

دوڑا

ارائه پوستر
نظر سنجی
نمره افغانی

بازمیجی :

۱۰ { میل نرم
پایان نرم

۲ ~~کاربرد~~ $\times ۲$

۸ تکرین

```

graph TD
    VL[vision learning] --> OD[object detection]
    OD --> RL[RL]
    RL --> SP[search policy]
    RL --> RF[reward function]
    RL --> A[action]
    RF --> DR[delayed reward]
    DR --> SP
    SP --> DE[dynamic exploration]
    DE --> SP
  
```

goal: learn $\Pi_\theta: S_t \rightarrow A_t$

to maximize $\sum_t r_t$ or $E[r_t]$

1x realtime
10000x simulation

generate
sample

 fit a model

$$J(\theta) = E_n \left[\sum_t f_t \right]$$

$$\approx \frac{1}{N} \sum_{t=1}^N \sum_i r_t^i$$

روشن مبنی بر مدل

improve policy

مادگر فن (S+αf)

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta)$$

متى $\nabla_{\theta} J(\theta)$ متساوي صفر؟

جستجوی متنی

Value-based RL

$$Q^\pi(s_t, a_t) = \sum_{t'=t}^T E_{\pi_\theta} [r(s'_t, a'_t) | s_t, a_t]$$

total reward from s_t, a_t

$V^\pi(s_t)$ fit model to predict it!

Deep Q learning

just update π

Direct

* Policy gradients \rightarrow برلن گاریان راحل، E گاریان

* Actor-Critic methods

where do rewards come from?

(\hookrightarrow from demonstration

copying (imitation learning)

inferring rewards (inverse RL)

from observing the world

predict (fe)

unsupervised

from other tasks

transfer

meta-learning

Optimization

Learn from exp

Generalization

Delayed consequence

Exploration

تجربه از میان

Data-driven AI

ذوق از میان و خلاصه

Richard Sutton

(AlphaGo) ! حلولیت طریقی RL لی

A bitter lesson

RLHF $\xrightarrow{\text{for}}$ search + exploration

markov process, MDP, Optimization

جلسہ ۱۰، ۱۱، ۱۴

marketing جو: Exploration کی
کسے suboptimal از جلوگیری از

RL برائی ساز - Compositional Generalization کی

RL by Sutton : آن ب محض /

متع

پیش نیاز ہے:

۲۰۱۳ (DQN) Atari Deepmind

- غذا گیری متصادی

۲۰۱۶ (TRPO) 2d locomotion

- بھینہ سازی

۲۰۱۵ Alpha Go (چوتھا) Deepmind Berkely

- یارگیری ڈرٹ

۲۰۱۹ (+GAE) 3d locomotion Berkely

۲۰۱۷ (GPS) Real robot manipulation :

۲۰۱۸ (PPO) Dota2

۲۰۱۸ DeepMimic

۲۰۱۸ Alphastar

:

RLHF, DPO, ..

* مُدَارِسَةٌ مَهْمَلَةٌ

یک تَجَزِّعٌ مَرْتَجَعٌ / outcome از

(پُلْمَ وَصْحَىٰ)

$$\{s_1, \dots, s_n, \dots\} : RL \rightarrow$$

$P[s_n | s_1, \dots, s_n] = P[s_n | s_{n-1}]$ \leftarrow شرط مارکوف \rightarrow فرض مارکوف *

معنی اگر دمیعت از روند بگیریم

$$P[s_n] := \pi_n$$

فریج اولیه

$$\pi_n = \pi_0 P^n \leq (s_j, \dots, s_1) P \text{ ماتریس اسما}$$

$$\lim_{n \rightarrow \infty} \pi_n \text{ نزدیکی!}$$

$$\pi^* = \pi^* P \quad (\pi^*) \text{ stationary } \leftarrow \text{این اسما}$$

و جدر رله در نزدیکی است \rightarrow که مارکوف است \rightarrow در جدر رله

مُدَارِسَةٌ مَهْمَلَةٌ \leftarrow دبرابر π^* است

left eigenvector من π^*

برابر ۱ eigenvalue !

* Finite State

* Irreducible (ین هر زر اس می بچست را برپا نماید)

* Aperiodic (بهم خود را می بند)

این دیگر اس حل نشود!

یک مُدَارِسَةٌ مَهْمَلَةٌ

markov reward process *

(MRP) (راهن حتم لایه) \rightarrow action بعد از

$$V(s_0) := \mathbb{E} \left[\sum_{n=0}^{\infty} \gamma^n R_n \mid S_0 = s_0 \right] \quad \text{Value function}$$

discount factor

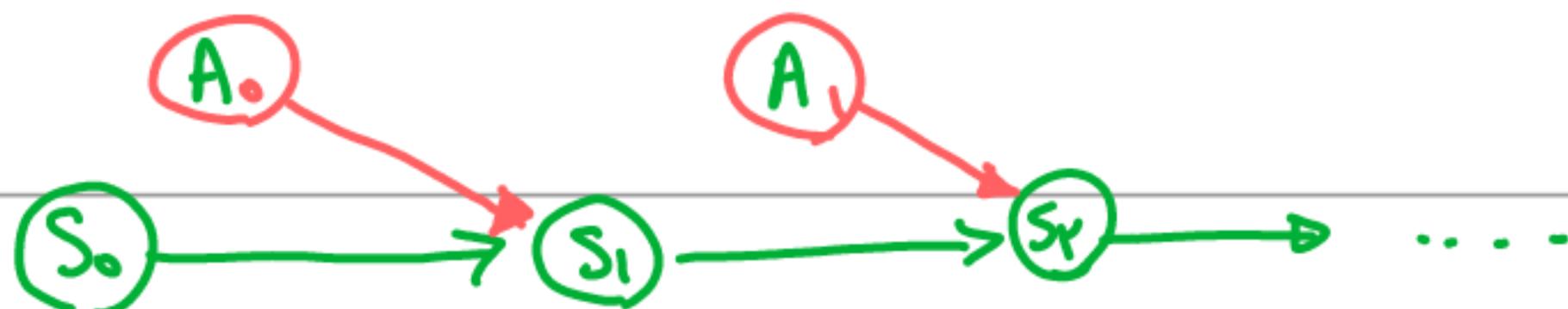
نهر اعشاری
عم تضیییت آیندگی
زد بدل رسیدن

$$= \mathbb{E} [R_0 + \gamma V(s_1) \mid S_0 = s_0]$$

Bellman Equation

Markov Decision Process

all random variable



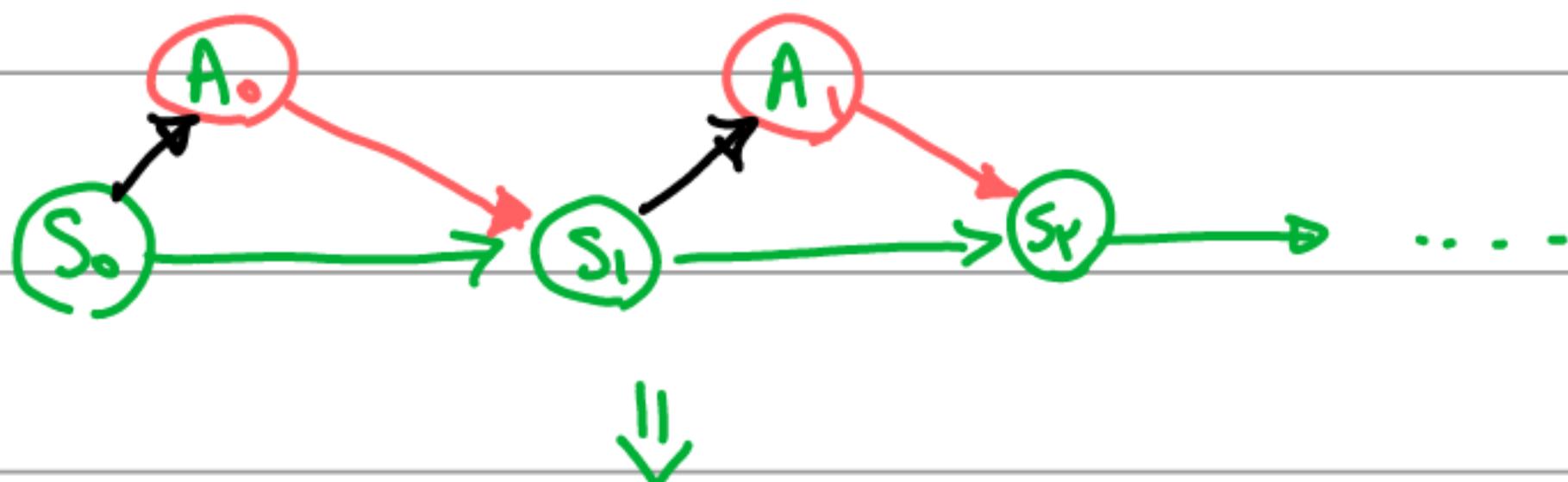
action
Action!

one node

parent by node, node
(Joint distribution ~ معین)

Reward: $\begin{cases} R(S_n, A_n, S_{n+1}) \\ \vdots \\ \text{عین مختلف} \end{cases} \rightarrow S_n \text{ طبیعت } A_n$

$\pi(A_0 | S_0)$



$$\mathbb{P}[S_0, A_0, S_1, A_1, \dots] = \mathbb{P}[S_0] \pi(A_0 | S_0) \mathbb{P}[S_1 | S_0, A_0] \pi(A_1 | S_1) \dots$$

$$\max_{\pi} \mathbb{E} \left[\sum_n \gamma^n R(S_n, A_n, S_{n+1}) \right] \quad \text{هدف: حفظ}$$

حول: اغراض حول استدلالات! E
خوب وسيلة دليل رسائل (Sample)

Constraint Optimization

$$\min_x f(x)$$

$$\text{s.t. } \forall i: g_i(x) \leq 0$$

Primal

$$\min_x \max_{\lambda_i \geq 0} f(x) + \sum_{i=1}^{\infty} \lambda_i g_i(x)$$

Duality gap

Dual

$$\max_{\lambda_i \geq 0} \min_x f(x) + \sum_{i=0}^{\infty} \lambda_i g_i(x)$$

Dual ≤ Primal

ابتدا نمی‌باشد

Slater condition باید برقرار شود. می‌توانیم Dual Gap که سریع

f, g کنار هم باشند. می‌توانیم Dual می‌توانیم داشت. * حالتی در نظر بگیریم که

Convex

interior point

و جریانه باش

اپنے جری

Dual \ Prime \ اداہ کر

$$\lambda^*; g(\lambda^*) = 0 \Leftarrow \left\{ \begin{array}{l} \lambda_i^* = 0 \Leftrightarrow g_i(x^*) < 0 \\ \text{or, } \lambda_i^* \leq g_i(x^*) = 0 \end{array} \right.$$

KKT جزء اسراط

(complementary slackness)

؟ رِحالے Bellman حُطُورِکیں نہیں؟
کہ یا موادِ مرتب
کیا مزدوجہ ریاضی

$x_1, \dots, x_n \sim f_{\theta}$ $\xrightarrow{\text{i.i.d}}$ point estimation

→ full posterior $p(\theta|x_1 \rightarrow x_n)$ → variation jika
explorasi \rightarrow θ

(Distributional RL) of RL tasks → integral

Point estimation پرنسٹن

$$MLE: \max_{\theta} P[x_1, \dots, x_n | \theta]$$

$$\underset{\theta \sim P_\theta}{MAP} : \max_{\theta} P[\theta | x_1 - x_n]$$

جَذْبَرِيَّةٌ بَرْزَانٌ Prior

* Conjugate prior

پریور (Prior) روسیہ میں ایک انتظامی افسوسنامہ کا نام ہے۔

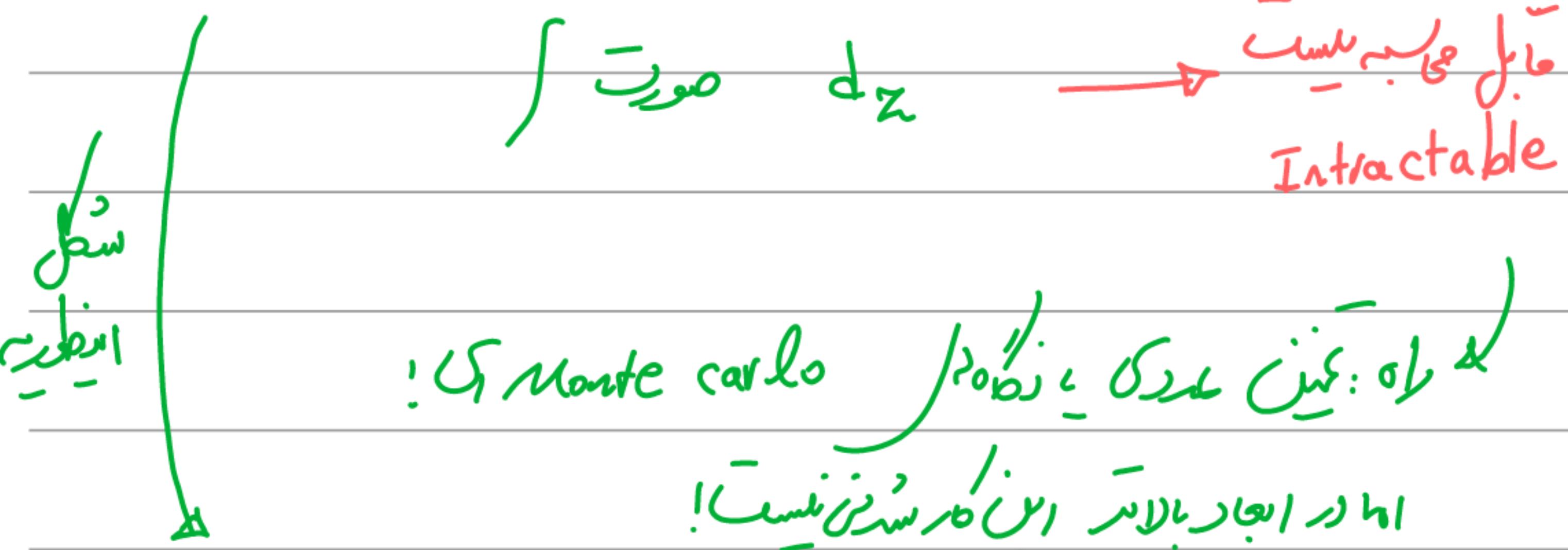
intractable posterior

!

$$x_1 \sim N(z, 1)$$

$$z \sim \text{Exp}(1)$$

$$P(z|x_1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_1-z)^2}{2}} \cdot e^{-z} \mathbb{1}_{(z>0)}$$

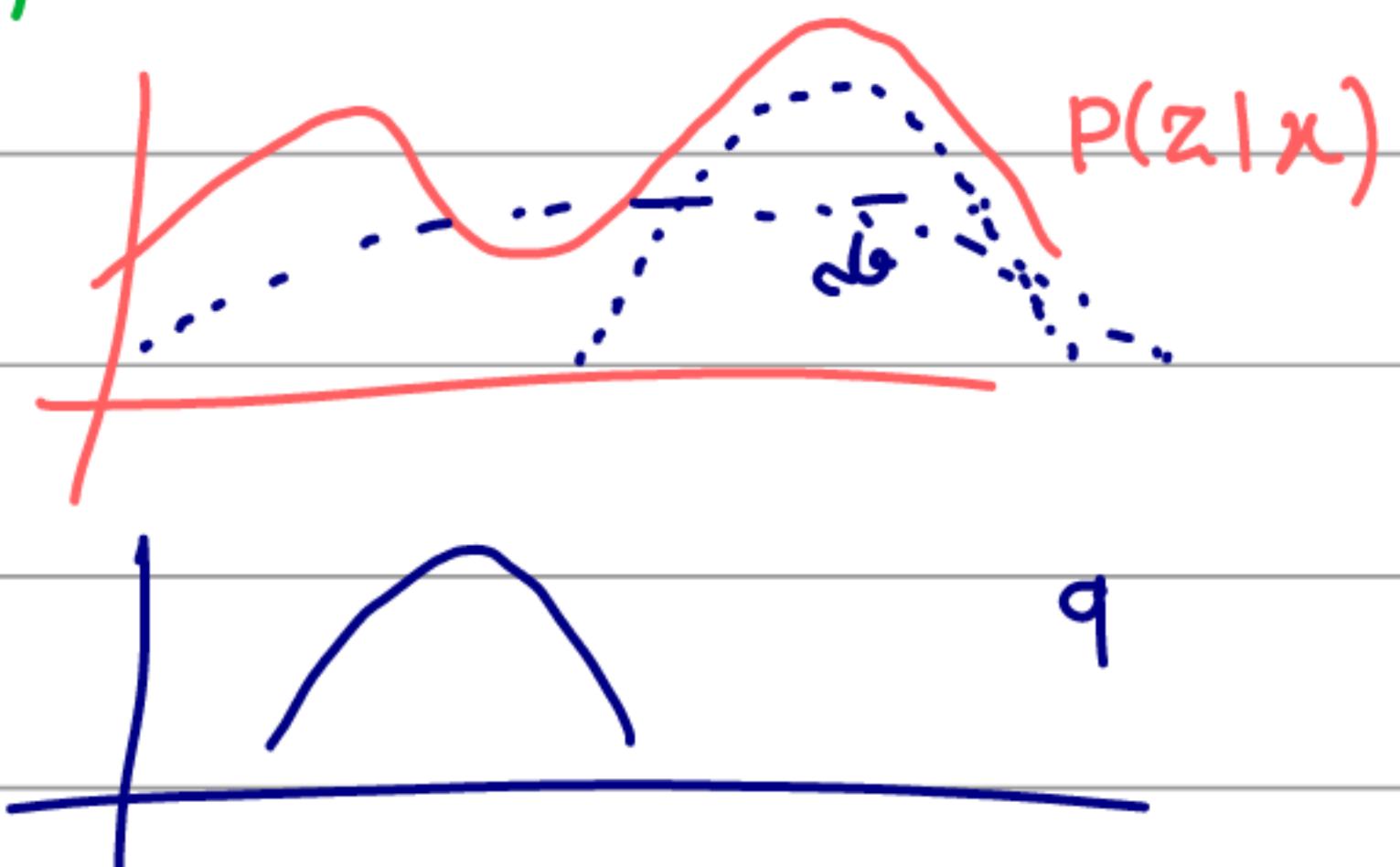


|| variational optimization ||

$$\min_q KL(q || P(z|x))$$

KL دivergence

$$KL = \mathbb{E}_{z \sim q} \left[\log \frac{q}{P(z|x)} \right]$$



mass ~ object

دلي بركان بـ تـ دـ سـ طـ مـ رـ اـ

$$D(q \parallel P(z|x)) = \mathbb{E}_{z \sim q} \left[\log \frac{q(z)}{P(z|x)} \right]$$

