エントロピー正則化付方策改善のための目的関数の補正

Objective Correction for Policy Improvement under Entropy Regularization

岩城 諒 *1　　浅田 稔 *1
Ryo Iwaki　　Minoru Asada

*1大阪大学大学院工学研究科 知能・機能創成工学専攻
Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University

Reinforcement learning aims to find a policy which maximizes long term future reward by interacting with unknown environment through trial and error. In this study, we propose an objective correction method for entropy regularized Markov decision process. After deriving a policy gradient under the regularization by the entropy and relative entropy, we propose an on-policy objective correction method for off-policy policy improvement under entropy regularization.

1. はじめに

強化学習は，未知の環境と試行錯誤的に相互作用しながら，意思決定則である方策を最適化することを目的とする．多くの場合，学習の目的関数は，設計者が報酬関数として定義した報酬の割引累積和である．環境と相互作用し学習サンプルを生成する挙動方策が，最適化したい推定方策と異なるとき，方策オフと呼ぶ．方策オフで学習可能であれば，過去の学習データの再利用が可能であり，サンプル効率よく学習できる [Lin 92, Mnih 15, Sugimoto 16]．

近年，エントロピーもしくは双対エントロピーによって目的関数を正則化する学習則が数多く研究されている．エントロピー正則化は，方策勾配法の興りとともに提案され [Williams 92]，深層強化学習においても有用性が示された [Mnih 16]．さらに近年，価値ベースの手法と方策ベースの手法が統一的に扱えることが示されてきた [O'Donoghue 16, Haarnoja 17, Nachum 17a, Schulman 17]．また，双対エントロピーによる正則化についても，様々な学習則が提案されてきた [Todorov 06, Todorov 10, Peters 10, Azar 12, Fox 16, Nachum 17b, Kozuno 17]．

一方で，(双対) エントロピーによる正則化を導入すると，目的関数と最適方策が元となるマルコフ決定過程と異なってしまうという問題がある．本研究では，エントロピーの正則化を利用した方策オフ型の学習則の一つである Path Consistency Learning (PCL) とその派生である trust-PCL[Nachum 17b] に着目し，これらの学習則を利用して元となるマルコフ決定過程の最適方策を求める手法を提案する．

2. 背景

2.1 マルコフ決定過程

マルコフ決定過程 (Markov Decision Process, MDP) における最適方策を獲得する問題を扱う．MDP は $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \rho_0, \gamma)$ の組によって特定される．$\mathcal{S}, \mathcal{A}$ はそれぞれ可能な状態と行動の集合である．離散時刻 $t \in \mathbb{N}_{\geq 0}$ において，エージェントは環境の状態 $s_t \in \mathcal{S}$ を観測し，行動 $a_t \in \mathcal{A}$ を選択する．環境の状態は，マルコフ性を有する状態遷移確率 $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ に従って次状態 s_{t+1} に遷移する．エージェントは，有界な報酬関数 $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ に従って環境から報酬を得る．ρ_0 は初期状態の分布，$\gamma \in [0, 1]$ は割引率である．エージェントの意思決定は，方策 π に従う．方策が決定論的であるとき，π は状態から行動への写像 $\pi : \mathcal{S} \to \mathcal{A}$ であり，方策が確率的であるとき，π はある状態である行動をとる確率 $\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ である．方策 π と初期状態の分布 ρ_0 に対し，将来の割引き状態分布 $\rho^\pi(s) = \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | \rho_0, \pi)$ が存在する．状態価値関数 $V^\pi(s) \triangleq \mathbb{E}_{\pi,\mathcal{P}}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s\right]$ は，方策 π のもとである状態 s に期待される割引収益である．同様に，行動価値関数 $Q^\pi(s, a) \triangleq \mathbb{E}_{\pi,\mathcal{P}}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, a_0 = a\right]$ は，ある状態 s において行動 a をとり，その後方策 π に従った場合に期待される割引収益である．さらに，アドバンテージ関数を $A^\pi(s, a) \triangleq Q^\pi(s, a) - V^\pi(s)$ によって定義する．強化学習は，目的関数

$$\eta(\pi) = \sum_{s \in \mathcal{S}} \rho_0 V^\pi(s) = \sum_{s \in \mathcal{S}} \rho^\pi(s) \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}(s, a) \tag{1}$$

