Policy-Based methods

Разворотнев Иван

18.06.2019

Сравнение с Value-based

Преимущества

- Имеет более сильные гарантии сходимости
- Эффективно работает в задачах с большим(непрерывным) множеством действий
- Оптимизирует стохастическую стратегию напрямую, не через exploration/explotation

Недостатки

- Обычно сходится к локальному, а не глобальному минимуму
- Оценки имею большую дисперсию

Ожидаемый возврат

Ожидаемый реворд:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] = \int_\tau P(\tau|\theta)R(\tau)d\tau$$

Суммарный реворд

Интеграл по пространству ВСЕХ возможных траекторий

Где вероятность траектории:

$$P(\tau|\theta) = p(s_0)\prod_{t=0}^{T} p(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t)$$

Тогда, оптимальная политика:

$$\pi^* = \underset{\pi}{arg\max} J(\theta)$$

И задача оптимизации:

$$\theta_{k+1} = \theta_k + \alpha\nabla_\theta J(\theta)$$

Но как найти $\nabla_\theta J(\theta)$?

Дифференцирование

$$\nabla_\theta P(\tau|\theta) = P(\tau|\theta)\underline{\nabla_\theta \log P(\tau|\theta)}$$

$$f(x)\,(\log f(x))' = f(x)\frac{f'(x)}{f(x)} = f'(x)$$

Избавляемся от произведения

$$\nabla_\theta \log P(\tau|\theta) = \underline{\nabla_\theta \log p(s_0)} + \sum_{t=0}^{T}[\nabla_\theta \log \pi_\theta(a_t|s_t) + \underline{\nabla_\theta \log p(s_{t+1}|a_t, s_t)}]$$

Не зависит от θ, обнуляется

Получим:

$$\nabla_\theta \log P(\tau|\theta) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)$$

Дифференцирование

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$$

$$= \nabla_\theta \int_\tau \underline{P(\tau|\theta)}R(\tau)d\tau$$

Перейдем к интегралу

$$= \int_\tau \nabla_\theta P(\tau|\theta)R(\tau)d\tau$$

Внесем градиент под интеграл

$$= \int_\tau \nabla_\theta \log P(\tau|\theta)\, \underline{P(\tau|\theta)}R(\tau)d\tau$$

Подставим из пред. слайда

$$= \mathbb{E}_{\tau\sim\pi_\theta}[P(\tau|\theta)R(\tau)]$$

Вернемся к мат. ожиданию

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, R(\tau)\right]$$

Развернем выражение

Аппроксимация мат. ожидания

Пусть есть множество траекторий $\mathcal{D} = \{\tau_i\}_{i=1..N}$, тогда:

$$\nabla_\theta J(\theta) \approx \frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t | s_t)\, R(\tau)$$

Обучение модели:

1. Сэмплирование траекторий
2. Обновление параметров $\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\theta)$

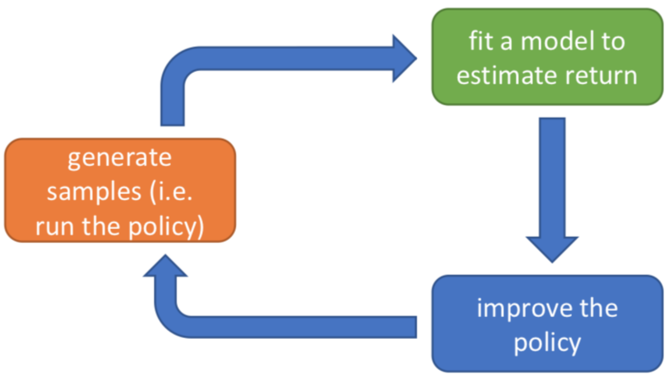
Алгоритм Reinforce

Псевдокод:

REINFORCE algorithm:

1. sample $\{\tau^i\}$ from $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ (run the policy)
2. $\nabla_\theta J(\theta) \approx \sum_i \left(\sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i)\right)\left(\sum_t r(\mathbf{s}_t^i, \mathbf{a}_t^i)\right)$
3. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

$$\hat{Q}^\pi(\mathbf{x}_t, \mathbf{u}_t) = \sum_{t'=t}^{T} r(\mathbf{x}_{t'}, \mathbf{u}_{t'})$$