$P(z|x)/P(z)$

$$= \mathbb{E}_{z \sim q} \left[\log \frac{q}{P(x,z)} + \log P(x) \right] = -\mathbb{E}_{z \sim q} \left[\log \frac{P(x,z)}{q(z)} \right] + \log P(x)$$

↙ evidence

KL ≥ 0

بذریغه میگیرد که داشتم

$$\mathbb{E}_{z \sim q} \left[\log \frac{P(x,z)}{q(z)} \right] \leq \log P(x)$$

↙ evidence

ELBO

=>

KL جکسون سازی

را بینیمه میکنم!

evidence lower bound

ELBO = evidence - KL

proposal
پیشنهاد

$$q_{\theta}(z) = \theta e^{-\theta z} \mathbf{1}_{z \geq 0}$$

مانند پرتو

باشد

برای validation

تولید کرد

و...

$$ELBO = \mathbb{E}_{z \sim q} \left[\log \frac{1}{\sqrt{\pi}} e^{-\frac{(x-z)^2}{r}} e^{-z} - \log \theta e^{-\theta z} \right]$$

ارجاع به زیر نبرد

$$= \mathbb{E}_{z \sim q} \left[-\frac{(x-z)^2}{r} - z - \log \theta + \theta z \right]$$

$$\left. \begin{aligned} E_{z \sim q}[z] &= \frac{1}{\theta} \\ \text{Var}(q) &= \frac{1}{\theta^2} \end{aligned} \right\} \Rightarrow \mathbb{E}_{z \sim q} \left[-\frac{x^2 - 2xz}{r} + zx - z - \log \theta + \theta z \right]$$

$$= \dots \Rightarrow \theta = \frac{1-x + \sqrt{(x-1)^2 + 1}}{r}$$

Importance Sampling

$$\underset{x \sim p}{E}[f(x)] = \int_{-\infty}^{\infty} f(x) P_x(x) dx$$

ماریس بالا: $f(x)$ $P_x(x)$

$$x_1, \dots, x_n \stackrel{i.i.d.}{\sim} P_x(x) \rightarrow \frac{1}{n} \sum f(x_i) \rightarrow \text{حرب نسبت!} \rightarrow$$

$$x_1, \dots, x_n \stackrel{i.i.d.}{\sim} q_x(x) \rightarrow \text{ازین دسته میکنیم که چون مبارہ است!} \rightarrow$$

محبہ سُر

$$\underset{x \sim q}{E}\left[\frac{f(x) P(x)}{q(x)}\right] = \int_{-\infty}^{\infty} f(x) P_x(x) dx$$

$$\underset{x \sim p}{E}[f(x)] \approx \frac{1}{n} \sum_{i=1}^n \frac{f(x_i) P(x_i)}{q(x_i)} \quad (x_1, \dots, x_n \stackrel{i.i.d.}{\sim} q)$$

$$\text{Var}(W) = \frac{1}{n} \text{Var}\left[\frac{f(x) P(x)}{q(x)}\right]$$

$$\text{Var}(X) = E[X^2] - \mu^2$$

$$\begin{aligned} &= \int \frac{f^2 P}{q} dx - \mu^2 \\ &\rightarrow \frac{f \cdot P}{\mu} \end{aligned}$$

$$\Rightarrow \text{Var}(W) = 0$$

اگر f متعین بندی نہ ہے!

ایسا ممکن نہ ہے وہ فرست خیز

Gibbs میں این طریقہ!

با روشن ہی مل

$$I(X; Y) = H(X) - H(X|Y)$$

حُوَدْر لا دار دعم تفعيل X

$$H(X) = - \sum_{x \in D} P(x) \log P(x)$$

ماكن من دعم

$$H(X|Y) = - \sum_{y \in D_y} P(y) \sum_{x \in D_x} P(x|y) \log P(x|y)$$

$$D(P_{x,y}(x,y) \| P_x(x)P_y(y))$$

حمر compress ١٪

تردن برای انتساب هسته بحثی دین نهاده عرضه بحثی اید explore فهم RL

$$I(X; Y | Z) \rightarrow \text{observation}$$

بحسب

نهاده آندرین سالور ایت

$$\max_{\pi: S \rightarrow A} E_{\pi} \left[\sum_{i=0}^{\infty} \gamma^i R(i) \right]$$

هدف: حفظ

$$\left. \begin{array}{l} P(s'|s,a) \\ s \rightarrow R(s) \end{array} \right\} \text{فراء MDP}$$

بع ازین: $V^{\pi}(s) := E_{\pi} \left[\sum_{i=0}^{\infty} \gamma^i R(i) | S_0 = s \right]$

$$V^{\pi}(s) = \sum_{s'} P(s'|s, \pi(s)) [R(s) + \gamma V^{\pi}(s')]$$

است s' deterministic \Rightarrow Bellman تحسن

بع بحثی ازین: $V^*(s) = \max_{\pi} V^{\pi}(s) = \max_{\pi} \sum_{s'} P(s'|s, \pi) [R(s) + \gamma V^*(s')]$

Policy, argmax

بس اورلن V^* به صورت V^* (Value Iteration) iterative

ايسو من براي $V_k(s)$

Bellman operator (T) سلسلة $V_k^*(s)$ بـ $(V_{k-1}^*(s))$ (برمجة)
یک سلسله را (V_k^*) داشت

$$\vec{V}_0 = \begin{bmatrix} \vdots \\ 0 \\ \vdots \end{bmatrix} \quad \vec{V}_k = T \vec{V}_{k-1}$$

#fixed point $V^* = TV^*$ اگر T محدود باشد و V^* نباشد

اگر T دو قدر کران دار باشد هر این سند به تبع ارزش بینی!

+ ابتدا contractive یعنی نفی

$$\|TV_1 - TV_2\|_\infty \leq \alpha \|V_1 - V_2\|_\infty \quad (\alpha < 1)$$

contractive $\Rightarrow T^n$ سیستم خوب

٢٢، ١٢، ١ جلسه

$$V_k^*(s) = \max_a \sum_{s'} P(s'|s,a) [R(s) + \gamma V_{k-1}^*(s')]$$

$$\vec{V}_k = T \vec{V}_{k-1}$$

روکی میں نوٹ، value iteration دو

Contraction Mapping

$$\|\vec{T}\vec{V}_0 - \vec{T}\vec{V}_1\|_\infty \leq \alpha \|\vec{V}_0 - \vec{V}_1\|_\infty \quad (\alpha < 1)$$

اندازہ اجرا کر جائے میں سے دو

دستور

$$\begin{aligned} |\vec{T}\vec{V}_1(s) - \vec{T}\vec{V}_0(s)| &\leq \max_a \left| \sum s' P(s'|s,a) \gamma (\vec{V}_1(s') - \vec{V}_0(s')) \right| \\ &= \gamma \|\vec{V}_1(s) - \vec{V}_0(s)\|_\infty \end{aligned}$$

تقریب متعین کرنے والے ضریب γ

$$\Rightarrow \|\vec{V}_k - \vec{V}^*\|_\infty \leq \gamma^k \|\vec{V}_0 - \vec{V}^*\|_\infty$$

وہی

Q-Values

$$Q(s,a) := \text{اصلی تجزیے کے سے شروع ہے}$$

$$Q^*(s,a) = \sum s' P(s'|s,a) [R(s) + \max_{a'} Q^*(s',a')]$$

Value Iteration چون؟ سرعتی میں؟

$\sum_{i=1}^n \max$

یست بین در میان این دو اگر رسم اراده مادر را در عذر آورده باشد

Policy Iteration

Policy Evaluation

Policy Improvement

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

بے رست اور دل پر محسوس

ادامه دادن ناچهراً سور و پیش نمایی

$$V_k^n(s) \leftarrow \sum P(s'|s, \pi(s)) (R(s, \pi(s), s') + \gamma V_{k+1}^n(s))$$

صیحتِ طمادل را چنین برداشت جن می بیند

یعنی رہنمائی کی سیاست policy ہے۔

Policy Implications

برحسب $\nabla \tilde{\omega}_k$ هنر ω_k است و قابل \int_0^t

$$\vec{V}^{\pi_{k+1}} \simeq \vec{T}_{\cdot \cdot \cdot}^{\pi_{k+1}} - \vec{T}^{\pi_{k+1}} \vec{V}^{\pi_k} > \vec{V}^{\pi_k}$$

$\max_{\theta} J(\theta)$

درایه بـ درایه نزدیک و کی است

* درایه بـ درایه بـزـلـکـرـمـ وـکـیـ است
*) اـنـرـعـارـ اـنـرـعـارـ

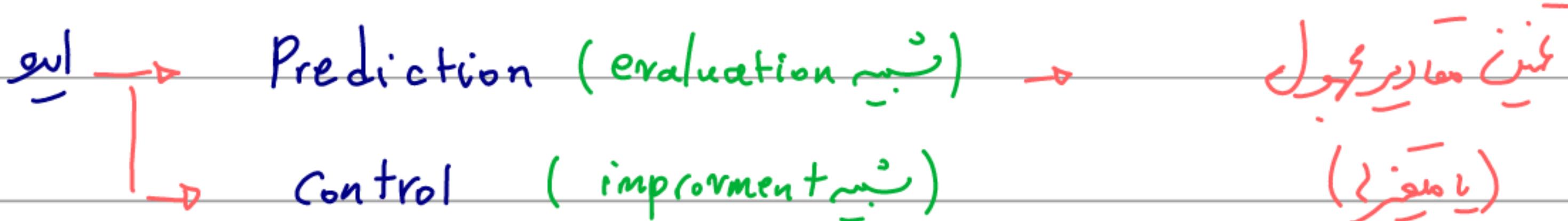
(Contraction بحدار / Bellman معنی +)

جذب

خطاب ششم

$P(s'|s, a), R$ world model بدلے با روشنی planning تبلو

اما (رطابت جو لاس)



Mars Rover JPL *

Monte Carlo prediction

نقش از خطای مجزایی

مسح از خطا هایی که داریم

$s_0, a_0, s_1, a_1, \dots \rightarrow s_0, r_0, a_0, \dots$

$$V^\pi(s) := \mathbb{E}[R(s) + \gamma V^\pi(s') | s_0 = s]$$

$$:= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | s_0 = s\right] \rightarrow \text{پیشنهاد بلند مدت episodic}$$

$$G_t = \sum_{s=t}^{\infty} \gamma^{s-t} R_s$$

برای یک چندین مرحله ای که در یک episode میگذرد

(یاد خواهد کرد) چندین مرحله ای که در یک episode میگذرد

نه فقط آخری بود است (دست نهادن)

(فعله ای) first-visit Monte Carlo

estimation error

(یاد خواهد کرد) every-visit Monte Carlo

یعنی طردی همانند ریکارڈ سود و میکن بخوبی سود

داریان نظر من میں دلیل بین مشترک میں

$$E[(\hat{\theta} - \theta^*)^2] = \text{bias}^2 + \text{variance}$$

k episodes } FV MC \xrightarrow{P} True Value
 $k \rightarrow \infty$ } EVMC \xrightarrow{P} True Value (سریع تر از FVMC)

سutton (Sutton) بحثیاً

راهنمایی نهادن حسابات: $V = \frac{1}{N} \sum G_{i,t}$

$$V^\pi(s) = \frac{N(s)-1}{N(s)} V^\pi(s) + \frac{G_{i,t}}{N(s)} = V^\pi(s) + \frac{1}{N(s)} (G_{i,t} - V^\pi(s))$$

(بهینه سازی کارکرد $G_{i,t}$)

MC در میان راهنمایی در حالت راضی نیست policy eval \rightarrow نهادن

برای زیاد تر راهنمایی در حالت راضی نیست و خطای خود را بینزین!

Policy evaluation diagram

Generalized Policy Improvement

$$\pi'(s) = \arg \max_a Q^\pi(s, a) \xrightarrow{\text{MG}} \text{Explore} \quad \text{نهادن}$$

state سیاست را تغییر دادن پس از یک مرحله

ϵ -Greedy شناسنی هم بگردد

جلسه هفتم ۱۳، ۱۴، ۱۵

(نیاز بزرگ) با بهترین ممکنیت از FVMC, EVMC برای Monte Carlo

طیف را میتوانیم (نیاز به طیف خود) EVMC

حصیت

خوب نیست زیرا بزرگترین سرعت (نیز مرکوز) FVMC

MSE کمتر است و این دلیل داشت EVMC

برای α -درست Incremental

نمایندگی خوب نیست

چگونه خوب است؟

$$\sum \alpha_n(s_j) = \infty, \quad \sum \alpha_n^*(s_j) < \infty$$

#kolmogorov law of large numbers

کوچکترین سرعت

بین C_1 و C_2 میباشد

طیف صفر نیست اما کاملاً نیست

حالات

propagate میشود

لر نیز به episodic برآیند

Policy improv ماباید خاص مراحت از

Q: قائل RL است

MC (On Policy) از سیسکو کامپانی تولید میشود!

+ خود بین اسرار طبقه بندی Off Policy >

Policy Improvement سطح e-greedy

$$\pi(a|s) = \arg \max_a Q(s, a) \text{ w.p } 1 - e + \frac{e}{|A|}$$

مرکوز

$$a' \neq \arg \max Q(s, a) \text{ w.p } \frac{e}{|A|}$$

$$\forall s: \mathbb{E}_{\pi \sim \pi'} [Q(s, a)] \geq V^{\pi_{old}}(s) \Rightarrow \forall s \quad V^{\pi}(s) \geq V^{\pi_{old}}(s) : \text{فم}$$

له نتیجه حکایت این است که از مجموعه ای از این اقدامات ممکن است انتخاب شود (جیسا که بحث تا پیش از اینجا مذکور شد)

$$Q^{\pi_i}(s, \pi_{i+1}(s)) = \frac{\epsilon}{|A|} \sum_{a \in A} Q^{\pi_i}(s, a) + (1-\epsilon) \max_{a \in A} Q^{\pi_i}(s, a)$$

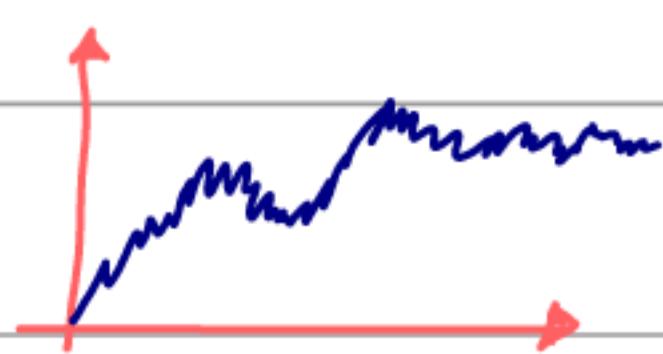
$$\sum_i \pi_i(a|s) - \left\{ \frac{\epsilon}{|A|} \right\} \geq 0 \quad \text{حالو}$$

$$\Rightarrow Q^{\pi_i}(s, \pi_{i+1}(s)) \geq \frac{\epsilon}{|A|} \sum_{a \in A} Q^{\pi_i}(s, a) + \sum_i (\pi_i(a|s) - \frac{\epsilon}{|A|}) Q^{\pi_i}(s, a)$$

$$= \mathbb{E}_{a \sim \pi_i} [Q^{\pi_i}(s, a)] = V^{\pi_i}(s)$$

حاله نیز با هر قدر که ϵ -greedy باشد

له اما Q را در حقیقت نمایم و در عمل بی خود داری این نرمی V را درست نماید

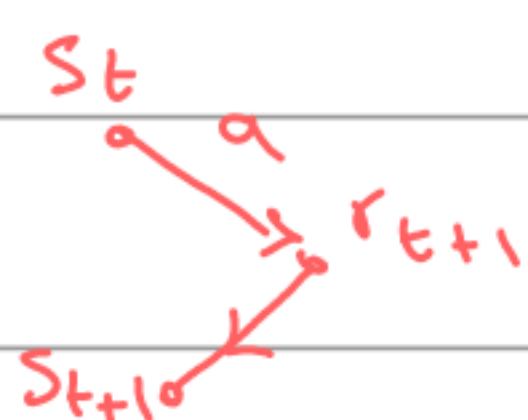


مثلثه عواید ایجاد کردن

من سود، حمل اینزد توجه را همچرید.

