GAE: 一般化アドバンテージ関数 [4]

菱沼 徹

2020 年 3 月 9 日

1 GAE: Generalized Advantage Function の定義 [4]

1.1 準備

初期状態 s_0 は，分布 ρ_0 からサンプルされる．軌道 $(s_0, a_0, s_1, a_1, \cdots)$ は，確率的定常方策 $a_t \sim \pi(a_t|s_t)$ に従って行動をサンプリングし，またダイナミクス $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$ に従って状態をサンプリングすることにより生成される．方策を特徴づけるパラメータを θ とする．報酬 $r_t = r(s_t, a_t, s_{t+1})$ は，各ステップにおいて受け取る．目標は，期待総報酬 $\sum_{t=0}^{\infty} r_t$ の最大化であり，全ての方策に対して有限であると仮定する．

価値関数を次のように定義する．

$$V^{\pi}(s_t) := \mathbb{E}_{s_{t+1:\infty},\, a_{t:\infty}} \left[\sum_{\ell=0}^{\infty} r_{t+\ell} \right], \quad Q^{\pi}(s_t, a_t) := \mathbb{E}_{s_{t+1:\infty},\, a_{t+1:\infty}} \left[\sum_{\ell=0}^{\infty} r_{t+\ell} \right]$$

また，アドバンテージ関数を次のように定義する．

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

方策勾配法は，勾配 $g := \nabla_{\theta} \mathbb{E}[\sum_{t=0}^{\infty} r_t]$ を繰り返し推定することにより，期待総報酬を最大化する．方策勾配に対する複数の関連する表現が存在する．

$$g = \mathbb{E}_{s_{0:\infty},\, a_{0:\infty}} \left[\sum_{t=0}^{\infty} \Psi_t \nabla_{\theta} \log \pi_{\theta}(a_t|s_t) \right]$$

$\Psi_t = A^{\pi}(s_t, a_t)$ とすることにより分散をほぼ最小にすることができる．実際にはアドバンテージ関数は未知で推定されなければならないため，これを推定する方法を以下で議論する．

1.2 割引関数

パラメータ γ を導入する．これにより，バイアスを導入してしまうことを引き換えにして，遠い報酬の重みを小さくすることにより分散を低くすることができる．このパラメータは，MDP の割引定式化において用いられる割引率に対応するが，この論文では割引問題における分散低減パラメータとして扱う．割引価値関数は，次により与えられる．

$$V^{\pi,\gamma}(s_t) := \mathbb{E}_{s_{t+1:\infty},\, a_{t:\infty}} \left[\sum_{\ell=0}^{\infty} \gamma^{\ell} r_{t+\ell} \right], \quad Q^{\pi,\gamma}(s_t, a_t) := \mathbb{E}_{s_{t+1:\infty},\, a_{t+1:\infty}} \left[\sum_{\ell=0}^{\infty} \gamma^{\ell} r_{t+\ell} \right]$$

また，割引アドバンテージ関数 $A^{\pi,\gamma}(s_t, a_t) := Q^{\pi,\gamma}(s_t, a_t) - V^{\pi,\gamma}(s_t)$ とする．方策勾配の割引近似は，次のように定義される．

$$g^\gamma = \mathbb{E}_{s_{0:\infty},\, a_{0:\infty}} \left[\sum_{t=0}^{\infty} A^{\pi,\gamma}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

以下では，$A^{\pi,\gamma}$ に対する推定をどのように得るかを議論する．

1.3 アドバンテージ関数

関数 V の TD 誤差を，$\delta_t^V := -V(s_t) + [r_t + \gamma V(s_{t+1})]$ と定義する．もし価値関数が $V = V^{\pi,\gamma}$ という条件を満たすなら，次が成り立つ．

$$\begin{aligned}
\mathbb{E}_{s_{t+1}} \left[\delta_t^V \right] &= -V^{\pi,\gamma}(s_t) + \mathbb{E}_{s_{t+1}} \left[r(s_t, a_t, s_{t+1}) + \gamma V^{\pi,\gamma}(s_{t+1}) \right] \\
&= -V^{\pi,\gamma}(s_t) + Q^{\pi,\gamma}(s_t, a_t) \\
&= A^{\pi,\gamma}(s_t, a_t)
\end{aligned}$$

