FrozenLake 4

UWAGA

Wczytaj do Colab plik **frozen_lake_slippery.py** lub **frozen_lake.py** (intrukcja w pliku **COLAB_instrukcja.pdf**).

Wczytaj też plik **plot_utils.py**.

In [1]:

```
from frozen_lake import FrozenLakeEnv
#from frozen_lake_slippery import FrozenLakeEnv
import numpy as np
from plot_utils import plot_values

env = FrozenLakeEnv()
```

Objaśnienie algorytmu

Algorytm wygląda następująco:

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   - $\Delta \leftarrow 0$
   - Loop for each $s \in \mathcal{S}$:
     - $v \leftarrow V(s)$
     - $V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))\big[r + \gamma V(s')\big]$
     - $\Delta \leftarrow \max(\Delta, |v - V(s)|)$

   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   - *old-action* $\leftarrow \pi(s)$
   - $\pi(s) \leftarrow \operatorname{argmax}_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$
   - If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*

   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

W algorymie można wyróżnić dwa główne bloki.

Blok 1

Algorytm wylicza **wartość zwrotów *V(s)*** dla wszystkich stanów ***s*** przy zadanej **polityce deterministycznej** oznaczonej ***pi*** (wyliczenie ***V*** odbywa się w **punkcie 2** algorytmu).

Blok 2

Algorytm znajduje **politykę deteministyczną *pi*** na podstawie wyliczonych wcześniej wartości ***V(s)***(**punkt 3** algorytmu). Warto zwrócić uwagę, że zarówno ***V(s)*** jak i polityka ***pi*** są początkowo dowolne (***punkt 1*** algorytmu).

To co ciekawe odbywa się w nastepującej **pętli**:

Algorytm wylicza ***V(s)*** dla zadanej **polityki deterministycznej *pi***, a następnie wyliczone wartości ***V(s)*** wykorzystuje **do modyfikacji polityki *pi***. Zmodyfikowana polityka służy do ponownego wyliczenia ***V(s)*** itd.

W efekcie takich wzajemnych modyfikacji ***V*** i ***pi*** znaleziona zostaje **polityka optymalna** (najlepsza - gwarantująca najlepsze zwroty) dla danego problemu.... ***MAGIA*** :)

Warto zwrócić uwagę na **punkt 3** w algorytmie:

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
      old-action $\leftarrow \pi(s)$
      $\boxed{\pi(s) \leftarrow \operatorname{argmax}_a \sum_{s',r} p(s',r \mid s,a)\big[r + \gamma V(s')\big]}$
      If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

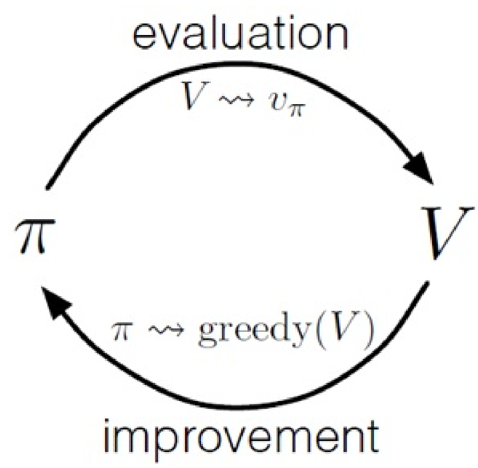
W formule w ramce użyta jest funkcja **argmax**. Jej definicja jest następująca:

$$\underset{x \in S \subseteq X}{\arg\max} f(x) := \{x \mid x \in S \wedge \forall y \in S : f(y) \leq f(x)\}.$$