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

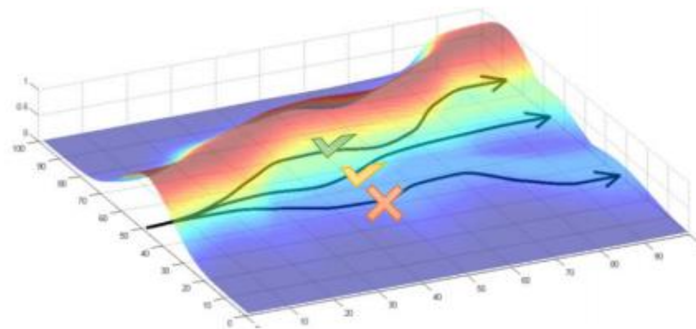
Имеет большую дисперсию и медленно сходится

Принцип работы

Policy Gradient baselines

$$\nabla_\theta J(\theta) \approx \frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\,(R(\tau) - b(s_t))$$

$$b(s_t) = \frac{1}{N} \sum_{i=1..N} R(\tau)$$



Докажем, что оценка не смещенная:

$$\mathbb{E}_{a_t \sim \pi_\theta}[\nabla_\theta log\pi_\theta(a_t|s_t) b(s_t)] = \int \overbrace{\pi_\theta(a_t|s_t) \nabla_\theta log\pi_\theta(a_t|s_t)}^{\nabla_\theta \pi_\theta(a_t|s_t)} b(s_t) da_t$$

$$= b(s_t) \nabla_\theta \int \pi_\theta(a_t|s_t) da_t = b(s_t) \nabla_\theta 1 = 0$$

Policy Gradient baselines

Можно найти baseline с минимальной дисперсией

$$\mathrm{Var}[x] = E[x^2] - E[x]^2$$

$$\nabla_\theta J(\theta) = E_{\tau \sim \pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau)(r(\tau) - b)]$$

$$\mathrm{Var} = E_{\tau \sim \pi_\theta(\tau)}[(\nabla_\theta \log \pi_\theta(\tau)(r(\tau) - b))^2] - \underbrace{E_{\tau \sim \pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau)(r(\tau) - b)]}_{E_{\tau \sim \pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau) r(\tau)]}{}^2$$

ДЗ: найти этот baseline

“Vanilla” Policy Gradient

Идея: Reinforce с baseline.

- $\nabla_\theta J(\theta) \approx \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, (R(\tau) - b(s_t))$
- Baseline-модель на каждом шаге обновляем $\|b(s_t) - R_t\|^2 \to min$

Algorithm 1 “Vanilla” policy gradient algorithm

Initialize policy parameter θ, baseline b
for iteration$=1, 2, \ldots$ **do**
  Collect a set of trajectories by executing the current policy
  At each timestep in each trajectory, compute
   the *return* $R_t = \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'}$, and
   the *advantage estimate* $\hat{A}_t = R_t - b(s_t)$.
  Re-fit the baseline, by minimizing $\|b(s_t) - R_t\|^2$,
   summed over all trajectories and timesteps.
  Update the policy, using a policy gradient estimate $\hat{g}$,
   which is a sum of terms $\nabla_\theta \log \pi(a_t \mid s_t, \theta) \hat{A}_t$
end for

Reward-To-Go

$$\nabla_\theta J(\theta) \approx \frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, R(\tau)$$

Сумма ревордов за эпизод

$$\nabla_\theta J(\theta) \approx \frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \overbrace{\sum_{t'=t} R(s_{t'}, a_{t'}, s_{t'+1})}^{\hat{Q}(s_t, a_t)}$$

Сумма ревордов за последующие шаги

- Более правильное взвешивает $\nabla_\theta \log \pi_\theta(a_t|s_t)$
- Уменьшает дисперсию

Актор-Критик

Идея:

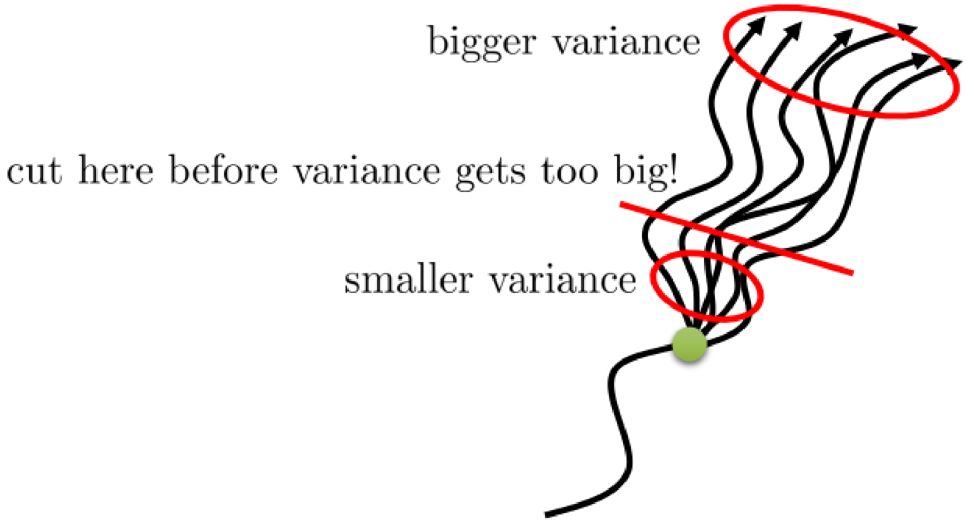
Совместить Q-Learning и Policy Gradient

Используем Q-Learning для выражения Q-функции

Обновляем PG с полученными Q-значениями

Актор – PG

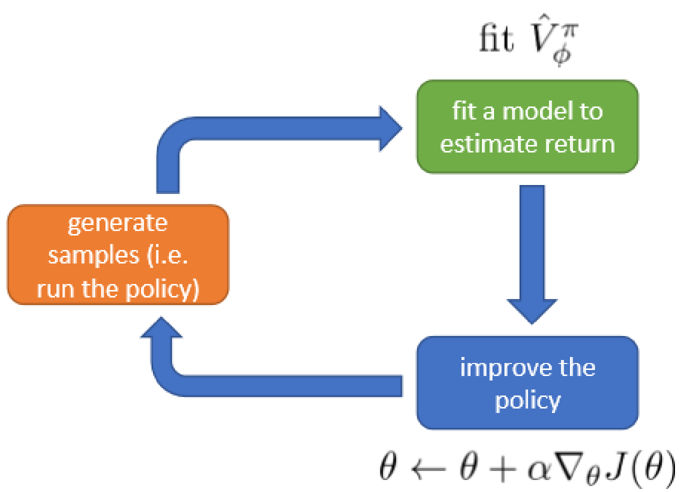
Критик - Q-Learning



Актор-Критик

Algorithm 1 Q Actor Critic

Initialize parameters s, θ, w and learning rates α_θ, α_w; sample $a \sim \pi_\theta(a|s)$.
for $t = 1 \dots T$: **do**
  Sample reward $r_t \sim R(s, a)$ and next state $s' \sim P(s'|s, a)$
  Then sample the next action $a' \sim \pi_\theta(a'|s')$
  Update the policy parameters: $\theta \leftarrow \theta + \alpha_\theta Q_w(s, a) \nabla_\theta \log \pi_\theta(a|s)$; Compute the correction (TD error) for action-value at time t:
    $\delta_t = r_t + \gamma Q_w(s', a') - Q_w(s, a)$
  and use it to update the parameters of Q function:
    $w \leftarrow w + \alpha_w \delta_t \nabla_w Q_w(s, a)$
  Move to a $\leftarrow a'$ and s $\leftarrow s'$
end for

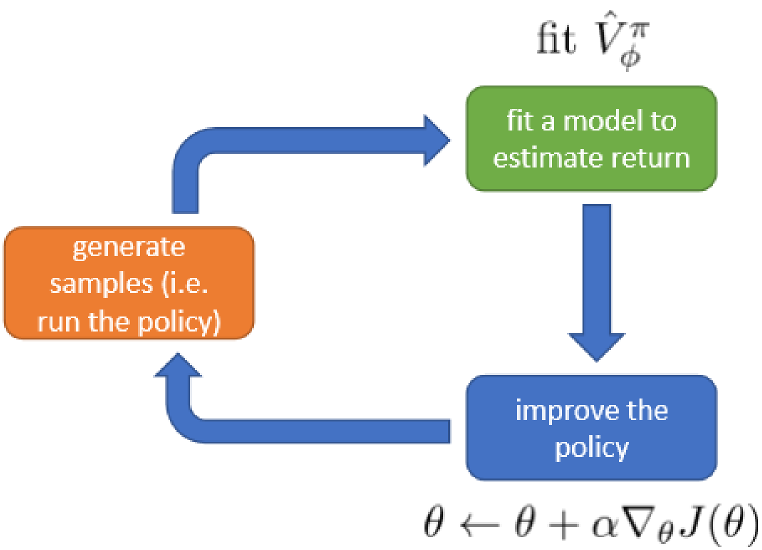



SAC
Actor-Critic
DDPG
TRPO
Online/Batch AC
PPO
A2C

Advantage Actor-Critic (A2C)

