:نمونه ای از بهینه سازی خط مشی مسئله Frozen lake:

خلاصه: ما از "روش تابع ارزش بهینه" که برای مسئله دریاچه یخ زده استفاده می شود، استفاده می کنیم، اما برای روشی متفاوت

هدف هدف ما بهینه سازی تولید نفت است. با وجود تفاوت های ظاهری، ما می خواهیم ببینیم که

دو مشکل (دریاچه یخ زده و بهینه سازی تولید نفت) را می توان با یک بهینه سازی مواجه کرد

.رویکرد

مقدمه: نمونه ای معمولی از حل مسئله با رویکرد یادگیری تقویتی است.

در این آموزش از "مسئله دریاچه یخ زده" به عنوان یک قیاس عملی برای شبیه سازی یک سناریوی نفت استفاده می کنیم.

بهینه سازی تولید، از جمله خطر هجوم آب (به چاه تولید) ناشی از تولید بیش از حد. برای این منظور، ما یک مثال کد قبلی در مورد کلی "مشکل دریاچه یخ زده" را اصلاح می کنیم.

(نمونه کد مرجع ما مجدداً از کد موجود در کتاب عالی راویچندیران، 2020 اقتباس شده است).

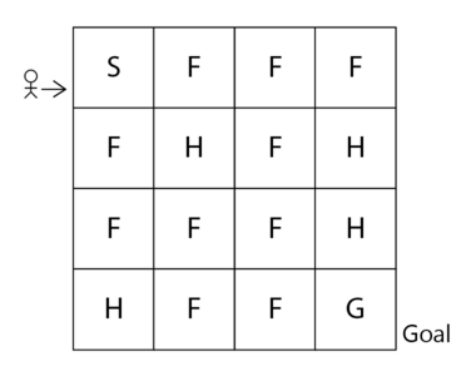
در قیاس ما، ما فرض می کنیم که اقدامات ممکن شامل افزایش یا کاهش تولید نفت است. اگر ما

چنین تولیدی را به درستی افزایش دهیم، در وضعیت خوبی باقی می‌مانیم (F=یخ زده). اگر بیش از حد افزایش دهیم، این می تواند

باعث هجوم آب (H= سوراخ) شود. چنین تشبیهی را پس از یادآوری اصول اولیه می توان بهتر درک کرد

"مشکل دریاچه یخ زده"، به شرح زیر.

محیط دریاچه یخ زده به صورت گرافیکی در زیر نشان داده شده است :



بیایید محیط دریاچه یخ زده را مرور کنیم. در محیط دریاچه یخ زده نشان داده شده در بالا، ما را در نظر می گیریم

حالات و نمادهای متناظر زیر:

S حالت شروع است

F حالت های یخ زده است

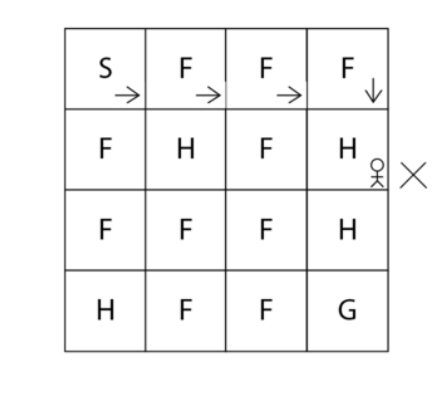
H حالت های سوراخ است

G حالت هدف است

در محیط دریاچه یخ زده، هدف ما رسیدن به حالت هدف G از حالت شروع S بدون بازدید از آن است

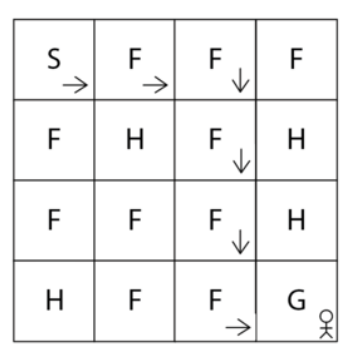
حالت های حفره H. در واقع، هنگام تلاش برای رسیدن به حالت هدف G از حالت شروع S، اگر عامل از سوراخ بازدید کند

H را حالت دهید سپس در سوراخ می افتد و مانند شکل زیر می میرد:



بنابراین، هدف عامل رسیدن به حالت G با شروع از حالت S بدون بازدید از حالت های سوراخ H است.