を最大化する決定論的な最適方策

$$\pi^* \in \arg\max_{\pi} \eta(\pi), \tag{2}$$

を獲得することである．方策の最適化としては様々な手法が提案されているが，パラメータ θ によって表現された方策 π_θ を考えるとき，方策勾配 $\nabla_\theta \eta(\pi)$ は以下のように与えられる [Sutton 99]:

$$\begin{aligned} \nabla_\theta \eta(\pi) &= \sum_{s \in \mathcal{S}} \rho^\pi(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) \\ &= \sum_{s \in \mathcal{S}} \rho^\pi(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) A^\pi(s, a). \end{aligned} \tag{3}$$

2.2 エントロピー正則化付マルコフ決定過程

本研究ではさらに，エントロピー正則化付マルコフ決定過程 (Entropy-Regularized Markov Decision Process, ERMDP) を扱う．§2.1 で導入した MDP に加え，$\tau \geq 0, \lambda \geq 0$ として，エントロピー正則化付状態価値関数を定義する:

$$\begin{aligned} V^\pi_{\tau,\lambda}(s) \triangleq \mathbb{E}_{\pi,\mathcal{P}}\Bigg[\sum_{t=0}^{\infty} \gamma^t \Bigg(\mathcal{R}(s_t, a_t) - \tau \ln \pi(a_t|s_t) \\ - \lambda \ln \frac{\pi(a_t|s_t)}{\bar{\pi}(a_t|s_t)} \Bigg) \Bigg| s_0 = s \Bigg]. \end{aligned} \tag{4}$$

連絡先: 岩城諒，ryo.iwaki@ams.eng.osaka-u.ac.jp

ただし，$\bar{\pi}$ は参照方策である．エントロピー正則化付状態価値関数は以下のベルマン方程式を満たす:

$$V^{\pi}_{\tau,\lambda}(s) = \sum_{a\in\mathcal{A}} \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma(\mathcal{P}V^{\pi}_{\tau,\lambda})(s,a) - \tau\ln\pi(a|s) - \lambda\ln\frac{\pi(a|s)}{\bar{\pi}(a|s)}\Big). \quad (5)$$

ただし，

$$(\mathcal{P}V^{*}_{\tau,\lambda})(s,a) = \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)V^{*}_{\tau,\lambda}(s')$$

とした．ERMDP での目的関数は，以下で与えられる:

$$\eta_{\tau,\lambda}(\pi) = \sum_{s\in\mathcal{S}} \rho_0(s)V^{\pi}_{\tau,\lambda}(s) = \eta(\pi) + \tau H_{\gamma}(\pi) - \lambda D_{\gamma}(\pi,\bar{\pi}). \quad (6)$$

ただし，

$$H_{\gamma}(\pi) = -\sum_{s\in\mathcal{S}} \rho^{\pi}(s)\sum_{a\in\mathcal{A}} \pi(\cdot|s)\ln\pi(\cdot|s),$$
$$D_{\gamma}(\pi,\bar{\pi}) = \sum_{s\in\mathcal{S}} \rho^{\pi}(s) D_{\mathrm{KL}}(\pi(\cdot|s),\bar{\pi}(\cdot|s))$$

である．

ERMDP において，最適価値 $V^{*}_{\tau,\lambda}(s)$ と最適方策 $\pi^{*}_{\tau,\lambda}$ に対し，以下の命題が成立する．

命題 1. [Nachum 17a, Nachum 17b, Kozuno 17] 目的関数 (6) の元で，最適価値 $V^{*}_{\tau,\lambda}(s)$ と最適方策 $\pi^{*}_{\tau,\lambda}$ は，全ての状態行動の組に対し以下の方程式を満たす:

$$\begin{aligned} V^{*}_{\tau,\lambda}(s) &= \mathcal{R}(s,a) + \gamma\sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)V^{*}_{\tau,\lambda}(s') \\ &\quad - (\tau+\lambda)\ln\pi^{*}_{\tau,\lambda}(a|s) + \lambda\ln\bar{\pi}(a|s) \\ &= \mathbb{E}_{s'\sim\mathcal{P}}\big[\mathcal{R}(s,a) + \gamma V^{*}_{\tau,\lambda}(s') \\ &\quad - (\tau+\lambda)\ln\pi^{*}_{\tau,\lambda}(a|s) + \lambda\ln\bar{\pi}(a|s)\big]. \end{aligned} \quad (7)$$

