Microsoft Research

NIPS Thanks Its Sponsors

amazon.com
Microsoft Research
Google
facebook
SKYTREE THE MACHINE LEARNING COMPANY
TWO SIGMA
United Technologies Research Center
D E Shaw & Co
xerox
IBM Research
YAHOO! LABS
TOYOTA
DRW TRADING GROUP
millionshort
criteo
PDT PARTNERS
Springer Machine Learning Journal
Disney Research

Actor-Critic Algorithms for Risk-Sensitive MDPs

Mohammad Ghavamzadeh

INRIA Lille – Team SequeL & Adobe Research

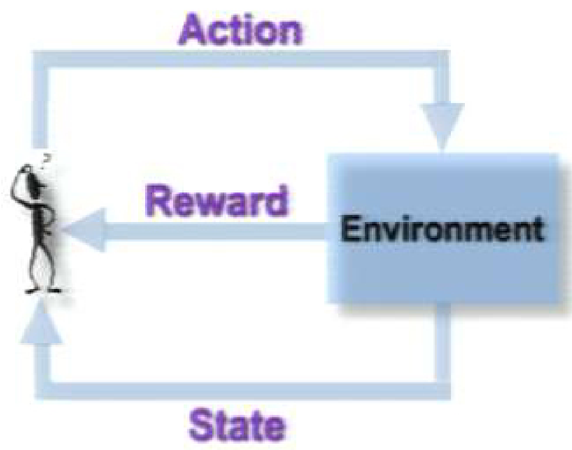
joint work with **Prashanth L.A.**

Sequential Decision-Making under Uncertainty

How Can I ... ?

- Move around in the physical world *(navigation)*
- Play and win a game
- Control the throughput of a power plant *(process control)*
- Manage a portfolio *(finance)*
- Medical diagnosis and treatment

Reinforcement Learning (RL)



- **RL:** A class of learning problems in which an agent interacts with a dynamic, stochastic, and incompletely known environment
- **Goal:** Learn an action-selection strategy, or ***policy***, to optimize some measure of its long-term performance
- **Interaction:** Modeled as a MDP

Markov Decision Process

MDP

- An MDP $\mathcal{M}$ is a tuple $\langle \mathcal{X}, \mathcal{A}, R, P, P_0 \rangle$.
- $\mathcal{X}$: set of states
- $\mathcal{A}$: set of actions
- $R(x, a)$: reward random variable, $\quad r(x, a) = \mathbb{E}[R(x, a)]$
- $P(\cdot|x, a)$: transition probability distribution
- $P_0(\cdot)$: initial state distribution

- **Stationary Policy:** a distribution over actions, conditioned on the current state $\mu(\cdot|x)$

Discounted Reward MDPs

For a given policy μ

Return

$$D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu$$

Risk-Neutral Objective

$$\mu^* = \arg\max_{\mu} \sum_{x \in \mathcal{X}} P_0(x) V^{\mu}(x)$$

where $V^{\mu}(x) = \mathbb{E}\left[D^{\mu}(x)\right]$.

Discounted Reward MDPs

For a given policy μ

Return

$$D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x, \; \mu$$

Risk-Neutral Objective

$$\mu^* = \arg\max_{\mu} \sum_{x \in \mathcal{X}} P_0(x) V^{\mu}(x)$$

where $V^{\mu}(x) = \mathbb{E}\left[D^{\mu}(x)\right]$.

Discounted Reward MDPs

For a given policy μ

Return

$$D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^{t} R(x_t, a_t) \mid x_0 = x,\ \mu$$

Risk-Neutral Objective *(for simplicity)*

$$\mu^{*} = \arg\max_{\mu} V^{\mu}(x^{0})$$

x^0 is the initial state, i.e., $P_0(x) = \delta(x - x^0)$.

Inria

Adobe

Average Reward MDPs

For a given policy μ

Average Reward

$$\rho(\mu) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] = \sum_{x,a} \pi^{\mu}(x,a)\, r(x,a)$$

$\pi^{\mu}(x,a)$: stationary dist. of state-action pair (x,a) under policy μ.

Risk-Neutral Objective

$$\mu^* = \arg\max_{\mu} \rho(\mu)$$

Average Reward MDPs

For a given policy μ

Average Reward

$$\rho(\mu) \;=\; \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] \;=\; \sum_{x,a} \pi^{\mu}(x,a)\, r(x,a)$$

$\pi^{\mu}(x,a)$: stationary dist. of state-action pair (x,a) under policy μ.

Risk-Neutral Objective

$$\mu^* = \arg\max_{\mu} \rho(\mu)$$

Average Reward MDPs

For a given policy μ

Average Reward

$$\rho(\mu) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] = \sum_{x,a} \pi^\mu(x,a)\, r(x,a)$$

$\pi^\mu(x,a)$: stationary dist. of state-action pair (x,a) under policy μ.

