|  |
| --- |
|  |
| Laboration 2 |
| Rapport |
|  |
|  |
|  |
|  |

**Innehåll**

[1 Instruktioner - 1 -](#_Toc57760901)

[2 BridgeGrid (Value Iteration) - 1 -](#_Toc57760902)

[3 DiscountGrid (Value Iteration) - 2 -](#_Toc57760903)

[4 BridgeGrid (Q-Learning) - 2 -](#_Toc57760904)

# Instruktioner

Denna rapporten hör till laboration 2 (MDPer och RL). I laborationsdokumnetet ombeds ni fylla i ett antal uppgifter i denna rapporten. När ni lämnar in laborationen skall denna ifyllda rapporten finnas med i arkivfilen.

# BridgeGrid (Value Iteration)

Byt ut värdet för antingen *discount* **eller** *noise* parametern i nedanstående tabell. Med det modifierade värdet skall value iteration agenten kunna följa den genererade policyn för att korsa bron och lämna Bridgeworld brädet via terminaltillståndet +10.00. Förklara också varför parameterinställning leder till önskad policy.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Värde** | **Förklaring** |
| discount | 0.9 | Vi justerade endast noise parametern och då till ett mycket litet värde eftersom noise avgör hur ofta en agent hamnar i ett oavsiktligt tillstånd. |
| noise | 0.00001 |
|  | |

Uppdateringsformeln för value iteration återges nedan§

och är ekvivalent med

där

# DiscountGrid (Value Iteration)

Fyll i de tre parametervärdena *discount*, *noise* och *livingReward* (*reward*) som leder till önskad policy och förklara även varför dessa parameterinställningar leder till önskad policy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Uppgift** | **Parameter** | **Värde** | **Förklaring** |
| a | discount | 0.0 |  |
| noise | 0.0 |
| reward | 0.0 |
|  | |
| b | discount | 0.0 |  |
| noise | 0.0 |
| reward | 0.0 |
|  | |
| c | discount | 0.0 |  |
| noise | 0.0 |
| reward | 0.0 |
|  | |
| d | discount | 0.0 |  |
| noise | 0.0 |
| reward | 0.0 |
|  | |
| e | discount | 0.0 |  |
| noise | 0.0 |
| reward | 0.0 |
|  | |

# BridgeGrid (Q-Learning)

Fyll i de två parametervärdena *epsilon* och *alpha* som leder till den optimala policyn under 50 träningsepisoder. Om den optimala policyn inte kan hittas med någon parameterkombination under 50 träningsepisoder, ange värdet *GÅR EJ* för de båda parametrarna. Motivera också erat svar, dvs varför hittas den optimala policyn med en viss parameterkombination eller varför går det inte att hitta den optimala policyn med någon parameterkombination?

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Värde** | **Förklaring** |
| epsilon | 0.0 |  |
| alpha | 0.0 |
|  | |