Методы градиентного спуска по стратегиям при обучении с подкреплением

Жирнов Михаил Денисович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: zhirnov.m.d@yandex.ru

Научный руководитель *— Ульянов Владимир Васильевич*

Методы градиентного спуска по стратегиям широко используются в прикладных задачах обучения с подкреплением. Они имеют многочисленные приложения в таких областях, как видеоигры, беспилотные автомобили и робототехника. К сожалению, эти подходы страдают от высокой дисперсии оцениваемого функционала. В настоящей работе исследуется задача снижения дисперсии оценки градиента с помощью линейной комбинации фиксированного набора базовых функций.

Пусть множества состояний $\mathcal{S} \subseteq \mathbb{R}^l$ и действий $\mathcal{A} \subseteq \mathbb{R}^l$ являются компактами, а стохастическая стратегия $\pi(\mathbf{a} \mid \mathbf{s}, \theta)$, $\forall \mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$, зависит от параметра $\theta \in \mathbb{R}^d$. Пусть также существует градиент стратегии $\pi(\mathbf{a} \mid \mathbf{s}, \theta)$ по θ для любых $\mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$ и $\|\nabla_\theta \log \pi(\mathbf{a} \mid \mathbf{s}, \theta)\|_2 \leq B$ равномерно по θ.

Для снижения дисперсии и сохранения несмещенности оценки предлагается использовать параметрическое семейство функций вида: $\{\mathbf{w}^\top \cdot \mathbf{\Psi}(\mathbf{s}, \mathbf{a}) : \ \mathbf{w} \in \mathbb{R}^K\}$, где $\mathbf{\Psi}(\mathbf{s}, \mathbf{a}) = (\Psi_1(\mathbf{s}, \mathbf{a}), \ldots, \Psi_K(\mathbf{s}, \mathbf{a}))$ — вектор базовых функций, удовлетворяющий следующему свойству:

$$\mathbb{E}_{\mathbf{a} \sim \pi(\cdot \mid \mathbf{s}, \theta)} \left[\Psi_i(\mathbf{s}, \mathbf{a})\right] = 0, \ i = \overline{1, K}. \tag{1}$$

Рассматривается случайный вектор вида:

$$\mathcal{J}(\mathbf{s}, \mathbf{a}, \theta, \mathbf{w}) = \nabla_\theta \log \pi(\mathbf{a} \mid \mathbf{s}, \theta) \cdot \left\{Q_{\pi_\theta}(\mathbf{s}, \mathbf{a}) - \mathbf{w}^\top \cdot \mathbf{\Psi}(\mathbf{s}, \mathbf{a})\right\}, \tag{2}$$

где пара $(\mathbf{s}, \mathbf{a})$ имеет некоторое совместное распределение $\rho_\theta(\cdot, \cdot)$.

Предметом исследования была задача минимизации по $\mathbf{w}$ теоретической дисперсии случайного вектора $\mathcal{J}(\mathbf{s}, \mathbf{a}, \theta, \mathbf{w})$, которая сводится к задаче минимизации по $\mathbf{w}$ следующего функционала:

$$\mathbb{V}(\mathbf{w}, \theta) = \mathbb{E}_{\rho_\theta(\cdot, \cdot)} \left\|\mathcal{J}(\mathbf{s}, \mathbf{a}, \theta, \mathbf{w})\right\|_2^2, \tag{3}$$

а также его эмпирического аналога $\mathbb{V}^{\mathrm{EV}}(\mathbf{w}, \theta, N)$ на N независимых траекториях. В результате были получены

Теорема 1. *При выполнении* (1) *вектор оптимальных коэффициентов* $\mathbf{w}^*$*, минимизирующий теоретическую дисперсию случайного вектора* $\mathcal{J}(\mathbf{s}, \mathbf{a}, \theta, \mathbf{w})$*, равен*

$$\mathbf{w}^* = \Big[\mathbb{E}_{\rho_\theta(\cdot,\cdot)} \Big\{\mathbf{A}(\mathbf{s},\mathbf{a})^\top \mathbf{A}(\mathbf{s},\mathbf{a})\Big\}\Big]^{-1} \cdot \mathbb{E}_{\rho_\theta(\cdot,\cdot)} \Big[\mathbf{b}(\mathbf{s},\mathbf{a})\Big], \tag{4}$$

