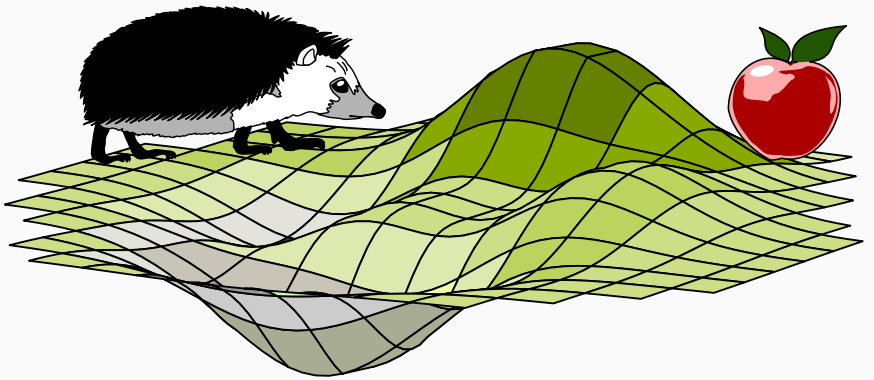
REINFORCEMENT LEARNING

CONVERGENCE II

Pavel Osinenko



Convergence II

Consider the following ∞-horizon optimal control problem:

$$\max_{K} J(x_0 \mid K) = \mathbb{E}\left[\int_0^{\infty} e^{-\gamma t} \rho(X_t, K(X_t))\, dt \mid X_0 = x_0 \right]$$

$$\text{s.t. } dX_t = f(X_t, U_t)\, dt + G(X_t, U_t)\, dB_t,$$

where $\{B_t\}_t$ is a d-dimensional Brownian motion.

Recall the HJB:

$$\max_{u \in U} \left\{ -\gamma V(x) + \mathcal{A}^u V(x) + \rho(x, u) \right\} = 0, \quad \forall x,$$

where the generator reads:

$$\mathcal{A}^u V(x) = \nabla V(x)^T f(x, u) + \frac{1}{2} tr\left(\sigma^T(x, u) \nabla^2 V(x) \sigma(x, u) \right)$$

Convergence II

The corresponding Hamiltonian reads:

$$\mathcal{H}(x, u \mid h) := \nabla h(x)^T f(x,u) + \frac{1}{2} tr\left(\sigma^T(x,u) \nabla^2 h(x) \sigma(x,u)\right) + \rho(x,u) - \gamma h(x)$$

for a generic function h of x.

If κ^* is the optimal policy, then the HJB can be rewritten as:

$$\mathcal{H}(x, \kappa^*(x) \mid V) = 0 \quad , \forall x$$

Our objective is to learn the value function under some behavior policy that generates $\{U_t\}_t$ rendering the closed loop stable in a suitable sense, but also being explorating sufficiently

Convergence II

We will use the Hamiltonian for this sake. First, let us introduce a critic in the following form:

$$\hat{V}_\theta(x) := \theta^T \varphi(x),$$

where θ is the weight vector and φ is the activation function.

Then, the value function function can be represented as

$$V(x) = \theta^{*T} \varphi(x) + \delta(x),$$

wher θ^* is the best weight vector and δ is the approximation error

Convergence II

On a compact, δ would be bounded, but we cannot expect the behavior policy to keep X_t bounded a.s., rather, it would ensure merely stochastic stability. Namely, in the following, we assume $\{U_t\}_t$ is stabilizing in 4th mean, i.e.,

$$\exists R > 0 \ \forall t \quad \mathbb{E}\left[\|X_t\|^4\right] \leq R^2.$$

This is a strong assumption, but it will be shown later why it is necessary

Convergence II

Let's work out the gradient and Hessian of $\hat{V}_\theta$.

First,

$$\nabla_x \hat{V}_\theta(x) = \nabla \varphi^T(x)\, \theta$$

then,

$$\nabla_x^2 \hat{V}_\theta(x) = \sum_{j=1}^{N_c} \theta_j \nabla^2 \varphi_j(x),$$

where N_c is the number of the critic's features and $\varphi_j(x)$ is the jth feature.

