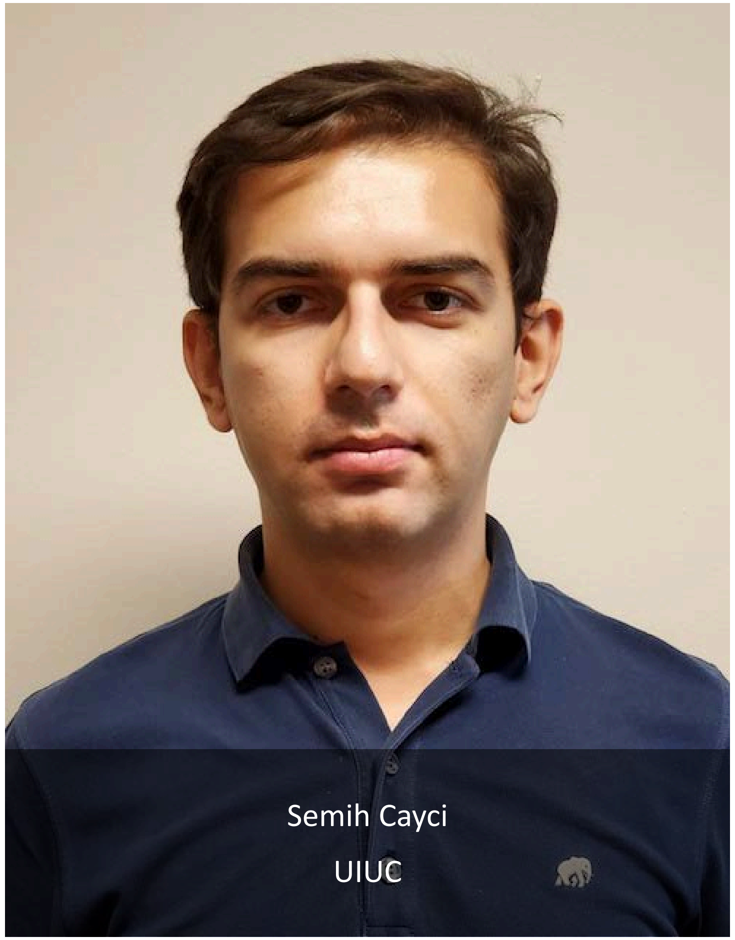
On the Convergence Rate of Entropy-Regularized Natural Policy Gradient with Linear Function Approximation

R. Srikant

c3.ai DTI/CSL/ECE

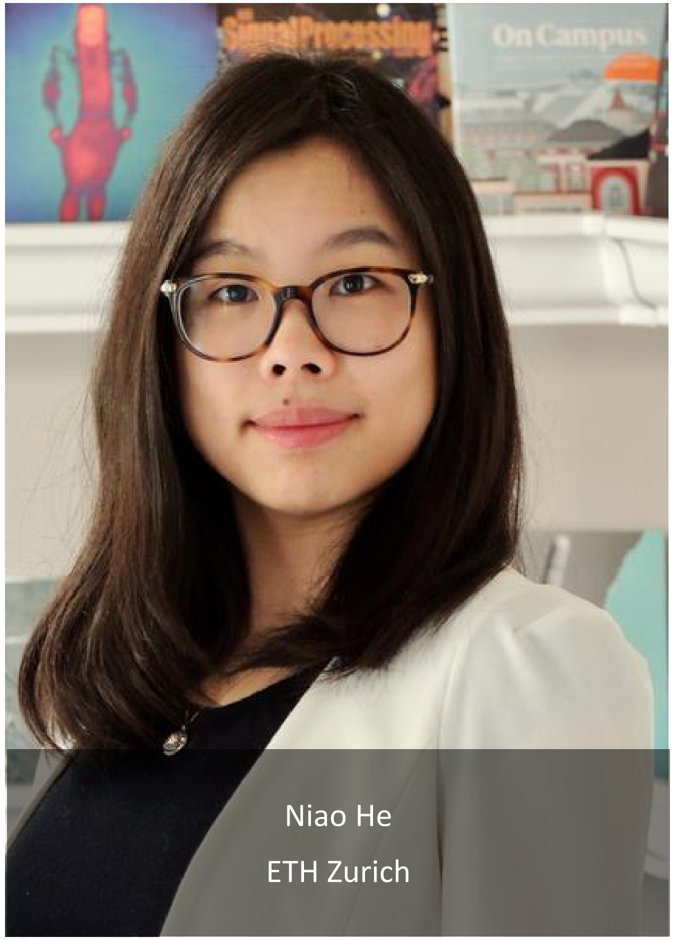
University of Illinois at Urbana-Champaign

