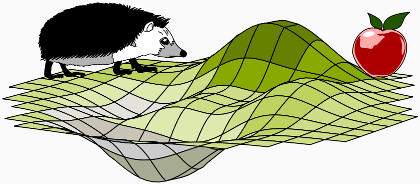
REINFORCEMENT LEARNING

CONVERGENCE I

Pavel Osinenko



Convergence I

First of all, convergence of policy gradient algorithms (in the sense of the policy PDF parameters) is the usual convergence of gradient algorithms* (in general, only to a local extremizer).
This is enabled by the policy gradient theorem which shows that the respective expression of the policy gradient (say, REINFORCE) is indeed a proper gradient of the objective.
Be aware that we are talking about PDF parameters for a definite initial state or a distribution thereof. In the following, we study convergence of tabular methods

* Provided the value is estimated sufficiently well

Convergence I

Let's consider the following ∞-horizon undiscounted* optimal control problem:

$$\min_{\mu} \mathbb{E}\left[\sum_{k=0}^{\infty} \rho(X_k, U_k) \right], \qquad \rho(x,u) \geq 0, \ \forall x, u$$

$$X_+ \sim P_X(x_+ | x, u)$$
$$U \sim P_U^{\mu(x)}(u|x)$$

This is cost minimization instead of reward maximization, but only chosen so for convenience.

We will use the notation X_+^u for the next state (as random variable) under the action u.

* We already tackled the discounted case in the DP Lecture (recall it exploits contraction)

Convergence I

We assume $\mu : \mathbb{X} \to \Pi$ where Π is the space of policy PDF parameters. Thus, we consider the parameters state-dependent. Compare this to the policy algorithms we discussed earlier.

Here, we will study convergence of the VI algorithm. Let's recall it: (iterate for i)

$$\mu_i(x) := \arg\min_{\vartheta} \left\{ \mathbb{E}_{U \sim \mathbb{P}_U^{\vartheta}} \left[\rho(x, U) + V_i(X_+^U) \right] \right\},$$

$$V_{i+1}(x) := \mathbb{E}_{U \sim \mathbb{P}_U^{\mu_i(x)}} \left[\rho(x, U) + V_i\left(X_+^U\right) \right],$$

where μ_i, V_i are the successive policy (PDF parameter) and value approximations

Convergence I

In the following, we will abuse the terminology somewhat and refer to functions $\mathbb{X} \to \Pi$ simply as policies.

So, let $\eta : \mathbb{X} \to \Pi$ be an arbitrary policy.

Arbitrary, but not quite ...

In general, the cost-to-go

$$J(x|\eta) = \mathbb{E}_{U_k \sim P_U^{\eta(X_k)}}\left[\sum_{k=0}^{\infty} \rho(X_k, U_k) \,\middle|\, X_0 = x \right]$$

may fail to converge (i.e., to be finite). We call those policies, for which it is finite, admissible. So let η be arbitrary, but admissible

Convergence I

Now, consider some generic iterations

$$\Lambda_{i+1}(x) := \mathbb{E}_{\upsilon \sim P_{\upsilon}^{\eta(x)}} \left[\rho(x, \upsilon) + \Lambda_i(X_+^{\upsilon}) \right]$$

starting from $\Lambda_0 \equiv 0$.

Let's compare them to our VI iteration by iteration.

So, at the first one we have:

$$V_1(x) = \mathbb{E}_{\upsilon \sim P_{\upsilon}^{\mu_0(x)}} \left[\rho(x, \upsilon) \right]$$

where

$$\mu_0(x) = \arg\min_{\vartheta} \mathbb{E}_{\upsilon \sim P_{\upsilon}^{\vartheta}} \left[\rho(x, \upsilon) \right]$$

Convergence I

At the same time,

$$\Lambda_1(x) = \mathbb{E}_{U \sim P_U^{\eta(x)}}\left[\rho(x,U)\right].$$

By optimality of μ_0, $V_1(x) \le \Lambda_1(x)$ and this holds for all x.