где $\mathbf{A}(\mathbf{s},\mathbf{a}) \in \mathbb{R}^{d\times K}$, $\mathbf{b}(\mathbf{s},\mathbf{a}) \in \mathbb{R}^K$*:*

$$\mathbf{A}(\mathbf{s},\mathbf{a}) = \mathbf{\Phi}(\mathbf{s},\mathbf{a}) \cdot \mathbf{\Psi}(\mathbf{s},\mathbf{a})^\top, \quad \mathbf{\Phi}(\mathbf{s},\mathbf{a}) = \nabla_\theta \log \pi(\mathbf{a} \mid \mathbf{s}, \theta), \tag{5}$$

$$\mathbf{b}(\mathbf{s},\mathbf{a}) = \mathbf{A}(\mathbf{s},\mathbf{a})^\top \cdot \mathbf{\Phi}(\mathbf{s},\mathbf{a}) \cdot Q_{\pi_\theta}(\mathbf{s},\mathbf{a}). \tag{6}$$

Теорема 2. *При выполнении* (1) *вектор оптимальных коэффициентов* $\mathbf{w}_N^*$*, минимизирующий эмпирическую дисперсию случайного вектора* $\mathcal{J}(\mathbf{s}, \mathbf{a}, \theta, \mathbf{w})$ *на* N *траекториях* $\{(\mathbf{s}_t^n, \mathbf{a}_t^n)\}_{t=1}^{T^n}$*, равен*

$$\mathbf{w}_N^* = \left[\sum_{n=1}^N \mathbf{A}_n^\top \mathbf{A}_n\right]^{-1} \cdot \left[\sum_{n=1}^N \mathbf{A}_n^\top \mathbf{b}_n\right], \tag{7}$$

где $\mathbf{A}_n = \sum_{t=1}^{T^n} \mathbf{\Phi}_{t,n} \mathbf{\Psi}_{t,n}^\top$, $\mathbf{b}_n = \sum_{t=1}^{T^n} \mathbf{\Phi}_{t,n} G_{t,n}$ *при* $\mathbf{\Psi}_{t,n} = \mathbf{\Psi}(\mathbf{s}_t^n, \mathbf{a}_t^n)$, $\mathbf{\Phi}_{t,n} = \nabla_\theta \log \pi(\mathbf{a}_t^n \mid \mathbf{s}_t^n, \theta)$*,* $G_{t,n}$ *— несмещенная оценка* $Q_{\pi_\theta}(\mathbf{s}_t^n, \mathbf{a}_t^n)$*.*

Также доказана асимптотическая несмещенность эмпирической дисперсии, рассчитанной на M независимых траекториях, с вектором оптимальных коэффициентов из (7), т.е.

$$\mathbb{E}_{\rho_\theta(\cdot,\cdot)} \left[\mathbb{V}^{\mathrm{EV}}(\mathbf{w}_N^*, \theta, M)\right] \underset{\substack{N\to\infty \\ M\to\infty}}{\xrightarrow{\mathbb{P}}} \mathbb{V}(\mathbf{w}^*, \theta), \tag{8}$$

где $\mathbf{w}^*$ — вектор оптимальных коэффициентов из (4).

Основная идея доказательства (8) заключается в применении теоремы 2.1 из работы [1], а также в использовании асимптотической несмещенности эмпирической оценки дисперсии.

Литература

1. Belomestny D., Iosipoi L., Paris Q., Zhivotovskiy N. Empirical Variance Minimization with Applications in Variance Reduction and Optimal Control // arXiv preprint arXiv:1712.04667. — 2017.