# AFU: Actor-Free critic Updates in off-policy RL for continuous control

**Nicolas Perrin-Gilbert** – Sorbonne Université, CNRS, ISIR, Paris, France nicolas.perrin-gilbert@cnrs.fr

When trying to design off-policy RL algorithms, using Q-learning as the starting point makes sense.
Among other possible approaches, off-policy Actor-Critic algorithms (DDPG, TD3, SAC, …) can be very efficient but they also have drawbacks. Arguably, the interweaving of the Critic and Actor updates has a negative impact on their off-policyness, while the clean separation between the Critic and the Actor in Q-learning is preferable.

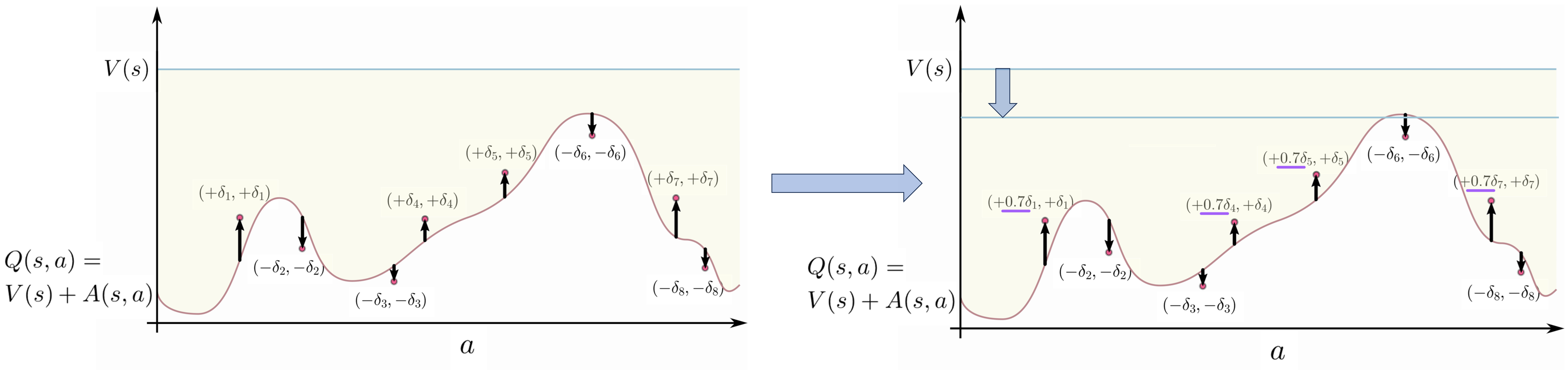
batch of transitions

$$L_Q(\psi) = \underset{(s,a,r,s')\in B}{\text{Mean}} \left[ \Big( Q_\psi(s,a) - r - \gamma \overbrace{\max_{a'} Q_\psi(s',a')}^{q^{s,a}_{target}} \Big)^2 \right]$$

state action reward new state discount factor

***Major issue:*** *to implement Q-learning, one needs to compute the maximum of the Q-function, which can be very hard if the actions belong to a high-dimensional continuous space.* This maximization is known as the **max-Q problem**.

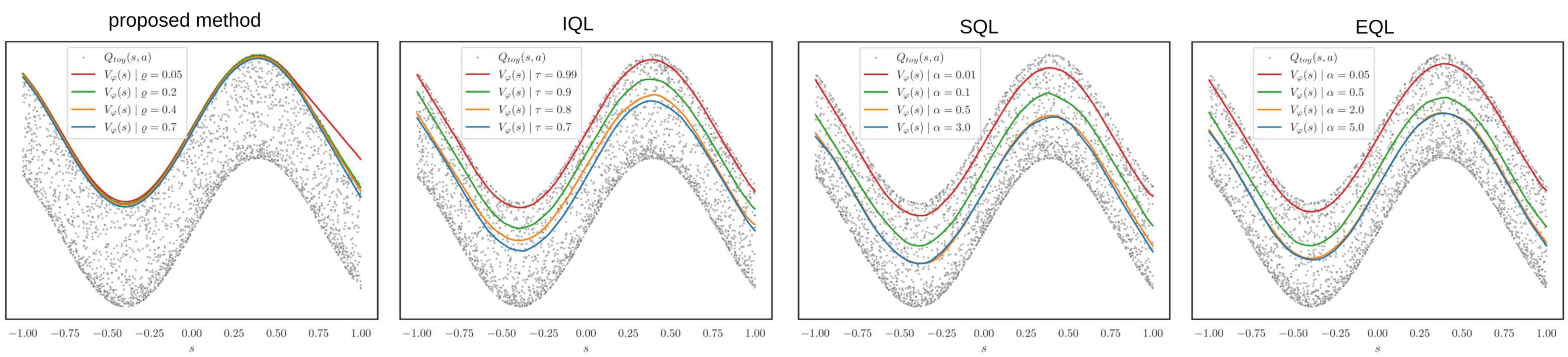
Implicit Q-learning (IQL) solves the max-Q problem by approximating the maximum with expectile regression. It leads to an efficient offline RL algorithm, **but IQL and similar algorithms (e.g. SQL, EQL) do not work well in online RL. Why?** The main reason is that in online RL, the max-Q problem is very dynamic: changes in Q affect the policy, and changes in the policy result in new data that affects Q, etc. So the approximation of the maximum has to be not only **accurate** but **fast** too, otherwise errors like overestimations of Q-values may lead to divergence.
**We propose a new approach to solve the max-Q problem that is both fast and accurate, based on a technique we call *conditional gradient rescaling*, which is a form of adaptive regularization.**