Risk-Neutral Objective

$$\mu^* = \arg\max_\mu \rho(\mu)$$

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^{t} R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

- a criterion that penalizes the ***variability*** induced by a given policy
- minimize some measure of ***risk*** as well as maximizing a usual optimization criterion

Risk-Sensitive Sequential Decision-Making

$$D^{\mu}(x) = \overbrace{\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return}\quad \textit{random variable}}$$

- a criterion that penalizes the ***variability*** induced by a given policy
- minimize some measure of ***risk*** as well as maximizing a usual optimization criterion

Risk-Sensitive Sequential Decision-Making

$$D^\mu(x) = \overbrace{\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

- a criterion that penalizes the ***variability*** induced by a given policy
- minimize some measure of ***risk*** as well as maximizing a usual optimization criterion

Inria

Adobe

Risk-Sensitive Sequential Decision-Making

Objective: to optimize a risk-sensitive criterion such as

- expected exponential utility *(Howard & Matheson 1972)*
- variance-related measures *(Sobel 1982; Filar et al. 1989)*
- percentile performance *(Filar et al. 1995)*

Open Question ???

construct conceptually meaningful and computationally tractable criteria

mainly negative results

(e.g., Sobel 1982; Filar et al., 1989; Mannor & Tsitsiklis, 2011)

Inria

Adobe

Risk-Sensitive Sequential Decision-Making

Objective: to optimize a risk-sensitive criterion such as

- expected exponential utility *(Howard & Matheson 1972)*
- variance-related measures *(Sobel 1982; Filar et al. 1989)*
- percentile performance *(Filar et al. 1995)*

Open Question ???

construct conceptually meaningful and computationally tractable criteria

mainly negative results

(e.g., Sobel 1982; Filar et al., 1989; Mannor & Tsitsiklis, 2011)

Risk-Sensitive Sequential Decision-Making

Objective: to optimize a risk-sensitive criterion such as

- expected exponential utility *(Howard & Matheson 1972)*
- variance-related measures *(Sobel 1982; Filar et al. 1989)*
- percentile performance *(Filar et al. 1995)*

Open Question ???

construct conceptually meaningful and computationally tractable criteria

mainly negative results

(e.g., Sobel 1982; Filar et al., 1989; Mannor & Tsitsiklis, 2011)

Inria

Adobe

Risk-Sensitive Sequential Decision-Making

long history in operations research

- most work has been in the context of MDPs *(model is known)*
- much less work in reinforcement learning (RL) framework

Risk-Sensitive RL

- expected exponential utility *(Borkar 2001, 2002)*
- several variance-related measures *(Tamar et al., 2012)*
  - policy gradient for the stochastic shortest path problem

Our Contributions

For ***discounted*** and ***average*** reward MDPs, we

1. define a measure of ***variability*** for a policy
   - a set of *(variance-related)* ***risk-sensitive criteria***
2. propose ***actor-critic algorithms*** to optimize the risk-sensitive criteria
   - define a ***class of parameterized stochastic policies***
   - ***estimate the gradient*** of the risk-sensitive criteria
   - update the policy parameters in the ascent direction
3. establish the ***asymptotic convergence*** of the algorithms
4. demonstrate the usefulness of the algorithms in a ***traffic signal control*** problem

Discounted Reward Setting

Discounted Reward MDPs

Return

$$D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x, \ \mu$$

Mean of Return *(value function)*

$$V^{\mu}(x) = \mathbb{E}\left[D^{\mu}(x)\right]$$

Variance of Return *(measure of variability)*

$$\Lambda^{\mu}(x) = \mathbb{E}\left[D^{\mu}(x)^2\right] - V^{\mu}(x)^2 = U^{\mu}(x) - V^{\mu}(x)^2$$

Discounted Reward MDPs

Risk-Sensitive Criteria

1. Maximize $V^{\mu}(x^0)$ s.t. $\Lambda^{\mu}(x^0) \le \alpha$
2. Minimize $\Lambda^{\mu}(x^0)$ s.t. $V^{\mu}(x^0) \ge \alpha$
3. Maximize the **Sharpe Ratio**: $V^{\mu}(x^0)/\sqrt{\Lambda^{\mu}(x^0)}$
4. Maximize $V^{\mu}(x^0) - \alpha\Lambda^{\mu}(x^0)$