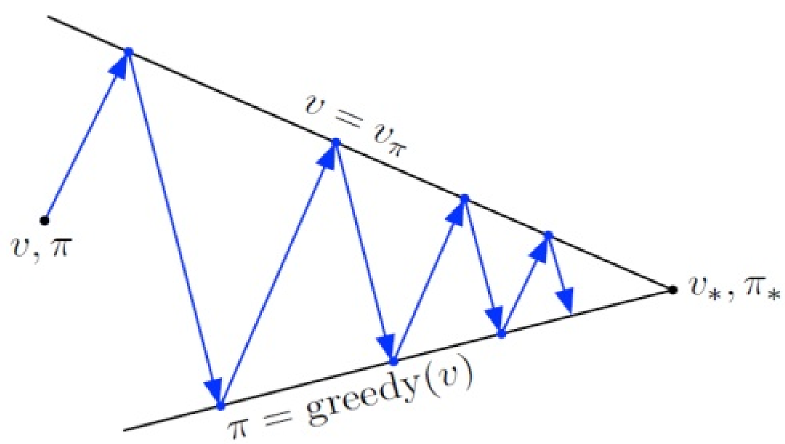
argmax f(x) zwraca wartość argumentu ***x*** dla którego funkcja ***f*** osiąga **wartość maksymalną**.

Linika oznaczona powyżej **czerwoną** ramką oznacza, że **nowa polityka *pi*** przypisuje stanowi ***s*** akcję, która (dla **aktualnych wartości *V***) daje **największy zwrot**! Czyli stosowana jest tutaj **strategia zachłanna (greedy)**.

Petlę w algotymie możemy zobrazować następująco:



Wynikiem działania algorytmu jest **optymalna polityka** (deterministyczna) **pi*** i **V*** dla tej polityki.



Polityka deterministyczna

W implemntacji algorytmu będziemy stosowali **politykę deteministyczną**. Jest to polityka, która każdemu stanowi ***s*** przypisuje akcję ***a***, która ma być wykonana w tym stanie. Możemy ją zdefiniować następująco (***env.nS*** to liczba stanów w środowisku ***env***):

In [2]:

```
pi = np.random.randint(0, env.nA, size=env.nS)

print(pi)
```

```
[0 3 2 2 0 3 1 3 0 0 0 2 2 3 1 2]
```

Polityka ta dla każdego stanu **s** określa akcję.

Przykład: akcja wykonana w stanie 4 (stany numerowane są od 0 do 15):

In [3]:

```
pi[4]
```

Out[3]:

```
0
```

Wyliczenie V dla zadanej polityki

Zajmijmy się **punktem 2** algorytmu czyli wyliczeniem ***V(s)*** dla zadanej **polityki deterministycznej *pi***. Problemu tego **dla polityki stochastycznej** dotyczyło **zadanie 3** z **RL_lab_4.pdf**.

Polecenie 1 (do uzupełnienia)

Uzupełnij poniższą definicję funkcji zwracającej ***V*** dla zadanej **polityki deterministycznej *policy***.

In [4]:

```
def det_policy_evaluation(env, policy, gamma=0.9, theta=1e-8):
    V = np.zeros(env.nS)

    while True:
        delta = .0
        for state in range(env.nS):
            Vs = 0
            for next_state in range(len(env.P[state][policy[state]])):

                prob, next_state, reward, done = env.P[state][policy[state]][next_state]
               #DO UZUPEŁNIENIA
                Vs += prob * (reward + gamma * V[next_state])


            delta = max(delta, np.abs(V[state] - Vs))
            V[state] = Vs
        if delta < theta:
            break

    return V
```

UWAGA: warto zwrócić uwagę, gdzie w powyższym kodzie użyta jest polityka deterministyczna **policy**:

```
for next_state in range(len(env.P[state][policy[state]])):
    #deterministic policy
```

```
    prob, next_state, reward, done = env.P[state][policy[state]][next_state]

    #DO UZUPEŁNIENIA
```

Testujemy działanie funkcji dla **polityki deterministycznej *pi***:

In [5]:

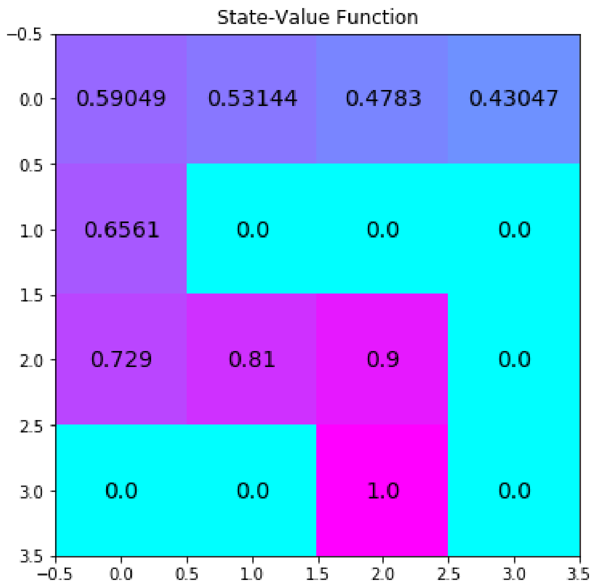
```
pi = np.array([1,0,0,0,1,0,0,0,2,2,1,0,0,0,2,0])
V = det_policy_evaluation(env,pi)
print(V)
```

```
[0.59049    0.531441   0.4782969  0.43046721 0.6561     0.
 0.         0.         0.729      0.81       0.9        0.
 0.         0.         1.         0.        ]
```

Wartości *V(s)* można przedstawić graficznie korzystając z funkcji **plot_values()** z **biblioteki plot_utils**:

In [6]:

```
plot_values(V)
```



Znalezienie polityki dla zadanego V metodą zachłanną

Zajmijmy się teraz **punktem 3** algorytmu czyli znalezieniem **polityki deterministycznej *pi*** dla danego ***V***. Interesuje nas następujący fragment kodu:

3. Policy Improvement
   $\textit{policy-stable} \leftarrow \textit{true}$
   For each $s \in \mathcal{S}$:
      $\textit{old-action} \leftarrow \pi(s)$
      $\pi(s) \leftarrow \arg\max_a \boxed{\sum_{s',r} p(s',r|s,a)\big[r + \gamma V(s')\big]}$
      If $\textit{old-action} \neq \pi(s)$, then $\textit{policy-stable} \leftarrow \textit{false}$
   If $\textit{policy-stable}$, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Warto zauważyć, że w powyższym kodzie nie tylko jest znajdowana nowa polityka, ale **także jest ona porównywana z dotychczasową**. Jeżeli obie są takie same, wówczas algorytm kończy działanie. W przeciwnym razie następuje powrót do **punktu**

dotychczasową. I jeżeli obie są takie same, wówczas algorytm kończy działanie. W przeciwnym razie następuje powrót do **punktu 2**.

Polecenie 2 (do uzupełnienia)

Znalezie polityki dla danego ***V*** można "zamknąć" w funkcji **det_policy_iteration**, która jako argumenty otrzyma ***env***, ***V*** i wartość ***gamma***, a zwróci nam wyliczoną politykę. Warto zwrócić uwagę na to, że **formuła oznaczona powyżej czerwoną ramką** była wykorzystywana w **Zadaniu 1 z części 5**. Czyli możemy wykorzystać funkcję już zdefiniowaną (w **FrozeLake_3.ipynb**). Uzupełnij jej definicję poniżej:

In [22]:

```
def Q_from_V(env, V, state,gamma=0.99):
    Q = np.zeros(env.nA)

    for action in range(env.nA):
        for next_state in range(len(env.P[state][action])):
            prob, next_state, reward, done = env.P[state][action][next_state]
            Q[action] += prob * (reward + gamma * V[next_state])

    return Q
```

Czas na funkcję **det_policy_iteration**:

In [23]:

```
def det_policy_iteration(env, V, gamma=0.99):
    policy = np.zeros([env.nS])

    for state in range(env.nS):
        #DO UZUPEŁNIENIA
        policy[state] = np.argmax(Q_from_V(env,V,state))


    return policy
```

Przetestuj działanie funkcji **det_policy_iteration** dla poniższego ***V***:

In [21]:

```
V = np.array([0.1,0.5,0.8,0.2,0.2,0.,0.4,0.,0.3,0.5,0.8,0.,0.,0.1,1.,0.])
pi = det_policy_iteration(env,V)
print(pi)
```

```
[2. 2. 3. 0. 1. 0. 1. 0. 2. 2. 1. 0. 0. 2. 2. 0.]
```

Implementacja algorytmu

W implementowanym algorytmie po znalezieniu nowej polityki następuje **porównanie** jej ze starą polityką:

$$\text{If } \textit{old-action} \neq \pi(s), \text{ then } \textit{policy-stable} \leftarrow \textit{false}$$

Polityki są u nas zapisane **w tablicach**. Musimy mieć zatem **metodę porównującą dwie tablice** i zwracającą **True/False** jeżeli tablice polityk **są/nie są** identyczne.