اما من کان درین راه عصی برگرد

Temporal Difference



$$V(s_t) \rightarrow V(s_t) + \alpha (R_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

طریقی را که موکنند از آنها که بجزئیات کننده را تبعیغ می کنند

اما با این طریق وسیع نمایند

SARSA: OnPolicy/Lo

greedy in the limit with infinite exploration

جیزت ۱۵/۱۲/۰۲

$$Q(s, a) \rightarrow Q^*(s, a)$$

GLIE کی تحریط e-greedy + MC (ضمیری تحریط)

(انهازی)

لے جائیں (s, a) را بخوبی تبریزی

کے خاتمہ سیستم حفظ کرنے کی طریقہ سیاست (جیز + بخوبی کرنے)

Temporal difference

$$V(s_t) \leftarrow V(s_t) + \alpha (R_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

immediate reward

estimated discounted reward

درین خیز حفظ تریکو (خط بھی اسکل مکانیزم طبقہ ترا)

sample
target

ریس بالا گردید

current

$$Q_{t+1}(s_t, a_t) \leftarrow Q_{t+1}(s_t, a_t) + \alpha (R_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t))$$

TD error

On Policy

SARSA

+ تطبیقی
MC
E-greedy

+ پسیت نسبت زنما حفظ کرنے کی طریقہ دینے بلانگز

$$\Rightarrow (\sum_{\alpha_t = \infty} + \sum_{\alpha_t < \infty}) \text{ باسے } \alpha \text{ بخوبی کرنے}$$

جیزت TD(0) حفظ MC و TD(1) حفظ (1)

بین خوبی از Value Function طریقہ داری داد

MC, TD حلقات بسیز

برای $n=1$ حفظ بریم (نفعی طام و نه افز)

+ یاری اسٹریکٹ کیم بین ۷۰ مختلف میلین نزد طاری بریم

+ مرض کنید $G^{(n)}$ مکار پرنسپس نیز n کم بجزی دست. حل بازنگرد می‌طنین نیز می‌کنم:

$$G_t^{\lambda} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

مختصر درس

دَفْعَتِي تَمَّى بِرَبِّمِ
(مُلْكٌ قَبِيلٌ مَارِسِي)

$$\text{TD}(\lambda) \leftarrow V(s_t) \leftarrow V(s_t) + \alpha (G_t^\lambda - V(s_t))$$

→ اندر رن بیسٹری ہے جو براں ان کو جب میں فتح کرے

backward-view 

forward-view 

جذبی را به تسلیم کردن reward حمله کرد

episode ۱۶

eligibility trace ewl

باپیوں تاریخی رانسریلست!

$$E_0(s) = 0, E_+(s) = \gamma \lambda E_+$$

یہ جوہ عمارتیں دیکھنے
بھروسہ ملے گے

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \rightarrow \text{جیئر TD لہو}$$

$$V(s) \leftarrow V(s) + \alpha \delta_t E_t(s)$$

سے Forward میں Backward ابیجا

جذر تولوی میگیرد

$$G_t^\lambda \leftarrow \text{انصرافی ریس} \rightarrow T_\lambda^n v(x_0) = (1-\lambda) E \left[\sum_{t=0}^n \lambda^t \left(\sum_{i=0}^t r_i^i \pi(x_i) + r^{t+1} v(x_{t+1}) \right) \right]$$

$$= E \left[\sum_{i=0}^n \lambda^i (r_i^n(x_i) + \delta^{i+1} V(x_{i+1}) - \delta^i V(x_i)) \right] + V(x_0)$$

(زیرا ۷ در مطلع بیان مرض رئیم)

forward = backward $\frac{d}{dt} \phi$ \rightarrow $\dot{\phi}$ \rightarrow $\ddot{\phi}$ \rightarrow offline obs \leftarrow update

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

نیز بر این روش online update می شود

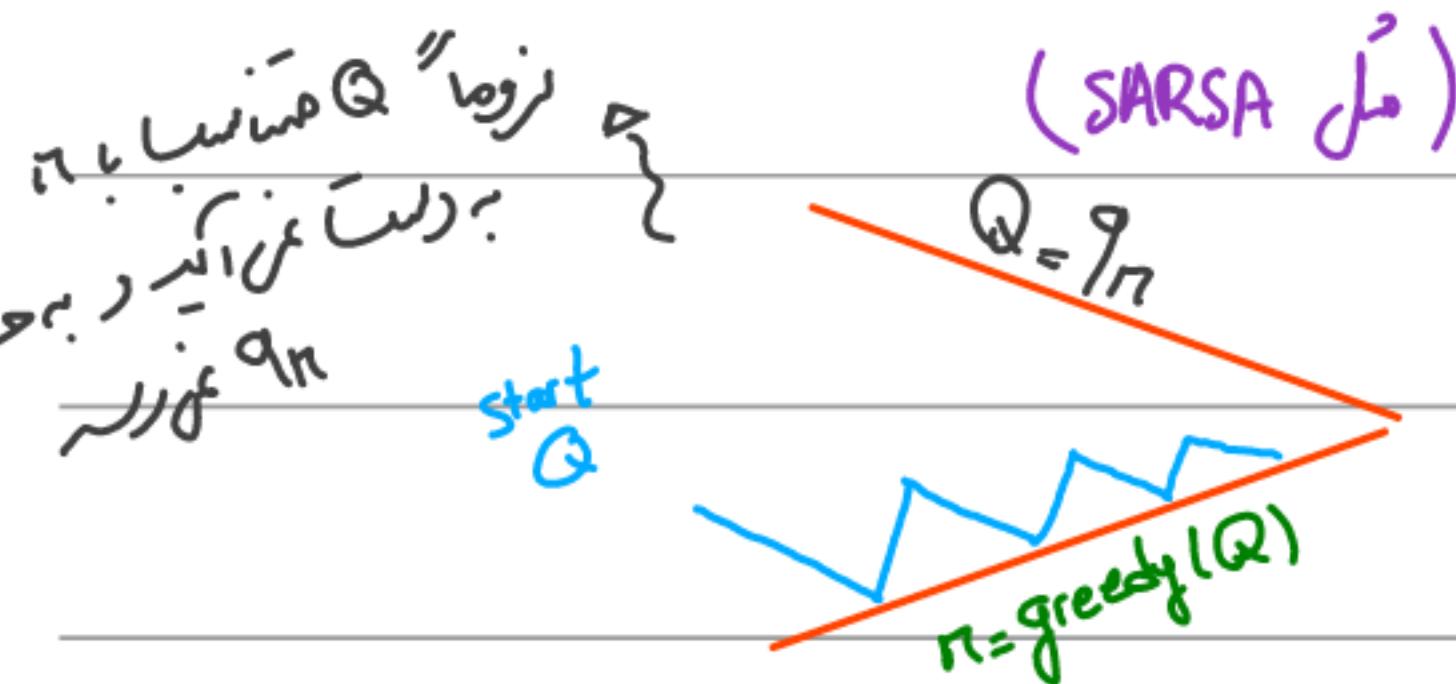
On-policy learning

Off-policy learning

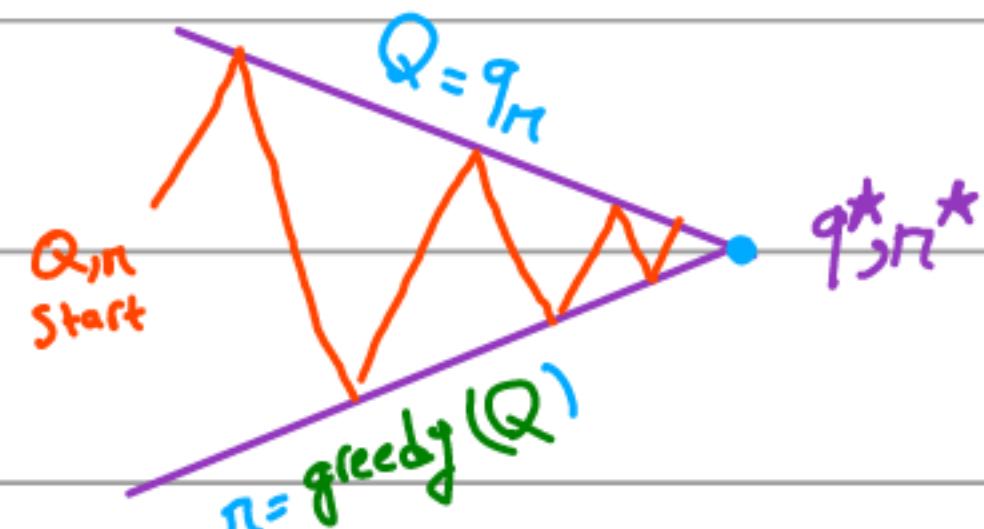
$$\pi_b = \pi_e$$

عن طريق

(ارقام زين)



$$\pi_b \neq \pi_e$$



→ مثال SARSA(λ)

$E(s,a)$ با مسیر backward

جهاز اینجا تابع off-Policy

Policy هو جسيم

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R + \gamma Q(s',a') - Q(s,a))$$

$$Q^*(s,a) \leftarrow \mathbb{E}_s [R(s) + \gamma \max_{a'} Q^*(s',a') - Q(s,a)]$$

Bellman operator

ازین طبقه کیم دیگر نیز کیم جیسے Policy چیزیں وانحریست میں لایکن را زدن اسکو در

$$q_{\text{tri}}(s,a) = q_t(s,a) + \alpha_t(R_{t+1} + \gamma \max_{a'} q_t(s_{t+1},a') - q_t(s_t,a_t))$$

با این دلیل Q-learning

ست، off-Policy

آخر بیشتر بازهم همین شد

(خوب خوب است)

البته بیشتر هر قدر من

جلسہ نمبر ۱۹، ۲۰ جولائی ۲۰۲۳ء

پیشگفتہ state کے DQL کا + TD, DP اور جو اس کا

Generalization ہے جو اس کا state کا پیشگفتہ چہ بایکرڈ؟



جس کا وارنر تابع باشد
(شبکہ ملبوس، یا جھپٹا یا ...)

$\pi = \epsilon\text{-greedy } \hat{Q}_w(s,a)$

خط میکنے برائی تک مجبوبہ ہے

(forgetting) جوں متعلق بہول نہیں؟

$(\phi(s_i, a_i), G_i)$

Monte Carlo ہے

لطف تیزی میں mini-batch

$(\phi(s_i, a_i), r_i + \gamma \hat{Q}(s_{i+1}, a_{i+1}))$ SARSA ہے

Replay Buffer

$(\phi(s_i, a_i), r_i + \gamma \max_a \hat{Q}(s_{i+1}, a))$ Q-learning ہے

* اپنے دل کا امنیت نہیں کیسے کرے؟

خوب تیزی سے دوسری بار train کرنے کا درجہ هر باری پہلی بار train کرنے کا درجہ بعد دارہ طرح جیسے

بس طبعاً = > تیری نہیں i.i.d. میں نہ

+ توجہ جو میں مارکوو مارکوفی مدل ایسے ہیں (Atari یا RNNs)

لہ براک ایتے زمان فریکر جو طرف دفعتہ

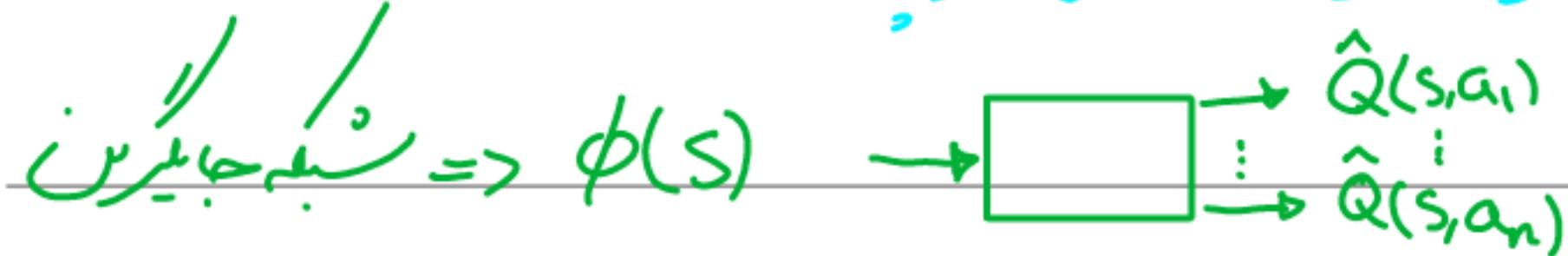
target ٹینک را زندھن لیں یا target کا جلو بڑا دفعہ

درستہ = > جائی \hat{Q}_w اس کا استدھار کرنے والے دلalloh بار

بلکہ بندھن لیں

وین ایجاد مقایم مورد DQN (با انری معنی نه) \rightarrow چون Q در جعبه می تواند بحذف از replay نیست

فقط آنچه نداشت ($\phi(s), a$) خوب نیست بخصوص وقت و منظر را که دارد است



+ باعث خطا: MSE سه مورد تراجی خطا را برگرداند

در زمانی به پاس بینندگان رسید اما به پاس خود می رسد

خواری الگوریتم (حالات را بدهی باش) \rightarrow همه خطا های این مورد را برگرداند

که درین داده i.i.d باشد ابتدا می تواند دلیل پیوسته باشد

$$\min_{\theta} L = \mathbb{E}_s \left[(V_{\theta}^{\pi} - y_t)^2 \right]$$

$y_t = r_t + \gamma V^{\pi}(s')$

$\delta \theta^T \phi(s')$ خطا بردن: i.i.d

$$\min_{\theta} L = \mathbb{E}_s \left[(V_{\theta}^{\pi} - y_t)^2 \right]$$

وین خط انتقالی برای snapshot است زیرا در اینجا از ترجیح میان

و اگر تمیز را نداشته باشد باید این رسم را دستگیری کنند (Mining)

و حل فیتلان θ_t را به دست آور

$$\lim_{t \rightarrow \infty} \theta_t = \theta^*$$

مثال:

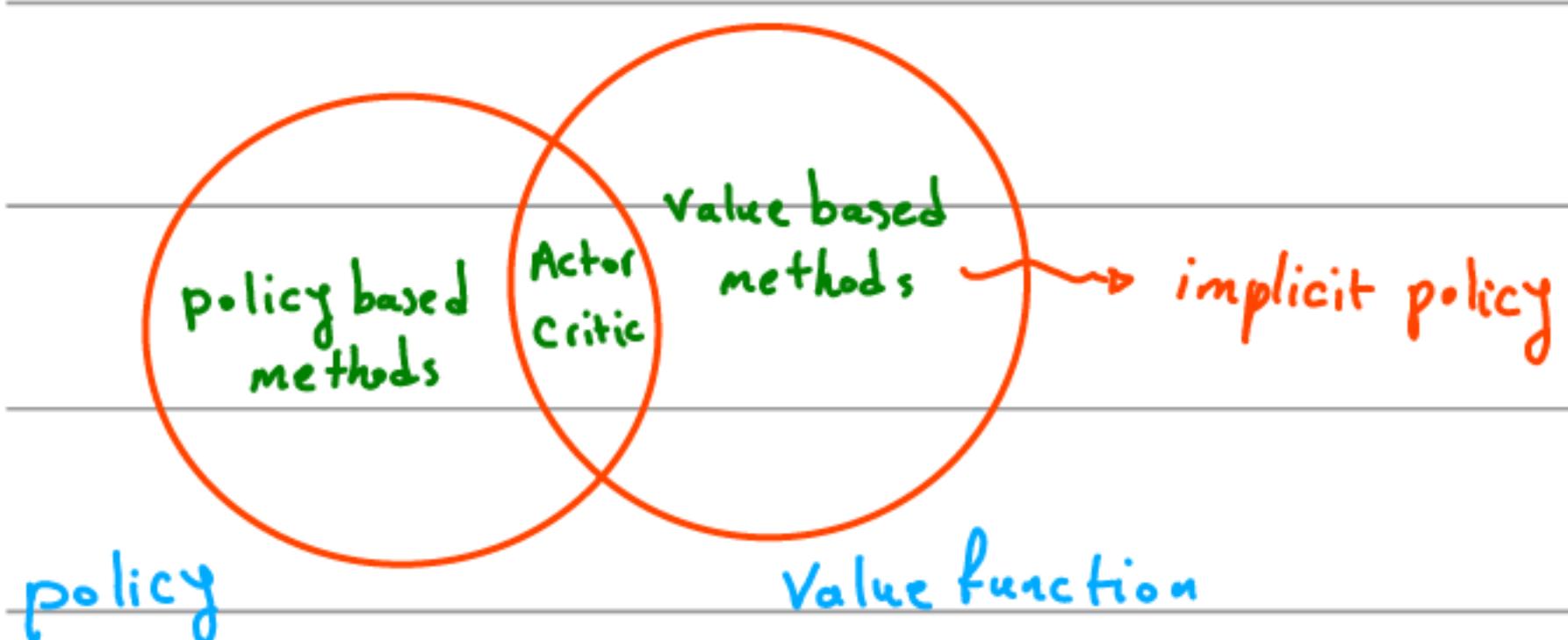
لین خطا داشته باشد

Policy Gradients

جلسہ ۲۰، ۱۲، ۲۲

چیزیں کیا Value Function ہے؟

کہ نہیں بھائی سیست ایسے دردار



دلیل گردنی مارکس اسکلی چون طبقہ

Value-Based Model-Free

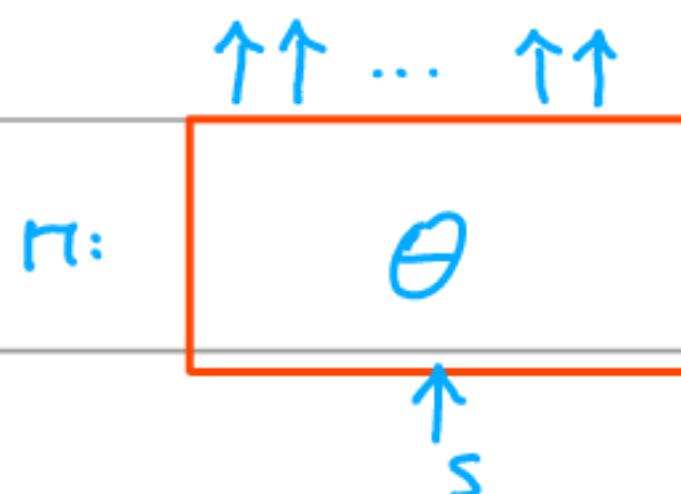
برحسب سیست ایسے درجہ حرارت

$$J(\theta) = E \left[\sum_t r(s_t, a_t) \right]$$

Policy optim...

