Fähigkeiten befindet.

Um diese Punkte zu realisieren, definieren wir zunächst die Zufallsvariablen  $S$  und  $A$  für Zustände und Aktionen. Sei nun  $Z \sim p(z)$  eine latente Variable, unter deren Bedingung die Strategie definiert wird. Bei fixem  $Z$  sprechen wir hier von einer ”Fähigkeit”. Entropie sowie Transinformation werden zur Basis  $e$  berechnet.

Nun soll die Transinformation zwischen Fähigkeiten und Zuständen,  $I(S; Z)$ , maximiert werden um sicherzustellen, dass die Fähigkeit die vom Agent durchlaufenen Zustände vorgibt. Anders gesagt wird sichergestellt, dass aus den besuchten Zuständen auf die Fähigkeit geschlossen werden kann.

Außerdem minimieren wir die Transinformation zwischen Fähigkeit und Aktionen bei gegebenen Zustand,  $I(A; Z|S)$ . So wird garantiert, dass nicht Aktionen, sondern Zustände für die Unterscheidung von Fähigkeiten verwendet werden.

Maximiert wird auch die Entropie  $H(A|S)$ .

Insgesamt maximieren wir nach [3] also

$$\mathcal{F}(\theta) \triangleq I(S; Z) + H(A|S) - I(A; Z|S) \quad (4)$$

$$\begin{aligned} &= (H(Z) - H(Z|S)) + H(A|S) - (H(A|S) - H(A|S, Z)) \\ &= H(Z) - H(Z|S) + H(A|S, Z) \end{aligned} \quad (5)$$

Der Term wurde unter Verwendung von (3) umgeformt.

Da sich  $p(z|S)$  nicht genau berechnen lässt, wird das Folgende mit einem discriminator  $q_\phi$  approximiert. Mit der Jensenschen Ungleichung haben wir laut [3] eine untere Schranke  $\mathcal{G}(\theta, \phi)$  für  $\mathcal{F}(\theta, \phi)$ :

$$\begin{aligned} \mathcal{F}(\theta) &= H(A|S, Z) - H(Z|S) + H(Z) \\ &= H(A|S, Z) + E_{z \sim p(z), s \sim \pi(z)}(\log p(z|s)) - E_{z \sim p(z)}(\log p(z)) \\ &\geq H(A|S, Z) + E_{z \sim p(z), s \sim \pi(z)}(\log q_\phi(z|s) - \log p(z)) \triangleq \mathcal{G}(\theta, \phi) \end{aligned} \quad (6)$$

In (6) wurden die Entropien nach (1) und (2) umgeformt.

Nach der unüberwachten Trainingsphase verfügt der Agent über eine Sammlung von verschiedenartigsten Fähigkeiten, von denen vermutlich einige nutzlos sind. Aufgrund der implementierten Diversity müssen jedoch Fähigkeiten existieren, welche sich stark von den unbrauchbaren unterscheiden. So ist es intuitiv einleuchtend, dass es sich hierbei um sinnvolle Fähigkeiten handelt.

### 4.3 Experimente und Beispiele

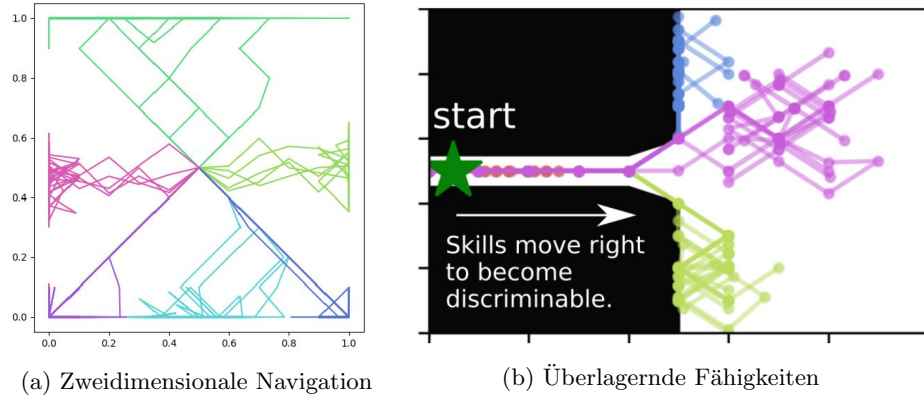


Abbildung 2: TODO

Quelle: [3]

Wie bereits mehrfach angemerkt ist das Ziel das eigenständige Erlernen von vielfältigen Fähigkeiten. [3] liefert zunächst das simple Beispiel einer zweidimen-

sionalen Navigation in einem leeren Raum 2a. Hier ist sehr anschaulich zu erkennen, wie die farblich kodierten Fähigkeiten unterschiedliche Zustandsräume abdecken.

Einen Schritt weiter gedacht zeigt das Beispiel 2b, dass sich Fähigkeiten unter Umständen auch zeitweilig überlagern können, solange sie schlussendlich unterscheidbare Zustandsräume abdecken.

In beiden Beispielen ist gut zu erkennen, wie sich die Fähigkeiten gegenseitig “abstoßen” und so automatisch ein Großteil des Zustandsraums abgedeckt wird. Die erkundeten Zustände liegen außerdem gleichmäßig über dem Zustandsraum verteilt.



Abbildung 3: Experiment “Half Cheetah”

Quelle: [2]

Interessanter wird es bei den komplexeren Versuchsaufbauten. Wir betrachten im Folgenden den vom Paper [3] als “Half Cheetah” betitelten Agenten. Hierbei handelt es sich um ein hundartiges, zweidimensionales Wesen, welches die Kontrolle über sein Vorder- und Hinterbein besitzt (siehe 3).