Coauthors



Semih Cayci

UIUC



Niao He

ETH Zurich

Markov Decision Processes

Dynamical system in discrete time:

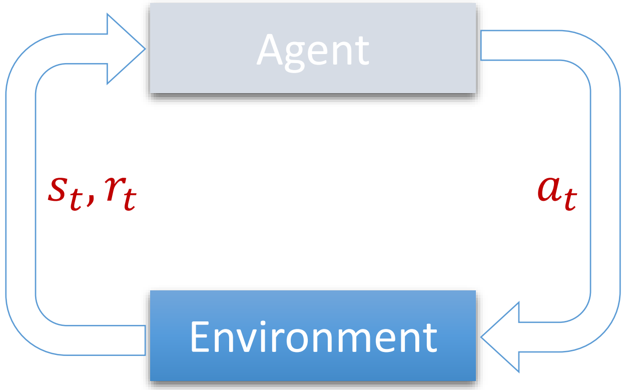
Control policy

$$s_{t+1} = f(s_t, a_t, w_t) \qquad a_t \sim \pi(\cdot\,|s_t) \qquad r_t = r(s_t, a_t)$$

s_t: state, a_t: control action, w_t: noise

Value function: For discount factor $\gamma \in (0, 1)$

$$V^\pi(s) = \sum_{t \geq 0} \gamma^t E_\pi[r(s_t, a_t)|s_0 = s]$$



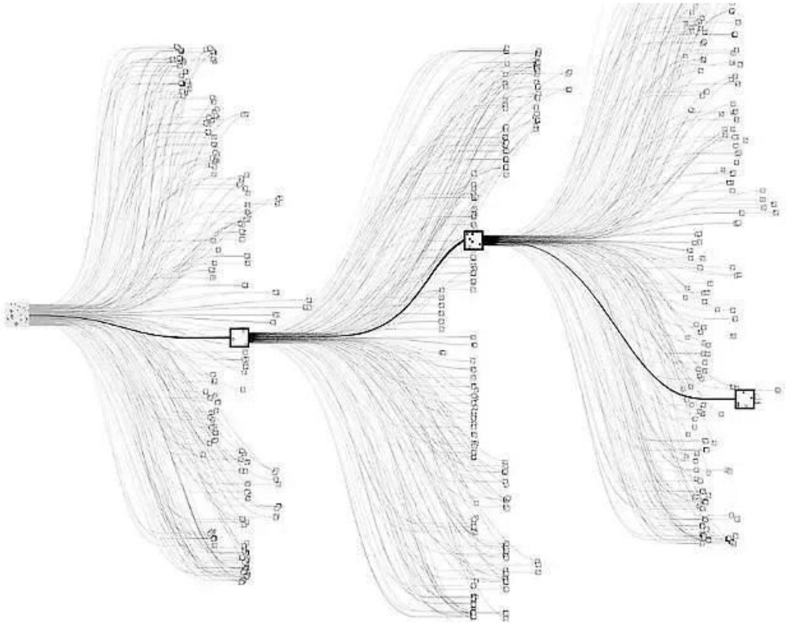
Goal: Find π^* to maximize $E_{s_0 \sim \mu} V^\pi(s_0)$

RL: Policy gradient

Main Challenge in RL: Large State Spaces

- Optimal policy can be found via tabular methods (in principle)

- Problem: High memory and time complexity for large state spaces



Source: Cheerla, '18

- Example: Go, Chess (Shannon number for Chess: 10^{120})

Entropy-Regularized NPG with LFA: Highlights

- **Function approximation:** Key element behind the success of RL in practice

Outline: Natural Policy Gradient with Linear Approximation

- Policy optimization with linear function approximation
- Entropy-regularization to encourage exploration and smoothen the optimization landscape
- Linear convergence rate up to compatible function approximation error
- O(1/T) convergence rate with weaker assumptions