Now, at the second iteration,

$$\begin{aligned} V_2(x) &= \mathbb{E}_{U \sim P_U^{\mu_1(x)}}\left[\rho(x,U) + V_1(X_+^U)\right] \\ &= \mathbb{E}_{U \sim P_U^{\mu_1(x)}}\left[\rho(x,U) + \mathbb{E}_{U \sim P_U^{\mu_0(X_+^U)}}\left[\rho(X_+^U, U) \mid X = x\right]\right] \\ &= \mathbb{E}_{\substack{U \sim P_U^{\mu_1(x)} \\ U_+ \sim P_U^{\mu_0(X_+^U)}}}\left[\rho(x,U) + \rho(X_+^U, U_+)\right] \end{aligned}$$

Convergence I

The policy reads:

$$\mu_1(x) = \arg\min_{\vartheta} \mathbb{E}_{\substack{U \sim \mathbb{P}_U^{\vartheta} \\ U_+ \sim \mathbb{P}_U^{\mu_0(X_+^U)}}} \left[\rho(x, U) + \rho(X_+^U, U_+) \right].$$

Observe that, again, by optimality of μ_1,

$$\mathbb{E}_{\substack{U \sim \mathbb{P}_U^{\mu_1(x)} \\ U_+ \sim \mathbb{P}_U^{\mu_0(X_+^U)}}} \left[\rho(x, U) + \rho(X_+^U, U_+) \right] \leq$$

$$\mathbb{E}_{\substack{U \sim \mathbb{P}_U^{\eta(x)} \\ U_+ \sim \mathbb{P}_U^{\mu_0(X_+^U)}}} \left[\rho(x, U) + \rho(X_+^U, U_+) \right]$$

Due to optimality of μ_0,

$$\mathbb{E}_{U\sim P_U^{\mu_0(x)}}\left[\rho(x,U)\right] \leq \mathbb{E}_{U\sim P_U^{\eta(x)}}\left[\rho(x,U)\right], \ \forall x,$$

which implies, in particular,

$$\mathbb{E}_{U_+\sim P_U^{\mu_0(X_+^U)}}\left[\rho(X_+^U, U_+)\right] \leq \mathbb{E}_{U\sim P_U^{\eta(X_+^U)}}\left[\rho(X_+^U, U_+)\right] \text{ a.s.}$$

no matter what U was.

Therefore,

$$\mathbb{E}_{\substack{U\sim P_U^{\eta(x)}\\ U_+\sim P_U^{\mu_0(X_+^U)}}}\left[\rho(x,U) + \rho(X_+^U, U_+)\right] \leq$$

$$\mathbb{E}_{\substack{U\sim P_U^{\eta(x)}\\ U_+\sim P_U^{\eta(X_+^U)}}}\left[\rho(x,U) + \rho(X_+^U, U_+)\right]$$

Convergence I

Combining all facts, conclude:

$$\mathbb{E}_{\substack{U \sim \mathbb{P}_U^{\mu_1(x)} \\ U_+ \sim \mathbb{P}_U^{\mu_0(X_+^U)}}} \left[\rho(x, U) + \rho(X_+^U, U_+) \right] \leq$$

$$\mathbb{E}_{\substack{U \sim \mathbb{P}_U^{\eta(x)} \\ U_+ \sim \mathbb{P}_U^{\eta(X_+^U)}}} \left[\rho(x, U) + \rho(X_+^U, U_+) \right]$$

which means $V_2(x) \leq \Lambda_2(x)$, $\forall x$

Convergence I

Now, suppose $V_i(x) \leq \Lambda_i(x), \forall x$.

Consider:

$$\mu_i(x) = \underset{\upsilon}{\operatorname{argmin}} \left\{ \mathbb{E}_{U \sim P_U^{\upsilon}} \left[\rho(x, U) + V_i(X_+^U) \right] \right\},$$

$$V_{i+1}(x) = \mathbb{E}_{U \sim P_U^{\mu_i(x)}} \left[\rho(x, U) + V_i\left(X_+^U\right) \right],$$

$$\Lambda_{i+1}(x) = \mathbb{E}_{U \sim P_U^{\eta(x)}} \left[\rho(x, U) + \Lambda_i\left(X_+^U\right) \right]$$

By optimality of μ_i,

$$\mathbb{E}_{U \sim P_U^{\mu_i(x)}} \left[\rho(x, U) + V_i\left(X_+^U\right) \right] \leq$$

$$\mathbb{E}_{U \sim P_U^{\eta(x)}} \left[\rho(x, U) + V_i\left(X_+^U\right) \right]$$

Convergence I

(cont.)

By the induction hypothesis,

$$\mathbb{E}_{v \sim P_v^{\eta(x)}}\left[\rho(x,v) + V_i(X_+^v)\right] \leq \mathbb{E}_{v \sim P_v^{\eta(x)}}\left[\rho(x,v) + \Lambda_i(X_+^v)\right].$$

Therefore, $V_{i+1}(x) \leq \Lambda_{i+1}(x),\ \forall x$. Proof by induction is complete.