Model-Based

الگوریتم

=> ہدف: $\max_{\theta} J(\theta)$ کے لئے دفعہ بڑھنے دیکھیں (یعنی)

؟ منہلان متنق راحاب کر (زیرا ارادہ ملکتے ہے Policy)

اجھی تینیں empirical دیکھیں، ابھی ایک ایسے دفعہ متنق کریں

دریں سعی کریں متنق ان رائکنیں بنیں

$$J(\theta) = E_{\tau \sim P_\theta} [r(\tau)] = \int P_\theta(\tau) r(\tau) d\tau \quad P_\theta(\tau) = P(s_0) P(s_1|s_0, a_0) \pi(a_1|s_1) \dots$$

episode ↪

$$\nabla J(\theta) = \nabla_\theta \int P_\theta(\tau) r(\tau) d\tau \stackrel{?}{=} \int \nabla_\theta P_\theta(\tau) r(\tau) d\tau$$

کہ τ دθ رابطہ فرم مارنے => اگر رجع نہیں پرسکھی خوبی دانتہ ہے درجہ حرارت

(Expected) P_θ کا دلایم دمنہلان اسکل کو طبعی سمجھ کر باید یہ فرم قابل تینیں سمجھیں کر (منڈبز) کر؟

$$\hookrightarrow \nabla J(\theta) = \int \nabla_\theta P_\theta(\tau) r(\tau) \frac{P_\theta(\tau)}{P_\theta(\tau)} d\tau = \int \nabla_\theta \log P_\theta(\tau) r(\tau) P_\theta(\tau) d\tau$$

$$= E_{\tau \sim P_\theta(\tau)} [\nabla_\theta \log P_\theta(\tau) r(\tau)] = \frac{1}{N} \sum_i \nabla_\theta \log P_\theta(\tau_i) r(\tau_i)$$

محضہ میں رات کا ہے

↳ باشد جمله مکن این طاری اطمینان معتبر $P_\theta(s_t | s_{t-1}, a_{t-1})$ معزز

$$\Rightarrow \nabla J(\theta) = E_{T \sim P_\theta} \left[\nabla_\theta \log P_\theta(T) r(T) \right] = E_{T \sim P_\theta} \left[(\nabla_\theta \log \pi_\theta(a_t | s_t) + \dots + \log \pi_\theta(a_T | s_T)) r(T) \right]$$

$$\approx \frac{1}{N} \sum_{i=1}^N \left(\sum_{t=1}^T \nabla_\theta \log \pi(a_t | s_t) \right) \sum_{j=1}^T r_j^{(i)}$$

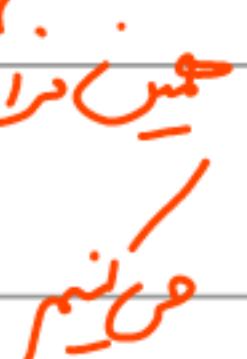
trajectory N :

حلاج نزدیک Backprop

$$\Rightarrow \theta^{(t+1)} \leftarrow \theta^{(t)} + \alpha \nabla_\theta J(\theta)$$

نیز MLE روی N میزد است و در این انتقال reward (حرس یعنی راهنمایی)

* عیز نایاب است

طی T را یافته هر دستگاه (یه مکن منزه / حرب)  

نزدیک دستگاه به امید

* پیش چرخی و خطایت

چنان Exploration وجود دارد و میزد آن را اتفاق برای این دویچ است؟

نماین الگوریتم: REINFORCE (مادر ۱۹۸۰)

* اگر action میزد باند حالت مطرد را ب تغییر نهاد فایده میزد میلین تغییر را میزد

نه میزد دوین الگوریتم مشتق را میزد

$\sum r_i$

Actor Critic چالش فاریاس (راه / محدودت برطین / رون) \rightarrow ابتداء دریز رون میزد Value به روش دویز

دنبت میزد به طور میلین صفر خواهد

Causality Trick

↳ پارس خصیت میلند به خط + مرتبط نیست و تأثیر یک اطاعت نداشند!

$$\Rightarrow \frac{1}{N} \sum_{i=1}^N \left(\sum_{t=1}^T \nabla_\theta \log \pi(a_t | s_t) \right) \sum_{j=1}^T r_j^{(i)} \xrightarrow{\text{num}} \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \nabla_\theta \log \pi(a_t | s_t) \left(\sum_{t'=t}^T r_{t'}^{(i)} \right) \xrightarrow{\text{reward to go}}$$

لهمایز جزئیات Q قابل توجه است

$$\nabla J(\theta) \approx \frac{1}{N} \sum \nabla_{\theta} \log P_{\theta}(t) [r(t) - b]$$

جُمِنْ نَارِيْبِ سَتْ

حَلْ:

ابْتَهَ رَتْرَكِ اِنْ مَعَكَ نَبْتَ نَسْتَ

$$E_{t \sim P_t} [\nabla_{\theta} \log P_{\theta}(t)] = 0$$

اَرْ بَلْنَادِمْ (r(T_i))

اَسْ طَارِيْسْ كَمْ مِنْ سُرْ دَرِيْسْ دَهْرَ اِنْ سُرْ

REINFORCE with baseline

بَانْ رَوْنْ

حبل يزدھم علارا

Policy Gradients مرور +

REINFORCE FOOT

لے جائیں رفع اصلان (I) اسَ!

راہگری حسن بن

لے بول بولنے کا ایک انتہائی نسبتی مفہوم (انٹریوپریسٹ) (discount factor)

یوں تسلیخ کرنے کا سب سے بڑا ترکیب (ہر کس میں دل کا حل دیکھئے) : Causality trick

$$\text{از روابط حلبی سچه من سُر دل مُنیرک است} \quad \rightarrow \quad \sum_{t'=t}^T \gamma^{t-t'} r_{t'}^{(i)}$$

$$④ \quad \sum_{t'=t}^T \gamma^{t'-1} r_{t'}^{(i)} \quad \hookrightarrow \text{reward to go}$$

لہ اسے b بیسٹ ہے تھیں بارہنے θ (انٹی چمچ جمل)

$$\frac{d \text{Var}}{db} = 0 \rightarrow b = \frac{E[g(\tau)^r r(\tau)]}{E[g(\tau)^r]}$$

$$g(\tau) := \nabla_{\theta} \log P_{\theta}(\tau) f(\tau)$$

ویریم کہ اس کے سوچ State dependent جو baseline ہے اور Actor critic association کا

Actor-Critic / π

$$\sum_{t'=t}^T r(s_{t'}^i, a_{t'}^i) \longleftrightarrow Q^\pi(s_t^i, a_t^i)$$

لے لذ اُن استعداد کیم در در دامع داریون / لم تمن سُر زیرا نہ یعنی امیر عین الٰت نہ فرن

trajectory مسار

۴ مَسْعِمَ إِنْ رَأَيْتَنِي بِرَبِّنِي هـ حَلَالِي سُرْد

: (is state dependent) previous baseline b_t

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t}|s_{i,t}) (Q(s_{i,t}, a_{i,t}) - b_t)$$

$$b_t = \underbrace{\frac{1}{N} \sum_i Q(s_{i,t}, a_{i,t})}_{V^{\pi}(s_t)} \quad a_{i,t} \sim \pi_{\theta}$$

بُلْدِي خفَق داشت سمت است b_t حم Q را علی سمت است

انظر Q در قسم تداریت و Actor در قسم تداریت

Advantage

- سطح

$$A^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) - V^{\pi}(s_t)$$

$$\Rightarrow \approx r(s_t, a_t) + V^{\pi}(s_{t+1}) \quad \text{Temporal difference}$$

$$\Rightarrow A^{\pi}(s_t, a_t) = r(s_t, a_t) + V^{\pi}(s_{t+1}) - V^{\pi}(s_t)$$

نقطه ∇ نایم راحلا سبب می شود که این را!

training data: $\{(s_{i,t}, \sum_{t'=t}^T r(s_{i,t'}, a_{i,t'}))\}$

بن سبیت در Monte Carlo

متداول داشت یاد نهاد

این سبیت critic

برست آرڈن $A^{\pi}(s, a) \Rightarrow \nabla_{\theta} J(\theta)$

$$\approx \sum_i \nabla_{\theta} \log \pi_{\theta}(a_{i,t}|s_i) A^{\pi}(s_i, a_i)$$

A2C

$\Leftarrow \Theta \Leftarrow$

(دیگر شکل) می خواهی!

جذبیت ۳، ۱، ۱۹

Policy gradients / مروج / محسن طبقیں درج +

Batch Actor-Critic Algorithm

$$\hookrightarrow \text{optimization:} \quad \min_{\phi} \|\hat{V}_{\phi}(s_t) - \text{target}\|_p^q \quad \xrightarrow{\text{خوبی ریکارڈ! بحث دلچسپی}}$$

$$\text{MC: } \sum_{i=1}^t r(s_{t+i}) \quad \leftarrow$$

$$\text{TD: } r(s_{t+1}) + \gamma \hat{V}_{\phi}(s_{t+1}) \quad \leftarrow$$

$$\hookrightarrow A^n(s_i, a_i) = r(s_i, a_i) + \hat{V}_{\phi}^n(s'_i) - \hat{V}_{\phi}^n(s_i)$$

Online Actor-Critic

4

$$\hookrightarrow \nabla_{\theta} J(\theta) \approx \sum_i \nabla_{\theta} \log \pi_{\theta}(a_i | s_i) \hat{A}^n(s_i, a_i)$$

جذبیت محسن شد!

$$\hookrightarrow \theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta)$$

 $\nabla_{\theta} J(\theta) \rightarrow A$ بهترین V را دریافت کرد!

اما طبقیں درج!

ایدیا / جذبیت در جریان و بازگشتی trajectory که worker چوی: ایدیا Δ

ایدیا / جذبیت asynchronous میباشد (A3C)

ایدیا / جذبیت vs. پیشنهادی از V Value Network DQN Δ ؟ایدیا / جذبیت Momentum Δ Off-Policy Δ ← Sample Inefficiency جذبیت Δ (s, a, s', r) Replay Buffer Δ \hookrightarrow (batch Δ) \hookrightarrow replay $\Delta \rightarrow \log \pi_{\theta}(a_i | s_i), y_i = r_i + \gamma \hat{V}_{\phi}^n(s'_i)$ ایدیا / جذبیت نیازمند بررسی داده هایی برای π_{θ} !

برکشنه V بودم است Q. نیوکلوزمینت از آن باید! Δ

حی جذب کننده بگیرم

$$y_i = r_i + \gamma \hat{Q}_\phi^*(s'_i, a'_i)$$

به بزرگتردن \hat{Q}_ϕ^* باعث

(MSE) (جهات خطا)

(میر با env تعامل نمایم و پیغام)

[... و میخواهیم را انگلندام نزدیک * حالت طبیعت زینت می‌فرمود

برکشنه (s_i, a_i, r_i, s'_{i+1}) بودم می‌نمایم (ینجایی که اینجا نمایم)

برحسب π از s_{i+1} Δ

[صلای ویدیویی بازگشایی شده تقطیعات :-]

+ سی پی‌ریس داده ریست یعنیدک (s) P(s) بود

دیگر ترجیح لست رو درست

A2C / مدل علیور

جلسه سیزدهم

(بیان نظر عصب بردن جبهه (درازهم))

من چه پیشنهاد

کار کن؟ (گرایین بطل کام مرتبط است دو)

که نسبت سیست بیش زیاد در این بین خوب نیست

که در چه راست خوب کار می کند؟

evaluation + improvement ← یا Policy iteration ← نسبت

advantage :

ارزش نرم سیست

کارکرد نسخه پیشین را در PG به طلایم می بینیم (مانند در بعد از دادن طلایم مسدات بالش)

$$J(\theta) = E_{\tau \sim P_\theta(\tau)} \left[\sum_t \gamma^t r(s_t, a_t) \right]$$

چه تغییری می کنیم و یا کرد؟

$$\tilde{J}(\theta') - J(\theta) = E_{\tau \sim P_\theta(\tau)} \left[\sum_t \gamma^t A^{\pi_\theta}(s_t, a_t) \right]$$

می خواهیم بگفت که $\theta' = \arg \max_{\theta'} \tilde{J}(\theta') - J(\theta)$ باشد و $\Delta(\theta'; \theta) < \epsilon$

که مسئله ایش کمین مین می کند احتمال نیم دو PG ربط دارد

برای $\theta' > \theta$ درست است

$$J(\theta) = E_{\tau \sim P_\theta(\tau)} \left[\sum_t \gamma^t r(s_t, a_t) \right]$$

$$= E_{\tau \sim P_\theta(\tau)} [V^{\pi_\theta}(s_0)]$$

خط احتسابی (روت بردن که نیست)

$$= E_{\tau \sim P_\theta'(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t V^{\pi_\theta}(s_t) - \sum_{t=1}^{\infty} \gamma^t V^{\pi_\theta}(s_t) \right]$$

$$\Rightarrow \tilde{J}(\theta') - J(\theta) = E_{\tau \sim P_\theta'(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) + \gamma^t (\gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t)) \right] = E_{\tau \sim P_\theta'(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t, a_t) \right]$$

اما زیرا P_θ' نایاب است importance sampling زاید است

$$\Rightarrow J(\theta') - J(\theta) = E_{\tau \sim P_{\theta'}} \left[\sum_t \gamma^t A^{\pi_\theta}(s_t, a_t) \right] = \sum_t E_{s_t \sim P_{\theta'}} \left[E_{a_t \sim \pi_\theta(a_t | s_t)} \left[\frac{\pi_\theta'(a_t | s_t)}{\pi_\theta(a_t | s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right]$$

لما $P_{\theta'}$ متعالٌ لـ π_θ ؟ ومتى θ حينئذ يُحدّد؟

$$\bar{A}(\theta') := \sum_t E_{S_t=p_{\theta'}} \left[E_{a_t \sim \pi'_{\theta}(a_t|S_t)} \left[\frac{\pi_{\theta}(a_t|S_t)}{\pi_{\theta'}(a_t|S_t)} r_t + \bar{A}_{(S_t, a_t)}^{\pi_{\theta}} \right] \right]$$

این رفع را بجا بینیم ایش

$$P_{\theta}'(s_t) = (1-G)^t P_{\theta}(s_t) + (1-(1-\epsilon)^t) P_{\text{mistake}}(s_t)$$

صلت طرح زیر مرسود (درستی)

$$\Rightarrow |P_0'(s_t) - P_0(s_t)| \leq \gamma \epsilon t \quad \rightsquigarrow \quad \epsilon \downarrow \text{oder}$$

$$E_{P_0'}[f(s_t)] = \sum_{s_t} (P_0' - P_0 + P_0) f \geq [P_0 f - |P_0 - P_0'| \max_{s_t} f(s_t)] \geq E_{P_0}[f(s_t)] - r \epsilon t \max_{s_t} |f|$$

لـ retC حلـ θ تـ θ $J(\theta) - j(\theta)$ بـ θ

$O(T_{\max})$ Advantage of implementation

$$\Theta' \leftarrow \operatorname{argmax}_{\Theta'} \sum_t E_{s_t \sim P_\Theta(s_t)} \left[E_{a_t \sim \pi_\Theta(a_t|s_t)} \left[\frac{\pi_{\Theta'}(a_t|s_t)}{\pi_\Theta(a_t|s_t)} \nabla_\Theta \pi_\Theta(s_t, a_t) \right] \right]$$

از کنکور برای آن دستگاه میتوانیم

Total Variation Distance

$$s.t \quad P_{\epsilon}(\pi_{\theta'}(a_t | s_t) || \pi_{\theta}(a_t | s_t)) \leq c \quad \text{---} \quad \text{الخطوة 4 // الـ } \epsilon$$

Δ s.t $D_{KL}(\pi_\theta'(a_t | s_t) || \pi_\theta(a_t | s_t)) < \epsilon \rightarrow$ الباي \in معاذري من سود

نیاں بہ عزیزیں جسیت Policy

$$\max_{\theta'} \min_{\lambda} L(\theta'; \lambda) \rightarrow$$

فیداری دارند برای همین خوب
کار کردن بگردان در دوستم شود و نه را بسیه فن ننم

$$\theta' = \arg\max_{\theta'} (\theta' - \theta)^T \nabla_{\theta} J(\theta)$$

$$\nabla_{\theta'} \bar{A}(\theta') = \dots \stackrel{\text{معظم } \theta}{\Rightarrow} \nabla_{\theta} \bar{A}(\theta) = \dots \text{PG} \rightarrow \nabla_{\theta} J(\theta) \quad \text{s.t. } D(\pi_{\theta'}, \pi_{\theta}) \in \mathbb{E}$$

objective \rightarrow سعى تجاه \rightarrow بالاتجاه

(٢٤, ٥٣) جبر چهارم

و در حبس ندش

$$\theta' \leftarrow \operatorname{argmax}_{\theta'} \sum_t E_{s_t \sim p_\theta(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_\theta'(a_t|s_t)}{\pi_\theta(a_t|s_t)} r^t A^{\pi_\theta}(s_t, a_t) \right] \right]. \text{ منتهی چنین کی.}$$

$$\text{s.t. } D_{KL}(\pi_\theta'(a_t|s_t) \| \pi_\theta(a_t|s_t)) < \epsilon$$

که حل: با این لگاریتمی بخوبی λ , α , θ' را بخوبی پیدا کنیم.

$$\lambda \leftarrow \lambda + \alpha(D_{KL}(\pi_\theta' \| \pi_\theta)) \rightarrow \text{Dual gradient descent}$$

اما از گرایون زارک بدست

$$\max_{s, g(\theta)} f(\theta) \quad f(\theta) \approx f(\theta) + \nabla f(\theta)^T (\theta' - \theta) \quad \text{از سطح تلور استفاده شود.} \quad f(\theta') = \text{اهدم.}$$

Trust Region برای تقریب خطی با θ نزدیک باشد. دنباله ایجاد شود.

