GRADIENT BANDIT BASELINES

David Krueger

ABSTRACT

We compare several baselines in the gradient bandit algorithm on the test bed presented in Sutton & Barto (2017). In the control case, past experience is off-policy, and so estimating the baseline from past experience seems suboptimal, and we propose and evaluate two ways of dealing with this issue. Our experiments don't show any convincing advantage for our proposed approaches. Our most interesting result is that the true value function is a suboptimal baseline in these experiments.

1 INTRODUCTION

We compare the following baselines:

From Sutton & Barto (2017):

1. no baseline
2. standard baseline (average of all rewards)

Idealized baselines (in practice we typically wouldn't actually have enough information to compute these):

1. True value-function (v^{π}) baseline
2. Optimal value-function (v^{*}) baseline

Our proposals (I didn't check if these are novel):

1. "Model-based" baseline: we track the mean reward of each arm, and average them according to the current policy π_t to estimate v^{π_t}. We use this estimate as the baseline and recompute it at every time-step.
2. Exponential moving average baseline: instead of a simple average of all rewards, we use an exponential moving average. We try decay hyperparameters $\gamma \in [0.5, 0.9, 0.99, 0.999, 0.9999]$.

Both of these proposals aim to address the problem of using off-policy data to estimate v^{π_t}, as in the average reward baseline.

The motivation for the model-based baseline is straightforward: to maximally reduce the variance of the updates, we'd like to use the true value-function, v^{π_t}, of our current policy π_t, as a baseline. If we just estimate the rewards of each arm a independently (as $\tilde{q}(a)$), then we can use our current policy to compute and unbiased estimate of v^{π_t} as $\tilde{v}^{\pi_t} = \mathbb{E}_{a \sim \pi_t} \tilde{q}(a)$.

An even simpler approach is just to weight more recently observed rewards more heavily in constructing the estimate of v^{π_t}. A natural way to do this is to use an exponential moving average.

2 RESULTS

We use the 10-armed bandit test-bed of Sutton & Barto (2017) to evaluate these algorithms. We shift the mean of the test-bed by 0,2,4, or 8, to create environments where baselines are important and plot the following:

1. Average reward

2. Percent of actions taken which were optimal
3. MSE of the baseline vs. the true v^{π_t}
4. The probability of the optimal action under π_t

As expected, no baseline is needed when the reward distributions aren't shifted.

For the shifted distributions, the simple average and exponential moving average ($\gamma = .99$) baselines give the best performance, with possibly a slight advantage for the exponential moving average. Somewhat surprisingly, the model-based baseline is outperformed by the simple average. Even more surprising is that the true value function is *also* outperformed by the averaging baselines (while doing somewhat better than the model-based algorithm which approximates it).

Reinforcing this result, we find that the MSE between the baseline and v^{π_t} has a characteristic *increase* around 100 steps for the averaging baselines (which are most successful), while the MSE of model-based baseline appears to be strictly decreasing. We don't have any interpretation of this at the moment, although we note that the best performing policies also explore less (see figure 2, right side).

In general, using v^* as a baseline doesn't seem to work well; this is to be expected since it ensures that the expected rewards are negative, as opposed to being centered at 0.

Comparing different values of γ, we see that the probability of the correct action is increased more quickly for large γ, but after more steps, smaller γ yields a higher probability of optimal action (see Figure 4).