در ذیل نشان داده شده است:



همانطور که در ابتدای این آموزش پیش بینی شد، ما سعی می کنیم "مشکل دریاچه یخ زده" را با تولید نفت تطبیق دهیم.

سناریو، از جمله خطر تهاجم آب. این می تواند با تولید بیش از حد ایجاد شود. برای هر حالت شروع

(یعنی نرخ مشخصی از تولید نفت)، هدف ما یافتن سیاست بهینه (تولید بهینه) است

مدیریت) برای تولید حداکثر مقدار روغن، بدون وارد شدن به شرایط نامطلوب آب

هجوم به چاه در این قیاس، ما فرض می کنیم که اقدامات ممکن شامل افزایش یا کاهش است

تولید نفت، با موجودیت متغیر. اگر تولید را به درستی افزایش دهیم، در وضعیت خوبی باقی می‌مانیم (F=یخ زده). اگر

ما بیش از حد افزایش می دهیم، این می تواند باعث هجوم آب شود (H = سوراخ). هدف ما رسیدن به یک حالت نهایی (G) است که در ما

شبیه سازی-قیاس، تولید نفت بهینه را نشان می دهد، که "بیشترین تولید برای آن چاه، بدون ریسک است.

تهاجم آب».

در قیاس ما، حالات دریاچه یخ زده به صورت زیر فرموله می شوند:

S دلالت بر حالت شروع دارد: برای مثال، این می تواند یک نرخ تولید روغن معین باشد، با جلو آب در یک

فاصله معینی از چاه تولید چنین فاصله ای را می توان برای مثال با استفاده از یک ژئوفیزیک تشخیص داد

روشی مانند توموگرافی مقاومتی الکتریکی گمانه (ERT) (برای جزئیات بیشتر به Dell'Aversana، 2021 مراجعه کنید.

بحث در مورد ERT همراه با یادگیری تقویتی). F دلالت بر حالات منجمد دارد: در قیاس ما، این

مربوط به حالت تولید با یک جبهه آب دور از چاه است. بنابراین هنوز یک حالت امن است، بدون هیچ گونه

خطر قابل توجه تهاجم گسترده آب H بر حالات حفره دلالت دارد: در قیاس ما، این با یک نرخ تولیدی مطابقت دارد

نرخ تولید در جایی که تهاجم آب قابل توجهی وجود دارد. G بر حالت هدف دلالت دارد: در قیاس ما، این

مربوط به نرخ تولید بهینه نفت است که در آن ما تولید بالا و خطر کم تهاجم آب داریم

همزمان. به عبارت دیگر، هدف ما رسیدن به یک وضعیت باثبات است که بدون آن مقدار زیادی نفت تولید کنیم

وارد شدن به وضعیت نامطلوب نفوذ آب به چاه تولیدی.

چگونه می توانیم به این هدف برسیم؟ یعنی چگونه بدون مراجعه به H از S به حالت G برسیم؟ که در

Reinforcement Learning، خط مشی بهینه به عامل می گوید که در هر حالت اقدام صحیح انجام دهد. بنابراین، اگر ما پیدا کنیم

سیاست بهینه، آنگاه می‌توانیم از S بدون بازدید از حالت H به حالت G برسیم. برای یافتن بهینه

خط مشی، ما می توانیم از "روش تکرار ارزش" برای یافتن خط مشی بهینه استفاده کنیم (این به خوبی توضیح داده شده است، به عنوان مثال، توسط

راویچاندیران، 2020).

همه حالت‌های ما (S تا G) از 0 تا 16 کدگذاری می‌شوند و هر چهار عملکرد (چپ، پایین، بالا، راست) کدگذاری می‌شوند.