従って，もし $V = V^{\pi,\gamma}$ なら，δ_t^V はアドバンテージ関数の不偏推定量である．

δ_t^V の k 個の割引和は，次のように書ける．

$$\begin{aligned}
\hat{A}_t^{(1)} &:= \delta_t^V = -V(s_t) + [r_t + \gamma V(s_{t+1})] \\
\hat{A}_t^{(k)} &:= \sum_{\ell=0}^{k-1} \gamma^\ell \delta_{t+\ell}^V = -V(s_t) + \left[\sum_{\ell=0}^{k-1} \gamma^\ell r_{t+\ell} + \gamma^k V(s_{t+k}) \right]
\end{aligned}$$

この式は，k ステップリターンの推定からベースライン $-V(s_t)$ を引いたものである．$\delta_t = \hat{A}_t^{(1)}$ との類似で，$\hat{A}_t^{(k)}$ をアドバンテージ関数の推定として考える．$k \to \infty$ とするにつれて，$\gamma^k V(s_{t+k})$ の項が小さくなっていくので，バイアスは小さくなる．なお，$k \to \infty$ では，

$$\hat{A}_t^{(\infty)} = -V(s_t) + \sum_{\ell=0}^{\infty} \gamma^\ell r_t$$

となり，経験リターン $\sum_{\ell=0}^{\infty} \gamma^\ell r_t$ から価値関数ベースライン $V(s_t)$ を引いたものが得られる．

■一般化アドバンテージ関数.　一般化アドバンテージ関数 (GAE) を，k ステップ推定量の指数重み付け平均として次のように定義する．

$$\hat{A}_t^{GAE(\gamma,\lambda)} := (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \hat{A}_t^{(k)} = \sum_{t=0}^{\infty} (\gamma\lambda)^\ell \delta_{t+\ell}^V \tag{1}$$

これは，価値関数の推定のための TD(λ) を，アドバンテージ関数に対応させたものである．$\lambda = 0,\ 1$ の時には，次が得られる．

$$\begin{aligned}
\hat{A}_t^{GAE(\gamma,0)} &= \hat{A}_t^{(0)} = -V(s_t) + r_t + \gamma V(s_{t+1}) \\
\hat{A}_t^{GAE(\gamma,1)} &= \hat{A}_t^{(\infty)} = -V(s_t) + \sum_{\ell=0}^{\infty} \gamma^\ell r_t
\end{aligned}$$

■注意.

- γ は，割引関数を定義するために導入された量である．
- λ は，推定のために追加で導入された量である．

Algorithm 1 PG with GAE

方策パラメータ θ_0 と価値価値関数パラメータ ϕ_0 を初期化する．
for $i = 0, 1, \cdots$ **do**
　N 時間ステップが得られるまで，現在の方策 π_θ をシミュレートする．
　$V = V_\phi$ を用いて，全ての時刻 $t \in \{1, 2, \cdots\}$ における δ_t^V を計算する．
　$\hat{A}_t = \sum_{\ell=0}^{\infty} (\gamma\lambda)^\ell \delta_{t+\ell}^V$ を全ての時間ステップに対して計算する．
　方策更新：θ_{i+1} を計算する．式 (2) や (3).
　価値関数近似更新：ϕ_{i+1} を計算する．式 (5).
end for

2　GAE の使い方

オリジナルの論文 [4] では TRPO [3] と組み合わせるやり方で書いてあるけど，ここでは特定の勾配法に拘らないやり方で書く．次のように近似器を置く．

- V_ϕ: 価値関数の近似器
- π_θ: 方策の近似器.