Roughly, we decompose the Q-function into $Q(s,a) = V(s) + A(s,a)$ , and force the advantage A to be negative.

Errors $\epsilon_a = (V(s) + A(s,a) - q^{s,a}_{target})^2$ contribute to similar variations δ on V and A, so we introduce asymmetry by scaling down the ***upward*** gradients on V.

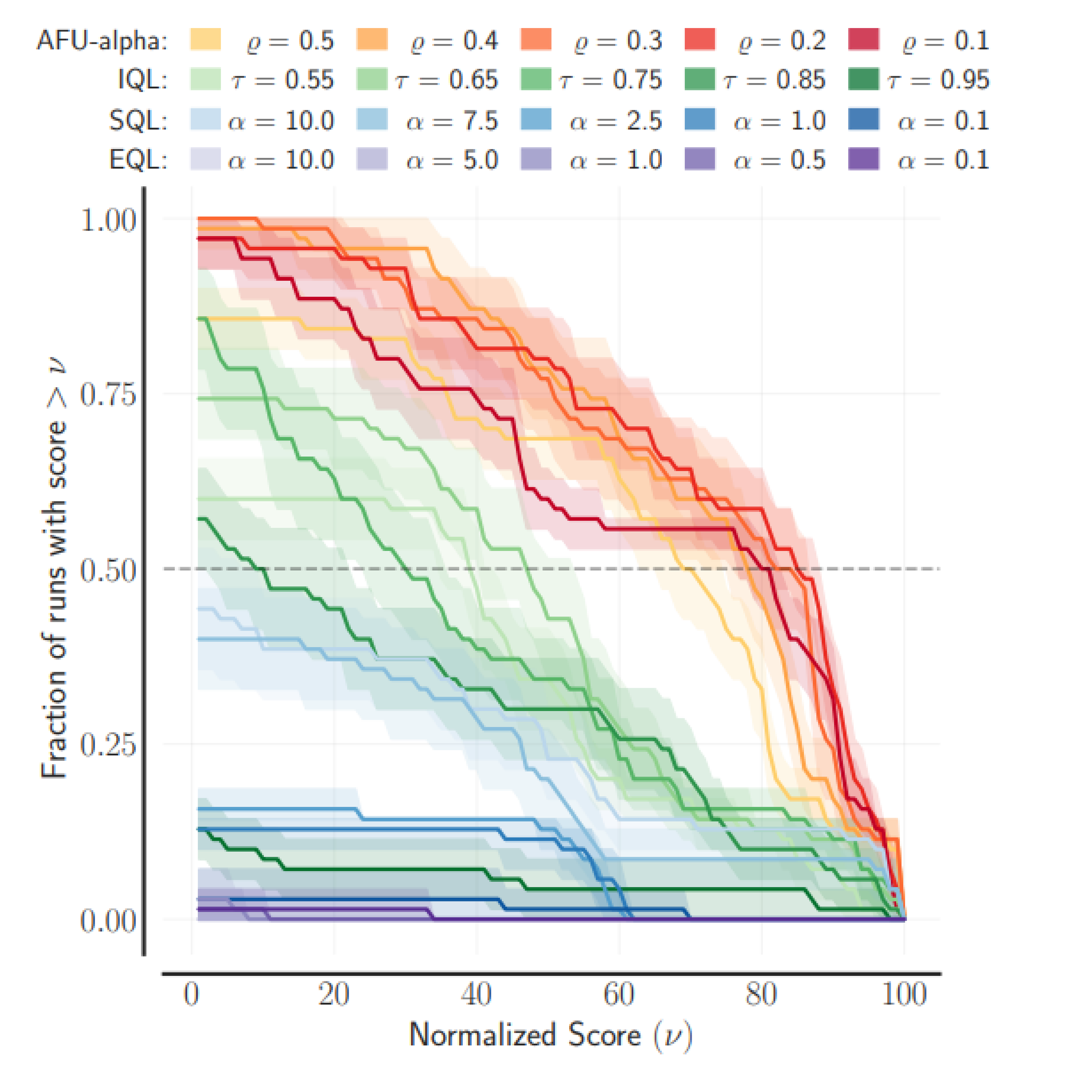
It results in an adaptive regularization of V, which quickly turns it into a precise approximation of the maximum of Q.