Problem: derive the above expressions

Convergence II

With this at hand, we can write down the Hamiltonian depending on the critic as:

$$\mathcal{H}(x,u \,/\, \hat{V}_\theta) = \theta^T \nabla \varphi(x) f(x,u) + \tfrac{1}{2}\theta^T \eta(x,u) + \rho(x,u) - \gamma \theta^T \varphi(x),$$

where

$$\eta(x,u) := \begin{bmatrix} \mathrm{tr}\left(\sigma^T(x,u)\, \nabla^2 \varphi_1(x)\, \sigma(x,u)\right) \\ \mathrm{tr}\left(\sigma^T(x,u)\, \nabla^2 \varphi_2(x)\, \sigma(x,u)\right) \\ \vdots \\ \mathrm{tr}\left(\sigma^T(x,u)\, \nabla^2 \varphi_{N_c}(x)\, \sigma(x,u)\right) \end{bmatrix}$$

Convergence II

Now, let us introduce a Hamiltonian approximation error for a generic action as:

$$\delta_{\mathcal{H}}(x,u) := \mathcal{H}(x,u \mid V) - \mathcal{H}(x,u \mid \hat{V}_{\theta^*})$$

$$= \nabla V(x)^T f(x,u) + \frac{1}{2} tr\left(\sigma^T(x,u) \nabla^2 V(x) \sigma(x,u)\right) + \rho(x,u) - \gamma V(x)$$

$$\theta^{*T} \nabla \varphi(x) f(x,u) - \frac{1}{2} \theta^{*T} \eta(x,u) - \rho(x,u) + \gamma \theta^{*T} \varphi(x)$$

$$= \nabla \delta(x)^T f(x,u) + \frac{1}{2} tr\left(\sigma^T(x,u) \nabla^2 \delta(x) \sigma(x,u)\right) - \gamma \delta(x)$$

Problem: verify this

Convergence II

Next, introduce a Hamiltonian TD as:

$$
\begin{aligned}
e_{\mathcal{H}}(\theta|x,u) &:= \mathcal{H}(x,u|\hat{V}_\theta) - \mathcal{H}(x,\kappa^*(x)|V) \\
&= \mathcal{H}(x,u|\hat{V}_\theta) \\
&= \theta^\intercal \nabla\varphi(x) f(x,u) + \tfrac{1}{2}\theta^\intercal \eta(x,u) + \rho(x,u) - \gamma\theta^\intercal\varphi(x)
\end{aligned}
$$

Then,

$$
\nabla_\theta e_{\mathcal{H}}(\theta|x,u) = \nabla\varphi(x) f(x,u) + \tfrac{1}{2}\eta(x,u) - \gamma\varphi(x)
$$

Convergence II

We proceed to constructing an experience replay.
First, introduce a data vector as follows:

$$w(x,u) := \nabla\varphi(x) f(x,u) + \frac{1}{2}\eta(x,u) - \gamma\varphi(x) .$$

Then,

$$e_{\mathcal{H}}(\theta|x,u) = \theta^T w(x,u) + p(x,u) .$$

Now, let's define the weight error:

$$\tilde{\theta} := \theta - \theta^* .$$

With this at hand, observe:

$$\begin{aligned} e_{\mathcal{H}}(\theta|x,u) &= \tilde{\theta}^T w(x,u) + \mathcal{H}(x,u|\hat{V}_{\theta^*}) \\ &= \tilde{\theta}^T w(x,u) + \mathcal{H}(x,u|V) - \delta_{\mathcal{H}}(x,u) \end{aligned}$$

Convergence II

Don't be tricked here:

$$e_H(\theta|x,u) = \tilde{\theta}^T \omega(x,u) + \mathcal{H}(x,u|V) - \delta_H(x,u)$$

$\mathcal{H}(x,u|V)$ is not zero!

This is because we are plugging in a generic action.

Convergence II

Back to our business with the experience replay. A common suggestion is to normalize by a factor

$$(w^T w + 1)^2$$

so that our critic loss now reads:

$$J^c(\theta \mid \{X_{t_k}, U_{t_k}\}_k^M) = \frac{1}{2} \sum_{k=1}^{M} \frac{e_H^2(\theta \mid X_{t_k}, U_{t_k})}{(W_{t_k}^T W_{t_k} + 1)^2},$$

where $\{X_{t_k}, U_{t_k}\}_k^M$ is the experience replay of size M such that $X_{t_M}, U_{t_M} = X_t, U_t$, the current state and action, respectively, and $W_t := w(X_t, U_t)$. There are options on how to update the experience replay. For instance, this can be done at equally distributed Δt-steps

Convergence II

This critic loss suggests to update the weights via stochastic gradient descent (effectively, on mini-batches):

$$\frac{\partial}{\partial t}\Theta_t := -\alpha \nabla_\theta J^c(\theta \mid \{X_{t_k}, U_{t_k}\}_k^M)$$

$$= -\alpha \sum_{k=1}^{M} \frac{e_{\mathcal{H}}(\theta \mid X_{t_k}, U_{t_k}) \nabla_\theta e_{\mathcal{H}}(\theta \mid X_{t_k}, U_{t_k})}{(W_{t_k}^T W_{t_k} + 1)^2},$$

where α is the learning rate.