Ohne jegliche externe Belohnung hat dieser Agent gelernt, nach vorne und hinten zu rennen beziehungsweise zu gehen. Außerdem besitzt er die Fähigkeit, einen Vorwärtssalto zu machen. Wir empfehlen zu besseren Veranschaulichung die Einsicht der Videos auf [2].

Eine Reward-Funktion, welche ein solches Verhalten hervorruft, manuell zu definieren ist nicht einfach.

#### 4.4 Verwendung der gelernten Fähigkeiten

Nachdem nun eine Sammlung von Fähigkeiten erlernt wurde stellt sich nun die Frage, wie sich diese effektiv nutzen lässt.

[3] führt hier als erstes an, dass der DIAYN Algorithmus als ein Vortraining genutzt werden kann. Danach extrahiert man die Fähigkeit mit dem größten Reward und passt entsprechend die Initialgewichte der eigentlichen Policy an. 4 zeigt für das Beispiel des Half Cheetahs, dass sich bei Initialisierung mit einer vorgelernten Fähigkeit (blaue Linie) schneller größere Rewards erzielen lassen als bei einer zufälligen Initialisierung (orangene Linie).

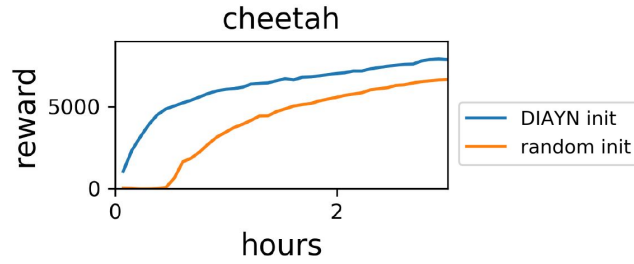


Abbildung 4: Policy Initialisierung

Quelle: [3]

Für uns interessanter ist allerdings die Verwendung für Hierarchisches Reinforcement Learning (HRL).

Nach [4] lernen HRL Methoden eine Policy, welche aus mehreren Schichten besteht. Jede Ebene kontrolliert hierbei eine andere temporäre Abstraktionsebene. Dies ermöglicht es dem Agenten, nicht nur Basisoperationen auszuführen, sondern auch komplexere Aktionen wie zum Beispiel Sequenzen von Operationen einer niedrigeren Ebene.

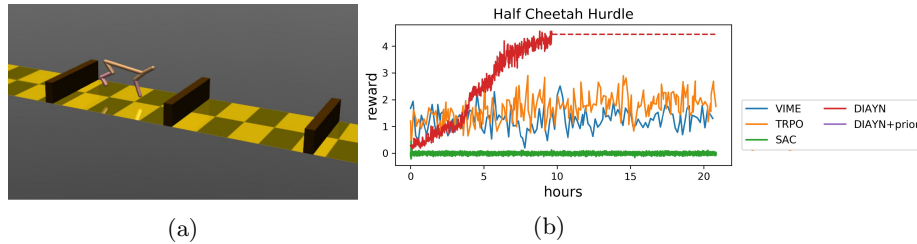


Abbildung 5: DIAYN für HRL

Quelle: [3]

Laut [3] eignet sich DIAYN hervorragend als Grundstein für HRL. Für ein Experiment diesbezüglich betrachten wir wieder den Half Cheetah. Dieser hat nun die Aufgabe, über Hürden zu springen (siehe 5a) und erhält Rewards für deren Überwinden.

Um die gelernten Fähigkeiten für HRL zu benutzen, wird DIAYN um einen "meta-controller" erweitert welcher kontrolliert, welche 10 Fähigkeiten jeweils als nächstes ausgeführt werden.

Das Experiment wird außerdem mit aktuellen Reinforcement Learning Algorithmen (VIME, TRPO und SAC) durchgeführt.

Wie die Grafik 5b zeigt, schneiden diese im Experiment sehr schlecht ab und es gibt quasi keine merkliche Steigerung des Rewards. Im Gegensatz dazu

zeigt der Ansatz mit dem DIAYN Algorithmus schnell deutliche Fortschritte und übertrifft merklich die anderen.

Aufgrund der Knappheit von externen Rewards gestaltet sich diese Aufgabe extrem schwierig für traditionelle RL Algorithmen. Diese müssen quasi zufällig eine Hürde übersteigen, um eine Belohnung von außen zu bekommen.

Dies zeigt nach [3], dass das eigenständige Lernen von Fähigkeiten einen effektiven Mechanismus bereit stellt, um bei Herausforderungen beim Erkunden und mit geringem externen Reward Lernerfolge zu erzielen.

## 5 Curiosity und Diversity im Vergleich

## 6 Verwandte Arbeiten

## 7 Schluss



## Autorenschaft

David Jonathan Müller hat die Abschnitte 1, 6 und ?? verfasst. Lukas Johannes Rieger hat die Abschnitte ?? und ?? verfasst. Den Abschnitt 7 haben beide Autoren gemeinsam verfasst. (In einer wissenschaftlichen Arbeit ohne Prüfungszweck stünde hier eine Danksagung!)

## Literatur

- [1] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [2] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. Eingesehen am 20.06.2020.
- [3] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- [4] Yannis Flet-Berliac. The promise of hierarchical reinforcement learning. *The Gradient*, 2019.
- [5] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [6] Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. *arXiv preprint arXiv:1810.02274*, 2018.
- [7] Juergen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes, 2008.
- [8] Martin Werner. *Information und Codierung: Grundlagen und Anwendungen*. Springer-Verlag, 2009.