از 0 تا 3 در "مجموعه ابزار بدنسازی" (که یک کیت ابزار یادگیری تقویتی معمولی است). در قیاس ما، افعال

که در بالا ذکر شد معانی زیر را دارند (تصور کنید از چرخی استفاده کنید که تولید را تنظیم می کند):

چپ = تولید متوسط کاهش پایین = تولید زیاد کاهش سمت راست = متوسط افزایش تولید بالا تر از ظ

کد نویسی:

مرحله اول شامل وارد کردن تمام ما های لازم است:

In [34]:

import gym

import numpy as np

In [35]:

:"Gym"از ایجاد می کنیم. در مرحله بعد، محیط دریاچه یخ زده را با استفاده

env = gym.make('FrozenLake-v0')

بیایید با استفاده از تابع رندر به محیط دریاچه یخ زده نگاه کنیم:

In [36]:

env.render()

SFFF

FHFH

FFFH

HFFG

می بینیم که عامل ما در حالت اولیه S است و باید بدون بازدید از حالت های H به حالت G برسد.

بیایید خط مشی بهینه را با استفاده از "روش تکرار ارزش" محاسبه کنیم. این روش محاسبه بهینه است

خط مشی MDP و ارزش آن (MDP نماد کوتاه "فرایند تصمیم مارکوف" است).

ابتدا بیایید نحوه محاسبه تابع مقدار بهینه را بیاموزیم و سپس نحوه استخراج بهینه را خواهیم دید

خط مشی از تابع مقدار بهینه محاسبه شده.

محاسبه تابع مقدار بهینه

تابعی به نام value\_iteration تعریف می کنیم که در آن تابع مقدار بهینه را به صورت تکراری محاسبه می کنیم

حداکثر تابع Q را می گیرد. یادآوری کنید که Q-Learning یک الگوریتم یادگیری تقویتی مبتنی بر ارزش است

که برای یافتن خط مشی انتخاب عمل بهینه با استفاده از تابع Q استفاده می شود. هدف ما به حداکثر رساندن ارزش است

تابع Q. جدول Q به ما کمک می کند تا بهترین عملکرد را برای هر حالت پیدا کنیم. تابع Q-value تابعی است که

یک جفت مشاهده-عمل را به یک مقدار اسکالر نشان می دهد که کل پاداش های بلند مدت مورد انتظار را نشان می دهد.

انتظار می رود عامل زمانی که از مشاهدات داده شده شروع می شود و عمل داده شده را اجرا می کند جمع شود. بعد،

تابع value\_iteration را تعریف می کنیم که محیط را به عنوان پارامتر می گیرد:

In [37]:

def value\_iteration(env):

#set the number of iterations

no\_of\_iterations= 1000

#set the threshold number for checking the convergence of the value function

threshold = 1e-20

#set the discount factor

gamma = 1.0

#REMINDER. The discount factor essentially determines how much the reinforcement learning

#cares about rewards in the distant future relative to those in the near future.

#If γ=0, the agent will be completely myopic and only learn about actions that produce

#If γ=1, the agent will evaluate each of its actions based on the sum total of all of its iteration.

#now, we will initialize the value table, with the value of all states to zero

value\_table = np.zeros(env.observation\_space.n)

#for every iteration

for i in range(num\_iterations)

#update the value table: every iteration, we use the updated value

#table (state values) from the previous iteration

updated\_value\_table = np.copy(value\_table)

#next, we compute the value function (state value) by taking the maximum of Q value

#thus, for each state, we compute the Q values of all the actions in the state and steps

#we update the value of the state as the one which has maximum Q value as shown below

for s in range(env.observation\_space.n):

Q\_values = [sum([prob\*(r + gamma \* updated\_value\_table[s\_])

for prob, s\_, r, \_ in env.P[s][a]])

for a in range(env.action\_space.n)]

value\_table[s] = max(Q\_values)

#after computing the value table, that is, value of all the states, we check whether it’s true or not.

#difference between value table obtained in the current iteration and previous iteration.

#less than or equal to a threshold value. If that conditon is true, then we break the operation.