Discounted Reward MDPs

Return

$$D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x, \ \mu$$

Mean of Return *(value function)*

$$V^{\mu}(x) = \mathbb{E}\left[D^{\mu}(x)\right]$$

Variance of Return *(measure of variability)*

$$\Lambda^{\mu}(x) = \mathbb{E}\left[D^{\mu}(x)^2\right] - V^{\mu}(x)^2 = U^{\mu}(x) - V^{\mu}(x)^2$$

Discounted Reward MDPs

Risk-Sensitive Criteria

1. Maximize $V^{\mu}(x^0)$ s.t. $\Lambda^{\mu}(x^0) \leq \alpha$
2. Minimize $\Lambda^{\mu}(x^0)$ s.t. $V^{\mu}(x^0) \geq \alpha$
3. Maximize the **Sharpe Ratio**: $V^{\mu}(x^0)/\sqrt{\Lambda^{\mu}(x^0)}$
4. Maximize $V^{\mu}(x^0) - \alpha\Lambda^{\mu}(x^0)$

Risk-Sensitive Discounted MDPs

Optimization Problem

$$\max_{\mu} V^{\mu}(x^0) \quad \text{s.t.} \quad \Lambda^{\mu}(x^0) \le \alpha$$

$$\Updownarrow$$

$$\max_{\lambda} \min_{\theta} L(\theta, \lambda) \stackrel{\triangle}{=} -V^{\theta}(x^0) + \lambda\big(\Lambda^{\theta}(x^0) - \alpha\big)$$

A class of parameterized stochastic policies

$$\big\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\big\}$$

One needs to evaluate $\nabla_{\theta} L(\theta, \lambda)$ and $\nabla_{\lambda} L(\theta, \lambda)$ to tune θ and λ

Risk-Sensitive Discounted MDPs

Optimization Problem

$$\max_{\mu} V^{\mu}(x^0) \quad \text{s.t.} \quad \Lambda^{\mu}(x^0) \le \alpha$$

$$\Updownarrow$$

$$\max_{\lambda} \min_{\theta} L(\theta, \lambda) \triangleq -V^{\theta}(x^0) + \lambda\big(\Lambda^{\theta}(x^0) - \alpha\big)$$

A class of parameterized stochastic policies

$$\big\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\big\}$$

One needs to evaluate $\nabla_{\theta} L(\theta, \lambda)$ and $\nabla_{\lambda} L(\theta, \lambda)$ to tune θ and λ

Why Estimating the Gradient is Challenging?

Computing the Gradient $\nabla_\theta L(\theta, \lambda)$

$$(1-\gamma)\nabla_\theta V^\theta(x^0) = \sum_{x,a} \pi_\gamma^\theta(x, a|x^0)\ \nabla_\theta \log \mu(a|x;\theta)\ Q^\theta(x, a)$$

$$(1-\gamma^2)\nabla_\theta U^\theta(x^0) = \sum_{x,a} \widetilde{\pi}_\gamma^\theta(x, a|x^0)\ \nabla_\theta \log \mu(a|x;\theta)\ W^\theta(x, a)$$

$$+ 2\gamma \sum_{x,a,x'} \widetilde{\pi}_\gamma^\theta(x, a|x^0)\ P(x'|x, a)\ r(x, a)\ \nabla_\theta V^\theta(x')$$

$\pi_\gamma^\theta(x, a|x^0)$ and $\widetilde{\pi}_\gamma^\theta(x, a|x^0)$ are γ and γ^2 discounted visiting state distributions of the Markov chain under policy θ

Simultaneous Perturbation (SP) Methods

Idea: Estimate the gradients $\nabla_\theta V^\theta(x^0)$ and $\nabla_\theta U^\theta(x^0)$ using two simulated trajectories of the system corresponding to policies with parameters θ and $\theta^+ = \theta + \beta\Delta$, $\beta > 0$.

Our actor-critic algorithms are based on two SP methods

1. Simultaneous Perturbation Stochastic Approximation (SPSA)
2. Smoothed Functional (SF)

Estimating the Gradient is Challenging

Computing the Gradient $\nabla_\theta L(\theta, \lambda)$

$$(1-\gamma)\nabla_\theta V^\theta(x^0) = \sum_{x,a} \pi_\gamma^\theta(x, a|x^0)\, \nabla_\theta \log \mu(a|x;\theta)\, Q^\theta(x,a)$$

$$(1-\gamma^2)\nabla_\theta U^\theta(x^0) = \sum_{x,a} \tilde{\pi}_\gamma^\theta(x, a|x^0)\, \nabla_\theta \log \mu(a|x;\theta)\, W^\theta(x,a)$$
$$+ 2\gamma \sum_{x,a,x'} \tilde{\pi}_\gamma^\theta(x, a|x^0)\, P(x'|x,a)\, r(x,a)\, \nabla_\theta V^\theta(x')$$

$\pi_\gamma^\theta(x, a|x^0)$ and $\tilde{\pi}_\gamma^\theta(x, a|x^0)$ are γ and γ^2 discounted visiting state distributions of the Markov chain under policy θ

Inria

Adobe

Simultaneous Perturbation (SP) Methods

Idea: Estimate the gradients $\nabla_\theta V^\theta(x^0)$ and $\nabla_\theta U^\theta(x^0)$ using two simulated trajectories of the system corresponding to policies with parameters θ and $\theta^+ = \theta + \beta\Delta$, $\beta > 0$.