Main message: Entropy regularization leads to improved convergence rates

Existing Work

- **NPG in function approximation regime:**

  (Agarwal, Kakade, Lee, Mahajan '20): Unregularized NPG and Q-NPG with log-linear parameterization
  $O\left(\frac{1}{\sqrt{T}}\right)$ convergence up to approximation errors

  (Chen, Khodadadian, Maguluri '21): $O(1/T)$ convergence with *off-policy* NPG

- **Entropy-Regularized NPG/PG:** Tabular setting

  (Cen, Cheng, Chen, Wei, Chi '20): Regularized NPG with direct softmax parameterization
  Linear convergence

  (Mei, Xiao, Szepesvari '20): Linear convergence of entropy-regularized PG with softmax
  $\Theta\left(\frac{1}{T}\right)$ convergence rate for unregularized PG

  (Lan '21): Linear convergence of NPG with general convex regularizers

  (Khodadadian , Jhunjhunwala, Varma, Maguluri '21): Linear convergence of unregularized NPG

Softmax Parameterization with LFA

Log-linear policy class $\Pi = \{\pi_\theta : \theta \in R^d\}$

$$\pi_\theta(a|s) = \frac{e^{\theta^\top \phi_{s,a}}}{\sum_{a' \in A} e^{\theta^\top \phi_{s,a'}}}$$

Basis vectors $\{\phi_{s,a} : s \in S, a \in A\}$

$$\max_{(s,a) \in S \times A} ||\phi_{s,a}||_2 \leq 1$$

- **Restricted policy class:** Does not contain all randomized policies
- **Convergence results** with respect to the best policy $\pi^* \in \Pi$

Entropy Regularization

Value function

$$V^{\pi}(\mu) = \sum_{t\geq 0} \gamma^t E_{\pi}[r(s_t, a_t) | s_0 \sim \mu]$$

Regularizer

$$H^{\pi}(\mu) = \sum_{t\geq 0} \gamma^t E_{\pi}[-\log \pi(a_t | s_t) \,| s_0 \sim \mu]$$

Entropy-Regularized Value function

$$V_{\lambda}^{\pi}(\mu) = V^{\pi}(\mu) + \lambda \cdot H^{\pi}(\mu)$$

$\lambda > 0 \Rightarrow$ Encourages exploration (Why? Max-entropy policy is $\pi(\cdot\,|s) \sim Unif(A), \forall s \in S$)

Avoids *near-deterministic* suboptimal policies (Haarnoja et al., ’17; Ahmed et al., ‘19)

Optimization Problem

Objective

$$\max_{\theta \in R^d} \quad V_\lambda^{\pi_\theta}(\mu)$$

Challenge: Highly non-convex optimization problem

$$\theta^* \in \arg\max_{\theta \in R^d} \; V_\lambda^{\pi_\theta}(\mu)$$

$$\pi^* := \pi_{\theta^*}$$

Algorithms

Entropy-Regularized Natural Policy Gradient

Entropy-Regularized Q-NPG with Gradient Clipping

- Deterministic setting: Sample-based estimation is not considered

- Questions: What is the role of entropy-regularization in convergence of NPG?
  How do approximation errors impact convergence?

- Methodology: Lyapunov analysis based on weighted-D_{KL} (Agarwal et al., '20)

Natural Policy Gradient Algorithm

Idea: Mirror descent with Bregman divergence $D(\theta,\theta') = (\theta - \theta')^\top G^{\pi_\theta}(\mu)(\theta - \theta')$

Fisher info matrix

$$G^{\pi_\theta}(\mu) = E_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s)\, \nabla_\theta^\top \log \pi_\theta(a|s)\right]$$

NPG Update

$$\theta^+ = \arg\max_\theta \left\{\nabla_\theta V_\lambda^{\pi_{\theta^-}}(\mu)(\theta - \theta^-) - \frac{1}{2\eta} D(\theta,\theta^-)\right\}$$

$$= \theta^- + \eta \cdot [G^{\pi_{\theta^-}}(\mu)]^{-1} \cdot \nabla_\theta \mathrm{V}_\lambda^{\pi_{\theta^-}}(\mu)$$