Now, since Θ_t does not explicitly depend on X_t, the application of the Ito rule trivially gives

$$d\Theta_t = \frac{\partial}{\partial t}\Theta_t \, dt$$

Convergence II

Now, recalling that

$$\nabla_\theta e_{\mathcal{H}}(\theta/x,u) = \nabla\varphi(x) f(x,u) + \frac{1}{2}\eta(x,u) - \gamma\varphi(x)$$

and

$$\omega(x,u) = \nabla\varphi(x) f(x,u) + \frac{1}{2}\eta(x,u) - \gamma\varphi(x)$$

we get:

$$d\Theta_t = -d\sum_{k=1}^{M} \frac{e_{\mathcal{H}}(\theta/X_{t_k}, U_{t_k}) W_{t_k}}{(W_{t_k}^T W_{t_k} + 1)^2} .$$

Since $\theta^* = \text{const}$, we may also write: $d\tilde{\Theta}_t = d\Theta_t$

Convergence II

Next, recalling that

$$e_{\mathcal{H}}(\theta|x,u) = \tilde{\theta}^T \omega(x,u) + \mathcal{H}(x,u|V) - \delta_{\mathcal{H}}(x,u)$$

write:

$$d\tilde{\theta}_t = -\alpha\left(\sum_{k=1}^{M} \frac{W_{t_k} W_{t_k}^T}{(W_{t_k}^T W_{t_k} + 1)^2}\tilde{\theta}_t + \sum_{k=1}^{M} \frac{W_{t_k}\left(\mathcal{H}(X_{t_k}, U_{t_k}|V) - \delta_{\mathcal{H}}(X_{t_k}, U_{t_k})\right)}{(W_{t_k}^T W_{t_k} + 1)^2}\right)dt$$

Convergence II

Define

$$\mathcal{E}_t := \sum_{k=1}^{M} \frac{W_{t_k} W_{t_k}^T}{\left(W_{t_k} W_{t_k}^T + 1\right)^2} .$$

Now, we need to figure out what kind of convergence of the weights we can achieve and under what conditions.

Let's consider the evolution of the norm of the weights. For this sake, we multiply the update rule by $\tilde{\Theta}_t^T$ on the left which yields:

$$d\|\tilde{\Theta}_t\|^2 = -\alpha\left(\|\tilde{\Theta}_t\|_{\mathcal{E}_t}^2 + \sum_{k=1}^{M} \frac{\tilde{\Theta}_t W_{t_k}\left(\mathcal{H}(X_{t_k}, U_{t_k} | V) - \delta_{\mathcal{H}}(X_{t_k}, U_{t_k})\right)}{\left(W_{t_k}^T W_{t_k} + 1\right)^2}\right) dt,$$

where $\|\tilde{\Theta}_t\|_{\mathcal{E}_t}^2 = \tilde{\Theta}_t^T \mathcal{E}_t \tilde{\Theta}_t$, the squared weighted norm

Convergence II

Let's inspect the SDE

$$d\|\tilde{\Theta}_t\|^2 = -\alpha\left(\|\tilde{\Theta}_t\|^2_{\mathcal{E}_t} + \sum_{k=1}^{M} \frac{\tilde{\Theta}_t W_{t_k}\left(\mathcal{H}(X_{t_k}, U_{t_k}|V) - \delta_{\mathcal{H}}(X_{t_k}, U_{t_k})\right)}{\left(W_{t_k}^T W_{t_k} + 1\right)^2}\right) dt .$$

First thing we require is that the behavior policy generating $\{U_t\}_t$ must ensure global existence and uniqueness of a strong solution to the environment SDE (with the initial condition $X_0 = x_0$ a.s.). This is not only a requirement on the behavior policy, but also on the environment itself. Literature on SDE commonly states various growth conditions on the drift and diffusion (see, e.g., Mao, X. (2007). Stochastic differential equations and applications)

Convergence II

The next thing we will need is persistence of excitation:

$$\mathcal{E}_t \succcurlyeq \varepsilon I_{N_c} \quad \text{a.s.},$$

where $\varepsilon = \text{const} > 0$ and I_{N_c} is $N_c \times N_c$ identity matrix.