Notice that η was arbitrary. Let's try the optimal policy μ^* in place of it

Convergence I

Recall the generic iteration:

$$\Lambda_{i+1}(x) := \mathbb{E}_{U \sim P_U^{\mu^*(x)}} \left[\rho(x, U) + \Lambda_i(X_+^U) \right].$$

Unwrapping the recursion gives:

$$\Lambda_i(x) = \mathbb{E}_{U_{+k} \sim P_U^{\mu^*(X_{+k})}} \left[\sum_{k=0}^{i-1} \rho(X_{+k}, U_{+k}) \middle| X = x \right]$$

which, by non-negativeness of ρ, is not greater than

$$\mathbb{E}_{U_{+k} \sim P_U^{\mu^*(X_{+k})}} \left[\sum_{k=0}^{\infty} \rho(X_{+k}, U_{+k}) \middle| X = x \right].$$

But this is just the optimum of the cost-to-go, or the value function $V(x)$

Convergence I

Thus, we may conclude that

$$\forall x, i \quad V_i(x) \leq V(x).$$

The next step is to compare V_{i+1} and V_i.
For this sake, recall:

$$V_{i+1}(x) = \mathbb{E}_{U \sim P_U^{\mu_i(x)}}\left[\rho(x, U) + V_i(X_+^U)\right]$$

$$V_i(x) = \mathbb{E}_{U \sim P_U^{\mu_{i-1}(x)}}\left[\rho(x, U) + V_{i-1}(X_+^U)\right]$$

Convergence I

By the virtue of μ_{i-1}, namely,

$$\mu_{i-1}(x) = \arg\min_{\vartheta} \left\{ \mathbb{E}_{U \sim P_U^{\vartheta}} \left[\rho(x, U) + \hat{V}_{i-1}(X_+^U) \right] \right\}$$

leads to the fact that

$$\mathbb{E}_{U \sim P_U^{\mu_{i-1}(x)}} \left[\rho(x, U) + \hat{V}_{i-1}(X_+^U) \right] \leq \mathbb{E}_{U \sim P_U^{\mu(x)}} \left[\rho(x, U) + \hat{V}_{i-1}(X_+^U) \right]$$

for any μ, in particular, μ_i even, whence

$$\mathbb{E}_{U \sim P_U^{\mu_{i-1}(x)}} \left[\rho(x, U) + \hat{V}_{i-1}(X_+^U) \right] \leq \mathbb{E}_{U \sim P_U^{\mu_i(x)}} \left[\rho(x, U) + \hat{V}_{i-1}(X_+^U) \right]$$

Convergence I

So far, we have thus:

$$V_i(x) \leq \mathbb{E}_{U \sim P_U^{\mu_i(x)}}\left[\rho(x, U) + V_{i-1}(X_+^U)\right] .$$

Now, let's do some unwrapping:

$$V_i(x) = \mathbb{E}_{U_{+k} \sim P_U^{\mu_{i-1-k}(X_{+k})}}\left[\sum_{k=0}^{i-1} \rho(X_{+k}, U_{+k}) \mid X = x\right] ,$$

$$V_{i+1}(x) = \mathbb{E}_{U_{+k} \sim P_U^{\mu_{i-2-k}(X_{+k})}}\left[\sum_{k=0}^{i} \rho(X_{+k}, U_{+k}) \mid X = x\right]$$

(verify this!)

Convergence I

For the starting state X_t^U, these read:

$$V_i(X_t^U) = \mathbb{E}_{U_{t+k} \sim \rho_U^{\mu_{i-k}(X_{t+k})}} \left[\sum_{k=1}^{i} \rho(X_{t+k}, U_{t+k}) \,\middle|\, X_t^U \right],$$

$$V_{i-1}(X_t^U) = \mathbb{E}_{U_{t+k} \sim \rho_U^{\mu_{i-1-k}(X_{t+k})}} \left[\sum_{k=1}^{i-1} \rho(X_{t+k}, U_{t+k}) \,\middle|\, X_t^U \right].$$

Let's add the ith stage cost under the policy μ_{i-1} to the last displayed expected sum

Convergence I

(cont.)

