Policy Gradient in practice

Don't become an alchemist :)

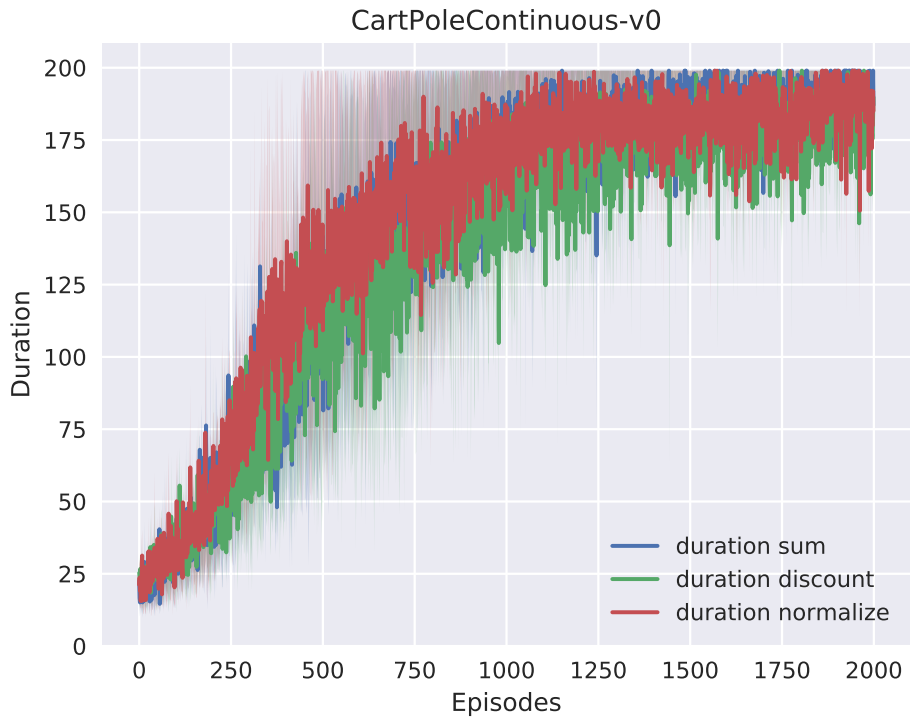
Olivier Sigaud

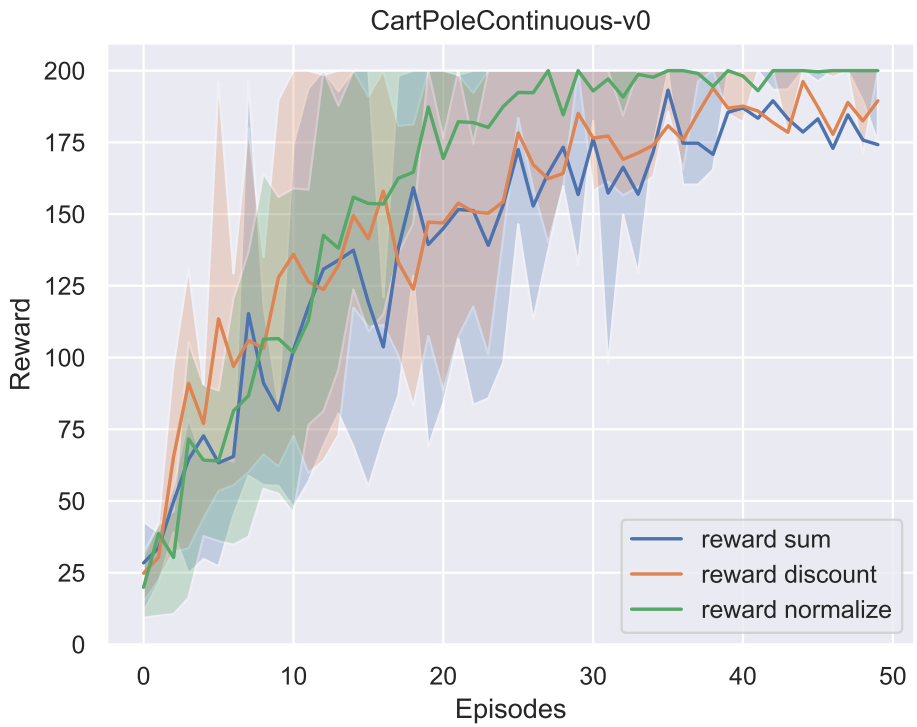
Sorbonne Université
http://people.isir.upmc.fr/sigaud