Policy Optimization (TRPO)

$$\nabla_{\theta'} \bar{A}(\theta') = \sum_t E_{s_t \sim p_\theta(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_\theta(a_t|s_t)}{\pi_\theta'(a_t|s_t)} r^t \nabla_{\theta'} \log \pi_\theta'(a_t|s_t) A^{\pi_\theta}(s_t, a_t) \right] \right]: \text{ مسیر}$$

$$\hookrightarrow \theta \text{ به این شکل: } \nabla_{\theta'} \bar{A}(\theta) = \dots \text{ PG} = \nabla_{\theta'} J(\theta)$$

بررسی close form J را هم بعد از خطی کنیم که D_{KL} است

وابدیهی ساخت مرتبه (نمای بزرگ): $\nabla_{\theta'} D_{KL}(\pi_\theta' \| \pi_\theta) = 0$

$$\nabla_{\theta'} \sum_a \pi_\theta'(a) \log \frac{\pi_\theta'(a)}{\pi_\theta(a)} = \sum_a \nabla_{\theta'} \pi_\theta'(a) \log \frac{\pi_\theta'(a)}{\pi_\theta(a)} + \pi_\theta' \frac{\nabla_{\theta'} \pi_\theta'}{\pi_\theta'}$$

$$\theta \text{ به این شکل: } \theta = \theta + \nabla_{\theta} \sum_a \pi_\theta(a) = 0$$

$$f(\theta') = f(\theta) + \nabla_{\theta} f|_{\theta=\theta} (\theta' - \theta) + \frac{1}{2} (\theta' - \theta)^T H_f(\theta' - \theta) + \dots \quad \text{درست:}$$

$$\sum_a H_{\pi_\theta'} \log \frac{\pi_\theta'}{\pi_\theta} + \nabla \pi_\theta' \frac{\nabla \pi_\theta'}{\pi_\theta} + H_{\pi_\theta(a)}|_{\theta=\theta} = \sum_a \frac{\nabla \pi_\theta' \nabla \pi_\theta^T}{\pi_\theta} \rightarrow \text{از اینجا لازم است که بجهت منفرد کنیم.}$$

$$\begin{cases} \max_{\theta'} \nabla_{\theta} J(\theta')^T (\theta' - \theta) \\ \text{s.t. } (\theta' - \theta)^T F(\theta' - \theta) \leq \epsilon \end{cases} \quad \text{::: \textcolor{blue}{\text{f}} \textcolor{red}{\text{f}} \textcolor{blue}{\text{f}}}$$

حلاصه دوستی Dual را زیر در جواب نوشته در آورده:

$$\nabla_{\theta'} J^T \Big|_{\theta'=\theta} (\theta' - \theta) - \lambda [(\theta' - \theta)^T F(\theta' - \theta) - c]$$

$$\nabla_{\theta'} f^T \Big|_{\theta'=\theta} - \alpha F(\theta' - \theta) = 0$$

يمكن العد من الأمام inverse

$$\text{حالات ممكنة} \Leftrightarrow \frac{1}{r}\alpha^*(F^{-1}\nabla f)^T F(F^{-1}\nabla f) \leq G \Rightarrow \alpha \leq \sqrt{\frac{r\epsilon}{\nabla f^T F^{-1} \nabla f}}$$

$\theta' = \theta + \alpha F^{-1} \nabla f$ میں سے ایک Policy Gradient میں خلوص *

دله میر دا سٽا

(جیسا کوئی وسیع نہ کر سکے گا)

[ایکٹ F دعیرہ دبرداری ریو F ورخ مام تازن (T)]

normalize θ بعدها نطبق TRPO على π_θ

Policy

* نظہ رج ۲ در RL عدم ترجیت / الیکٹریک بولاریزیشن
نہ اسکے بعد مختلف

Conjugate direction (σ_{ad})

حروف من كنور

میں بھرنا رہے گے اسی کا انت F^{-1}

Conjugate gradient \leftarrow

100

(π_1 , δ_{crit}) (Proximal Policy Opt) PPO π_{target} δ^*

يتم راحـت نسبـت $\frac{\pi_{\theta'}(als)}{\pi_{\theta}(als)}$ ما بين $1-\epsilon$ و $1+\epsilon$

جذب پارامتر (۰۳، ۱، ۲۸)

$$P_{KL}(\pi_{\theta'} || \pi_{\theta}) \approx (\theta' - \theta)^T F(\theta' - \theta)$$

برای F کو توان فرم ایده ریاضی برداشت آوردن (اول جذب طبقه بندی شده)

$$\begin{cases} \max_{\theta'} \nabla_{\theta} J(\theta')^T (\theta' - \theta) \\ \text{s.t. } (\theta' - \theta)^T F(\theta' - \theta) \leq \epsilon \end{cases}$$

صوت معلم:

$$\theta' = \theta + \alpha \nabla_{\theta} J \quad (\alpha \text{ پارامتری } \alpha)$$

close form

$$\alpha \leq \sqrt{\frac{2\epsilon}{\nabla_{\theta}^T F^{-1} \nabla_{\theta}}}$$

میتوانیم $\| \theta' - \theta \|$ را کم کنیم و بهینه سازی کنیم: $F = I$

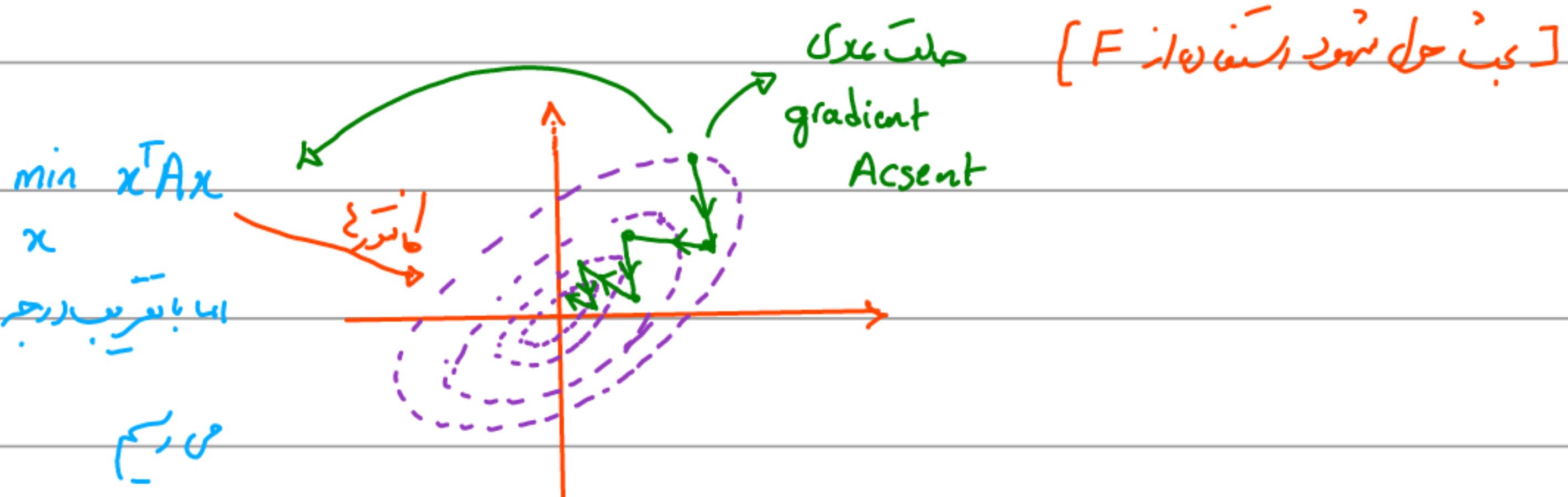
$$\theta' \leftarrow \theta + \alpha \nabla_{\theta} J \quad \alpha = \sqrt{\frac{2\epsilon}{\| \nabla_{\theta} J \|^2}} \rightarrow$$

steepest ascent

inf scale TRPO تحسیس از آن است که در راستای برداری ورودی چتربان را نظریه جریانی

$$F = \sum_i \frac{1}{\lambda_i} v_i v_i^T$$

روزی روزانه استم من کنم



نمای پارامتر = نمای خودکشیدگار

PPO میتواند

Proximal Policy Optimization

$\max_{\pi_\theta} \mathbb{E}_{s \sim \pi_\theta, a \sim \pi_\theta} \left[\frac{\pi_\theta'(a|s)}{\pi_\theta(a|s)} \mathcal{A}^{\pi_\theta}(s, a) \right]$

$$\arg\max_{\theta} \mathbb{E}_{s \sim \pi_\theta, a \sim \pi_\theta} \min \left\{ \frac{\pi_\theta'(a|s)}{\pi_\theta(a|s)} \mathcal{A}^{\pi_\theta}(s, a), \text{clip}\left(\frac{\pi_\theta'(a|s)}{\pi_\theta(a|s)}, 1-\epsilon, 1+\epsilon\right) \mathcal{A}^{\pi_\theta}(s, a) \right\}$$

منطقی
متناهی
متین

نهایی حزب حل محدود دارد

جستجوی حرکت از

TRPO و PPO

اینها در پیشنهاد می‌کنند: این بقیه همچنان وارد گزینه‌های مختلف می‌شوند

[کلید در اینجا]

این replay buffer وی بسیار sample inefficient است و حتی onpolicy است

SAC و DDPG
Soft actor-critic

این off policy Actor-critic کیمی را دارند

جستجوی نزدیک

مردحه کننده (تغیر لفظ مبحث)

$$D_{KL}(\pi_{\theta'}, \pi_{\theta}) \leq \epsilon$$

$$\xrightarrow{(\theta' - \theta)^T F (\theta' - \theta) \leq \epsilon} = [\lambda_i v_i v_i^T]$$

$$\xrightarrow{\sum \lambda_i \langle \theta' - \theta, v_i \rangle^2 \leq \epsilon}$$

$$\xrightarrow{\sum \lambda_i \Delta \theta_{(i)} \leq \epsilon} \text{تغییر } v_i$$

محاده برای سیمین نزدیک

و تغییر با عبارت از $\frac{1}{\lambda_i}$

برای طبقه بندی Gradient Ascent

$$\Delta \theta = \eta \frac{\nabla J}{b}$$

$$\Delta \theta^T F \Delta \theta \leq \epsilon$$

$$\xrightarrow{\sum \lambda_i (\langle v_i, \eta b \rangle)^2 \leq \epsilon} \eta \sum \lambda_i \langle v_i, b \rangle^2 \leq \epsilon$$

اگر b بزرگ است نزدیکی به مقدار b را داشته باشیم در حریث بقیه ابعاد منظم باشند این!حالا اگر b برویم لایم! natural gradient!

$$\Delta \theta = \eta F^{-1} \frac{\nabla J}{b}$$

$$\xrightarrow{\eta \sum \lambda_i (v_i^T F^{-1} b)^2 \leq \epsilon} \eta \sum \frac{1}{\lambda_i} \langle v_i, b \rangle^2 \leq \epsilon$$

$$\underbrace{v_i^T \sum_{j=1}^n \lambda_j^{-1} v_j v_j^T}_{b} b$$

$$\frac{1}{\lambda_i} \langle v_i, b \rangle$$

اگر b براي λ_{max} حرکت را بروی آن حدیث نهاده کنیم در این آنچه تر تغییر حریث می شود

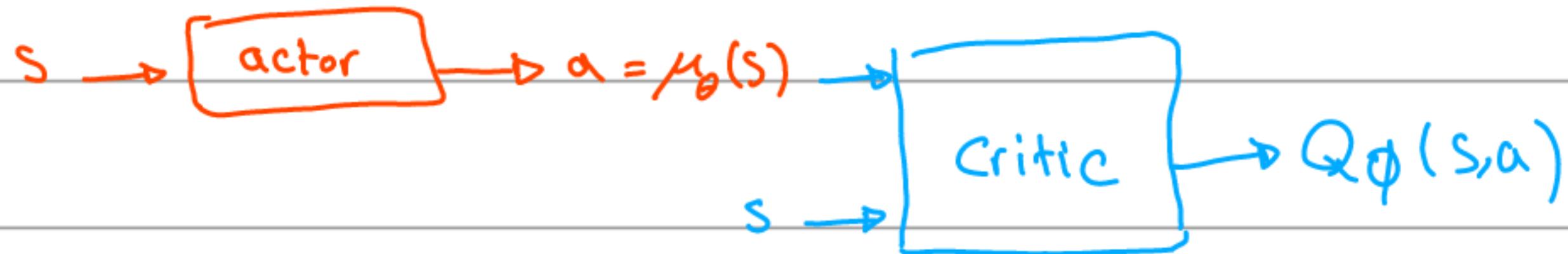
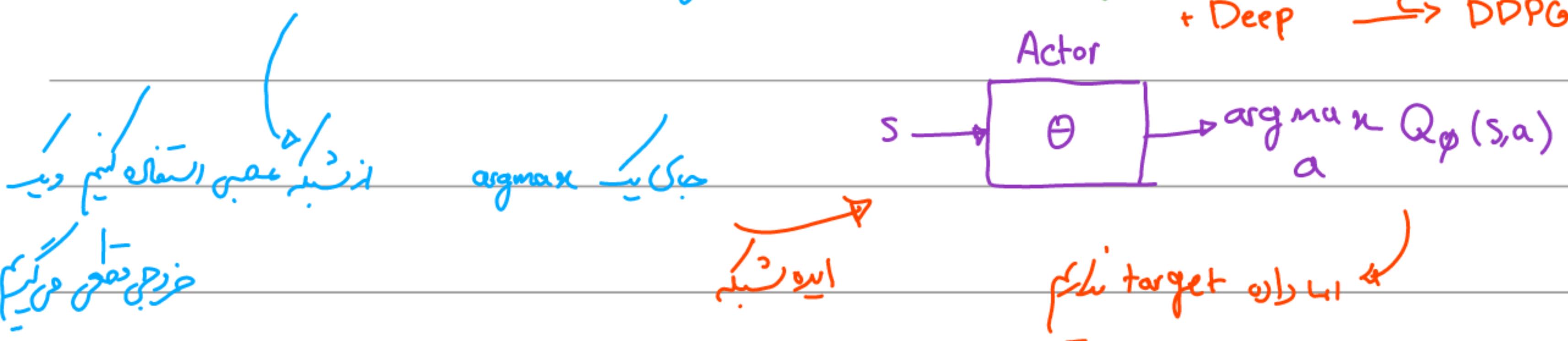
هیئت Off Policy بولن

(less sample efficient)

جلسن نورگری دنمون زیاد

سیم، action یا Q learning یا (Deterministic Policy Gradient) DPG

+ Deep \hookrightarrow DDPG



متغیرهای Hyperparameter هستند که می‌توانند تأثیراتی بر عملکرد داشته باشند.

کوچک کردن γ و ϵ برای اینکه در آینده اقدامات را بپیش‌بینی کنیم.

$$\text{target } y_i = r_i + \gamma Q_\phi(s'_i, M_\phi(s_i))$$

از شبکه‌ی جیال استفاده کنیم به هر دو دسته بینی برخیز

DQN خواهد

خط احتمال را انتخاب خواهد

exploration \leftarrow جلسن همان را انتخاب خواهد کرد

$$a_t = \mu(s_t | \theta^t) + N_t$$

که مخفف با اینکه نورگری را انتخاب کرد

جلسن که یعنی

Soft Policies

$$J(\pi) = \sum_{t=0}^T E_{(s_t, a_t) \sim \pi_\theta} [r(s_t, a_t) + \alpha H(\pi(\cdot | s_t))]$$

Entropy

له مزاعم عدم تفاصيل دقيقه بين سيستم بالبيان

تحت يبرازهم جنگنه برای اثاب را استثنیم

جذب exploration

مکن است خوب را استثنیم

$$Q^\pi(s, a) = r(s, a) + \gamma E_{s' | s, a} [V^\pi(s')]$$

عادل چشم نورد؟ Bellman

$$V^\pi(s) = E_{a \sim \pi(a|s)} [Q^\pi(s, a) - \alpha \log \pi(a|s)]$$

هدف: فرم همه سیاست (التصویر)

$$\pi^*(a|s) = \frac{\exp Q^*(s, a)}{\sum_{a'} \exp Q^*(s, a')}$$

جی softmax، max

طبع ممکن

درایه: حدود

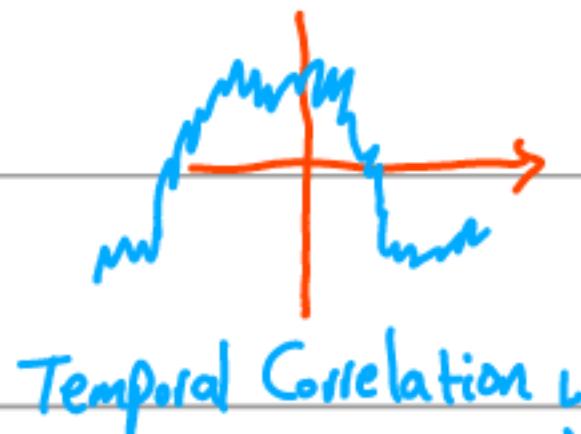
نمودار Q از Off Policy در مورد π

حیث معمم π_{target}

که زیرا مقدار از Policy مولان حالت را در

نیز کن i.i.d نیز داشت DDPG +

راسته بین مولان Temporal Correlation و Ornstein Uhlenbeck



ویژگی از در میان زیرا مولان explore میگردند

جز learning rate α

Critic i.e. actor $\nabla_{\theta} \text{LR}$

جزی از این

"Soft Policies"

$$\pi^*(a|s) = \frac{\exp Q^*(s,a)/\alpha}{\sum \exp Q^*(s,a')/\alpha}$$

درویں مطابق

$$\max_{\pi} V(s) = E_{a \sim \pi(\cdot|s)} [Q(s,a) - \alpha \log \pi(a|s)]$$

$$\text{s.t. } \sum \pi(a|s) = 1 \forall \pi(a|s) > 0.$$

$$\mathcal{L} = \sum_a \pi(a|s) Q(s,a) - \alpha \pi(a|s) \log \pi(a|s) - \lambda (\sum \pi(a|s) - 1)$$

$$\frac{\partial \mathcal{L}}{\partial \pi(a|s)} = Q(s,a_0) - \alpha \log \pi(a_0|s) - \alpha - \lambda = 0$$

$$\Rightarrow \pi(a_0|s) = e^{-\frac{\lambda + \alpha}{\alpha} Q(s,a_0)/\alpha}$$

برای این ساده تر نظر کنید

$$\pi(a_0|s) = \frac{e^{Q(s,a_0)/\alpha}}{\sum e^{Q(s,a)/\alpha}}$$

رایج میگیم با این نظر داشت این را reward r داشت : Policy Iteration

نماین با این برای میگردید s_{t+1} داشت $\pi(s_t, a_t) = r$

soft actor-critic

$$Q(s,a) \leftarrow r(s,a) + \mathbb{E}_{s' \sim p_{s,a}, \pi} [Q(s',a') - \log \pi(a'|s')] \quad \cdot Q \text{ جنگل}$$

عمران شد $\pi \leftarrow$

$$\pi(a|s) \leftarrow \operatorname{argmax}_{\pi} \left(\pi || \frac{e^{Q(s,a)}}{\sum_{a'} e^{Q(s,a')}} \right) \quad \cdot \pi \text{ بزرگان}$$

ریوند V و Q مینیمیز

$$J_V(\gamma) = \sum_{\substack{s \\ \text{minibatch}}} V(s) - \left(\sum_a (Q(s,a) \pi_\theta(a|s) - \alpha \pi_\theta(a|s) \log \pi_\theta(a|s)) \right)$$

$$J_Q(\omega) = \sum_{\substack{s,a \\ \text{minibatch}}} \left(Q_\omega(s,a) - (r(s,a) + \gamma V_{\omega}(s')) \right)^2$$

$$J_\pi(\theta) = D_{KL} (\pi_\theta || \frac{\exp(Q_\omega(s,a)) / \alpha}{k})$$

بنچادری فتنہ

"Model Based RL"

حدت مردیک

پارسیکن environment choice

+ اگر دینکن با جسم بیو چشم آن را لایک کریم:

$Q, V \rightarrow$ Policy π

→ محدودت دینکن ساخت

(پلین) Planning ساخت

$$P(s_1, a_1, \dots, s_T, a_T) = P(s_1) \prod_{i=1}^{T-1} T(s_i | a_i) P(s_{i+1} | a_i, s_i)$$

Model-Free RL: ساخت

فرضیه کریم ہیں رائی دیں

اممداد رہیں آن رائی دیں و ...