Our actor-critic algorithms are based on two SP methods

1. Simultaneous Perturbation Stochastic Approximation (SPSA)
2. Smoothed Functional (SF)

Simultaneous Perturbation Methods

SPSA Gradient Estimate

$$\partial_{\theta^{(i)}} \widehat{V}^{\theta}(x^0) \approx \frac{\widehat{V}^{\theta+\beta\Delta}(x^0) - \widehat{V}^{\theta}(x^0)}{\beta\Delta^{(i)}}, \qquad i = 1, \ldots, \kappa_1$$

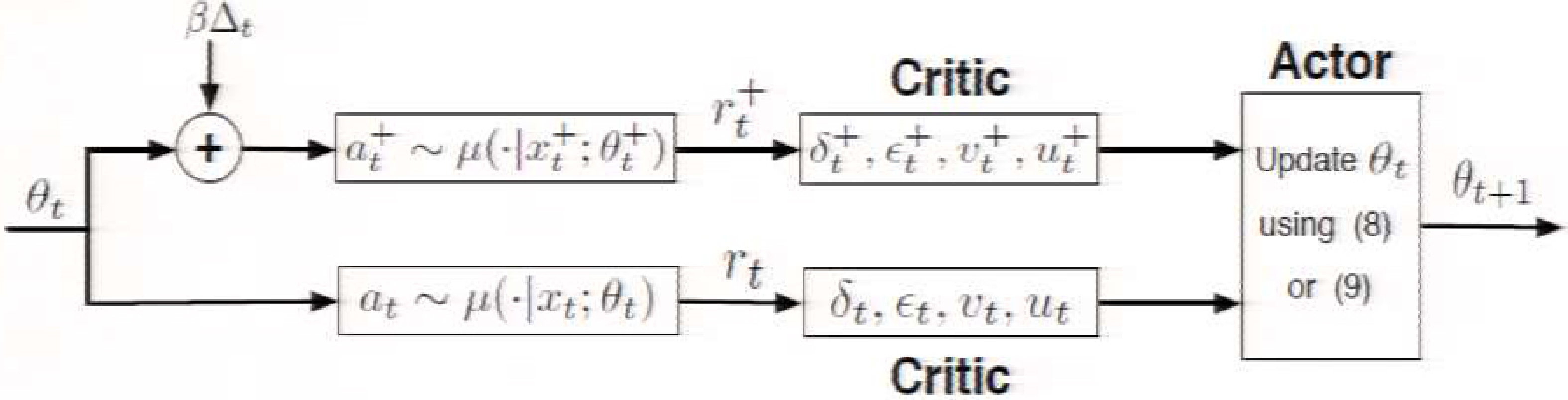
Δ is a vector of independent Rademacher random variables

SF Gradient Estimate

$$\partial_{\theta^{(i)}} \widehat{V}^{\theta}(x^0) \approx \frac{\Delta^{(i)}}{\beta}\left(\widehat{V}^{\theta+\beta\Delta}(x^0) - \widehat{V}^{\theta}(x^0)\right), \qquad i = 1, \ldots, \kappa_1$$

Δ is a vector of independent Gaussian $\mathcal{N}(0, 1)$ random variables

Risk-Sensitive Actor-Critic Algorithms



Trajectory 1 take action $a_t \sim \mu(\cdot|x_t;\theta_t)$, observe reward $r(x_t, a_t)$ and next state x_{t+1}

Trajectory 2 take action $a_t^+ \sim \mu(\cdot|x_t^+;\theta_t^+)$, observe reward $r(x_t^+, a_t^+)$ and next state x_{t+1}^+

Critic update the critic parameters v_t, v_t^+ for value and u_t, u_t^+ for square value functions in a TD-like fashion

Actor estimate $\nabla V^\theta(x^0)$ and $\nabla U^\theta(x^0)$ using SPSA or SF and update the policy parameter θ and the Lagrange multiplier λ