Normalization issue

- Normalize each local return with $\frac{r_t^{(i)} - \bar{r}}{std(r)} \rightarrow \sim \mathcal{N}(0, 1)$
- In CartPole and CartPoleContinuous, $r = 1$ for all steps before failure
- Thus, at all steps, $r - \bar{r} = 0$ and std $= 0$
- Cannot be applied
- By discounting the reward, we avoid this
- In the sum case, longer trajectories are more rewarded
- Globally, poorly informative gradient

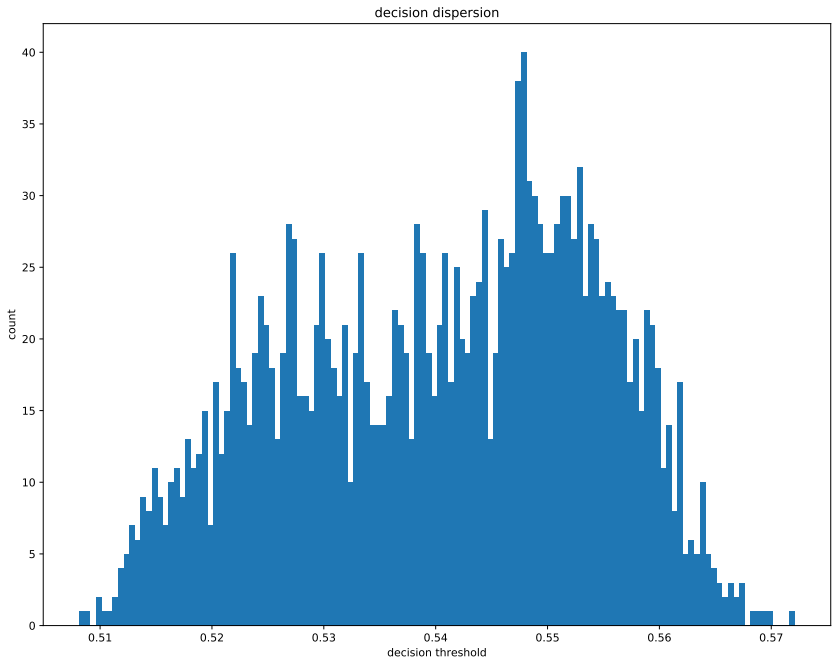
Deterministic vs stochastic evaluation

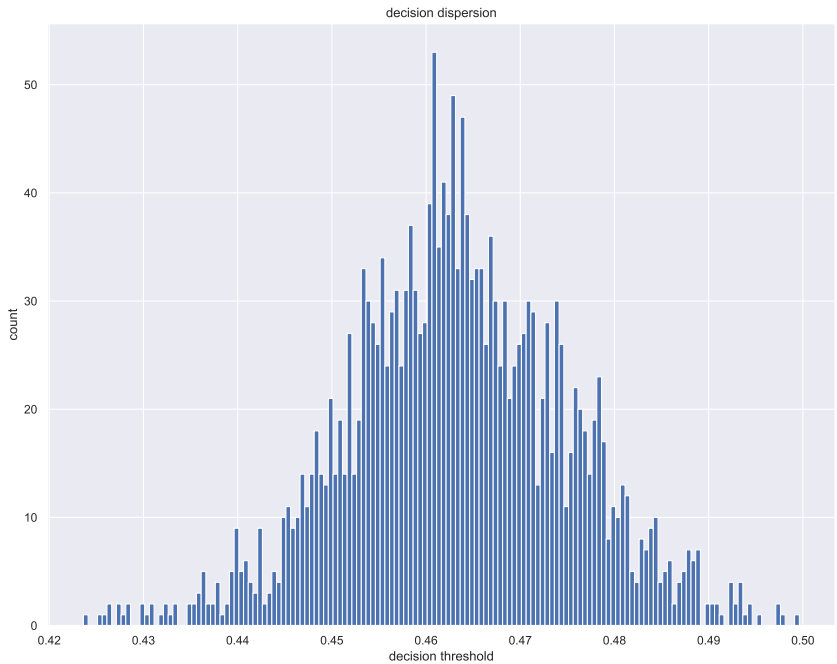




- Separating training and evaluation epochs is a good practice
- Evaluating a deterministic version of the policy leads to less noisy results
- Less episodes because only evaluation episodes are displayed

