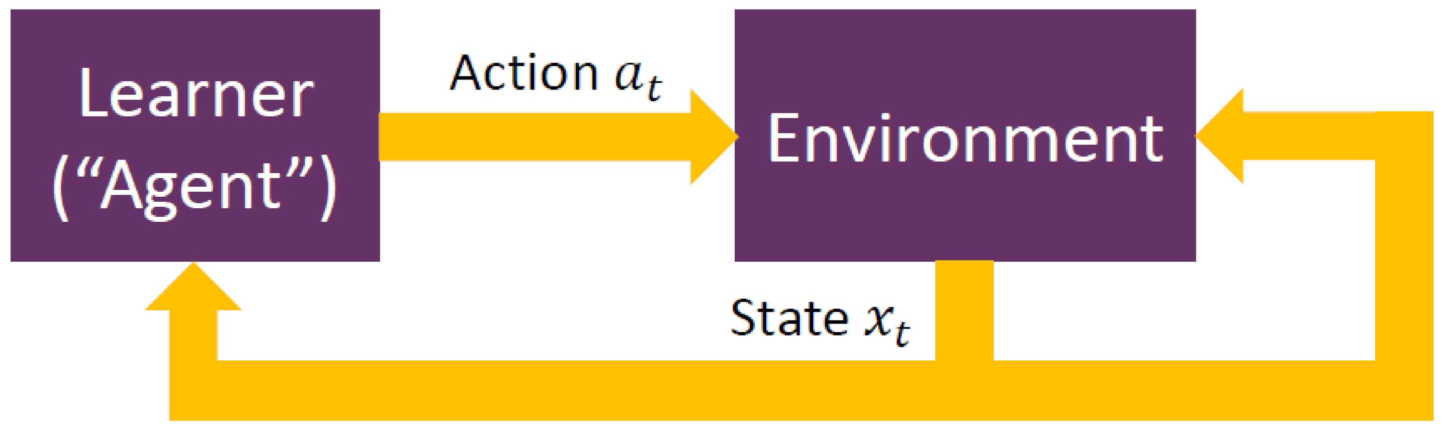
A unified view of entropy-regularized Markov decision processes

Gergely Neu Anders Jonsson Vicenç Gómez

Universitat Pompeu Fabra, Barcelona, Spain

Markov decision processes



Repeat for $t = 1, 2, \ldots$:

- LEARNER
  - observes state x_t and plays action a_t
  - obtains reward $r(x_t, a_t)$,
- ENVIRONMENT generates next state

$$x_{t+1} \sim P(\cdot|x_t, a_t).$$

GOAL: maximize long-term rewards!

- Average-reward criterion:

$$\liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r(x_t, a_t)\right] \to \max.$$

- Basic fact: enough to consider stationary policies

$$\pi(a|x) = \mathbb{P}[a_t = a | x_t = x].$$

- (Mild) Assumption: every π induces stationary distribution μ_π:

$$\mu_\pi(x, a) = \lim_{t\to\infty} \mathbb{P}[x_t = x, a_t = a].$$

- Every feasible stationary distribution μ induces a policy:

$$\pi_\mu(a|x) = \frac{\mu(x, a)}{\sum_b \mu(x, b)}.$$

The LP formulation for average-reward MDPs

Primal LP

$$\rho^* = \max_{\mu\in\Delta} \langle \mu, r\rangle$$

$$\Delta = \left\{ \text{distribution } \mu : \sum_b \mu(y, b) = \sum_{x,a} P(y|x, a)\mu(x, a) \quad (\forall y) \right\}$$

Dual "LP" ≡ The Bellman equations

$$V^*(x) = \max_a \left(r(x, a) - \rho^* + \sum_y P(y|x, a)V^*(y) \right) \quad (\forall x, a)$$

Optimal policy:

$$\pi(a|x) = \mathbb{I}\left\{ a = \arg\max_b \left(r(x, b) - \rho^* + \sum_y P(y|x, b)V^*(y) \right) \right\}$$

Regularized Markov decision processes

Primal convex program

$$\rho^*_\eta = \max_{\mu\in\Delta} \left(\langle \mu, r\rangle - \frac{1}{\eta}\sum_{x,a} \mu(x, a) \log \pi_\mu(a|x) \right)$$

Dual "convex program" ≡ Regularized Bellman equations

$$V^*_\eta(x) = \frac{1}{\eta} \log \sum_a \exp\left(\eta \left(r(x, a) - \rho^*_\eta + \sum_y P(y|x, a)V^*_\eta(y) \right) \right)$$

Optimal regularized policy:

$$\pi(a|x) \propto e^{\eta\left(r(x,a)+\sum_y P(y|x,a)V^*_\eta(y)\right)}$$

Theorem

The two convex programs are connected by Lagrangian duality.