Open-Loop

Deterministic چک

$$a_1, \dots, a_T = \operatorname{argmax}_{a_1, \dots, a_T} \mathbb{E}_r[r(s_i, a_i) \text{ s.t. } s_{t+1} = f(s_t, a_t)]$$

Stochastic چک

$$\operatorname{argmax}_{a_1, \dots, a_T} E\left[\sum_t r(s_t, a_t) | a_1, \dots, a_T\right]$$

اممداد حالت تعداد منظم طور نسبت دیگر

→ نزد فیکر در صفت کل میز

نیز جزوی

Global

local linear چک

Local

$$k + s_t + k_t$$

Stochastic Optimizations

Open loop حل

نهایی زمانی نهایی محدودیت را در

$$a_1, \dots, a_T = \operatorname{argmax} J(a_1, \dots, a_T)$$

guess & check تنبیه و انتبار کن : Random shooting

اهلکردن : دهن : از Cross-Entropy را انتبار کن. حل تابع بازرسان

برترین کن دین زنگ (ترکیب CEM)

(episode) ابجاق احتمال (action) و طبقه بندی هر دسته (عاد)

کتابت مطابق با نتایج

Model Predictive Control MPC : close loop چگونه کرد ؟

برای T ممکن بین کن قطعیت \hat{x}_T بر جای درود به replan

کردن ؟ دین دین برای تعیین از خوب نیست !

Time-Varying Linear روش مبتدا بر

Discrete: Monte Carlo Tree Search

نهایی جایی کنیم از انتبار کن مسیر را فرخیم اما این دسته دهن نیست

(Replan) MPC

MCTS

نهایی جایی بود (ستاد) ایستادن را بفرمود، دهن آن را منتظر بفرمود (یعنی go)

؟ اما برای حالت تغیری خود؟ بفرمود مسیر را شرک کن دریک یک بروت همین

حدید کنم چیز که بفرمود و آنرا را میگیرم (UCB1)

جذب در گراف مغلق کیم؟

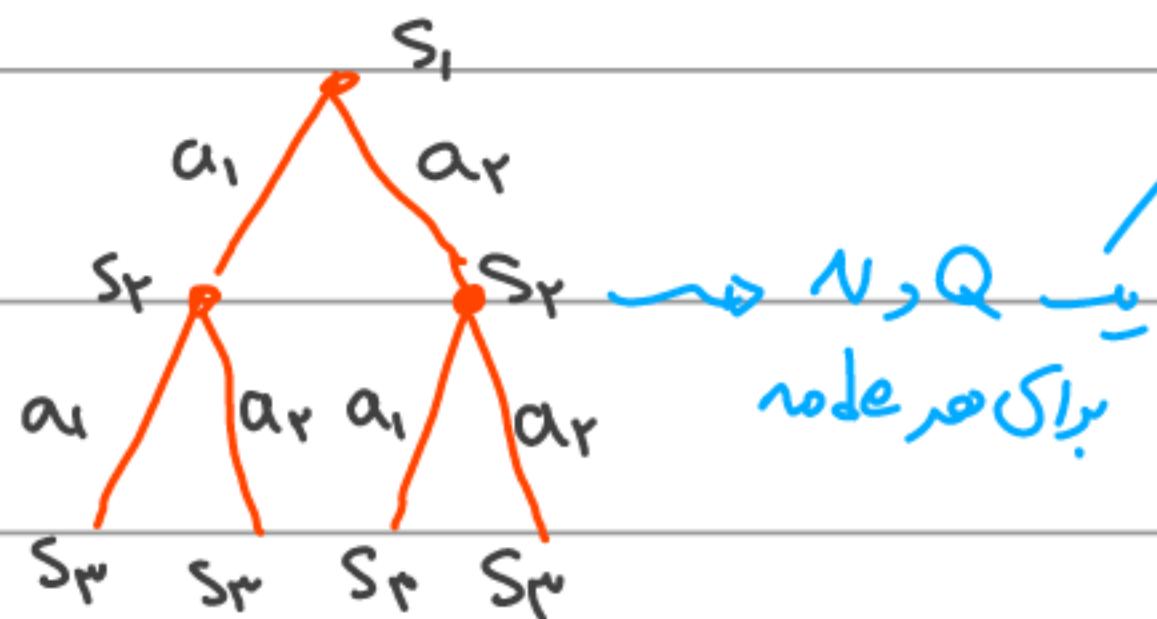
دایرکت reward rollout

$$\text{Score}(s_t) = \frac{Q(s_t)}{N(s_t)} + \gamma C \sqrt{\frac{\ln N(s_{t-1})}{N(s_t)}} \quad \text{Exploration}(s_t)$$

VCT → جدید action بررسی اگر جویید

وگزینه بررسی بهترین Score

در برداشتن سرعت دادن به صفت منحصر



جذع اندازی کیم بهترین راه را بخواهیم

حالاً اگر همان قطعیت کام نداشتم کیم دیگر از این replan

نهیم کیم از شروع معتبر استفاده کیم بدل کنیم

روشنی کیم

Monte Carlo Tree Search برعکس AlphaGo *

* برای برداشتن سرعت دادن (البته بازیگران فکر نمی‌کردند)

min-max ab*



و مدل داریم

$$\max \sum_{t=0}^{\infty} r(s_t, a_t)$$

$$\text{s.t. } s_{t+1} = f(s_t, a_t)$$

planning (با برای تخمین از نظر سعی کنیم)

از مدل داده شده محاسبه

(system identification)

نهایت مل ریزی طریق

نهایت

(Distribution Shift) [بی جایی]

که نیز در نظر گرفته شود

پس

Model Predictive Control (MPC)

: replanning از خطا راست نشیم

replan باز بحسب حسنه خود می دهند - : دیر تر باز replanning

زمان (T) کو افزایش داده تا آنرا برآورد کنند (برای برآورد کردن آن)

(cheetah) هر چند مدل-بازی مدل-بازی

! Overfitting

فرایند overshoot (برآورد) overfit (برآورد) replan

کرن بوسن؟

- مدل مطابقت ندارد

سرد آنها نشود

f # [f] می باشد
باشد

confidence interval

مکالمہ حسب علم تھیت!

لہ از ایہ آئریں ہستھاں سیم بدل کر رعنے میں مدد ملتے

لہ! خوب سستے زیرا خطا ری اکارک نہستے، سُنْتَرَنَتَ (مِلَّا لِلَّهِ بَعْدَ مِنْهُ)

دیجیتال پارسیانی $P(\theta | D)$ باشد؛ Bayesian معنی دارد.

$$\int P(\text{str}_i | S_{\text{true}}, \theta) P(\theta | D) d\theta$$

تُسمى هذه بـ *BNN* (Bayesian Neural Network)

$P(\theta | D) = \pi(\theta | D)$

$$P(\theta|D) \approx \frac{1}{N} \sum_i \delta(\theta_i)$$

$$P(\theta|D) \approx \frac{1}{N} \sum \delta(\theta_i)$$

$$\int P(s_{t+1} | s_t, a_t, \theta) P(\theta | D) d\theta \approx \frac{1}{N} \sum_i P(s_{t+1} | s_t, a_t, \theta_i)$$

(تَسْهِيْلِي اِخْرَجْي)

! SGD, init w randomness. ! resample w/o l

$$(ii) : \bar{J}(a_1, \dots, a_H) = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^H r(s_{t+1,i}, a_t), \text{ where } s_{t+1,i} = f_i(s_{t,i}, a_t)$$

جواب دادن با $\hat{\theta}$ که از نمونه i در مجموعه N انتخاب شده باشد، با استفاده از معادله (۱) می‌باشد.

بانک ایدها زیر مبنای مدل-free می باشد

لہ زندگانی High Dimension خوب جواب نہیں دے سکتے

بہ حوالہ ریڈر : planning خیلے زندگی کی مسالوں کی سرسری۔ دستِ فرنزی / میر کی مسالوں کی زندگی حواس

دیجیٹل Planner - سچے کام کا دنارہ

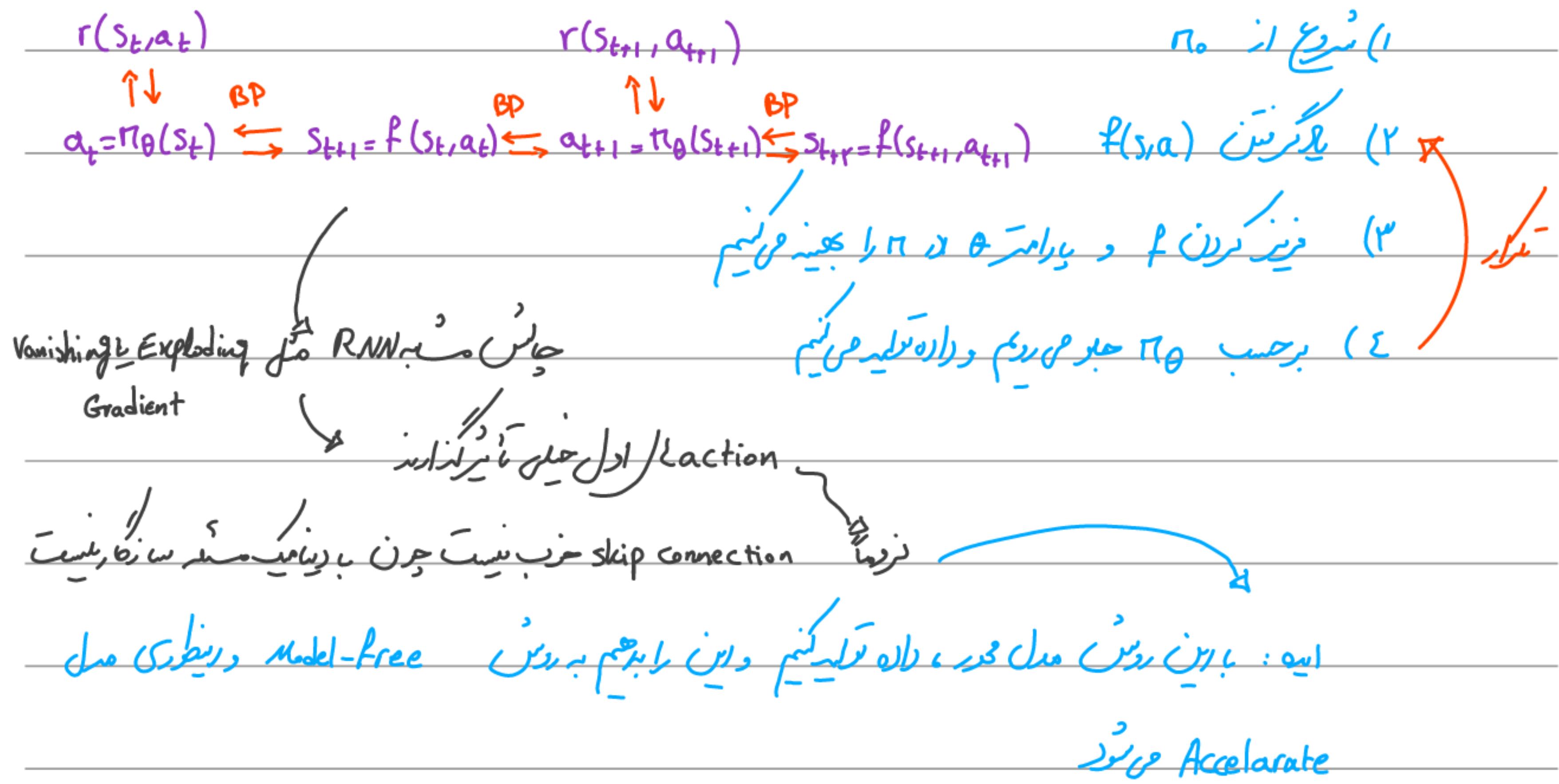
△ معین نصف open loop است. مدل فرایندتری از عملکارن یا نیازمندی با MPC به وقایع مسلط است

$\max_{a_1, \dots, a_H} \mathbb{E}[C_{t+1}]$ vs. $\max_{\pi} \mathbb{E}_{\pi}[\sum_t r_t]$

لے ایہ: ستارڈنگ ہی میسن بر مل یا گیرنے Policy برمیں

نیکوں کے لئے

Backpropagation Through Time



لکه در کامپیوٽر جسیکا BP از f طبقه بندی می‌کند و آن را به صورت model-free می‌دانیم.

بررسی short rollout شرکت میکنیم که این state چه تأثیری بر trajectory دارد

Dyna \rightarrow حل Q-learning

آخر حرف (و) يكتب كـ و (و) + دهان (و) \rightarrow butter; دھان

مودل فری آئرنیشن میڈیم Model-free learning میڈیم short rollout اور

خوب پیش میں نہیں
این درس کا خرچ

حصہ جوہم (۲۹، ۳۰)

دست نہستات!

soft actor critic ارادات

جیسے ... Q^{π_t} ... بیان حفظ ... $\alpha H(\pi(.|s_t))$... افکار

soft actor critic الگوریتم ... میں سے softmax ... حالت ہیئت میں

(کرنسی)

? D_{KL} چاہے خوشی Policy Improvement پر پسند

$$\pi_{\text{new}}(.|s_t) = \arg\min_{\pi'} D_{KL}(\pi'(.|s_t) || \exp(Q^{\pi_{\text{old}}}(s_t, \cdot) - \log Z^{\pi_{\text{old}}}(s_t)))$$

$$= \arg\max_{\pi'} \bar{D}_{KL}(\pi'(.|s_t))$$

اندازہ جیسے کرنے کی طرف میرے ... $\pi_{\text{new}} \geq \pi_{\text{old}}$

$$\arg\max_{\pi'} \sum_a \pi'(a|s) (Q^{\pi_{\text{old}}}(s, a) - \log \pi'(a|s))$$

سبسے $V^{\pi}(s)$

حالت خود پر زیر احمد π_{old} نہ π_{new}

$$E_{a \sim \pi_{\text{new}}(.|s)} [Q^{\pi_{\text{old}}}(s_t, a_t) - \log \pi_{\text{new}}(a_t|s_t)] \geq V^{\pi_{\text{old}}}(s_t) : \text{کوئی تغیر نہیں}$$

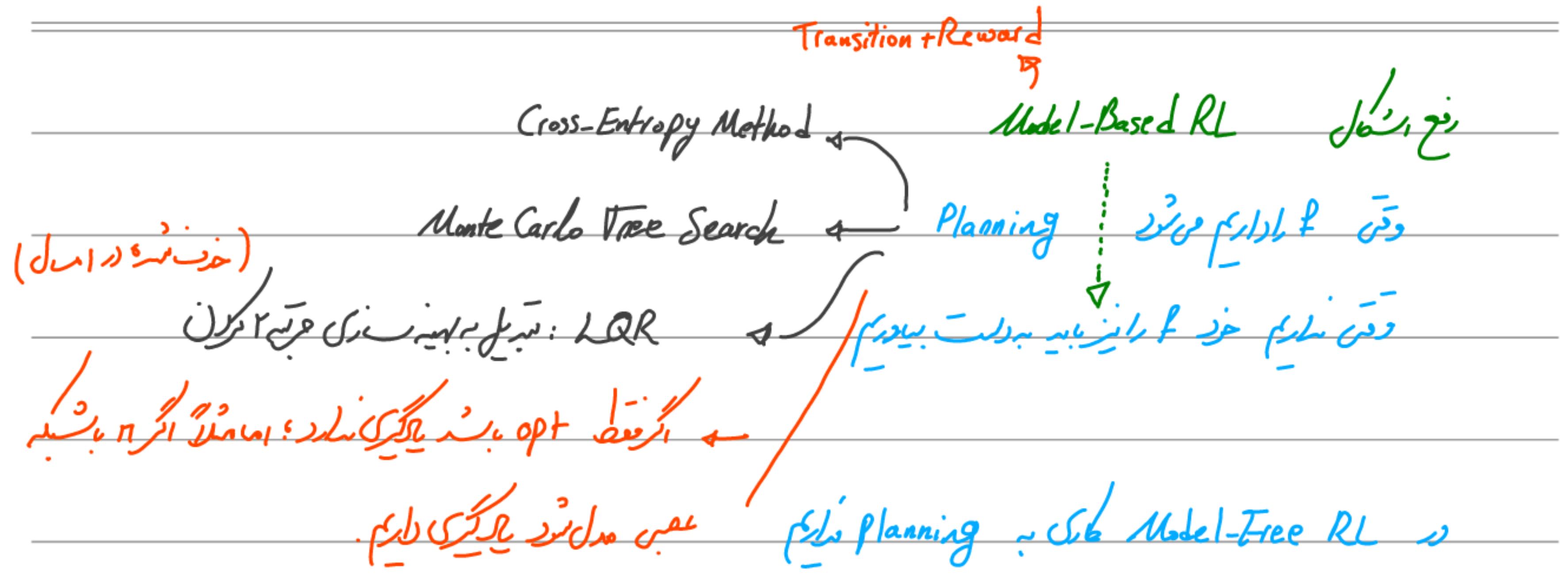
$$\Rightarrow Q^{\pi_{\text{old}}}(s_t, a_t) \leq Q^{\pi_{\text{new}}}(s_t, a_t)$$

$$(a_t) \pi_{\text{new}}(a_t|s_t), Q^{\pi_{\text{old}}}(s_t, a_t) \leq V^{\pi_{\text{old}}}(s_t) + E[X] \leq E[Y] \times Y \quad (\text{ایسا ہے})$$

π_{old} اور حسب π_{new} بر حسب action

Policy Improvement کیم!

جتنی نسبت بحالات تبلی (آئندہ اکٹھونے کے لئے) متعال و ملائیں اور تغیر کرنے کا Q +



حالاتی که در اینجا Δ stochastic می‌باشد، در اینجا Δ action می‌باشد. اگر باید پیش‌بینی کرد که در اینجا Δ action می‌باشد، ممکن است Δ CEM را با Δ GMM مترادف ندانید.

The diagram illustrates the flow of information and control in two types of systems:

- Open-Loop**: Represented by a purple arrow pointing from **perception** to **control**.
- Close-Loop**: Represented by a purple arrow pointing from **control** back to **perception**.

A blue dashed arrow points from the **Open-Loop** to the **Close-Loop**, indicating a transition or comparison between them.