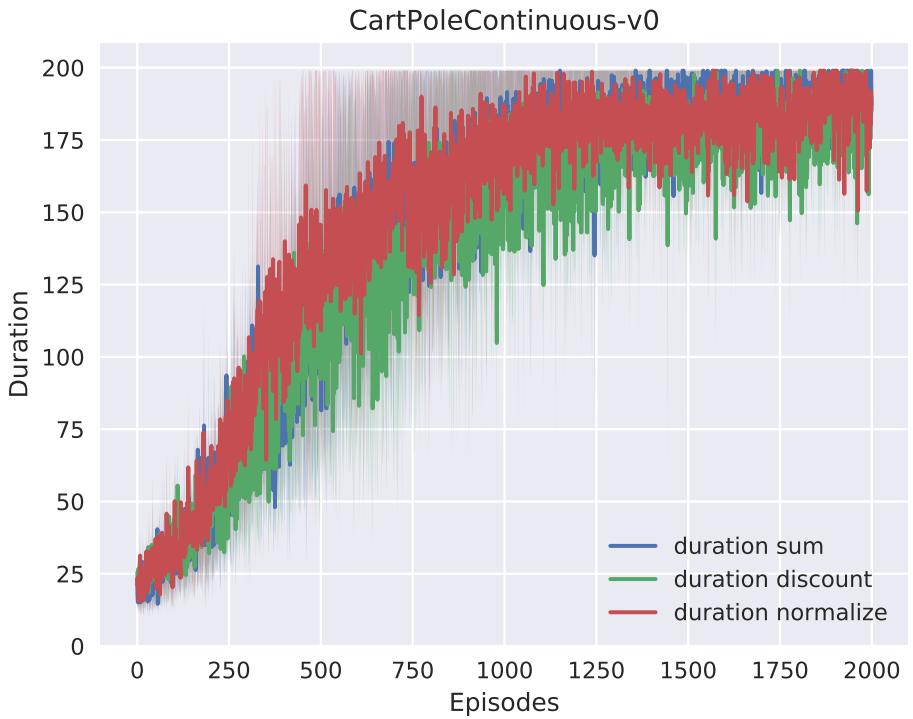
Two Initial Bernoulli policies

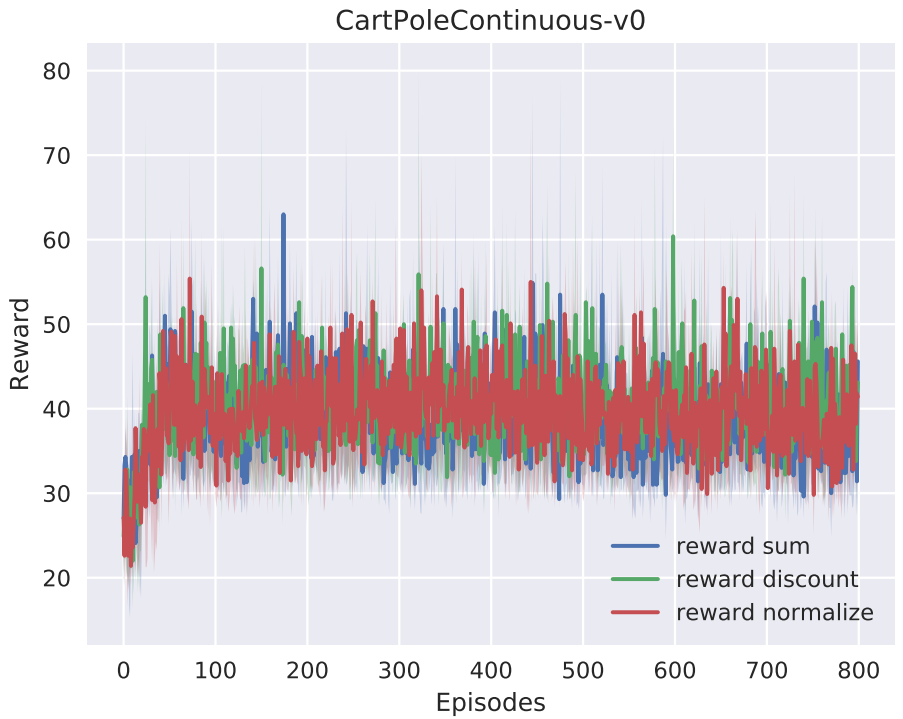




- To make deterministic policy, choice if threshold > 0.5 or < 0.5
- With the default initialization
- Initial decision thresholds are often all above 0.5, or all below 0.5
- Thus initial deterministic policies always take the same action!