The reason is that without such a condition, no convergence can be ensured

Convergence II

The stated conditions on global solution and persistence of excitation allow us to utilize some stochastic calculus (stopping times, Fatou's lemma, dominated convergence) to derive:

$$\mathbb{E}\left[\|\tilde{\Theta}_t\|^2\right] \leq e^{-\alpha\varepsilon t}\|\tilde{\theta}(0)\|^2 + \frac{\alpha}{2}e^{-\alpha t}\mathbb{E}\left[\int_0^t e^{\alpha\tau}\|\tilde{\Theta}_\tau\|\sum_{k=1}^{M}\left(H(X_{\tau_k}, U_{\tau_k}|V) - \delta_H(X_{\tau_k}, U_{\tau_k})\right)d\tau\right]$$

observing that $\frac{\|w\|}{(w^Tw+1)^2} \leq \frac{1}{2}$ always

Convergence II

Now, we inspect

$$\mathbb{E}\left[\|\tilde{\Theta}_t\|^2\right] \leq e^{-\alpha\varepsilon t}\|\tilde{\theta}(0)\|^2 + \frac{\alpha}{2}e^{-\alpha t}\mathbb{E}\left[\int_0^t e^{\alpha\tau}\|\tilde{\Theta}_\tau\|\sum_{k=1}^{M}\left(\mathcal{H}(X_{\tau_k}, U_{\tau_k}|V) - \delta_{\mathcal{H}}(X_{\tau_k}, U_{\tau_k})\right)d\tau\right].$$

An immediate observation is the random variable

$$\sum_{k=1}^{M}\left(\mathcal{H}(X_{\tau_k}, U_{\tau_k}|V) - \delta_{\mathcal{H}}(X_{\tau_k}, U_{\tau_k})\right)$$

has to be mean-square bounded for otherwise no progress can be made at this point

Convergence II

Recall:

$$\mathcal{H}(x,u/V) = \nabla V(x)^T f(x,u) + \frac{1}{2} tr\left(\sigma^T(x,u) \nabla^2 V(x) \sigma(x,u)\right) + \rho(x,u) - \gamma V(x),$$

$$\delta_{\mathcal{H}}(x,u) = \nabla \delta(x)^T f(x,u) + \frac{1}{2} tr\left(\sigma^T(x,u) \nabla^2 \delta(x) \sigma(x,u)\right) - \gamma \delta(x).$$

Let's call the behavior policy $\mu : \mathcal{X} \longrightarrow \mathcal{U}$ and assume u is generated by it.

We assume:

- f, σ of linear growth (as it is usually done in SDE analyses) under μ and $\mu(0) = 0$, $f(0,0) = 0$, $\sigma(0,0) = 0$
- δ of quadratic growth with $\delta(0) = 0$
- ρ, V of quadratic growth

Convergence II

These assumptions allow us to state:

$$\exists C > 0 \quad |H(x,u|V) - \delta_H(x,u)| \leq C\|x\|^2.$$

Since the behavior policy was assumed 4th-mean stabilizing, we have:

$$\forall t \quad \mathbb{E}\left[|H(X_t,U_t|V) - \delta_H(X_t,U_t)|^2\right] \leq C^2R^2$$

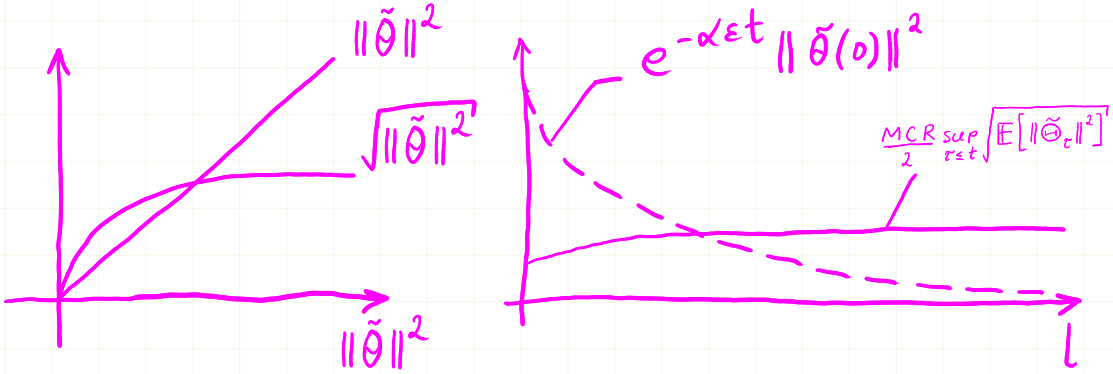
Convergence II

Therefore, using the Fubini's lemma and the Cauchy-Schwartz inequality yields:

$$\mathbb{E}\left[\|\tilde{\Theta}_t\|^2\right] \leq e^{-\alpha\varepsilon t}\|\tilde{\theta}(0)\|^2 + \frac{MCR}{2}\sup_{\tau \leq t}\sqrt{\mathbb{E}\left[\|\tilde{\Theta}_\tau\|^2\right]} \; .$$