すなわち，最適方策は以下で与えられる:

$$\pi^{*}_{\tau,\lambda}(a|s) = \frac{\bar{\pi}(a|s)^{\frac{\lambda}{\tau+\lambda}}\exp\left(\left(\mathcal{R}(s,a) + \gamma(\mathcal{P}V^{*}_{\tau,\lambda})(s,a)\right)/(\tau+\lambda)\right)}{\exp\left(V^{*}_{\tau,\lambda}(s)/(\tau+\lambda)\right)}. \quad (8)$$

さらに，全ての状態行動の組に対し，価値関数 V と方策 π が方程式 (7) を満たすとき，その価値関数と方策はそれぞれ最適価値と最適方策である: $\pi = \pi^{*}_{\tau,\lambda}, V = V^{*}_{\tau,\lambda}$.

さらに，命題 1 は次のように複数ステップ予測 (multi-steps prediction) に拡張できる．

系 2. [Nachum 17b] $n \in \mathbb{N}_{\geq 1}$ とする．目的関数 (6) の元で，最適価値 $V^{*}_{\tau,\lambda}(s)$ と最適方策 $\pi^{*}_{\tau,\lambda}$ は，任意の状態 s_t から開始される全ての状態行動系列に対し以下の方程式を満たす:

$$V^{*}_{\tau,\lambda}(s_t) = \mathbb{E}_{s_{t+i}\sim\mathcal{P}}\Big[\gamma^n V^{*}_{\tau,\lambda}(s_{t+n}) + \sum_{i=0}^{n-1}\gamma^i\big(\mathcal{R}(s_{t+i},a_{t+i}) - (\tau+\lambda)\ln\pi^{*}_{\tau,\lambda}(a_{t+i}|s_{t+i}) + \lambda\ln\bar{\pi}(a_{t+i}|s_{t+i})\big)\Big]. \quad (9)$$

式 (7), (9) ともに，期待値は状態遷移確率に従う後続状態のサンプリングのみに対して定義されていて，方策に依存しない．よって，Path Consistency Learning (PCL)[Nachum 17a] と Trust-PCL[Nachum 17b] は，Q 学習 [Watkins 89, Watkins 92] と同様に，式 (9) の両辺の残差を誤差信号としてそれを最小化するよう学習する．すなわち，それぞれパラメータ θ と ϕ で表現された方策 $\pi_{\theta}(a|s)$ と価値関数 $V_{\phi}(s)$ に対し，以下のようにパラメータを更新する:

$$\begin{aligned} \delta_{t:t+n}(\theta,\phi) &= -V_{\phi}(s_t) + \gamma^n V_{\phi}(s_{t+n}) + \sum_{i=0}^{n-1}\gamma^i\big(\mathcal{R}(s_{t+i},a_{t+i}) \\ &\quad -(\tau+\lambda)\ln\pi_{\theta}(a_{t+i}|s_{t+i}) + \lambda\ln\bar{\pi}(a_{t+i}|s_{t+i})\big), \\ \Delta\theta &\propto -\frac{1}{2}\nabla_{\theta}\delta^2_{t:t+n}(\theta,\phi) \\ &= \delta_{t:t+n}(\theta,\phi)\sum_{i=0}^{n-1}\gamma^i\nabla_{\theta}\ln\pi_{\theta}(a_{t+i}|s_{t+i}), \\ \Delta\phi &\propto -\frac{1}{2}\nabla_{\phi}\delta^2_{t:t+n}(\theta,\phi) \\ &= \delta_{t:t+n}(\theta,\phi)\left(\nabla_{\phi}V_{\phi}(s_t) - \gamma^n\nabla_{\phi}V_{\phi}(s_{t+n})\right). \end{aligned} \quad (10)$$

ただし，$\tau \neq 0 \wedge \lambda = 0$ のときが PCL，$\tau \neq 0 \wedge \lambda \neq 0$ のときが Trust-PCL である．