Risk-Sensitive Actor-Critic Algorithms

Critic Updates *(Tamar et al., 2013)*

$$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t) \qquad v^+_{t+1} = v^+_t + \zeta_3(t)\delta^+_t\phi_v(x^+_t)$$

$$u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t) \qquad u^+_{t+1} = u^+_t + \zeta_3(t)\epsilon^+_t\phi_u(x^+_t)$$

where the TD-errors $\delta_t, \delta^+_t, \epsilon_t, \epsilon^+_t$ are computed as

$$\delta_t = r(x_t, a_t) + \gamma v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$$

$$\delta^+_t = r(x^+_t, a^+_t) + \gamma {v^+_t}^\top \phi_v(x^+_{t+1}) - {v^+_t}^\top \phi_v(x^+_t)$$

$$\epsilon_t = r(x_t, a_t)^2 + 2\gamma r(x_t, a_t) v_t^\top \phi_v(x_{t+1}) + \gamma^2 u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$$

$$\epsilon^+_t = r(x^+_t, a^+_t)^2 + 2\gamma r(x^+_t, a^+_t) {v^+_t}^\top \phi_v(x^+_{t+1}) + \gamma^2 {u^+_t}^\top \phi_u(x^+_{t+1}) - {u^+_t}^\top \phi_u(x^+_t)$$

Risk-Sensitive Actor-Critic Algorithms

Actor Updates

$$\theta_{t+1}^{(i)} = \Gamma_i \left[\theta_t^{(i)} + \frac{\zeta_2(t)}{\beta \Delta_t^{(i)}} \Big(\big(1 + 2\lambda_t v_t^\top \phi_v(x^0)\big)(v_t^+ - v_t)^\top \phi_v(x^0) - \lambda_t (u_t^+ - u_t)^\top \phi_u(x^0) \Big) \right]$$

$$\lambda_{t+1} = \Gamma_\lambda \left[\lambda_t + \zeta_1(t) \Big(u_t^\top \phi_u(x^0) - \big(v_t^\top \phi_v(x^0)\big)^2 - \alpha \Big) \right]$$

step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ are chosen such that the critic, policy parameter, and Lagrange multiplier updates are on the fastest, intermediate, and slowest time-scales, respectively.

three time-scale stochastic approximation algorithm

Risk-Sensitive Actor-Critic Algorithms

Actor Updates

$$\theta_{t+1}^{(i)} = \Gamma_i\left[\theta_t^{(i)} + \frac{\zeta_2(t)}{\beta\Delta_t^{(i)}}\Big(\big(1+2\lambda_t v_t^\top \phi_v(x^0)\big)(v_t^+ - v_t)^\top \phi_v(x^0) - \lambda_t(u_t^+ - u_t)^\top \phi_u(x^0)\Big)\right]$$

$$\lambda_{t+1} = \Gamma_\lambda\left[\lambda_t + \zeta_1(t)\Big(u_t^\top \phi_u(x^0) - \big(v_t^\top \phi_v(x^0)\big)^2 - \alpha\Big)\right]$$

step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ are chosen such that the critic, policy parameter, and Lagrange multiplier updates are on the fastest, intermediate, and slowest time-scales, respectively.

three time-scale stochastic approximation algorithm

Risk-Sensitive Actor-Critic Algorithms

Critic Updates *(Tamar et al., 2013)*

$$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t) \qquad v^+_{t+1} = v^+_t + \zeta_3(t)\delta^+_t\phi_v(x^+_t)$$

$$u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t) \qquad u^+_{t+1} = u^+_t + \zeta_3(t)\epsilon^+_t\phi_u(x^+_t)$$

where the TD-errors $\delta_t, \delta^+_t, \epsilon_t, \epsilon^+_t$ are computed as

$$\delta_t = r(x_t, a_t) + \gamma v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$$

$$\delta^+_t = r(x^+_t, a^+_t) + \gamma {v^+_t}^\top \phi_v(x^+_{t+1}) - {v^+_t}^\top \phi_v(x^+_t)$$

$$\epsilon_t = r(x_t, a_t)^2 + 2\gamma r(x_t, a_t) v_t^\top \phi_v(x_{t+1}) + \gamma^2 u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$$

$$\epsilon^+_t = r(x^+_t, a^+_t)^2 + 2\gamma r(x^+_t, a^+_t) {v^+_t}^\top \phi_v(x^+_{t+1}) + \gamma^2 {u^+_t}^\top \phi_u(x^+_{t+1}) - {u^+_t}^\top \phi_u(x^+_t)$$