$$\mathbb{E}_{U_{+k} \sim P_U^{\mu_{i-1-k}(X_{+k})}} \left[\sum_{k=1}^{i-1} \rho(X_{+k}, U_{+k}) \,\middle|\, X_+^U \right] \leqslant$$

$$\mathbb{E}_{\substack{U_{+k} \sim P_U^{\mu_{i-1-k}(X_{+k})} \\ U_{+i} \sim P_U^{\mu_{i-1}(X_{+i})}}} \left[\sum_{k=1}^{i-1} \rho(X_{+k}, U_{+k}) + \rho(X_{+i}, U_{+i}) \,\middle|\, X_+^U \right] =$$

$$\mathbb{E}_{U_{+k} \sim P_U^{\mu_{i-k}(X_{+k})}} \left[\sum_{k=1}^{i} \rho(X_{+k}, U_{+k}) \,\middle|\, X_+^U \right] =$$

$$V_i(X_+^U)$$

Convergence I

Therefore,

$$\begin{aligned} V_i(x) &\leq \mathbb{E}_{U \sim P_U^{\mu_i(x)}}\left[\rho(x, U) + V_{i+1}(X_+^U) \right] \\ &\leq \mathbb{E}_{U \sim P_U^{\mu_i(x)}}\left[\rho(x, U) + V_i(X_+^U) \right] \\ &= V_{i+1}(x). \end{aligned}$$

Thus, we have :

$$\forall i, x \qquad V_i(x) \leq V_{i+1}(x)$$

$$V_i(x) \leq V(x) < \infty$$

Convergence I

So, V_i is point-wise increasing and bounded above whence, by the monotone convergence theorem, it point-wise convergent.

Let's denote the limit V_∞.

By the virtue of the value function, $V(x) \leq V_\infty(x)$.

On the other hand, $\forall i, x \quad V_i(x) \leq V(x)$, whence $V_\infty(x) \leq V(x)$.

So, $V_\infty(x) = V(x), \forall x$

Convergence I

We have:

$$V_\infty(x) = \mathbb{E}_{U \sim P_U^{\mu_\infty(x)}}\left[\rho(x,U) + V_\infty(X_+^U)\right] =$$

$$V(x) = \mathbb{E}_{U \sim P_U^{\mu(x)}}\left[\rho(x,U) + V(X_+^U)\right],$$

$$\mu^*(x) = \arg\min_{\vartheta} \mathbb{E}_{U \sim P_U^{\vartheta}}\left[\rho(x,U) + V(X_+^U)\right],$$

$$\mu_\infty(x) = \arg\min_{\vartheta} \mathbb{E}_{U \sim P_U^{\vartheta}}\left[\rho(x,U) + V_\infty(X_+^U)\right].$$

Assuming $\mu^*(x)$ is unique, we conclude $\mu_\infty = \mu^*$

Discussion

We saw that the convergence result worked thanks to the policy being state-dependent. The argument collapses if it's made fixed over states. Next, the VI „algorithm" is only a real algorithm when the state space is finite. Otherwise, a compact and a finite mesh over it must be taken. This necessarily introduces approximation errors. The convergence result considered herein is not inherently robust to such errors

(cont.)

Works are known in this direction, e.g.,

Heydari, A. (2016). Theoretical and numerical analysis of approximate dynamic programming with approximation errors. Journal of Guidance, Control, and Dynamics, 39(2), 301-311.

But they assume the errors to tend to zero sufficiently fast along iterations, for otherwise their accumulation effect destroys convergence.
There are other sources of errors, such as non-exact optimization in finding μ_is.
In general, we cannot expect any kind of errors to vanish. What makes things worse is that no rate of convergence was provided above

Convergence of PI

$$V_i(x) := \mathbb{E}_{U \sim P_U^{\mu_i(x)}} \left[\rho(x, U) + V_i(X_+^U) \right], \quad \forall x$$

$$\mu_{i+1}(x) := \arg\min_{\vartheta} \left\{ \mathbb{E}_{U \sim P_U^{\vartheta}} \left[\rho(x, U) + V_i(X_+^U) \right] \right\}, \quad \forall x$$

is somewhat similar.

First of all, let's look at the value update. It resembles the HJB, but with a policy μ_i. A particular way to calculate it is to pretend it's a fixed-point equation and iterate over j as per:

$$\forall x \quad V_i^{j+1}(x) := \mathbb{E}_{U \sim P_U^{\mu_i(x)}} \left[\rho(x, U) + V_i^j(X_+^U) \right]$$

Convergence I

Let's say $V_i^0 \equiv 0$.

Then, V_i^∞ is just the cost-to-go under μ_i (show it).

For the value update thus to be meaningful, μ_i needs to be admissible.

In fact, the very first candidate μ_0 also needs to be admissible and it makes a lot of difference to VI. In other words, PI is an order of magnitude stricter on the requierements than VI

Convergence I

But, PI has some serious advantage compared to VI.