Polecenie 3 (do uzupełnienia)

Uzupełnij poniższą funkcję pozwalającą porównywać dwie polityki (tablice) będące jej argumentami i zwracającą **True/False**:

In [15]:

```
def compPolicy(p1,p2):
    comp = (p1 == p2)
    equal_arrays = comp.all()
    return equal_arrays

    #DO UZUPEŁNIENIA
```

Przetestuj działanie funkcji dla dwóch przykładowych polityk:

In [16]:

```
pi_1 = np.array([3,2,0,3,0,3,3,0,1,3,3,0,0,3,0,0])
pi_2 = np.array([3,2,0,3,0,0,1,0,1,3,3,0,0,3,0,0])

print(compPolicy(pi_2,pi_1))
```

```
False
```

Mamy już wszystkie elementy wymagane do implementacji całego **algorytmu**.

Polecenie 4 (do uzupełnienia)

Uzupełnij poniższą petlę tak, aby otrzymać **pełną implementację algorytmu**:

In [17]:

```
pi = np.array([1,0,0,0,1,0,0,0,2,2,1,0,0,0,2,0])
V = np.array([0.1,0.5,0.8,0.2,0.2,0.,0.4,0.,0.3,0.5,0.8,0.,0.,0.1,1.,0.])

while True:
    #DO UZUPEŁNIENIA
    V = det_policy_evaluation(env,pi)
    pi = det_policy_iteration(env,V)
    pi_temp = pi
    if compPolicy(pi_temp, pi):
        break
```

Po wykonaniu powyższego kodu w zmiennej ***pi*** będzie zapisana **optymalna polityka**. Wypiszmy ją:

In [18]:

```
print(pi)
```

```
[1. 0. 0. 0. 1. 0. 1. 0. 2. 2. 1. 0. 0. 2. 2. 0.]
```

Polecenie 5 (do uzupełnienia)

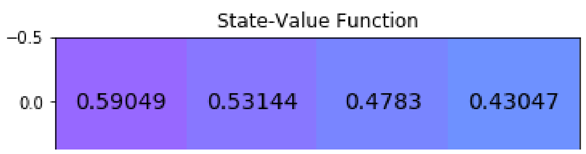
Sprawdź czy **znaleziona polityka** jest optymalna. Odpowiedź uzasadnij.

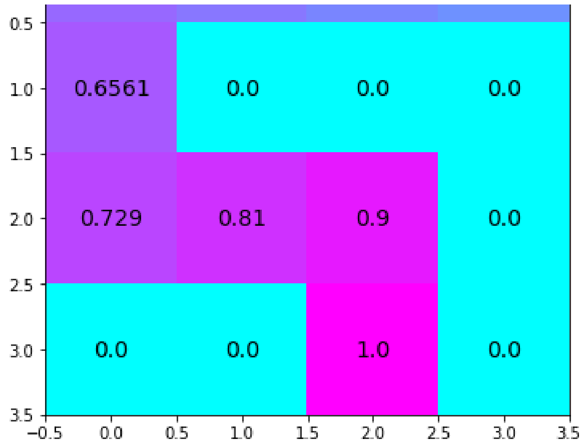
TUTAJ WPISZ ODPOWIEDŹ:

Zobaczmy jak wyglądają **wartości oczekiwane zwrotów V** dla znalezionej polityki:

In [19]:

```
plot_values(V)
```





Polecenie 6 (do uzupełnienia)

Sprawdź czy znaleziona powyżej **polityka optymalna** zmienia się jeżeli przyjmiemy dwie różne wartości parametru **gamma 0.1** i **0.99**.

TUTAJ WPISZ ODPOWIEDŹ:

In [ ]:

```
Z moich obserwacji wynika ze polityka sie nie zmienia
```