Algorithm 1. NPG with Entropy Regularization

Initialization: $\theta_0 = 0$

for $t = 0, 1, \ldots, T-1$

$$w_t = [G^{\pi_t}(\mu)]^{-1} \cdot \nabla_\theta \mathrm{V}_\lambda^{\pi_t}(\mu)$$

$$\theta_{t+1} = \theta_t + \eta \cdot w_t$$

Compatible Function Approximation

Q-function

$$Q^{\pi}(s,a) = r(s,a) + \gamma E_{s' \sim P(\cdot|s,a)}[V_{\lambda}^{\pi}(s')]$$

Fisher info matrix

$$G^{\pi_\theta}(\mu) = E_{s \sim d_{\mu}^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)}[\nabla_\theta \log \pi_\theta(a|s)\, \nabla_\theta^{\top} \log \pi_\theta(a|s)]$$

Compatible Function Approximation (Kakade, '02)

$$L(w,\theta) = E_{s \sim d_{\mu}^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)}\left[\left(w^{\top}\nabla_\theta \log \pi_\theta(a|s) - \{Q_{\lambda}^{\pi_\theta}(s,a) - \lambda \log \pi_\theta(a|s)\}\right)^2\right]$$

$$w^* = \arg\min_{w \in R^d} L(w,\theta)$$

$$w^* = \frac{1}{1-\gamma}[G^{\pi_\theta}(\mu)]^{-1} \cdot \nabla_\theta V_{\lambda}^{\pi_\theta}(\mu)$$

Assumptions

1. Approximation Error

$$\sup_{t\geq 1} \min_{w\in R^d} L(w,\theta_t) \leq \epsilon_{approx}$$

2. Concentrability Coefficient

$$C_t = E_{s\sim d_\mu^{\pi_t}, a\sim\pi_t(\cdot|s)}\left[\left(\frac{d_\mu^{\pi^*}(s)\pi^*(a|s)}{d_\mu^{\pi_t}(s)\pi_t(a|s)}\right)^2\right] \leq C^* < \infty, \forall t$$

3. Regularity of the Basis Vectors

$$F(\mu) = E_{s\sim\mu, a\sim Unif(A)}\left[\left(\phi_{s,a} - E_{a'\sim Unif(A)}\phi_{s,a'}\right)\left(\phi_{s,a} - E_{a'\sim Unif(A)}\phi_{s,a'}\right)^\top\right]$$

$$\sigma_{min}\big(F(\mu)\big) \geq \sigma > 0$$

NPG: Convergence Results

NPG with constant step-size: $\eta = \frac{(1-\gamma)^2 \sigma^2 r_{min}}{(r_{max} + \lambda \log|A|)^2}$

Potential function:

$$\Phi(\pi) = \sum_s d_\mu^{\pi^*}(s) D_{KL}\big(\pi^*(\cdot\,|s)||\pi(\cdot\,|s)\big)$$

Linear convergence:

$$\Phi(\pi_T) \leq (1-\eta\lambda)^T \log|A| + \frac{\sqrt{C^* \epsilon_a}}{\lambda}$$

$$V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu) \leq \frac{(1-\eta\lambda)^T}{\eta(1-\gamma)} \log|A| + \frac{\sqrt{C^* \epsilon_a}}{\lambda\eta(1-\gamma)}$$

Lyapunov Drift Analysis

Lyapunov Function

$$\Phi(\pi) = \sum_s d_\mu^{\pi^*}(s) D_{KL}\big(\pi^*(\cdot\,|s)||\pi(\cdot\,|s)\big)$$

Lyapunov Drift

(Agarwal et al., '20) for unregularized Q-NPG

Entropy Regularization

$$\Phi(\pi_{t+1}) - \Phi(\pi_t) \le \underbrace{-\eta\lambda\Phi(\pi_t)}_{\text{Entropy Regularization}} - \eta(1-\gamma)\left(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\right)$$

$$-\eta E_{s\sim d_\mu^{\pi^*}, a\sim\pi^*(.|s)}\left[w_t^\top \nabla_\theta \log \pi_t(a|s) \; - Q_\lambda^{\pi_t}(s,a) + \lambda \log \pi_t(a|s)\right]$$

$$-\eta E_{s\sim d_\mu^{\pi^*}}\left[V_\lambda^{\pi_t}(s)\right] + \frac{1}{2}\eta^2 \|w_t\|_2^2$$