$$\nabla_\theta J(\theta) \approx \frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \overbrace{[\hat{Q}(s_t,a_t) - V(s_t)]}^{\mathrm{A}(s_t,a_t)}$$

Берем в качестве baseline, V-значение



$$A(s_t,a_t) = \hat{Q}(s_t,a_t) - V(s_t)$$
$$= r(s_t,a_t) - \gamma V(s_t) - V(s_t)$$

$$\hat{A}^\pi(\mathbf{s}_i,\mathbf{a}_i) = r(\mathbf{s}_i,\mathbf{a}_i) + \gamma \hat{V}^\pi_\phi(\mathbf{s}'_i) - \hat{V}^\pi_\phi(\mathbf{s}_i)$$

Фактический реворд, V-предсказание критика

Advantage Actor-Critic (A2C)

batch actor-critic algorithm:

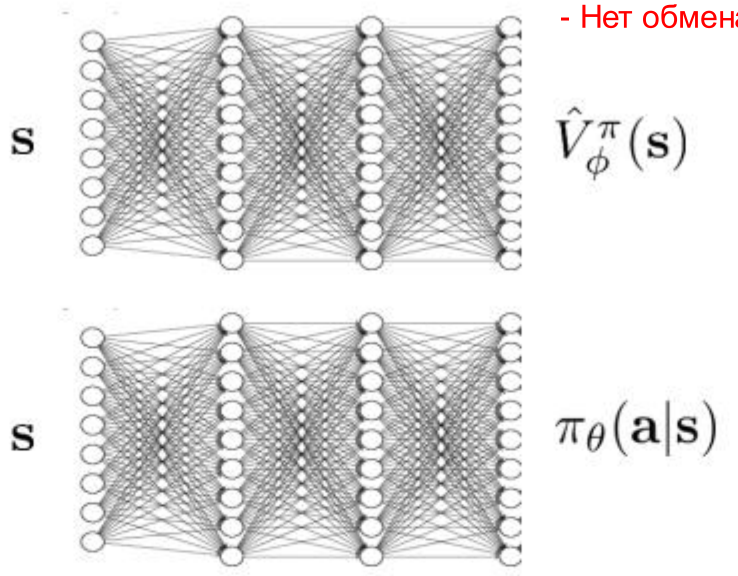
1. sample $\{\mathbf{s}_i, \mathbf{a}_i\}$ from $\pi_\theta(\mathbf{a}|\mathbf{s})$ (run it on the robot)
2. fit $\hat{V}_\phi^\pi(\mathbf{s})$ to sampled reward sums
3. evaluate $\hat{A}^\pi(\mathbf{s}_i, \mathbf{a}_i) = r(\mathbf{s}_i, \mathbf{a}_i) + \gamma \hat{V}_\phi^\pi(\mathbf{s}_i') - \hat{V}_\phi^\pi(\mathbf{s}_i)$
4. $\nabla_\theta J(\theta) \approx \sum_i \nabla_\theta \log \pi_\theta(\mathbf{a}_i|\mathbf{s}_i) \hat{A}^\pi(\mathbf{s}_i, \mathbf{a}_i)$
5. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

online actor-critic algorithm:

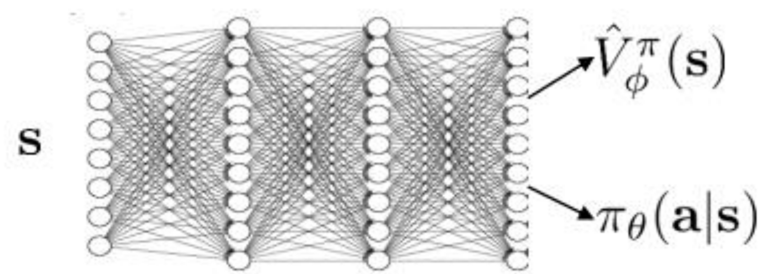
1. take action $\mathbf{a} \sim \pi_\theta(\mathbf{a}|\mathbf{s})$, get $(\mathbf{s}, \mathbf{a}, \mathbf{s}', r)$
2. update $\hat{V}_\phi^\pi$ using target $r + \gamma \hat{V}_\phi^\pi(\mathbf{s}')$
3. evaluate $\hat{A}^\pi(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \gamma \hat{V}_\phi^\pi(\mathbf{s}') - \hat{V}_\phi^\pi(\mathbf{s})$
4. $\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s}) \hat{A}^\pi(\mathbf{s}, \mathbf{a})$
5. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