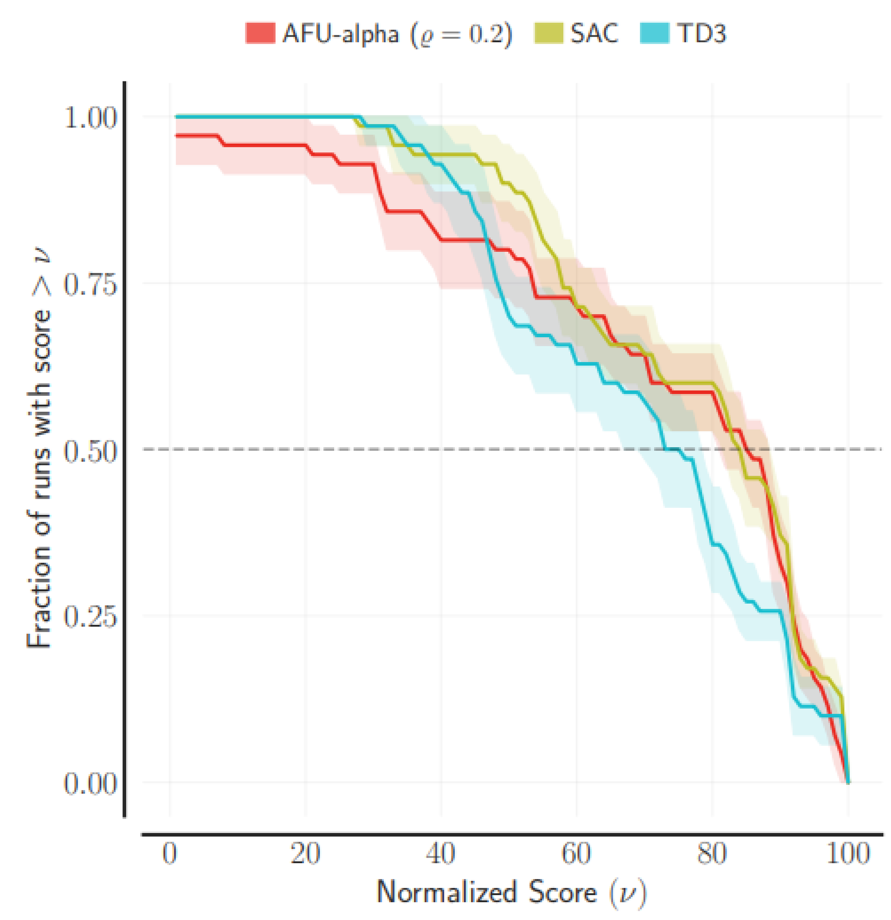

We show on a toy problem why our method is a better choice than expectile regression (IQL) or similar approaches (SQL & EQL) to solve the max-Q problem.
We use this method to design a new off-policy RL algorithm: **AFU-alpha**, which works well both for offline and online RL (the focus of the paper). On a benchmark of 7 MuJoCo tasks, we show that its efficiency is comparable to that of SAC and TD3:

We also exhibit a simple example of environment in which SAC fails by getting stuck in a local optimum, and we propose a small modification of AFU-alpha (**AFU-beta**) that does not have the same problem and easily converges towards the true optimal policy.

**To the best of our knowledge, AFU (AFU-alpha & AFU-beta) is the first model-free off-policy algorithm that is competitive with state-of-the-art Actor-Critic methods while departing from the Actor-Critic perspective.**




(a) AFU-alpha works best with $\varrho \in \{0.2, 0.3\}$. Using the IQL, SQL and EQL baselines to solve the max-Q problem results in a clear performance deterioration.




(b) AFU-alpha is competitive with SAC and TD3 on a benchmark of 7 diverse tasks, with action space dimensions ranging from 1 (InvertedDoublePendulum) to 17 (Humanoid), and observation space dimensions ranging from 11 to 376 (Humanoid).