Lyapunov Drift Analysis

Lyapunov Function

$$\Phi(\pi) = \sum_s d_\mu^{\pi^*}(s) D_{KL}\big(\pi^*(\cdot\,|s)||\pi(\cdot\,|s)\big)$$

Lyapunov Drift

(Agarwal et al., '20) for unregularized Q-NPG

$$\begin{aligned}\Phi(\pi_{t+1}) - \Phi(\pi_t) \le & -\eta\lambda\Phi(\pi_t) - \underbrace{\eta(1-\gamma)\left(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\right)}_{\boxed{\text{(PDL)}}} \\ & -\eta E_{s\sim d_\mu^{\pi^*}, a\sim\pi^*(.|s)}\left[w_t^\top \nabla_\theta \log \pi_t(a|s) - Q_\lambda^{\pi_t}(s,a) + \lambda \log \pi_t(a|s)\right] \\ & -\eta E_{s\sim d_\mu^{\pi^*}}\left[V_\lambda^{\pi_t}(s)\right] + \frac{1}{2}\eta^2 \|w_t\|_2^2\end{aligned}$$

Lyapunov Drift Analysis

Lyapunov Function

$$\Phi(\pi) = \sum_{s} d_{\mu}^{\pi^*}(s) D_{KL}\big(\pi^*(\cdot\,|s)||\pi(\cdot\,|s)\big)$$

Lyapunov Drift

$$\Phi(\pi_{t+1}) - \Phi(\pi_t) \leq -\eta\lambda\Phi(\pi_t) - \eta(1-\gamma)\left(V_{\lambda}^{\pi^*}(\mu) - V_{\lambda}^{\pi_t}(\mu)\right)$$

$$+\eta \cdot \sqrt{C_t} \cdot \sqrt{L(w_t, \theta_t)}$$

(CFA + Assumptions 1 & 2)

$$-\eta E_{s\sim d_{\mu}^{\pi^*}}\left[V_{\lambda}^{\pi_t}(s)\right] + \frac{1}{2}\eta^2 \|w_t\|_2^2$$

Lyapunov Drift Analysis

Lyapunov Function

$$\Phi(\pi) = \sum_s d_\mu^{\pi^*}(s) D_{KL}\big(\pi^*(\cdot\,|s)||\pi(\cdot\,|s)\big)$$

Lyapunov Drift

$$\Phi(\pi_{t+1}) - \Phi(\pi_t) \le -\eta\lambda\Phi(\pi_t) - \eta(1-\gamma)\left(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\right)$$

$$+\eta \cdot \sqrt{C^*} \cdot \sqrt{\epsilon_{approx}}$$

$$-\eta E_{s\sim d_\mu^{\pi^*}}\left[V_\lambda^{\pi_t}(s)\right] + \frac{1}{2}\eta^2 \|w_t\|_2^2$$

$$\boxed{\begin{array}{c} w_t = \left(G^{\pi_t}(\mu)\right)^{-1} \nabla V_\lambda^{\pi}(\mu) \\ \sigma_{min}\left(G^{\pi_t}(\mu)\right) \ge \sigma > 0 \end{array}}$$

Lyapunov Drift Analysis

Lyapunov Function

$$\Phi(\pi) = \sum_s d_\mu^{\pi^*}(s) D_{KL}\big(\pi^*(\cdot\,|s)||\pi(\cdot\,|s)\big)$$

Lyapunov Drift

$$\begin{aligned}\Phi(\pi_{t+1}) \le\ & (1-\eta\lambda)\Phi(\pi_t) - \eta(1-\gamma)\left(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\right) \\ & +\eta \cdot \sqrt{C^*} \cdot \sqrt{\epsilon_{approx}} \\ & -\eta E_{s\sim d_\mu^{\pi^*}}\left[V_\lambda^{\pi_t}(s)\right] + \frac{1}{2}\eta^2 \|w_t\|_2^2\end{aligned}$$

Linear convergence under our assumptions

Regularity of Random Features

Question When is the regularity condition (Assumption 3) satisfied?