Архитектура нейросети

- просто и стабильно
- Нет обмена информацией между актором и критиком



Отдельные сети



Совместная сеть

Loss-функция

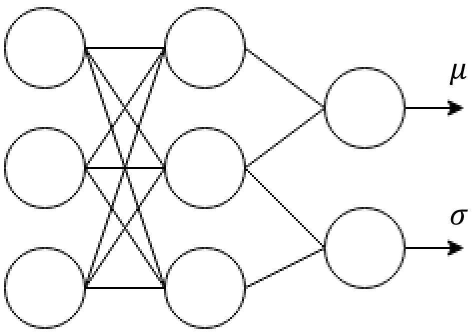
$$\mathcal{L}^0\ (\pi_\theta, \mathcal{D}) \approx \frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \sum_{t=0}^{T} \log \pi_\theta(a_t|s_t)\, R(\tau) \to max$$ (Взвешенная максимизация правдоподобия)

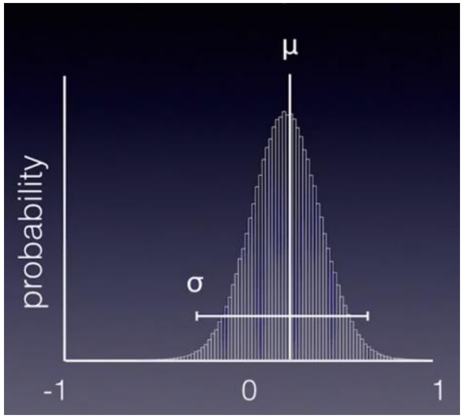
$$\boxed{\nabla_\theta \mathcal{L}\ \ (\pi_\theta, \mathcal{D}) \approx \frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, R(\tau)\ \approx \nabla_\theta J}$$

$$\boxed{\mathcal{L}\ \ (\pi_\theta, \mathcal{D}) \approx -\frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \sum_{t=0}^{T} \log \pi_\theta(a_t|s_t)\, R(\tau) \to min}$$

Непрерывное пространство действий

Actor





сэмплирование

(клип)

Действие

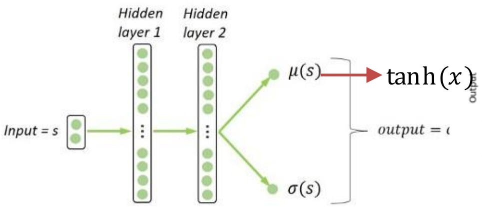
Тогда, ошибка актора:

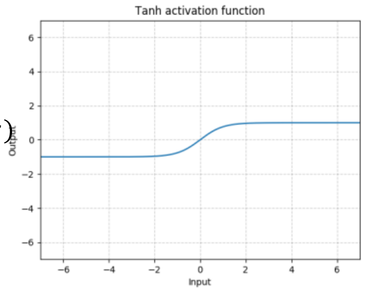
$$\mathcal{L}_\theta = -\underline{\log(\mathcal{N}(a|\mu_\theta(s_t), \sigma_\theta(s_t))}A(s_t))$$

$$\log \pi_\theta(a|s) = \frac{(x-\mu)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}$$

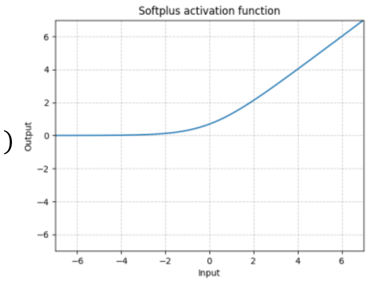
$$H = \log\sqrt{2\pi\sigma^2}$$

Архитектура сети:





$softplus(x) = \log(1 + e^x)$



Конец

Спасибо за внимание