Lemma: The conditional entropy of $(A|X) \sim \mu$

$$R(\mu) = \sum_{x,a} \mu(x, a) \log \pi_\mu(a|x)$$

is convex in μ and the associated Bregman divergence is

$$D(\mu\|\mu') = \sum_{x,a} \mu(x, a) \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)} \geq 0.$$

Algorithmic framework for regularized RL

Mirror descent

$$\mu_{t+1} = \arg\max_{\mu\in\Delta} \left(\langle \mu, r\rangle - \frac{1}{\eta} D(\mu\|\mu_t) \right)$$

- TRPO = Mirror Descent with

$$D_{\text{TRPO}}(\mu\|\mu_{\text{old}}) = \sum_{x,a} \nu_{\text{old}}(x)\pi_\mu(a|x) \log \frac{\pi_\mu(a|x)}{\pi_{\text{old}}(a|x)}.$$

- NEW RESULT: TRPO converges to the optimal policy!
- Other methods: Dynamic Policy Programming, Ψ-learning,...

Dual Averaging

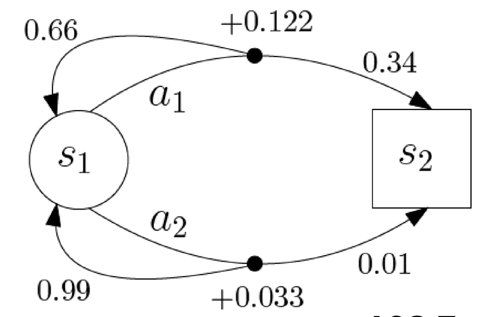
$$\mu_{t+1} = \arg\max_{\mu\in\Delta} \left(\langle \mu, r\rangle - \frac{1}{\eta_t} R(\mu) \right)$$

- A3C = Dual averaging with

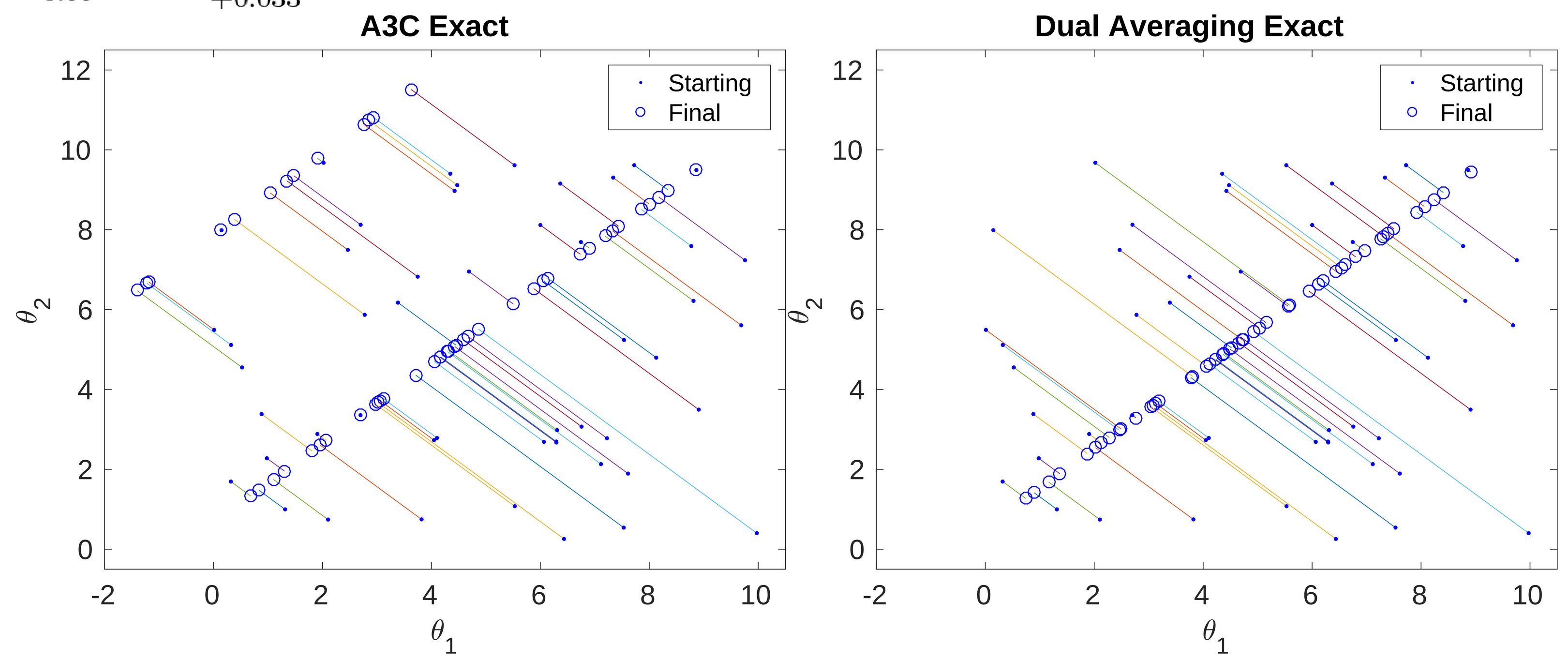
$$R_{\text{A3C}}(\mu) = \sum_{x,a} \nu_{\text{old}}(x)\pi_\mu(a|x) \log \pi_\mu(a|x).$$

- DIVERGENCE ALERT!! A3C optimizes a non-stationary objective with no underlying mirror space!!!
- Other methods: "Energy-based RL", "MellowMax RL", G-learning, "path-consistency learning",...