方策と価値関数を反復的に更新する全体的なアルゴリズムは, Algorithm 1 に与えられる．方策更新 $\theta_i \to \theta_{i+1}$ は価値関数 V_{ϕ_i} を用いて行われて，$V_{\phi_{i+1}}$ ではないことに注意する．もし $V_{\phi_{i+1}}$ を用いて方策更新してしまうと，次のような不具合が起こる：価値関数がオーバーフィットした状況だとを考えると，Bellman 残差 $r_t + \gamma V(s_{t+1}) - V(s_t)$ は全ての時間ステップにおいてゼロになり，したがって方策勾配もゼロになり，方策は改善されない．

2.1　アドバンテージ関数推定を用いた方策勾配

状態遷移と報酬の観測系列 $\{s_t, a_t\}_{t=0}^N$ と $\{r_t\}_{t=0}^N$ が与えられた場合，現在の価値関数近似 V_ϕ を用いて，アドバンテージ関数を次のように推定する．

$$\hat{A}_t^{GAE(\gamma,\lambda)} = \sum_{t=0}^{N} (\gamma\lambda)^\ell \delta_{t+\ell}^{V_\phi} = \sum_{t=0}^{N} (\gamma\lambda)^\ell \left[-V_\phi(s_{t+\ell}) + r_{t+\ell} + \gamma V_\phi(s_{t+1+\ell})\right]$$

■通常の方策勾配.　通常の方策勾配法では，$\min_\theta \mathbb{E}[\sum_{t=0}^{\infty} r_t]$ を解くために，各反復で勾配 $\nabla_\theta \mathbb{E}[\sum_{t=0}^{\infty} r_t]$ を考える．方策勾配の割引近似を，次のように計算する．

$$\begin{aligned} g^\gamma &= \mathbb{E}_{s_{0:\infty},\, a_{0:\infty}} \left[\sum_{t=0}^{\infty} A^{\pi,\gamma}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right] \\ &\approx \sum_{t=0}^{N} \hat{A}_t^{GAE(\pi,\gamma)} \nabla_\theta \log \pi_\theta(a_t|s_t) \end{aligned} \tag{2}$$

■TRPO.　雑に言うと，TRPO の場合，各反復での更新量を制限するために，次の問題を解くことを考える．

$$\begin{array}{ll} \min_\theta & \mathbb{E}[\sum_{t=0}^{\infty} r_t] \\ \text{s.t.} & ||\theta - \theta_{old}|| \le \epsilon \end{array}$$

ちゃんと書くと次のようになる（[4] や [3] を参照）.

$$\begin{array}{ll} \min_\theta & \frac{1}{N}\sum_{n=1}^{N}\frac{\pi_\theta(a_n|s_n)}{\pi_{\theta_{old}}(a_n|s_n)}\hat{A}_n \\ s.t. & \frac{1}{N}\sum_{n=1}^{N} D_{KL}\left(\pi_{\theta_{old}}(\cdot|s_n)||\pi_\theta(\cdot|s_n)\right) \leq \epsilon \end{array} \tag{3}$$

2.2 価値関数推定

サンプルされた報酬の割引総和を，次のようにおく.

$$\hat{V}_t = \sum_{\ell=0}^{N}(\lambda\gamma)^\ell r_{t+\ell}$$

価値関数を非線形関数近似する場合，最も単純なアプローチは非線形回帰問題を解くことである.

$$\min_\phi \sum_{n=1}^{N} ||V_\phi(s_n) - \hat{V}_n||^2$$

論文 [4] では，各反復法において価値関数を，信頼領域法 [2] を用いて最適化する（これは，価値関数推定の非線形最適化問題に対して信頼領域法を使っているのであり，方策最適化に対して信頼領域法を使う TRPO とは違うという事に注意！）信頼領域問題を定式化するために，まず $\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}||V_{\phi_{old}}(s_n) - \hat{V}_n||^2$ を計算する．ここで，ϕ_{old} は，最適化前のパラメータである．そうして，次の制約付き最適化問題を解く.

$$\begin{array}{ll} \min_\phi & \sum_{n=1}^{N}||V_\phi(s_n) - \hat{V}_n||^2 \\ s.t. & \frac{1}{N}\sum_{n=1}^{N}\frac{||V_\phi(s_n) - V_{\phi_{old}}(s_n)||^2}{2\sigma^2} \leq \epsilon \end{array} \tag{4}$$

この拘束は，価値関数は平均 $V_\phi(s)$ で分散 σ^2 を持つ条件付きガウス分布をパラメトライズするように取るとした場合の，以前の価値関数と新しい価値関数の間の平均 KL ダイバージェンスが ϵ よりも小さくなる拘束と等価である.