Policy Gradient with Normal Policies





- Coded with adaptive variance
- Does not reach optimal performance
- What's happenning???

The NormalPolicy python class

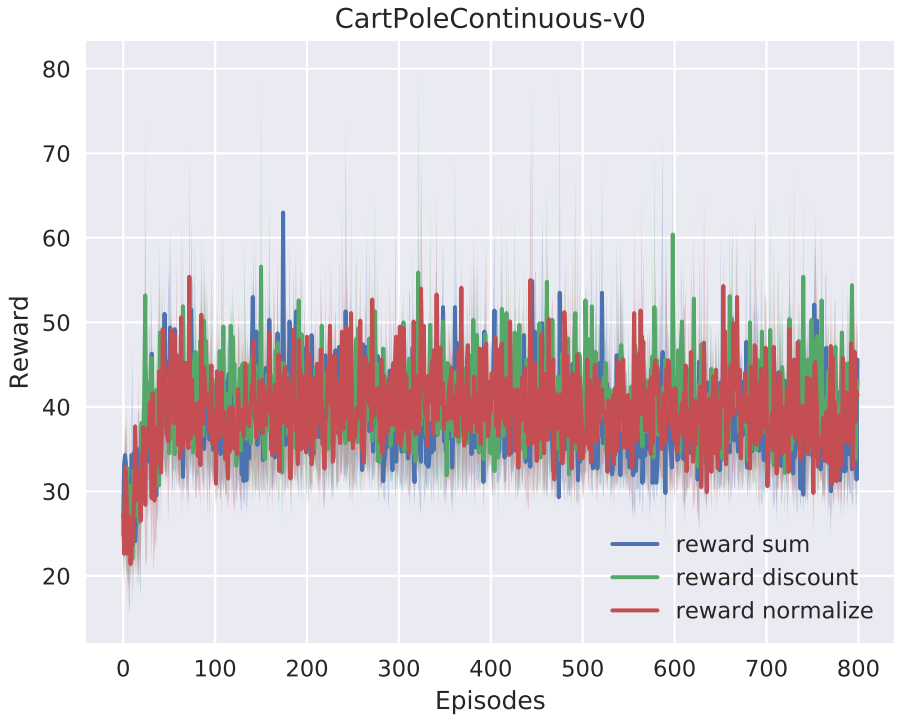
```
class NormalPolicy(GenericNet):
    def __init__(self, l1, l2, l3, l4, learning_rate):
        super(NormalPolicy, self).__init__()
```