3. エントロピー正則化付方策勾配

エントロピー正則化付行動価値関数を，以下のように定義する:

$$Q^{\pi}_{\tau,\lambda}(s,a) \triangleq \mathbb{E}_{\pi,\mathcal{P}}\left[\sum_{t=0}^{\infty}\gamma^t\Big(\mathcal{R}(s_t,a_t) - \tau\ln\pi(a_t|s_t) - \lambda\ln\frac{\pi(a_t|s_t)}{\bar{\pi}(a_t|s_t)}\Big)\middle| s_0 = s, a_0 = a\right]. \quad (11)$$

エントロピー正則化付行動価値は以下のベルマン方程式

$$Q^{\pi}_{\tau,\lambda}(s,a) = \mathcal{R}(s,a) + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s,a)\sum_{a'\in\mathcal{A}}\pi(a'|s')Q^{\pi}_{\tau,\lambda}(s',a') - \tau\ln\pi(a|s) - \lambda\ln\frac{\pi(a|s)}{\bar{\pi}(a|s)} \quad (12)$$

を満たし，エントロピー正則化付状態価値との間に以下の関係が成立する:

$$V^{\pi}_{\tau,\lambda}(s) = \sum_{a\in\mathcal{A}}\pi(a|s)Q^{\pi}_{\tau,\lambda}(s,a), \quad (13)$$

$$Q^{\pi}_{\tau,\lambda}(s,a) = \mathcal{R}(s,a) + \gamma(\mathcal{P}V^{\pi}_{\tau,\lambda})(s,a) - \tau\ln\pi(a|s) - \lambda\ln\frac{\pi(a|s)}{\bar{\pi}(a|s)}. \quad (14)$$

さらに，エントロピー正則化付アドバンテージを以下のように定義する:

$$A^{\pi}_{\tau,\lambda}(s,a) = Q^{\pi}_{\tau,\lambda}(s,a) - V^{\pi}_{\tau,\lambda}(s).$$

MDP におけるアドバンテージ A^π と同様に，方策 π についての $A^\pi_\tau(s,a)$ の期待値は 0 である:

$$\begin{aligned}\sum_{a\in\mathcal{A}}\pi(a|s)A^\pi_{\tau,\lambda}(s,a) &= \sum_{a\in\mathcal{A}}\pi(a|s)\left(Q^\pi_{\tau,\lambda}(s,a)-V^\pi_{\tau,\lambda}(s)\right)\\ &= \sum_{a\in\mathcal{A}}\pi(a|s)Q^\pi_{\tau,\lambda}(s,a)-V^\pi_{\tau,\lambda}(s)\\ &= 0.\end{aligned}$$

最後の等号は (13) によって成立する．さらに，式 (5) に基づき，エントロピー正則化付 Temporal Difference (TD) 誤差 δ^π_τ を，以下のように定義する:

$$\begin{aligned}\delta^\pi_{\tau,\lambda}(s,a,s') = \mathcal{R}(s,a)-\tau\ln\pi(a|s)-\lambda\ln\frac{\pi(a|s)}{\bar\pi(a|s)}\\ +\gamma V^\pi_{\tau,\lambda}(s')-V^\pi_{\tau,\lambda}(s).\end{aligned}$$

TD 誤差 δ^π_τ の期待値は，アドバンテージ $A^\pi_\tau(s,a)$ に一致する:

$$\begin{aligned}\bar\delta^\pi_{\tau,\lambda}(s,a) &\triangleq \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[\delta^\pi_{\tau,\lambda}|s,a\right]\\ &= \mathcal{R}(s,a)-\tau\ln\pi(a|s)-\lambda\ln\frac{\pi(a|s)}{\bar\pi(a|s)}\\ &\quad+\gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[V^\pi_{\tau,\lambda}(s')\right]-V^\pi_{\tau,\lambda}(s_t)\\ &= Q^\pi_{\tau,\lambda}(s,a)-V^\pi_{\tau,\lambda}(s)\\ &= A^\pi_{\tau,\lambda}(s,a).\end{aligned}$$