Risk-Sensitive Actor-Critic Algorithms

Actor Updates

$$\theta_{t+1}^{(i)} = \Gamma_i\left[\theta_t^{(i)} + \frac{\zeta_2(t)}{\beta\Delta_t^{(i)}}\Big(\big(1+2\lambda_t v_t^\top \phi_v(x^0)\big)(v_t^+ - v_t)^\top \phi_v(x^0) - \lambda_t(u_t^+ - u_t)^\top \phi_u(x^0)\Big)\right]$$

$$\lambda_{t+1} = \Gamma_\lambda\left[\lambda_t + \zeta_1(t)\Big(u_t^\top \phi_u(x^0) - \big(v_t^\top \phi_v(x^0)\big)^2 - \alpha\Big)\right]$$

step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ are chosen such that the critic, policy parameter, and Lagrange multiplier updates are on the fastest, intermediate, and slowest time-scales, respectively.

three time-scale stochastic approximation algorithm

Risk-Sensitive Actor-Critic Algorithms

Actor Updates

$$\theta^{(i)}_{t+1} = \Gamma_i\Big[\theta^{(i)}_t + \frac{\zeta_2(t)}{\beta\Delta^{(i)}_t}\Big(\big(1+2\lambda_t v_t^\top\phi_v(x^0)\big)(v_t^+ - v_t)^\top\phi_v(x^0) - \lambda_t(u_t^+ - u_t)^\top\phi_u(x^0)\Big)\Big]$$

$$\lambda_{t+1} = \Gamma_\lambda\Big[\lambda_t + \zeta_1(t)\Big(u_t^\top\phi_u(x^0) - \big(v_t^\top\phi_v(x^0)\big)^2 - \alpha\Big)\Big]$$

step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ are chosen such that the critic, policy parameter, and Lagrange multiplier updates are on the fastest, intermediate, and slowest time-scales, respectively.

three time-scale stochastic approximation algorithm

Risk-Sensitive Actor-Critic Algorithms

Actor Updates

$$\theta_{t+1}^{(i)} = \Gamma_i\Big[\theta_t^{(i)} + \frac{\zeta_2(t)}{\beta\Delta_t^{(i)}}\Big(\big(1+2\lambda_t v_t^\top \phi_v(x^0)\big)(v_t^+ - v_t)^\top \phi_v(x^0) - \lambda_t(u_t^+ - u_t)^\top \phi_u(x^0)\Big)\Big]$$

$$\lambda_{t+1} = \Gamma_\lambda\Big[\lambda_t + \zeta_1(t)\Big(u_t^\top \phi_u(x^0) - \big(v_t^\top \phi_v(x^0)\big)^2 - \alpha\Big)\Big]$$

step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ are chosen such that the critic, policy parameter, and Lagrange multiplier updates are on the fastest, intermediate, and slowest time-scales, respectively.

three time-scale stochastic approximation algorithm

Average Reward Setting

Average Reward MDPs

Average Reward

$$\rho(\mu) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] = \sum_{x,a} \pi^\mu(x,a)\, r(x,a)$$

Long-Run Variance ***(measure of variability)***

$$\Lambda(\mu) = \sum_{x,a} \pi^\mu(x,a)\left[r(x,a) - \rho(\mu)\right]^2 = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \left(R_t - \rho(\mu)\right)^2 \mid \mu\right]$$

The frequency of visiting state-action pairs, $\pi^\mu(x,a)$***, determines the variability in the average reward.***

Average Reward MDPs

Average Reward

$$\rho(\mu) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] = \sum_{x,a} \pi^{\mu}(x,a)\, r(x,a)$$

Long-Run Variance *(measure of variability)*

$$\Lambda(\mu) = \sum_{x,a} \pi^{\mu}(x,a)\left[r(x,a) - \rho(\mu)\right]^2 = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \left(R_t - \rho(\mu)\right)^2 \mid \mu\right]$$

$$= \eta(\mu) - \rho(\mu)^2, \qquad \textbf{where} \qquad \eta(\mu) = \sum_{x,a} \pi^{\mu}(x,a)\, r(x,a)^2$$

Risk-Sensitive Average Reward MDPs

Optimization Problem

$$\max_{\mu} \rho(\mu) \quad \text{s.t.} \quad \Lambda(\mu) \leq \alpha$$

$$\Updownarrow$$

$$\max_{\lambda} \min_{\theta} L(\theta, \lambda) \stackrel{\triangle}{=} -\rho(\theta) + \lambda(\Lambda(\theta) - \alpha)$$

One needs to evaluate $\nabla_\theta L(\theta, \lambda)$ and $\nabla_\lambda L(\theta, \lambda)$ to tune θ and λ

Computing the Gradients

Computing the Gradient $\nabla_\theta L(\theta, \lambda)$

$$\nabla \rho(\theta) = \sum_{x,a} \pi(x, a; \theta) \nabla \log \mu(a|x; \theta) Q(x, a; \theta)$$

$$\nabla \eta(\theta) = \sum_{x,a} \pi(x, a; \theta) \nabla \log \mu(a|x; \theta) W(x, a; \theta)$$