Experiment: the convergence of A3C



$$\pi(a_1|s_1) = \frac{\exp(\theta_1)}{\exp(\theta_1)+\exp(\theta_2)}$$



A3C has multiple stationary points
(≡ fixed points of softmax value iteration)

Patching A3C: Gradient descent on the objective regularized with

$$R(\mu) = \sum_{x,a} \nu_\mu(x)\pi_\mu(a|x) \log \frac{\pi_\mu(a|x)}{\pi_{\text{old}}(a|x)}.$$

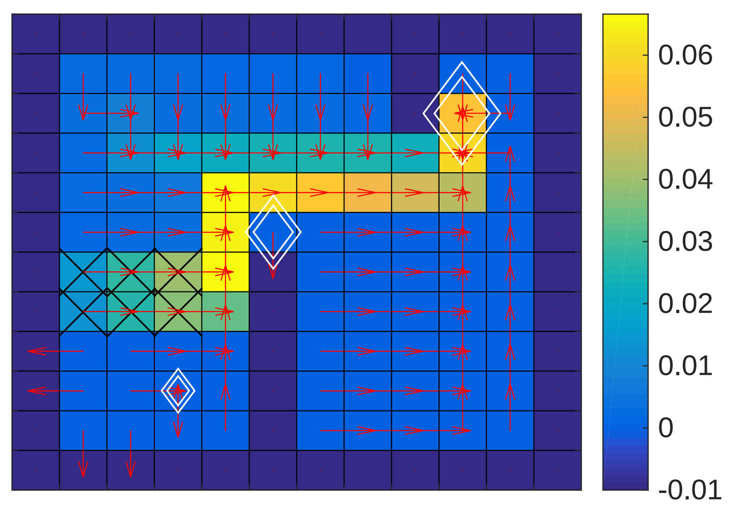
Regularized Policy Gradient Theorem

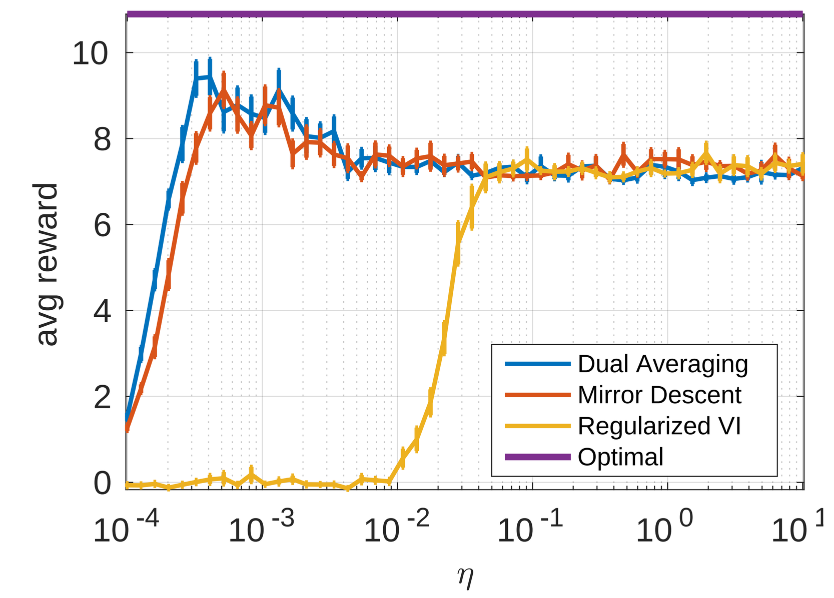
$$\nabla_\theta \left(\langle \mu_\theta, r\rangle - \tfrac{1}{\eta}R(\mu_\theta) \right) = \mathbb{E}_{(x,a)\sim\mu_\theta}\left[\nabla_\theta \log \pi_\theta(a|x) A^\pi_\eta(x, a) \right],$$

where A^π_η is the regularized advantage function satisfying

$$A^\pi_\eta(x, a) = r(x, a) - \tfrac{1}{\eta}\log \pi(a|x) + \sum_y P(y|x, a)V^\pi_\eta(y) - V^\pi_\eta(x)$$

Experiment: model-based RL





"Regularization curve":

- η too large: convergence to suboptimal goal ↔ overfitting
- η too small: policy too close to uniform ↔ underfitting

Dual Averaging perspective seems essential!

- DA theory suggests $\eta_t = t \cdot \eta_0$
- Regularized Value Iteration with constant η is bad