次の 2 次計画問題を解ことにより，信頼領域問題の近似解を計算する.

$$\begin{array}{ll} \min_\phi & g^T(\phi_- \phi_{old}) \\ s.t. & (\phi_- \phi_{old})^T\left[\frac{1}{N}\sum_n j_n j_n^T\right]_{\phi_{old}}(\phi_- \phi_{old}) \leq \epsilon \end{array} \tag{5}$$

ここで，g は目的関数 $\sum_{n=1}^{N}||V_\phi(s_n) - \hat{V}_n||^2$ の勾配であり，また $j_n = \nabla_\phi V_\phi(s_n)$ である.

3　自分のための補足

■KL ダイバージェンス．　連続確率変数の場合，一般に，KL ダイバージェンスは次のように定義される．

$$KL(q(\boldsymbol{u})||p(\boldsymbol{u})) = \int q(\boldsymbol{u}) \ln \frac{q(\boldsymbol{u})}{p(\boldsymbol{u})} d\boldsymbol{u} = \int q(\boldsymbol{u})[\ln q(\boldsymbol{u}) - \ln p(\boldsymbol{u})] d\boldsymbol{u}$$

正規分布 $q(u) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$，$p(u) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ とすると，

$$\begin{aligned}
&KL(q(\boldsymbol{u})||p(\boldsymbol{u})) = \int q(\boldsymbol{u})[\ln q(\boldsymbol{u}) - \ln p(\boldsymbol{u})] d\boldsymbol{u} \\
&= \frac{1}{2} \ln \frac{\det \boldsymbol{\Sigma}_p}{\det \boldsymbol{\Sigma}_q} + \int q(\boldsymbol{u}) \left[-\frac{1}{2} (\boldsymbol{u} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q^{-1} (\boldsymbol{u} - \boldsymbol{\mu}_q) - \frac{1}{2} (\boldsymbol{u} - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{u} - \boldsymbol{\mu}_p) \right] d\boldsymbol{u} \\
&= \frac{1}{2} \ln \frac{\det \boldsymbol{\Sigma}_p}{\det \boldsymbol{\Sigma}_q} - \frac{1}{2} \int q(\boldsymbol{u}) \mathrm{tr} \left[\boldsymbol{\Sigma}_q^{-1} (\boldsymbol{u} - \boldsymbol{\mu}_q)(\boldsymbol{u} - \boldsymbol{\mu}_q)^T - \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{u} - \boldsymbol{\mu}_p)(\boldsymbol{u} - \boldsymbol{\mu}_p)^T \right] d\boldsymbol{u} \\
&= \frac{1}{2} \ln \frac{\det \boldsymbol{\Sigma}_p}{\det \boldsymbol{\Sigma}_q} - \frac{1}{2} \mathrm{tr}[\boldsymbol{\Sigma}_q^{-1} \boldsymbol{\Sigma}_q] + \frac{1}{2} \int q(\boldsymbol{u}) \mathrm{tr} \left[\boldsymbol{\Sigma}_p^{-1} (\boldsymbol{u}\boldsymbol{u}^T - 2\boldsymbol{\mu}_p \boldsymbol{u}^T + \boldsymbol{\mu}_p \boldsymbol{\mu}_p^T) \right] d\boldsymbol{u} \\
&= \frac{1}{2} \ln \frac{\det \boldsymbol{\Sigma}_p}{\det \boldsymbol{\Sigma}_q} - \frac{D}{2} + \frac{1}{2} \mathrm{tr} \int q(\boldsymbol{u}) \left[\boldsymbol{\Sigma}_p^{-1} \boldsymbol{u}\boldsymbol{u}^T \right] d\boldsymbol{u} + \frac{1}{2} \mathrm{tr} \left[\boldsymbol{\Sigma}_p^{-1} (-2\boldsymbol{\mu}_p \boldsymbol{\mu}_q^T + \boldsymbol{\mu}_p \boldsymbol{\mu}_p^T) \right] d\boldsymbol{u} \\
&= \frac{1}{2} \left\{ \ln \frac{\det \boldsymbol{\Sigma}_p}{\det \boldsymbol{\Sigma}_q} - D + \mathrm{tr} \left[\boldsymbol{\Sigma}_p^{-1} (-2\boldsymbol{\mu}_p \boldsymbol{\mu}_q^T + \boldsymbol{\mu}_p \boldsymbol{\mu}_p^T) \right] + \int q(\boldsymbol{u}) \mathrm{tr} \left[\boldsymbol{\Sigma}_p^{-1} [(\boldsymbol{u} - \boldsymbol{\mu}_q)(\boldsymbol{u} - \boldsymbol{\mu}_q)^T + 2\boldsymbol{\mu}_q \boldsymbol{u}^T - \boldsymbol{\mu}_q \boldsymbol{\mu}_q^T] \right] d\boldsymbol{u} \right\} \\
&= \frac{1}{2} \left\{ \ln \frac{\det \boldsymbol{\Sigma}_p}{\det \boldsymbol{\Sigma}_q} - D + \mathrm{tr} \left[\boldsymbol{\Sigma}_p^{-1} (-2\boldsymbol{\mu}_p \boldsymbol{\mu}_q^T + \boldsymbol{\mu}_p \boldsymbol{\mu}_p^T) \right] + \mathrm{tr} \left[\boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\Sigma}_q + 2\boldsymbol{\mu}_q \boldsymbol{\mu}_q^T - \boldsymbol{\mu}_q \boldsymbol{\mu}_q^T) \right] \right\} \\
&= \frac{1}{2} \left\{ \ln \frac{\det \boldsymbol{\Sigma}_p}{\det \boldsymbol{\Sigma}_q} - D + \mathrm{tr} \left[\boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\Sigma}_q + \boldsymbol{\mu}_p \boldsymbol{\mu}_p^T - 2\boldsymbol{\mu}_p \boldsymbol{\mu}_q^T + \boldsymbol{\mu}_q \boldsymbol{\mu}_q^T) \right] \right\} \\
&= \frac{1}{2} \left\{ \ln \frac{\det \boldsymbol{\Sigma}_p}{\det \boldsymbol{\Sigma}_q} - D + \mathrm{tr} \left[\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_q \right] + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q) \right\}
\end{aligned}$$