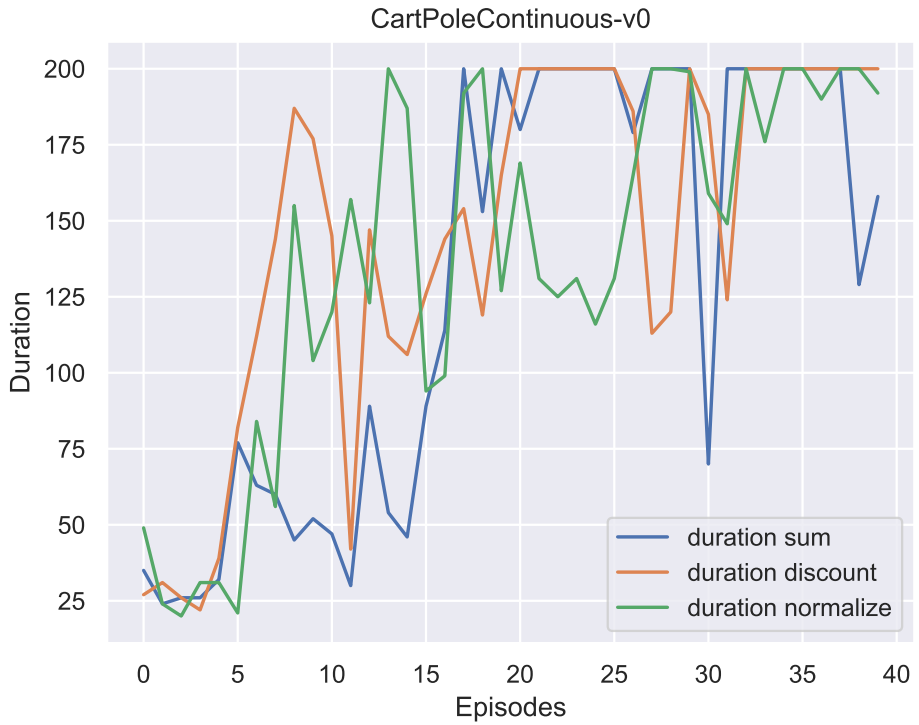
[...]

```
def forward(self, state):
    state = torch.from_numpy(state).float()
    state = self.relu(self.fc1(state))
    state = self.relu(self.fc2(state))
    mu = self.tanh(self.fc_mu(state))
    std = 2  # self.softplus(self.fc_std(state))
    return mu, std
```

- Due to init, the variance is very small
- All trajectories keep the same
- Bug fix: fixed variance. Tuning std helps

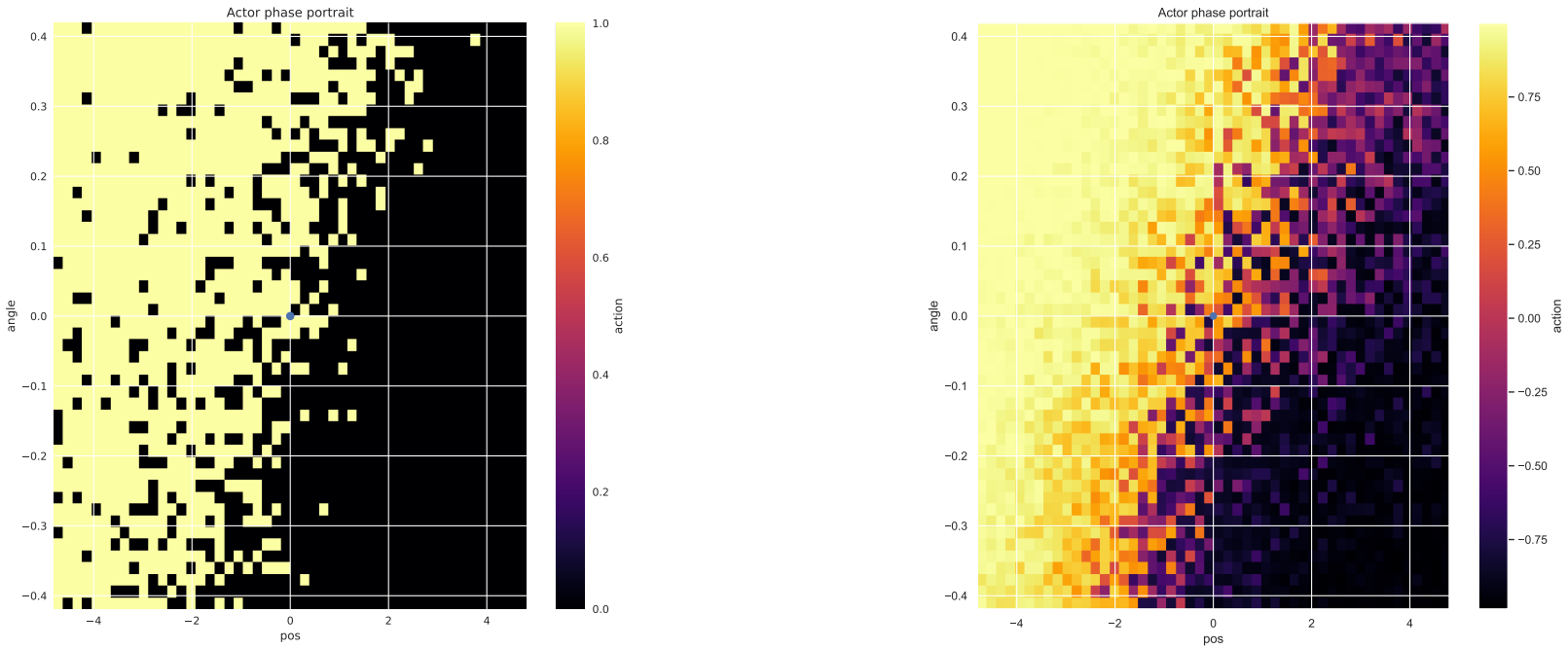
Policy Gradient with Normal Policies: bug fixed