We know that

$$V_i(x) = \mathbb{E}_{U_{+k} \sim P_U^{\mu_i(X_{+k})}} \left[\sum_{k=0}^{\infty} \rho(X_{+k}, U_{+k}) \,\middle|\, X = x \right] < \infty \quad \forall i, x.$$

If ρ is positive-definite, deduce that

$$\lim_{k \to \infty} \mathbb{E}_{U_{+k} \sim P_U^{\mu_i(X_{+k})}} \left[\rho(X_{+k}, U_{+k}) \right] = 0 \quad \text{a.s.}$$

Let's assume, furthermore, that ρ is convex in $\|x\|, \|u\|$. For instance,

$$\exists c > 0 \ \forall x, u \quad \rho(x, u) \geq c(\|x\|^2 + \|u\|^2)$$

Convergence I

Then, $\rho(\mathbb{E}[\|X\|], \mathbb{E}[\|U\|]) \leq \mathbb{E}[\rho(X, U)]$ by the Jensen's inequality.

Let's abuse the notation a bit and write:

$$\mathbb{E}_{U_{+\infty} \sim P_U^{\mu_i(X_{+\infty})}}\left[\rho(X_{+\infty}, U_{+\infty})\right] = 0 \quad a.s.$$

Then, also

$$\rho\left(\mathbb{E}[\|X_{+\infty}\|], \mathbb{E}_{U_{+\infty} \sim P_U^{\mu_i(X_{+\infty})}}[\|U_{+\infty}\|]\right) = 0 \quad a.s.$$

whence $\mathbb{E}[\|X_{+\infty}\|] = 0$ a.s.

$$\mathbb{E}_{U_{+\infty} \sim P_U^{\mu_i(X_{+\infty})}}[\|U_{+\infty}\|] = 0 \quad a.s.$$

Convergence I

Therefore, to be admissible, U must be zero-mean if the current state is zero.

Furthermore, it must asymptotically stabilize the state in mean.

Evidently,

$$\mathbb{E}_{U_{+k} \sim \mathbb{P}_U^{\mu_i(X_{+k})}} \left[\rho(X_{+k}, U_{+k}) \right] < \infty, \quad \forall k,$$

whence, by the same token, the state has to be mean-stable. This is a strong conclusion: the environment under PI is (stochastically) stable, at every iteration.

Under VI, it is only guaranteed at the limit $i \to \infty$

Convergence I

Let's study the PI further. Recall:

$$V_i(x) = \mathbb{E}_{U \sim P_U^{\mu_i(x)}} \left[\rho(x, U) + V_i(X_+^U) \right], \quad \forall x$$

$$\mu_{i+1}(x) = \underset{\vartheta}{\arg\min} \left\{ \mathbb{E}_{U \sim P_U^{\vartheta}} \left[\rho(x, U) + V_i(X_+^U) \right] \right\}, \quad \forall x$$

So far, we know how the value update works. Let's study convergence iteration-to-iteration. First of all, by the virtue of the value function,

$$\forall i, x \quad V(x) \leq V_\infty(x)$$

Convergence I

Let's compare ith and i+1st iterations:

$$V_i(x) = \mathbb{E}_{U \sim P_U^{\mu_i(x)}} \left[\rho(x, U) + V_i\left(X_+^U\right) \right],$$

$$V_{i+1}(x) = \mathbb{E}_{U \sim P_U^{\mu_{i+1}(x)}} \left[\rho(x, U) + V_{i+1}\left(X_+^U\right) \right]$$

Convergence I

(cont.)

By optimality of

$$\mu_{i+1}(x) = \operatorname*{argmin}_{\upsilon} \left\{ \mathbb{E}_{U \sim \mathbb{P}_U^{\upsilon}} \left[\rho(x, U) + V_i(X_+^U) \right] \right\},$$

and the fact that each V_i is a cost-to-go, deduce

$$V_{i+1}(x) = \mathbb{E}_{U_{+k} \sim \mathbb{P}_U^{\mu_{i+1}(X_{+k})}} \left[\sum_{k=0}^{\infty} \rho(X_{+k}, U_{+k}) \,\middle|\, X = x \right] \leqslant$$

$$\mathbb{E}_{U_{+k} \sim \mathbb{P}_U^{\mu_{i}(X_{+k})}} \left[\sum_{k=0}^{\infty} \rho(X_{+k}, U_{+k}) \,\middle|\, X = x \right] =$$

$$V_i(x)$$

Convergence I

We have:

$$\forall i, x \qquad V_{i+1}(x) \leq V_i(x)$$
$$V(x) \leq V_i(x).$$

Therefore, again, by the monotone convergence theorem and optimality of V,

$$\forall x \qquad \lim_{i\to\infty} V_i(x) = V(x)$$
$$\lim_{i\to\infty} \mu_i(x) = \mu^*(x).$$