From this, we finally can draw the conclusion that the weight error decays in mean-square norm until the square-root term starts to dominate $e^{-\alpha\varepsilon t}\|\tilde{\theta}(0)\|^2$

Illustration:



Convergence II

If the environment were deterministic or, alternatively, was driven by a.s. bounded noise, the assumptions would be much weaker.

We could proceed by the same token as before, but utilize the fact that X_t stay in a compact a.s. whence, exploiting continuity, various uniform bounds could be derived.

That would greatly simplify the analysis.

We state this as an exercise for the reader:

Problem: analyze convergence of the critic weights when the environment is $\dot{x} = f(x, u)$

Convergence II

Notice that the learning was performed in an off-policy manner. You could interpret the corresponding phase as exploration. After the weights converged (in mean-square), optimal actions could be computed by Hamiltonian maximization

$$u := \arg\max_{u} \mathcal{H}(x, u \mid \hat{V}_\theta)$$

at a state x

Convergence II

For instance, if:

$$f(x,u) \mapsto f(x) + g(x)u \quad \text{(control-affine drift)}$$
$$G(x,u) \mapsto G(x) + D(x)u\zeta^T(x) \quad \text{(control-affine diffusion)}$$
$$\rho(x,u) = q(x) + u^T R u \quad \text{(reward quadratic in action)}$$

then, the stationarity condition on the Hamiltonian

$$\nabla_u \mathcal{H}(x,u \mid \hat{V}_\theta) = 0$$

yields (omitting the x argument for brevity)

$$\hat{u}^* = -\left(\frac{1}{2}\zeta^T\zeta R^{-1} D^T \nabla_x^2 \hat{V}_\theta D + I_m\right)^{-1} R^{-1}\left(g^T \nabla_x \hat{V}_\theta + D^T \nabla_x^2 \hat{V}_\theta G\zeta\right).$$

Problem: derive this!

Convergence II

We may introduce an actor neural network

$$K_\theta(x) := \vartheta^T \psi(x)$$

by analogy with the critic and learn its weights via stochastic gradient descent on an actor loss formed from the actor error

$$e_a(\vartheta | x, u) := \vartheta^T \psi(x) - \hat{u}^* ,$$

where $\hat{u}^*$ is obtained as on the previous slide.

Notice $\nabla_u H(x, K^*(x) | V) = 0$ whereas $\hat{u}^*$ mimics $K^*(x)$ hoping that the critic $\hat{V}_\theta$ be close to the value function

Convergence II

Convergence analysis of the actor weights towards their best values may be conducted in the same manner as with the critic and is an exercise.

A big question is:

> can we train the critic and the actor online and simultaneously as per described update rules?

The answer is: in general, unfortunately, no.
We can't be sure in advance that thereby trained actor establishes necessary stability properties of the environment. But this is a large separate subject that goes beyond this lecture

Convergence II

Further reading:

- Vamvoudakis, K. G., Miranda, M. F., & Hespanha, J. P. (2015). Asymptotically stable adaptive-optimal control algorithm with saturating actuators and relaxed persistence of excitation.
  IEEE transactions on neural networks and learning systems, 27(11), 2386-2398.

  Deterministic control-affine CT environment
  Made some serious mistakes: assumed $H(x,u \mid V) = 0$ incorrectly; tacitly used an assumption $g(x) \neq 0$ in a robustifying term without stating it

- Sokolov, Y., Kozma, R., Werbos, L. D., & Werbos, P. J. (2015). Complete stability analysis of a heuristic approximate dynamic programming control design.
  Automatica, 59, 9-18.

  Deterministic DT environment
  Ignored effects of simultaneous actor-critic learning on the environment

- H. Zhang, L. Cui, X. Zhang and Y. Luo, "Data-Driven Robust Approximate Optimal Tracking Control for Unknown General Nonlinear Systems Using Adaptive Dynamic Programming Method,"
  in IEEE Transactions on Neural Networks, vol. 22, no. 12, pp. 2226-2236, Dec. 2011

  Deterministic CT environment, considered model learning along
  Made a flawed assumption on the model error