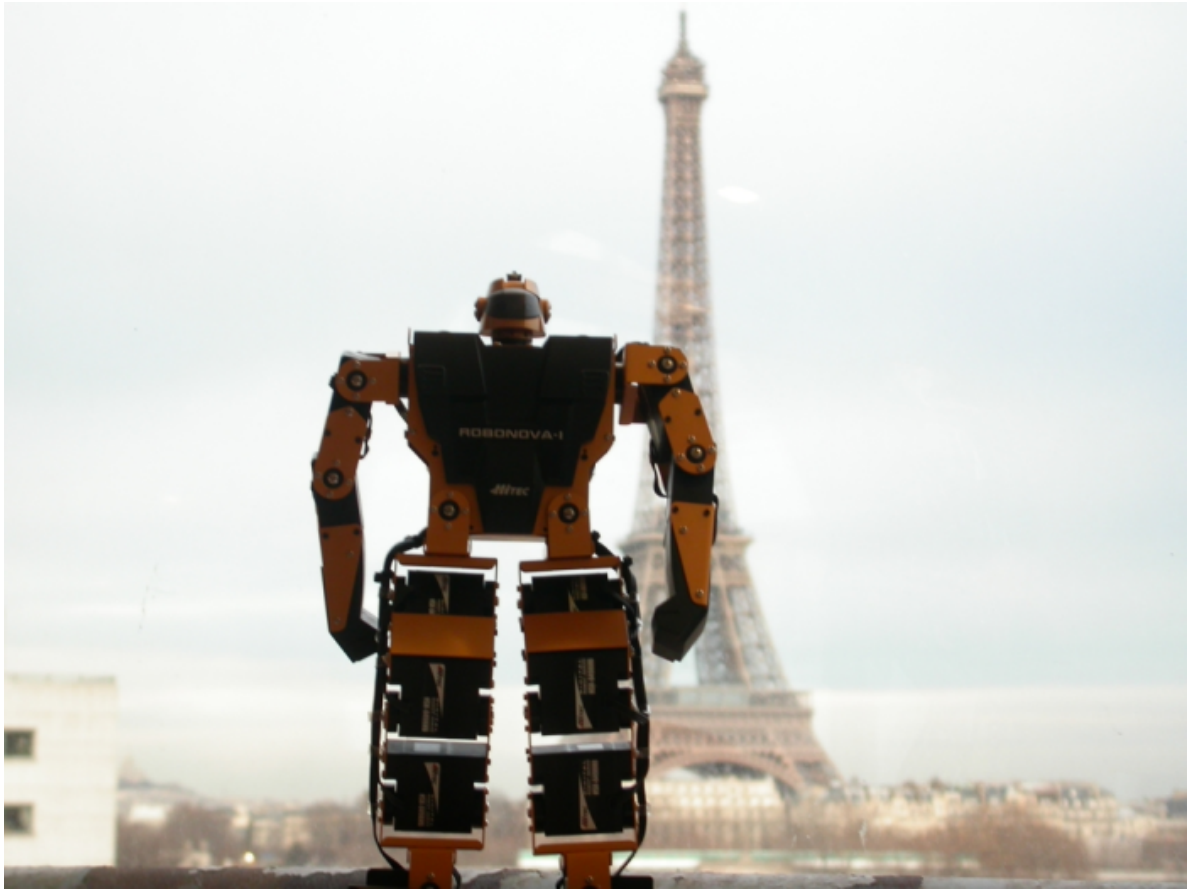
- One might rather use a squashed Gaussian (see SAC video)

Normal Policy



- Bernoulli (left) and Normal (right) policies
- Actions in a smaller range, and more continuous

Any question?



Send mail to: Olivier.Sigaud@upmc.fr