Below the loops, the text "in open-loop systems" is written in black ink.

On the left side, there is handwritten text in black ink: "control" above "ناریم" (we will), and "perception" above "حواس" (senses). Below these, a bracket groups "control" and "perception" with the text "is sub-optimal" in black ink below it.

On the right side, there is handwritten text in blue ink: "control" above "ناریم" (we will), and "perception" above "حواس" (senses). Below these, a bracket groups "control" and "perception" with the text "is optimal" in blue ink below it.

بررسی مفهوم مترید و دیرید + Partial Observable + روش‌های
Bayesian RL بحث زیادی مطرح نموده بگزینه کن

۴، ۲، ۱۱ جذبیت

 ϵ -greedy حملن "Exploration and"

بلان حینی بے پایان باری نزدیکی میں میں: Montezuma's revenge game

(sub task) Temporally extended task

جزئی task میں چرخنا

 $(1-\epsilon) \in \Omega(k) \Omega(1)$: explore کرو بہم درختی exploit بھی در $k-1$ خط قطبی pick task بڑی

کوئی خوبی نہیں → explore وہ کو انجام دلے → Sub Optimal

خطابی؟

مخصوص بنتی حینی exploration میں ہے

$$\text{Reg}(T) = TE[r(a^*)] - \sum_{t=1}^T r(a_t)$$

اویسیف یعنی بدل سمت ہے

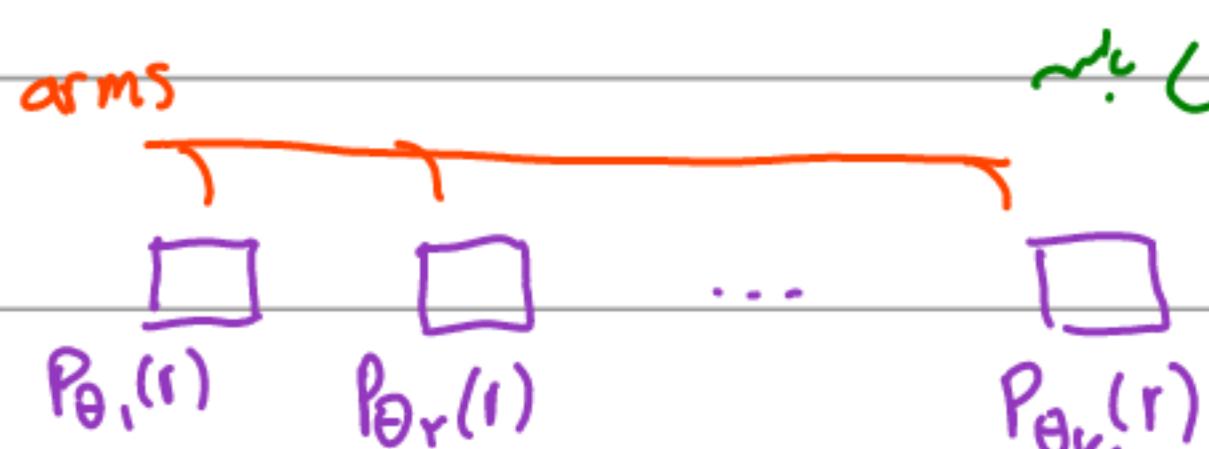
$$E[\sum_t r_t^{a^*}]$$

یا یعنی Regret ہونم

جیسا کہ پیشہ لیا ہے تو

رسنگ خاص میں تسلیم ہے اور درست تحسین میں کرکن باری $\lg(T)$ بدل سمتکرکن پین $O(\lg T)$ کا مامن

کوئی براک مادھلات کرو ہم کسی بھی



Multi Armed Bandit Problem

تو stateless میں

هر باری دست رکھنے کا منظور تین چیزوں میں

Optimistic Exploration

Upper Confidence Bound

($\hat{\mu}_a$) میکنیں جیسے bandit هر جو reward میں مبنی ہے

$$a = \operatorname{argmax} \hat{\mu}_a + C \sigma_a \quad \text{std}(\hat{\mu}_a) \sim \frac{1}{\sqrt{n(a)}}$$

لہ انتہا جو چند انتہا راستہ نہیں کہ جو ایسا نہ رہتا ازدھے

بندھوں میں سے ایسا جو انتہا خوب بارہ تسلیم کرے

جیسا کہ انتہا میں مبنی std میں مبنی ہے

جیسا کہ $O(\log T)$ برا کی پیشہ نہیں میں مبنی

Concentration Inequalities

Markov Inequality: $\forall X \geq 0 : P[X \geq \epsilon] \leq \frac{E[X]}{\epsilon}$

Cheinoff Bound: $P[X \geq \epsilon] = P[e^{tX} \geq e^{t\epsilon}] \leq \frac{E[e^{tX}]}{e^{t\epsilon}}$

$\Rightarrow P[X \geq \epsilon] \leq \inf_{t > 0} \frac{E[e^{tX}]}{e^{t\epsilon}}$

ایسا جو f کا انتہا میں سے ایسا جو $f(E(X)) \leq E(f(X))$

$$\begin{aligned} E[e^{tX}] &= E[e^{t(a+b)}] = e^{ta} E[e^{tb}] = e^{ta} \left(\underbrace{\frac{e^{tb}}{b-a} + \frac{b}{b-a}}_{g(u)} \right) \\ &\leq E[\alpha e^{tb} + (1-\alpha)e^{ta}] = \frac{\alpha e^{tb} + (1-\alpha)e^{ta}}{b-a} = e^{ta} \underbrace{\left(\frac{-\alpha}{b-a} e^{tb} + \frac{b}{b-a} \right)}_{g(u)} \end{aligned}$$

$$\Rightarrow g(u) = \tau u + \log(-\tau e^u + 1 + \tau) \quad g'(u) = g''(u) = 0$$

$$g''(u) = \frac{P(1-P)}{u^2} \Rightarrow g(u) = \frac{u + \alpha u^2}{2!} + \frac{g''(\xi)u^2}{2!} < \frac{u^2}{2!} \Rightarrow$$

$$E[e^{tx}] \leq e^{g(u)} \leq e^{\frac{1}{\lambda} t^r (b-a)}$$

$$x \in \mathbb{R} \Rightarrow P[X \geq \epsilon] \leq \inf_{t > 0} e^{\frac{1}{\lambda} t^r (b-a) - t\epsilon} \Rightarrow t^* = \frac{\epsilon}{(b-a)^r}$$

(\Rightarrow Hoeffding Bound)

$$P\left[\frac{x_1 + \dots + x_n - E[X]}{\sqrt{n}} \geq \epsilon\right] \leq \inf_{t > 0} \frac{E[e^{tZ}]}{e^{t\epsilon}} \rightarrow e^{-\frac{t^r(b-a)}{\lambda}}$$

$$\dots \leq e^{-\frac{\epsilon^r n c}{C}} \quad C = (b-a)^r$$

$$\epsilon = \sqrt{\frac{\log \delta}{n}} \quad (\text{Concentrate})$$

برای اینجا بندی می‌شود \Rightarrow

(07, 1, 19) حسنه

Bandit $\Rightarrow +$

Optimistic Estimate: $\arg\max_a \hat{\mu}_a + C\sigma_a$

$\hat{\mu}_a + \sqrt{\frac{r \log T}{N(a)}}$

↑ Std of $\hat{\mu}_a$

از این روش مطمئنیت زدن نمی‌کند

$$\frac{1}{T} \text{Regret} = \frac{1}{T} \sum_{t=1}^T Q(a^*) - Q(a_t)$$

sub-linear

ویرایش ترکیبی

Concentration bounds $\Rightarrow +$

$$P\left[\frac{x_1 + \dots + x_n}{n} - E[X] > \epsilon\right] \leq e^{-\frac{r n \epsilon^2}{c}} \rightarrow (b-a)^r$$

w.h.p $\hat{\mu} \leq E[X] + \sqrt{\frac{r \log(t/\delta)}{n}}$

$(b-a=1 \text{ بازی})$

t^r بهتر

Union Bound

★ $\forall a, V_t: Q(a) - \hat{Q}(a) \leq \sqrt{\frac{r \log t/\delta}{N(a)}} \quad \epsilon$

$$V(a_t) = \hat{Q}(a_t) + \sqrt{\frac{r \log t/\delta}{n}}$$

$$a^* = \arg\max Q(a), a_t = \arg\max V(a)$$

$$\text{Regret} = \sum_{t=1}^T Q(a^*) - V(a_t) + V(a_t) - Q(a_t)$$

w.h.p ≤ 0

a_t, a^* برابری

$$\Rightarrow \text{Regret} \leq \sum_{t=1}^T V(a_t) - Q(a_t) \leq \sum_{t=1}^T \sqrt{\frac{r \log(t/\delta)}{N(a_t)}}$$

Union Bound

$$\Delta \geq 1 - m \sum_{t=1}^T \frac{\delta}{t^r} \leq 1 - m\delta \rightarrow$$

* مستقیم

و از هر کدام خوب نمی‌شوند

$$\Rightarrow \text{Regret} \leq \sum_{t=1}^T \sqrt{\frac{r \log t/\delta}{N(a_t)}} \leq r \sqrt{\log T/\delta} \sum_{t=1}^T \sqrt{\frac{1}{N(a_t)}}$$

w.h.p
1- $\gamma_m \delta$

$$\sum_{a=a_1}^{a_m} \sum_{i=1}^{N(a)} \sqrt{\frac{1}{i}} \leq m \sum_{i=1}^{T_m} \sqrt{\frac{1}{i}} \leq m \sqrt{\frac{T_m}{m}}$$

وَهُنَّ مُحْسِنُونَ بِرَأْيِهِنَّ مُحْسِنُونَ

$$\Rightarrow \text{Regret} \leq \sum \sqrt{m T \log(T/\delta)}$$

! This is tight to this

Exploration in RL is obviously

Contextual Bandit

$Q(s, a)$ "قيمة قصوى"

ةٰذٰلُ

(! This is per Episode only)

«Thompson Sampling»

$\theta_1, \dots, \theta_n$

bandit مينيوز داعر

Bayesian

مكتن بصرت بيزي

$\hat{p}(\theta_1, \dots, \theta_n)$
(Posterior)

بريليم $\theta_1, \dots, \theta_n$ هـ

$P(\theta_1, \dots, \theta_n | r_1, \dots, r_{t+1})$
برحسب قدر

$= \frac{P(r_{t+1} | \theta_1, \dots, \theta_n, r_1, \dots, r_t) P(\theta_1, \dots, \theta_n | r_1, \dots, r_t)}{C}$

كمه تبني

بسن را بارداهم action a

(argmax)

دست نهاد (فقط نهاد كرسن)

UCB كل خبر است مثل

$\theta_{t+1} (1 - \theta_{t+1})$

$s_{t+1} | s_t, a_t \leftarrow$ next state

Information Gain (IG) : عویض

$$IG(z, y) = E_y [H(\hat{P}(z)) - H(\hat{P}(z)|y)]$$

مقدار عویض

$$\underset{a}{\operatorname{argmax}} \quad IG(z, y|a)$$

تکنیک برای تعیین a^*

$$\Rightarrow g(a) = IG(\theta_a, r_a | a)$$

$$\Delta(a) = E[r(a^*) - r(a)]$$

$$\Rightarrow \text{argmin } \frac{\Delta(a)}{g(a)}$$

تکنیک برای تعیین a^* Deep Reinforcement Learning +

Deep RL "Exploration from scratch"

جذب خودکار (M, R, N)

ابهاد دیگر میان این تئوری

نمایع خود باید الگوریتم سوچ شود

اگر MDP باشد (متغیر و پیوست) \rightarrow UCB

$$r^*(s, a) = r(s, a) + \beta(N(s))$$

$$\frac{1}{\sqrt{n}} = \frac{1}{n} + \frac{1}{x}$$

در الگوریتم همان بدلی میان

Pseudo Count را نسبت دهم. از density

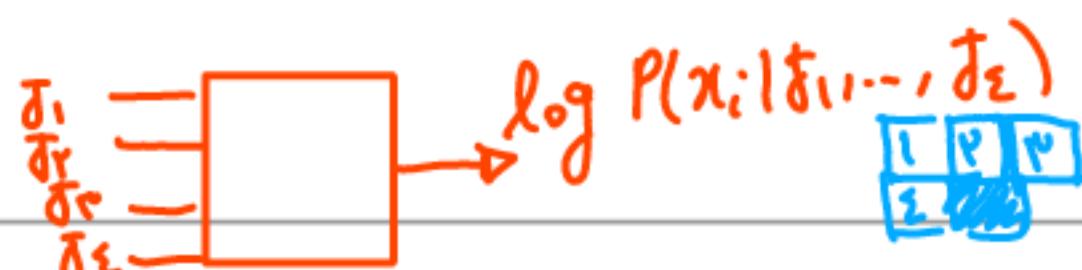
$$P_\theta(s) \leq P_\theta(s, a) \rightarrow \hat{N}(s) \rightarrow \hat{N}(s) = n P_\theta(s)$$

اینها کاربری autoregressive

$$P_\theta'(s, a) = P_\theta(s) + 1$$

$$\hat{n} = \frac{1 - P_\theta(s)}{P_\theta'(s) - P_\theta(s)} \rightarrow \text{Pseudo count}$$

Montezuma در بازی



RL، CTS و CTS class

\rightarrow log likelihood

است!

با خطا زدن Pixel خوب است \rightarrow Quantize \rightarrow میانگین مخفی تر \rightarrow حالت دست داد \rightarrow TRPD

Super state

و colision \rightarrow state (Hash & Clustering)

لهم بتوجه به Autoencoder

\rightarrow Q.O.D Generalization

و اینجایی: از classifier برای عین (s) \rightarrow $P_\theta(s)$ دارد و خوب است/مرغی را از جید جدا کند

$$D_s^*(s) = \frac{1}{1 + P_\theta(s)}$$

$$P_\theta(s) = \frac{1 - D_s(s)}{D_s(s)}$$

shared backbone encoding

در مکانیزم مبتذل بر bonus (این جایزه) این طور است +
intrinsic reward

خر خوبی همیشگی

$$\epsilon(s,a) = \|\hat{f}_0(s,a) - f^*(s,a)\| \quad \text{بسیار بزرگ} \hat{f}(s,a) \quad \text{این نیست} \quad \checkmark$$

$$\hookrightarrow \text{یک } f^* \text{ خوب} \rightarrow f^*(s_t, a_t) = s_{t+1} \rightarrow \text{یک طبع طرد}$$

Bootstrapped DQN

مبتذل بر

bootstrapping برای دریافت دنباله $\theta_0, \dots, \theta_n$ از Q function خوب، احتمالی Posterior $P(Q)$ حون $\bar{P}(Q)$ قابل پیش‌بینی هست
با استفاده از Particles Q_0, \dots, Q_n و "fixed" "replay" "buffer" که از آنها برای اینجا استفاده شدند
آنها انتخاب می‌کنند و آنرا با Q را در آن episode uniform می‌دانند
(برای آخرین باره بسیار بخوبی)

بنابراین دیگر داشتن

(٢٥,٢/٢٣) جلسه

Bootstrapped DQN

Thompson

Sampling

مروج

$$\text{intractable } \mathbb{E}_w P(Q) \rightarrow P(Q|h_t) \xrightarrow{\text{دقت}} Q_1^{(+)}, \dots, Q_k^{(+)} \xrightarrow{\text{دقت}} Q_1^{(+)}, \dots, Q_k^{(+)}$$

(Exploration) انتخاب مناسب (برای حالت Q_i در episode i)
 (Exploitation) خوب طبق Montezuma (برای حل مسئله)

دسته دوم: جریده خود (Intrinsic Reward) IG

$$IG(z, y | a) \xrightarrow{\text{action}} \text{all Information gain} \xrightarrow{\text{جذب}} IG$$

دسته ثالث: جمله خود (Intrinsic reward)

$$P(s) \leftarrow P(s'|s, a) \xrightarrow{\text{in Policy}} \text{Intrinsic reward}$$

دسته رابع: جمله خود (Intrinsic reward)

$$z = P(s'|s, a) \rightarrow (z = \theta \leftrightarrow P_\theta(s'|s, a))$$

$$I(z; y | a) \xrightarrow{\text{KL}} \text{VIME}_{\text{out}}$$

$$r_a = r + \eta D_{KL}(P(\theta | h, s_t, a_t, s_{t+1}) \| P(\theta | h))$$

$$\theta \xrightarrow{\text{history}} y \xrightarrow{\text{history}} \frac{1}{\text{تکرار}} \xrightarrow{\text{تکرار}} \theta \xrightarrow{s_t \rightarrow \theta} s_{t+1}$$

$$q(\theta | \phi) = \prod_{i=1}^d q(\theta_i | \phi_i) \xrightarrow{\text{جذب}} q(\theta | \phi) \sim \mathcal{N}(\mu, \sigma^2)$$

$$\xrightarrow{\text{جذب}} q(\cdot | (\mu_i, \sigma_i^2)) \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

$$D_{KL}(P(\theta|h, \{s_t, a_t, s_{t+1}\}) || P(\theta|h)) \xrightarrow{\text{Replay Pool}} \text{دریش} \quad \text{(FIFO Replay)}$$

دربش (FIFO Replay)

حراجی داشتی بدریش
میکنیم

جیسے VI میں خواهد
بین (θ|φ) ↓

$q(\theta|\phi')$

ELBO + variational inference

کم کردن ELBO کا نتیجہ کیم
کم کردن ELBO کا نتیجہ کیم

$E_{z \sim q} \left[\log \frac{P(x, z)}{q(z)} \right] \leftarrow D(q || P(z|x))$

(ELBO)

دربش Z کا نتیجہ

دربش Z کا نتیجہ

دربش Z کا نتیجہ

دربش Z کا نتیجہ

D-Hessian

برای VI کا نتیجہ (Replay gradient) کا طور پر فراہم کیا جاتا ہے

(Hessian) کا نتیجہ کیا جاتا ہے

Reward جیسے VI میں RL میں

Low Rank Approximation (جیسے VI میں RL میں)

VI-M: Variational Information Maximization Exploration

Theoretical Beauty

دینی ایجاد

Interest in RL

(Monte Carlo) میں خوب اسٹے (ملے)

کھلپے لینے، متن sparse Exploration

لٹریک لینے، حتم منزد

Model-Based RL

(o^t, r, r_a)

Close-loop vs Open-loop

non-parametric ω] \rightarrow (elite) \rightarrow (نیتی مجده سنت) \rightarrow Cross-Entropy Method -
 parallelizable]

$$\left[\begin{array}{c} J(A_1) \\ \vdots \\ J(A_N) \end{array} \right] \rightarrow CE\left(Z_i \left[\begin{array}{c} e^{J(A_1)} \\ \vdots \\ e^{J(A_N)} \end{array} \right], P(A)\right) = \sum e^{J(A_i)} \log P(A_i)$$

! در Cross Entropy می توان $e^{J(A_i)}$ باید $P(A_i)$ باشد

(action-space S_t) \leftarrow (Monte Carlo Tree Search) MCTS

+ دست رفته تا مرحله (جی تی ای) نشان داده شد

δ^N \leftarrow $\delta^N + VCB$ (برای N نشان داده شد) episode

$$\sqrt{\frac{\log N(s_{t+1})}{N(s_t)}}$$

$$\min_{u_1, \dots, u_T} c(x_1, u_1) + c(f(x_1, u_1), u_2) + c(f(f(x_1, u_1), u_2), u_3) + \dots \quad \text{LQR}$$

$$f(x_t, u_t) = F_t \begin{bmatrix} x_t \\ u_t \end{bmatrix} + f_t$$

Time Dependent

c (خواص برای f) فرم خاص برای f

Linear Quadratic
Regulator

$$c(x_t, u_t) = \frac{1}{2} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^T C_t \begin{bmatrix} x_t \\ u_t \end{bmatrix} + \begin{bmatrix} x_t \\ u_t \end{bmatrix}^T c_t$$

جلسہ میت روپ (MCTS)

MCTS & CEM جو لیں؟

? Uncertainty

? Model-based + Model-free

(CB) planning & Policy

$$f(s_t, a_t) = s_{t+1}$$

کوچانہ نفع از $\pi_f \rightarrow \pi_0$

Support حکمی مقاومت!