以下の命題は，ERMDP における目的関数 (6) を最大化するための最急勾配方向を与える．

命題 3. ERMDP における目的関数 (6) についての方策勾配は以下で与えられる:

$$\nabla_\theta\eta_{\tau,\lambda}(\pi)=\sum_{s\in\mathcal{S}}\rho^\pi(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)\left(Q^\pi_{\tau,\lambda}(s,a)-b(s)\right).$$

命題 3 は，MDP における方策勾配定理 [Sutton 99] と同様の手順で，式 (12)-(14) を利用して証明できる．さらに，従来の方策勾配法と同様に，$b(s)=V^\pi_{\tau,\lambda}(s)$ とおくことで以下を得る:

$$\begin{aligned}\nabla_\theta\eta_{\tau,\lambda}(\pi) &= \sum_{s\in\mathcal{S}}\rho^\pi(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)\left(Q^\pi_{\tau,\lambda}(s,a)-V^\pi_{\tau,\lambda}(s)\right)\\ &= \sum_{s\in\mathcal{S}}\rho^\pi(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)A^\pi_{\tau,\lambda}(s,a)\\ &= \sum_{s\in\mathcal{S}}\rho^\pi(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)\bar\delta^\pi_{\tau,\lambda}(s,a).\end{aligned}\tag{15}$$

4. 目的関数の補正

エントロピーによって正則化する場合，PCL などの方策オフ学習が可能である一方で，目的関数 (6) は元の MDP の目的関数 (1) と異なる．さらに得られる最適方策 (8) はソフトマックス関数であるため，元の MDP の決定論的な最適方策 (2) とは異なり，ERMDP での最適方策の”良さ”は，参照方策 $\bar\pi$ の選び方に依存する．本節では，ERMDP で定式化される学習則を MDP での目的関数最大化に利用する方法について議論する．

命題 4. MDP における方策勾配 (3) は，ERMDP におけるエントロピー正則化付方策勾配を利用して以下のように表現できる:

$$\begin{aligned}&\nabla_\theta\eta(\pi)\\ &=\sum_{s\in\mathcal{S}}\rho^\pi(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)\bar\delta^\pi_{\tau,\lambda}(s,a)\\ &\quad+\sum_{s\in\mathcal{S}}\rho^\pi(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)\big((\tau+\lambda)\ln\pi_\theta(a|s)-\lambda\ln\bar\pi(a|s)\big)\\ &=\mathbb{E}_{\pi,\rho^\pi}\left[\nabla_\theta\ln\pi_\theta(a|s)\bar\delta^\pi_{\tau,\lambda}(s,a)\right]\\ &\quad+\mathbb{E}_{\pi,\rho^\pi}\left[\nabla_\theta\ln\pi_\theta(a|s)\big((\tau+\lambda)\ln\pi_\theta(a|s)-\lambda\ln\bar\pi(a|s)\big)\right].\end{aligned}\tag{16}$$

命題 4 は，式 (6) の両辺を θ について偏微分し，命題 4 を利用することで容易に示される．式 (16) の初項はエントロピー正則化における方策勾配 (15) であり，第二項は ERMDP における方策勾配を MDP における方策勾配へと補正する．さらに，初項の $\nabla_\theta\ln\pi_\theta(a|s)\bar\delta^\pi_{\tau,\lambda}(s,a)$ は PCL による方策の更新方向 (10) において，状態価値関数を現在の方策に関するエントロピー正則化付状態価値 $V^\pi_{\tau,\lambda}$ に置き換え，さらに $n=1$ とした場合と等価である．

β を挙動方策とする．$\bar\delta^\pi_{\tau,\lambda}$ を $\delta_{t:t+1}(\theta,\phi)$ で置き換え，PCL と同様に，式 (16) の初項の期待値を β によって生成されたサンプルから計算することで，以下の方策オフと方策オンのサンプリングが混合した方策勾配を得る:

$$\begin{aligned}&\nabla_\theta\eta_\beta(\pi)\\ &=\mathbb{E}_{\beta,\rho^\beta}\left[\nabla_\theta\ln\pi_\theta(a|s)\delta_{t:t+1}(\theta,\phi)\right]\\ &\quad+\mathbb{E}_{\pi,\rho^\pi}\left[\nabla_\theta\ln\pi_\theta(a|s)\big((\tau+\lambda)\ln\pi_\theta(a|s)-\lambda\ln\bar\pi(a|s)\big)\right].\end{aligned}\tag{17}$$