U^μ and W^μ are the differential value and action-value functions associated with the square reward, satisfying the following Poisson equations:

$$\eta(\mu) + U^\mu(x) = \sum_{a} \mu(a|x) \left[r(x, a)^2 + \sum_{x'} P(x'|x, a) U^\mu(x') \right]$$

$$\eta(\mu) + W^\mu(x, a) = r(x, a)^2 + \sum_{x'} P(x'|x, a) U^\mu(x')$$

Risk-Sensitive Actor-Critic Algorithm

Input: policy $\mu(\cdot|\cdot;\theta)$ and value function feature vectors $\phi_v(\cdot)$ and $\phi_u(\cdot)$
Initialization: policy parameters $\theta = \theta_0$; value function weight vectors $v = v_0$ and $u = u_0$; initial state $x_0 \sim P_0(x)$
for $t = 0, 1, 2, \ldots$ **do**

Draw action $a_t \sim \mu(\cdot|x_t;\theta_t)$ and observe reward $R(x_t, a_t)$ and next state x_{t+1}

Average Updates:
$$\widehat{\rho}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\rho}_t + \zeta_4(t)R(x_t, a_t)$$
$$\widehat{\eta}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\eta}_t + \zeta_4(t)R(x_t, a_t)^2$$

TD Errors:
$$\delta_t = R(x_t, a_t) - \widehat{\rho}_{t+1} + v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$$
$$\epsilon_t = R(x_t, a_t)^2 - \widehat{\eta}_{t+1} + u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$$

Critic Update:
$$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t), \qquad u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t)$$

Actor Update:
$$\theta_{t+1} = \Gamma\Big(\theta_t - \zeta_2(t)\big(-\delta_t\psi_t + \lambda_t(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)\big)\Big)$$
$$\lambda_{t+1} = \Gamma_\lambda\Big(\lambda_t + \zeta_1(t)(\widehat{\eta}_{t+1} - \widehat{\rho}_{t+1}^2 - \alpha)\Big)$$

end for
return policy and value function parameters θ, λ, v, u

Risk-Sensitive Actor-Critic Algorithm

Input: policy $\mu(\cdot|\cdot;\theta)$ and value function feature vectors $\phi_v(\cdot)$ and $\phi_u(\cdot)$
Initialization: policy parameters $\theta = \theta_0$; value function weight vectors $v = v_0$ and $u = u_0$; initial state $x_0 \sim P_0(x)$
for $t = 0, 1, 2, \dots$ **do**
Draw action $a_t \sim \mu(\cdot|x_t;\theta_t)$ and observe reward $R(x_t, a_t)$ and next state x_{t+1}

Average Updates:
$$\widehat{\rho}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\rho}_t + \zeta_4(t)R(x_t, a_t)$$
$$\widehat{\eta}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\eta}_t + \zeta_4(t)R(x_t, a_t)^2$$

TD Errors:
$$\delta_t = R(x_t, a_t) - \widehat{\rho}_{t+1} + v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$$
$$\epsilon_t = R(x_t, a_t)^2 - \widehat{\eta}_{t+1} + u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$$

Critic Update:
$$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t), \qquad u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t)$$

Actor Update:
$$\theta_{t+1} = \Gamma\Big(\theta_t - \zeta_2(t)\big(-\delta_t\psi_t + \lambda_t(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)\big)\Big)$$
$$\lambda_{t+1} = \Gamma_\lambda\Big(\lambda_t + \zeta_1(t)(\widehat{\eta}_{t+1} - \widehat{\rho}^2_{t+1} - \alpha)\Big)$$

end for
return policy and value function parameters θ, λ, v, u

three time-scale stochastic approximation algorithm

Traffic Signal Control MDP

Problem Description

State: vector of queue lengths and elapsed times
$x_t = (q_1, \ldots, q_N, t_1, \ldots, t_N)$

Action: feasible sign configurations

Cost:

$$h(x_t) = r_1 * \Big[\sum_{i \in I_p} r_2 * q_i(t) + \sum_{i \notin I_p} s_2 * q_i(t)\Big] + s_1 * \Big[\sum_{i \in I_p} r_2 * t_i(t) + \sum_{i \notin I_p} s_2 * t_i(t)\Big]$$

Aim: find a risk-sensitive control strategy that minimizes the total delay experienced by road users, while also reducing the variations

Experimental Results

Traffic Signal Control MDP

Problem Description

State: vector of queue lengths and elapsed times
$x_t = (q_1, \ldots, q_N, t_1, \ldots, t_N)$