جنہیں فرمایا جائے کہ اسے دیکھنے کا

برحسب کوئی جریب

Drift کوچانہ

MPC

کوچانہ (دینے) نفع

(replan or recompute)

خطوٹ کو تحریر کرنے!

منسد Saturate

overshoot مل بزرگ

saturation مل کریں

درتھرگتن بازہ اطمینان

کنن سعی تعیین (Uncertainty)

Entropy انتروپی

خوبیست! زیرا مل بتعیین

بال اطمینان

کنن مل قطعہ مل

 $P(\theta | D)$

init (یا اسٹارٹ)

Bootstrapping

کوچانہ a_{t+1}, \dots, a_T

میانگن بچوندہ اسٹارٹ

حاسیب کریں!

ساعی Ensemble

TIME $\int P(s_{t+1} | s_t, a_t, \theta) P(\theta | D) d\theta$

Bayesian Neural Net با تربیت

Meanfield

نحوه آر انلاین

Offline RL و نفع Model-Based

جنه بیت پنجم (۱۴۰۳، ۱۴۰۴)

+ مروج برگزینش عمل در برنامه ریزی کو ... MPC

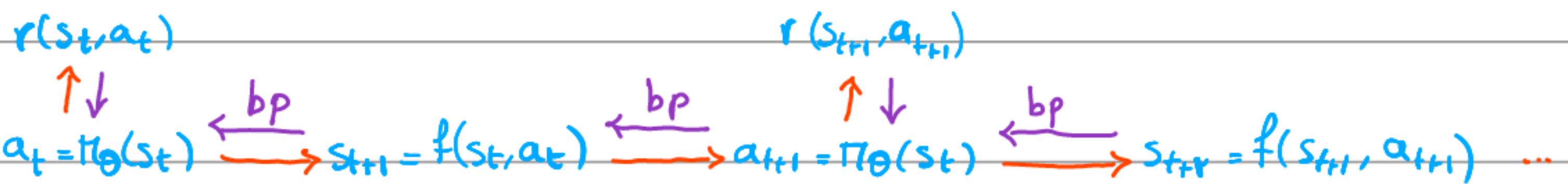
نحوه آر انلاین Ensemble + مجموعه ای از مدل های مربوط به حالت (باشد بزرگتر)

Posterior پوزیشن

نحوه آر انلاین با احتمال اولیه برآورد و توزیع احتمالی مدل های مربوط به حالت (باشد بزرگتر) میان مدل های مربوط به حالت

Policy Learning و نفع Planning

(Computation Graph) ترجیح را در شبکه



نحوه آر انلاین با احتمال اولیه برآورد و توزیع احتمالی مدل های مربوط به حالت

Actor-Critic و RNN

skip connection، LSTM یا ... (که میتواند گرانی و Vanishing or Exploding Gradient را باعث نماید)

از نظر زیاده تر dynamic (است برداشتی)

Model-free

نحوه آر انلاین با احتمال اولیه برآورد و توزیع احتمالی مدل های مربوط به حالت

Trajectory

نحوه آر انلاین با احتمال اولیه برآورد و توزیع احتمالی مدل های مربوط به حالت



Trajectory / Rollout

state چون که از این دستور کار نماید، short rollout

Replay Buffer

off-policy RL

Dyna-Style

Dyna Style

Buffer :

$\hat{P}(s'|s,a) \rightarrow \hat{Q}(s,a)$ در مجموعه

$$Q(s,a) \leftarrow Q(s,a) + \alpha E_{s,r} [r + \max_{a'} Q(s',a') - Q(s,a)]$$

(1) باطری trajectory داده شده است

P جزو

Model-Based Learning مبنی بر اینکه مدل



"Offline RL"

Model Reuse \rightarrow generalization

foundation model

مشخصاتی را که باید داشته باشد

On-Policy RL — Off-Policy RL

(Buffer :)

Offline RL

در راستای Offline RL درست است. پس است

خبر در مورد کدام دیدگاه است

Data Generation برای Buffer

شبیه سازی داده های Deep learning

(Exploration)

جواب ناگفته باشند اگر راه دخیل های پرست

که از آنها برای تطمیع دارند (رسانه ای خبر را تحقیق نمود)

(Q-learning \rightarrow offline \rightarrow Overestimation) حالت

(Maximization Bias

(less log Q نیز)

(QOD) ! فرمول قیمتی Q را که تنظیم

Offline RL

حاجه بینت داشتم بودم

full off-Policy

کام می توانیم درینجا

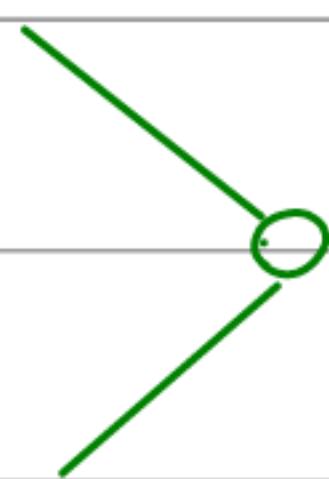
$$D = \{(s_i, a_i, s'_i, r_i)\}$$

OPE off-Policy Evaluation

سیستم از کنترلر مارکوف زبان

جی تین $J(\pi)$

پس از آن چیزی که من



بلنم نتایج خوب را بخواهم (رسانید / نظریه خوب بقدر برای این

stich

استیج (هد)

من درین دریب برداشتن از کسو

Q-Learning

از قدرمی باز

RL دادارک را

دستوراتی که آنها

نه سعی و خطا نیست

$$Q(s, a) = r + \gamma \max_a Q(s', a')$$

از این راه بخواهیم

برآورد می کنیم

برآورد می کنیم

و تئوری می شود

* نویشتم یعنی سیستم از زیر گشتم (در این قیاس نمی شود)

Masnirki این را بفرمود

Adversarial example

، Distribution shift

$$Q(s,a) \leftarrow r(s,a) + E_{\pi \sim \pi_{\text{new}}} [Q(s,a)]$$

Policy Constraint Method

$$\pi_{\text{new}}(a|s) = \underset{\pi}{\operatorname{argmax}} E_{a \sim \pi(a|s)} [Q(s,a)] \text{ s.t. } D_{KL}(\pi || \pi_{\beta}) \leq \epsilon$$

ابن باعیت دلخواه از کنترلر (سیستم را داده از اینکنترلر تولید کن)

له این پیشترایلر

خرد خوبی KL

$$(UMD) \quad \pi_{\beta}(a|s) \geq \epsilon \quad \pi_{\text{new}}(a|s) > 0$$

Explicit Policy Constraint Method

$$D_{KL}(\pi || \pi_{\beta}) = -E_{\pi}[\log \pi_{\beta}(a|s)] - H(\pi) \quad : \text{actor} \quad \text{تغیرهای}$$

$$\theta \leftarrow \underset{\theta}{\operatorname{argmax}} E_{s \sim D} [E_{a \sim \pi_{\theta}(a|s)} [Q(s,a) + \lambda \log \pi_{\beta}(a|s)] + \gamma H(\pi(a|s))]$$

$$\bar{r}(s,a) = r(s,a) - D(\pi, \pi_{\beta}) \quad : \text{reward} \quad \text{برحسب}$$

negative bias

Implicit Policy Constraint Method

$$\pi^*(a|s) = \frac{1}{Z(s)} \pi_{\beta}(a|s) \exp\left(\frac{1}{2} Q(s,a)\right) \quad \text{پنج بحث مذکور برحسب}$$

لذتگویی scale \rightarrow SAC

جذب مذکور برحسب دلخواه از کنترلر $D(\pi || \pi^*)$. علاوه بر این دلخواه از کنترلر برحسب

\downarrow cross-entropy

$$\pi_{\text{new}} \leftarrow \underset{\pi}{\operatorname{argmax}} E_{(s,a) \sim \pi_{\beta}} \left[\log \pi(a|s) \frac{1}{Z(s)} e^{\frac{1}{\lambda} A^{\text{old}}(s,a)} \right] \rightarrow Q$$

Policy Gradient \leftarrow پیشوند
Gradient \leftarrow مازدنز گزین

میتوانیم actor-critic \leftarrow

جذب مجاز و خصم

٢٣٨

» Offline RL

الاتجاه

Advantage Weighted Regression

Implicit Policy Constraint
Methods

جذب مجاز و خصم

ایجاد: از ۰۰۰ در برداشت Q اجتناب کنیم

$$Q(s, a) \leftarrow r(s, a) + V(s')$$

$$V \leftarrow \arg \min \frac{1}{n} \sum l(V(s_i), Q(s_i, a_i))$$

استایل

 π_{new}

ل

$$\rightarrow \text{MSE} \text{ از } (V(s_i) - Q(s_i, a_i))^2$$

استایل

ین مرحله را تبدیل کنیم

استایل است اس بزرگ داشتیم

$$\begin{aligned} & 1. \begin{cases} (1-\tau)x^r & \text{if } x > 0 \\ \tau x^r & \text{else} \end{cases} \\ & \text{جذب} \end{aligned}$$

Implicit Q-learning (IQL)

لکچر ۲: جذب مجاز و خصم

Gradient

Conservative Q-Learning (CQL)

$$\hat{Q}^n = \arg \min_Q \max_{\mu} \alpha E_{s \sim D, a \sim \mu(a|s)} [Q(s, a)] - \alpha E_{(s, a) \sim D} [Q(s, a)]$$

برای اینجا درست است

+ خطی می باشد

(بدون متفق نه)

$$E[\hat{Q}^n] < E[Q^n]$$

$$\hat{Q}^n < Q^n \rightsquigarrow \text{Overestimate}$$

افزایش درست از $R = E_{s \sim D} [H(\mu(s))]$ برای μ ایجاد می شود
و ضعیفی می کند که \hat{Q}^n

Model Based Offline RL

نیازی به loop نیست، Offline الگوریتم خود را درست می کند

$$\tilde{r}(s,a) = r(s,a) - \lambda u(s,a)$$

: 106

Combo

CQL (جی
و جی)

CQL + Bellman
Backup

دسته دینه می کند

$$\arg\max_Q \beta [E_{\pi^{\text{old}}}(Q(s,a)) - E_{\pi^{\text{old}}}[\hat{Q}(s,a)]]$$

+ Simulation

کارایی Offline RL را در برداشت

□ در بعضی مدل ها ممکن است این روش سعی خطی را جبرید کرد (Overfit) یا در آنها بسیار آسان شود (آفلاین برداشت)

(Safety Critical)

برای Offline RL باید

→ حذف خطی باش!

Inverse RL + Imitation Learning

Generalization

simulation vs. robotics

ـ نتائج، احصائيات ...

ـ فتح sparse ـ مكافأة reward مع الـ

ـ خروقات ذات تعريف محدود

expert ـ تدريب (demonstrations) ـ انتداب / اكتشاف خوب بـ

ـ trajectory ـ تسلسل درن

↓
Imitation

$$\underset{\theta}{\operatorname{argmin}} \mathbb{E}_{s \sim p(s|\pi_\theta)} [L(\pi^*(s), \pi_\theta(s))]$$

ـ trajectory ـ بحسب *

(

ـ از π^* خوب طریق

ـ دستیابی به مكافأة reward و مكافأة

ـ فتح مكافأة reward

ـ expert ـ تدريب Offline RL

ـ انتداب / اكتشاف بالذات

Behavioral Cloning ـ (direct) ـ مفهوم انتداب / اكتشاف

Inverse RL

ـ بررسی برای سیاست انتداب / اكتشاف ـ سیاست انتداب / اكتشاف ـ reward ـ فتح مكافأة

ـ انتداب / اكتشاف من اجل حمل مزدوج

Behavioral Cloning

$$\underset{\theta}{\operatorname{argmin}} \mathbb{E}_{(s,a)} \left[-p^* [L(a^*, \pi_\theta(s^*))] \right] : \text{supervised loss}$$

distributional mismatch \rightarrow این توزیع داده خروجی

دستورات OOD نسبت به π^* : ایشون
فیلم، خود را recover کنید
و فحیت دبرگرد

recovery (ذیلی) \rightarrow Demonstration Augmentation

Augment کرن وی تغیر Adjust

چالش: نمونه های i.i.d

و مسأله خطا کنند چالش منزد

☒ دست مدل خطا را بیندازد و آن را باز خواهد شد از مدار خوب است

Dataset Aggregation \rightarrow DAgger

برآجع آنکه باید بخوبی از "عمر طبیعت" داشلان استفاده کنیم.

$\nabla \varphi$ برای P^i میانگین مولدهم

expert recovery action *

و آنرا باعث می کند سیستم را درست کنم

show motion \rightarrow expert react

expert \rightarrow visual feedback

safety measures \rightarrow چالش های ایمنی (وقتی در...)

Inverse RL

expert trajectory \rightarrow reward function \rightarrow يكمل المقدمة

expert trajectory \rightarrow π^* \rightarrow underspecified #

pb trajectory \rightarrow π^* , reward \rightarrow

$$r_\psi(s, a) = \gamma^T f(s, a)$$

(use feature extractor) feature Extractor \rightarrow

is Pre-trained, f \rightarrow

$$E_{\pi^*}[f(s, a)] = E_{\pi^*}[f(s, a)]$$

$$\text{رسالة} \rightarrow \gamma^T E_{\pi^*}[f(s, a)], \gamma \in \mathbb{S}$$

مهم جعله در مركز تقوير

(use feature matching \rightarrow π^*)

feature matching

$$\sum_{a \in A} \gamma^T E_{\pi^*}[f(s, a)]$$

$$\gamma^T E_{\pi^*}[f(s, a)]$$

margin

$$\geq \max_{a \in A} \gamma^T E_{\pi^*}[f(s, a)] + m$$

m جانب در γ \rightarrow margin \rightarrow margin

$$\min_{\gamma} \frac{1}{t} \|\gamma\|^2 \text{ s.t. } \gamma^T E_{\pi^*}[f(s, a)] \geq \max_{a \in A} \gamma^T E_{\pi^*}[f(s, a)] + D_{KL}(\pi, \pi^*) : \text{ SVM } \rightarrow$$

{0, 1}

use quadratic loss

$s_t, a_t \rightarrow O_t$

O_t

$O_t \rightarrow$

Optimality Variable

Deep reward

a_t

a_t

a_t

جذب المقدمة \rightarrow SubOptimal \rightarrow $P(O_t | s_t, a_t) = \exp(r(s_t, a_t))$

جزء باردة

π^* suboptimal

جذب المقدمة

جذب المقدمة

$$P(T | O_{1:T}) \propto P(T) \exp \left(\sum_t r(s_t, a_t) \right)$$

$\{O_t\}$: hidden RV

$$\hat{r}_\psi(s_t, a_t) \leftarrow \frac{1}{n} \sum_{i=1}^n r_\psi(s_i, a_i)$$

$$\text{Log likelihood: } \max_{\gamma} \frac{1}{N} \sum_{i=1}^N \log P(\tau_i | O_{1:T}, \gamma) = \max_{\gamma} \frac{1}{N} \sum_{i=1}^N r_\gamma(\tau_i) - \log Z_\gamma$$

$$Z = \int P(\tau) \exp(E_p(\tau)) d\tau$$

partition function ↩

partition function

(normalization)

$$\nabla_{\gamma} L = \frac{1}{N} \sum_i \nabla_{\gamma} r_{\gamma}(\tau_i) - \frac{1}{z} \int p(\tau) \exp(r_{\gamma}(\tau)) \nabla_{\gamma} r_{\gamma}(\tau) d\tau$$

$$= E_{\tau \sim \pi^*} [\nabla_\gamma r_\gamma(\tau)] - E_{\tau \sim p(\tau | a_{:T}, \gamma)} [\nabla_\gamma r_\gamma(\tau_i)]$$

expert samples

soft optimal policy under current reward

$$f(\tau_i)$$

بادشاہ ملک سید علی گریم دہنے بلیغ
→ SAC مڈل

حربدار بیک SAC میتواند جمهوری روسیه را بخواهد

$$w_i = \frac{P(\tau) \exp(r_p(\tau))}{\pi(\tau_j)} = \frac{\exp(\sum_t r_p(s_t, a_t))}{\pi(a_t | s_t)} \rightarrow \text{Guided Cost Learning Algorithm}$$

گانڈی

دیگر اخراج نمایند و درین لوب را دارند

$$\nabla_y L \approx \frac{1}{n} \sum_i \nabla_y f_y(t_i) - \frac{1}{\sum w_j} \sum_j w_j \nabla_y f_y(t_j)$$