Random features

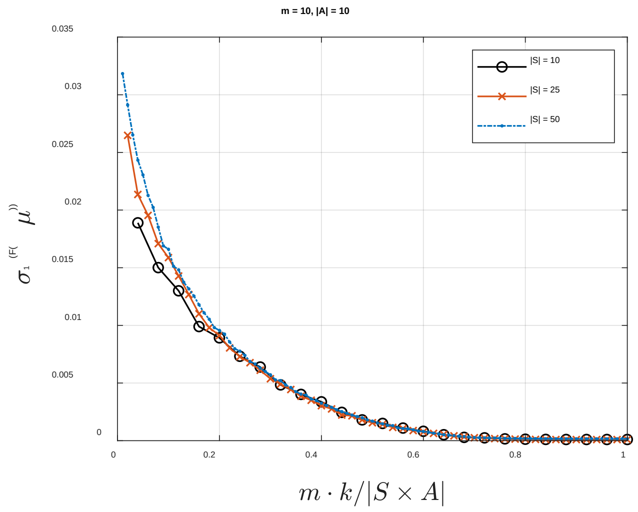
1. Gaussian ensemble $\phi_{s,a} \sim N(0, I_d)$ iid for all $(s,a) \in S \times A$

$$|A| = 2, x \in (0,1) \qquad \sigma_{min}(F(\mu)) \geq \frac{x^2}{4}\left(1 - x - \sqrt{\frac{\log \delta^{-1}}{2|S|}} - \sqrt{\frac{d \cdot \log(|S|+1)}{|S|}}\right) \text{ for any } w.p. \geq 1 - \delta \in (0,1)$$

2. Neural Tangent Kernel

$$\phi_{s,a} = \left[\frac{1}{\sqrt{m}} c_i \psi(s,a) \mathbf{1}\{W_i^\top \psi(s,a) \geq 0\}\right]_{i \in [m]}$$

$$c_i \sim Rademacher; \;\; W_i \sim N(0, I_k); \quad d = k \times m$$



Assumption 3 holds if $d \ll |S \times A|$

NPG: Convergence Results

NPG with constant step-size: $\eta = \frac{(1-\gamma)^2\sigma^2 r_{min}}{(r_{max} + \lambda \log|A|)^2}$

Potential function:

$$\Phi(\pi) = \sum_s d_\mu^{\pi^*}(s) D_{KL}\big(\pi^*(\cdot\,|s)||\pi(\cdot\,|s)\big)$$

Linear convergence:

$$\Phi(\pi_T) \leq (1-\eta\lambda)^T \log|A| + \frac{\sqrt{C^*\epsilon_a}}{\lambda}$$

$$V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu) \leq \frac{(1-\eta\lambda)^T}{\eta(1-\gamma)} \log|A| + \frac{\sqrt{C^*\epsilon_a}}{\lambda\eta(1-\gamma)}$$

Assumptions

1. Approximation Error

$$\sup_{t\geq 1} \min_{w\in R^d} E_{s\sim d_\mu^{\pi_{\theta_t}},\, a\sim\pi_{\theta_t}(\cdot|S)}\left[\left(w^\top \nabla_\theta \log \pi_{\theta_t}(a|s) - \left\{Q_\lambda^{\pi_{\theta_t}}(s,a) - \lambda \log \pi_{\theta_t}(a|s)\right\}\right)^2\right] \leq \epsilon_{approx}$$

2. Concentrability Coefficient

3. Regularity of the Basis Vectors

Q-NPG with Entropy Regularization

Q-NPG: variant of algorithm originally proposed in (Agarwal et al., '20)

$$w_t = \arg\min_{w:||w||_2 \le R} E_{s,a}\left[\left(w^\top \phi_{s,a} - Q_\lambda^{\pi_t}(s,a)\right)^2\right]$$

$$\theta_{t+1} = \theta_t(1 - \eta_t\lambda) + \eta_t w_t \qquad \text{with step-size } \eta_t = \frac{1}{\lambda(t+1)}$$