このとき，式 (17) 右辺初項は，ERMDP での最適方策へ向かう勾配を計算している．式 (17) に基づき，方策オフのサンプルからエントロピー正則化付目的関数を最適化し，方策オンのサンプルから目的関数を補正することで，サンプル効率よく元の MDP における最適方策を学習できる．

5. おわりに

本研究では，エントロピー正則化付方策勾配定理を導出し，さらに MDP での目的関数を ERMDP での更新則を利用して最大化することを提案した．今後の課題として，複数ステップ予測（$n>1$）への拡張，提案法の実験的な評価，収束性に関する議論などが挙げられる．

謝辞

本研究は，JST，CREST，JPMJCR17A4 の支援を受けたものである．

参考文献

[Azar 12] Azar, M. G., Gómez, V., and Kappen, H. J.: Dynamic Policy Programming, *Journal of Machine Learning Research*, Vol. 13, (2012)

[Fox 16] Fox, R., Pakman, A., and Tishby, N.: Taming the Noise in Reinforcement Learning via Soft Updates, in *Uncertainty in Artificial Intelligence* (2016)

[Haarnoja 17] Haarnoja, T., Tang, H., Abbeel, P., and Levine, S.: Reinforcement Learning with Deep Energy-Based Policies, in *International Conference on Machine Learning* (2017)

[Kozuno 17] Kozuno, T., Uchibe, E., and Doya, K.: Unifying Value Iteration, Advantage Learning, and Dynamic Policy Programming, in *arXiv* (2017)

[Lin 92] Lin, L.-J.: Self-improving reactive agents based on reinforcement learning, planning and teaching, *Machine Learning*, Vol. 8, No. 3/4, pp. 69–97 (1992)

[Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015)

[Mnih 16] Mnih, V., Mirza, A. P. B. M., Graves, A., Harley, T., Lillicrap, T. P., Silver, D., and Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning, in *International Conference on Machine Learning*, pp. 1928–1937 (2016)

[Nachum 17a] Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D.: Bridging the Gap Between Value and Policy Based Reinforcement Learning, in *Advances in Neural Information Processing Systems* (2017)

[Nachum 17b] Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D.: Trust-PCL: An Off-Policy Trust Region Method for Continuous Control, in *International Conference on Learning Representations* (2017)

[O'Donoghue 16] O'Donoghue, B., Munos, R., Kavukcuoglu, K., and Mnih, V.: Combining Policy Gradient and Q-Learning, in *International Conference on Learning Representations* (2016)

[Peters 10] Peters, J., Mülling, K., and Altün, Y.: Relative Entropy Policy Search, in *AAAI* (2010)

[Schulman 17] Schulman, J., Chen, X., and Abbeel, P.: Equivalence Between Policy Gradients and Soft Q-Learning, in *arXiv* (2017)

[Sugimoto 16] Sugimoto, N., Tangkaratt, V., Wensveen, T., Zhao, T., Sugiyama, M., and Morimoto, J.: Trial and Error: Using Previous Experiences as Simulation Models in Humanoid Motor Learning, *IEEE Robotics & Automation Magazine*, Vol. 23, No. 1, pp. 96–105 (2016)

[Sutton 99] Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y.: Policy Gradient Methods for Reinforcement Learning with Function Approximation, in *Advances in Neural Information Processing Systems*, Vol. 12, pp. 1057–1063 (1999)

[Todorov 06] Todorov, E.: Linearly-solvable Markov decision problems, in *Advances in Neural Information Processing Systems* (2006)

[Todorov 10] Todorov, E.: Policy gradients in linearly-solvable MDPs, in *Advances in Neural Information Processing Systems* (2010)

[Watkins 89] Watkins, C. J. C. H.: *Learning from Delayed Rewards*, PhD thesis, King ' s College, Cambridge, England (1989)

[Watkins 92] Watkins, C. J. C. H. and Dayan, P.: Q-learning, *Machine Learning*, Vol. 8, pp. 279–292 (1992)

[Williams 92] Williams, R. J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning, *Machine Learning*, Vol. 8, pp. 229–256 (1992)