（注：トレースの性質 $\mathrm{tr}(X + Y) = \mathrm{tr}(X) + \mathrm{tr}(Y)$ と $\mathrm{tr}(XY) = \mathrm{tr}(YX$ を使った）（よく使う式で，例えばノート [1] の最後の節に導出が乗っている）

1 変数の場合は，$q(u) = \mathcal{N}(u|\mu_q, \sigma_q^2)$ と $p(u) = \mathcal{N}(u|\mu_p, \sigma_p^2)$ として，

$$\begin{aligned}
D_{KL}(q(u)||p(u)) &= \int_{-\infty}^{\infty} q(u) \ln \frac{q(u)}{p(u)} du \\
&= \int_{-\infty}^{\infty} q(u) \left(-\frac{(u - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2} \ln(2\pi\sigma_q^2) + \frac{(u - \mu_p)^2}{2\sigma_p^2} + \frac{1}{2} \ln(2\pi\sigma_p^2) \right) du \\
&= -\left(\frac{(\mu_q^2 + \sigma_q^2) - 2\mu_q^2 + \mu_q^2}{2\sigma_q^2} \right) + \left(\frac{(\mu_q^2 + \sigma_q^2) - 2\mu_q\mu_p + \mu_p^2}{2\sigma_p^2} \right) + \frac{1}{2} \ln \frac{\sigma_p^2}{\sigma_q^2} \\
&= -\frac{1}{2} + \left(\frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} \right) + \frac{1}{2} \ln \frac{\sigma_p^2}{\sigma_q^2} \\
&= \frac{1}{2} \left\{ \ln \frac{\sigma_p^2}{\sigma_q^2} - 1 + \left(\frac{\sigma_q^2}{\sigma_p^2} \right) + \left(\frac{(\mu_p - \mu_q)^2}{\sigma_p^2} \right) \right\}
\end{aligned}$$

平均 μ_i を更新することだけを考える場合には，KL ダイバージェンスの中で $\frac{(\mu_q - \mu_p)^2}{2\sigma_p^2}$ の項以外は関係ない．式 (4) は，更新するパラメータに関係ない項を ϵ の中に押し込んで書いてある．

参考文献

[1] John Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3, 2007.

[2] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, second edition, 2006.

[3] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[4] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *International Conference on Learning Representation*, 2016.