Algorithm 2. Q-NPG with Entropy Regularization

Initialization: $\theta_0 = 0$

for $t = 0, 1, \ldots, T-1$

$$w_t = \arg\min_{w:||w||_2 \le R} E_{s,a}\left[\left(w^\top \phi_{s,a} - Q_\lambda^{\pi_t}(s,a)\right)^2\right]$$

$$g_t = w_t - \lambda \cdot \theta_t$$

$$\theta_{t+1} = \theta_t + \eta_t \cdot g_t$$

Assumptions

1. Approximation Error

$$\sup_{\theta} \min_{w:||w||\le R} E\left[\left(w^\top \phi_{s,a} - Q_{\lambda}^{\pi_\theta}(s,a)\right)^2\right] \le \epsilon(R)$$

2. Concentrability Coefficient

$$M_t = E_{s\sim d_\mu^{\pi_t}}\left[\left(\frac{d_\mu^{\pi^*}(s)}{d_\mu^{\pi_t}(s)}\right)^2\right] \le M < \infty, \forall t$$

What Does Entropy Regularization Do?

1. Regularization

$$\sup_{t\geq 0} \|\theta_t\|_2 \leq \frac{R}{\lambda}$$

$$\sup_{t\geq 0} \|g_t\|_2 \leq 2R$$

2. Persistence of Excitation

$$\inf_{t\geq 0} \min_{s,a} \pi_t(a|s) \geq p_{min} \geq \frac{e^{-2R/\lambda}}{|A|} > 0$$

Q-NPG: Convergence Results

NPG with adaptive step-size: $\eta_t = \dfrac{1}{\lambda(t+1)}$

O(1/T) convergence:

$$\Phi(\pi_T) \le \frac{\sqrt{M\epsilon(R)}\left(1 + p_{min}^{-1}\right)}{\lambda} + \frac{2R^2}{\lambda^2} \cdot \frac{\log T}{T}$$

$$\min_{0 \le t < T} \left\{ V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu) \right\} \le \frac{\sqrt{M\epsilon(R)}\left(1 + p_{min}^{-1}\right)}{1-\gamma} + \frac{2R^2}{(1-\gamma)} \cdot \frac{\log T}{\lambda T}$$

Q-NPG: Convergence Results

O(1/T) convergence:

$$\Phi(\pi_T) \leq \frac{\sqrt{M\epsilon(R)}\left(1 + p_{min}^{-1}\right)}{\lambda} + \frac{2R^2}{\lambda^2} \cdot \frac{\log T}{T}$$

$$\min_{0 \leq t < T} \left\{ V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu) \right\} \leq \frac{\sqrt{M\epsilon(R)}\left(1 + p_{min}^{-1}\right)}{1 - \gamma} + \frac{2R^2}{(1 - \gamma)} \cdot \frac{\log T}{\lambda T}$$

- No regularity conditions
- $\lambda = \frac{1}{\log T}$ can be chosen, convergence to $\max_\theta V^{\pi_\theta}(\mu)$

- Sublinear convergence and not last iterate convergence
- Higher approximation error $\epsilon(R)$ due to “gradient clipping”

Q-NPG: Lyapunov Analysis

$$\Phi(\pi_t) - \Phi(\pi_{t-1}) \leq -\frac{t-1}{t}\Phi(\pi_{t-1}) - \eta_{t-1}(1-\gamma)\left(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_{t-1}}(\mu)\right)$$

$$+\, \eta_{t-1} \cdot \sqrt{M}\left(1 + p_{min}^{-1}\right)\sqrt{E_{s,a}\left[\left(w_{t-1}^{\top}\phi_{s,a} - Q_\lambda^{\pi_{t-1}}(s,a)\right)^2\right]} + \; 2\eta_{t-1}^2 R^2$$

Induction:

$$\Phi(\pi_T) \leq -\frac{1-\gamma}{\lambda T}\sum_{t<T}\left(V_\lambda^{\pi^*}(\mu) - V_\lambda^{\pi_t}(\mu)\right) + \frac{\sqrt{M\epsilon(R)}\left(1 + p_{min}^{-1}\right)}{\lambda} + \frac{2R^2}{\lambda^2} \cdot \frac{\log T}{T}$$

Conclusions

NPG with entropy-regularization achieves linear convergence up to CFA error

Regularity conditions (basis & concentrability) are required

Works well in function approximation regime $d \ll |S \times A|$

Last iterate convergence

Q-NPG with entropy-regularization achieves O(1/T) rate up to approximation error

Much milder conditions

Best iterate convergence

Extensions to neural network based actor-critic algorithms