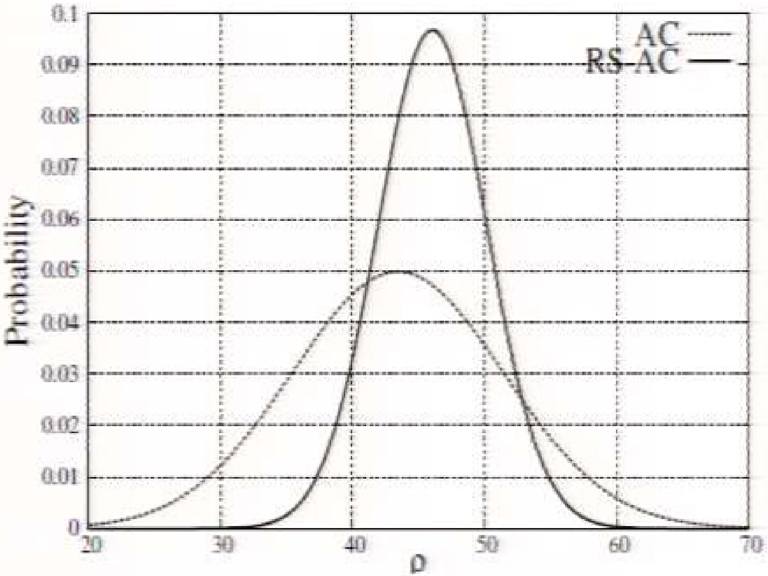
Action: feasible sign configurations

Cost:

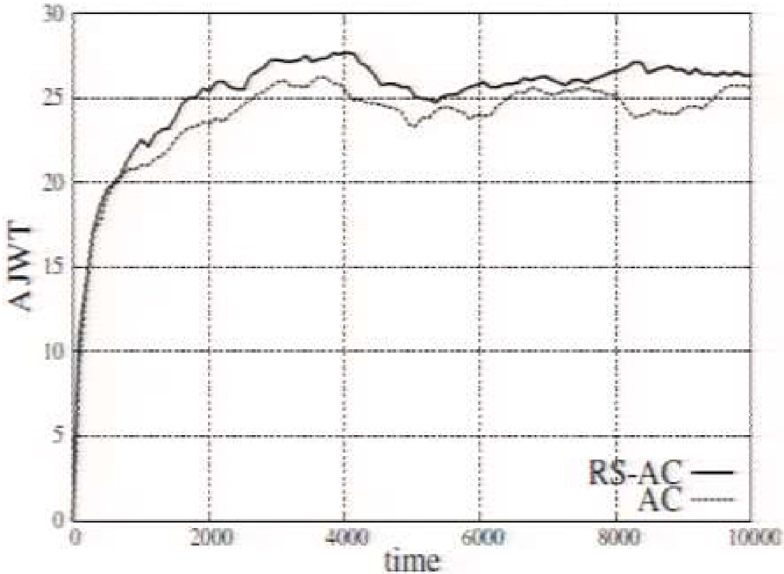
$$h(x_t) = r_1 * \Big[\sum_{i \in I_p} r_2 * q_i(t) + \sum_{i \notin I_p} s_2 * q_i(t)\Big] + s_1 * \Big[\sum_{i \in I_p} r_2 * t_i(t) + \sum_{i \notin I_p} s_2 * t_i(t)\Big]$$

Aim: find a risk-sensitive control strategy that minimizes the total delay experienced by road users, while also reducing the variations

Results - Average Reward Setting



(a) Distribution of ρ



(b) Average junction waiting time

RS-AC vs. Risk-Nutral AC: higher return with lower variance

Conclusions

For ***discounted*** and ***average*** reward MDPs, we

- define a set of *(variance-related)* ***risk-sensitive criteria***
- show how to ***estimate the gradient*** of these risk-sensitive criteria
- propose ***actor-critic algorithms*** to optimize these risk-sensitive criteria
- establish the ***asymptotic convergence*** of the algorithms
- demonstrate their usefulness in a ***traffic signal control*** problem

Future Work

For ***discounted*** and ***average*** reward MDPs,

- study other ***(more sophisticated)*** risk-sensitive criteria
- develop algorithms to ***(approximately)*** optimize these risk-sensitive criteria
- obtain finite-time bounds on the quality of solution of actor-critic ***(risk-neutral and risk-sensitive)*** algorithms

Thank You !

?

Adobe Research hiring *interns* and *researchers*

NIPS Thanks Its Sponsors

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

amazon.com

Microsoft Research

Google

facebook

SKYTREE
THE MACHINE LEARNING COMPANY

TWO SIGMA

United Technologies Research Center

D E Shaw & Co

YAHOO! LABS

IBM Research

xerox

TOYOTA

DRW TRADING GROUP

millionshort

criteo

PDT PARTNERS

Springer
Machine Learning Journal

Disney Research