#value table as our optimal value function as shown below:

if (np.sum(np.fabs(updated\_value\_table - value\_table)) <= threshold):

break

return value\_table

استخراج خط مشی بهینه از تابع مقدار بهینه:

سپس، اجازه دهید نحوه استخراج خط مشی بهینه از تابع مقدار بهینه محاسبه شده را ببینیم.

ابتدا تابعی به نام extract\_policy تعریف می کنیم که مقدار\_table را به عنوان پارامتر می گیرد:

In [86]:

def extract\_policy(value\_table):

#set the discount factor.

gamma = 1

#first, we initialize the policy with zeros, that is, first, we set the actions for all

#be zero

policy = np.zeros(env.observation\_space.n)

#Next, we compute the Q function using the optimal value function that was obtained from

#previous step. After computing the Q function, we can extract policy by selecting action

#maximum Q value. Since we are computing the Q function using the optimal value

#function, the policy extracted from the Q function will be the optimal policy.

#As shown below, for each state, we compute the Q values for all the actions in the states

#then we extract policy by selecting the action which has maximum Q value.

#for each state

for s in range(env.observation\_space.n):

#compute the Q value of all the actions in the state (again, we apply one of the Best options.)

Q\_values = [sum([prob\*(r + gamma \* value\_table[s\_])

for a in range(env.action\_space.n)]

#extract policy by selecting the action which has maximum Q value

policy[s] = np.argmax(np.array(Q\_values))

return policy

اکنون، خواهیم دید که چگونه سیاست بهینه را در محیط دریاچه یخ زده خود استخراج کنیم

همه اش را بگذاریم کنار هم:.

یادآوری می کنیم که در محیط دریاچه یخ زده، هدف ما یافتن سیاست بهینه ای است که درست را انتخاب کند

In [87]:

optimal\_value\_function = value\_iteration(env=env)

در مرحله بعد، مطابق شکل، خط مشی بهینه را از تابع مقدار بهینه با استفاده از تابع extract\_policy استخراج می کنیم

زیر:

optimal\_policy = extract\_policy(optimal\_value\_function)

ما می توانیم خط مشی بهینه به دست آمده را چاپ کنیم:

In [89]:

print(optimal\_policy)

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

همانطور که می توانیم مشاهده کنیم، خط مشی بهینه ما به ما می گوید که در هر حالت اقدام صحیح را انجام دهیم. به دنبال تشبیه ما،

این به معنای تنظیم صحیح تولید نفت، با تنظیم آن از طریق چرخ تولید است. به منظور. واسه اینکه. برای اینکه

ما باید اعداد فوق را به یک سیاست تصمیم گیری عملی با هدف بهینه سازی تولید نفت تبدیل کنیم

روابط بین اعداد و اعمال فوق را اعمال کنید: در قیاس ما، 0 به معنای «تولید متوسط» است

کاهش"؛ 1 به معنای "کاهش تولید زیاد"؛ 2 به معنای "افزایش تولید متوسط"؛ 3 به معنای "بالا بودن تولید" است.

افزایش تولید.» البته در سناریوهای واقعی تر، می توان آن تعاریف کیفی را جایگزین کرد با.

ارزش های کمی تولید.

نتیجه گیری و اظهارات پایانی:

با استفاده از راهکار «بازی دریاچه یخ زده» توانستیم سیاستی کارآمد برای تنظیم نفت تعیین کنیم

تولید چاه ما به روش بهینه این بدان معناست که: تولید با حداکثر نرخ ممکن بدون هزینه

به وضعیت نامطلوب آب عظیم در خود چاه. مثال ما بسیار ساده است، زیرا چنین است

اهداف نمایشی این فقط یک شبیه سازی پیش پا افتاده است. با این حال، رویکرد دارای اعتبار کلی و

همین روش را می توان برای سناریوهای بسیار پیچیده اعمال کرد. برای اعمال آن در موارد واقعی، باید واقعی داشته باشیم

داده های تولید و اندازه گیری های واقعی فاصله جبهه آب با توجه به چاه. ممکنه

برای مثال با استفاده از گمانه ERT (توموگرافی مقاومتی الکتریکی). ما این روش را به تفصیل شرح داده ایم

آثار قبلی (Dell'Aversana، 2021).