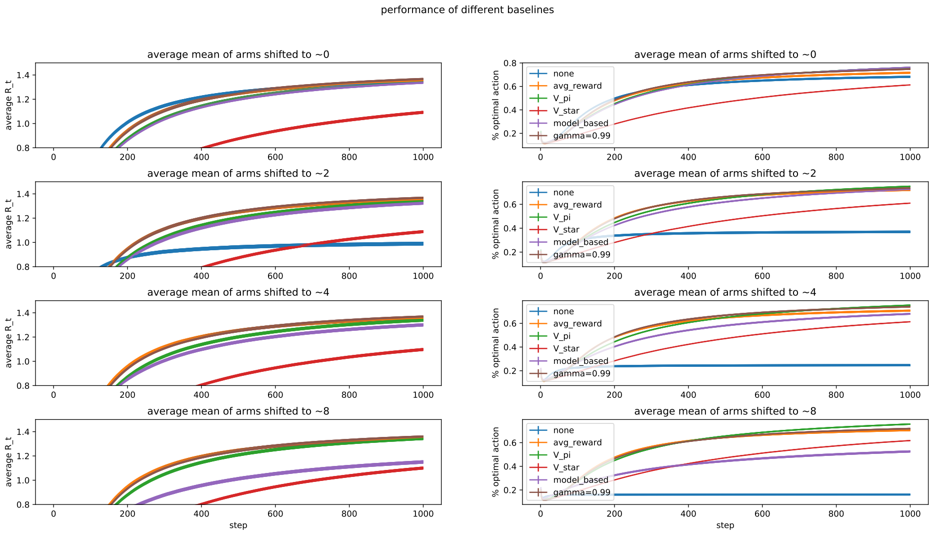

Figure 1: Average reward and % optimal action for all baseline algorithms. Plots include mean and standard error bars over all environments. For the exponential moving average, the value of γ which maximized (average) total rewards is used.

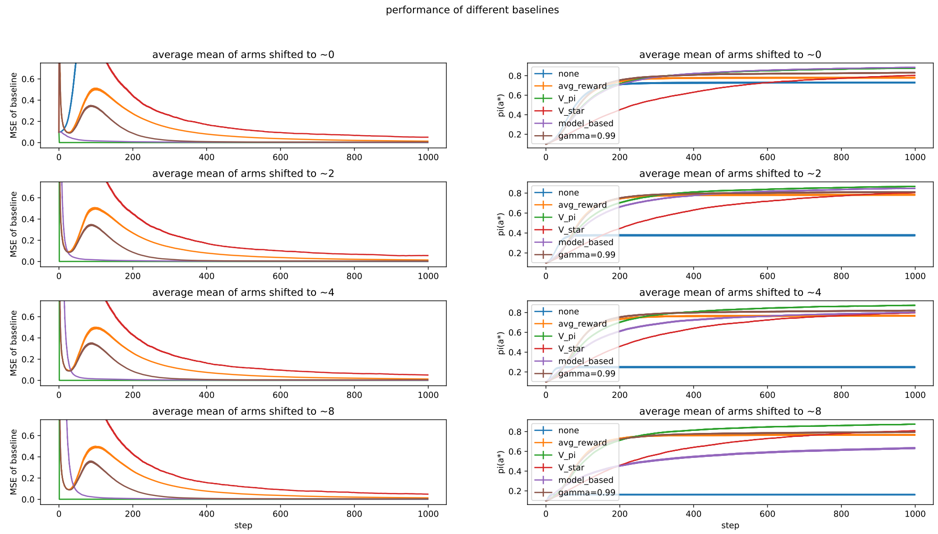

Figure 2: MSE of the baseline and probability of the optimal action under the learned policy at each step for all baseline algorithms. Plots include mean and standard error bars over all environments. For the exponential moving average, the value of γ which maximized (average) total rewards is used.

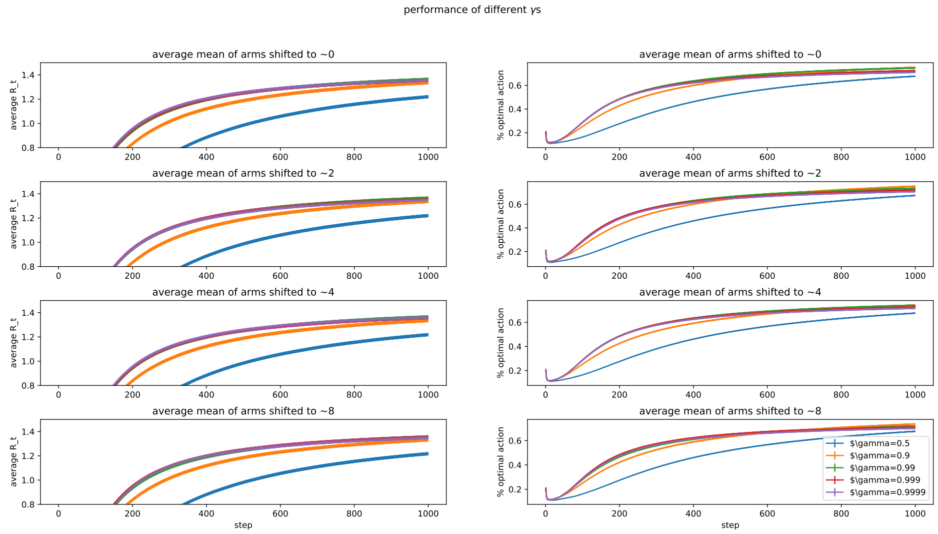

Figure 3: Average reward and % optimal action for all values of γ considered. Plots include mean and standard error bars over all environments.

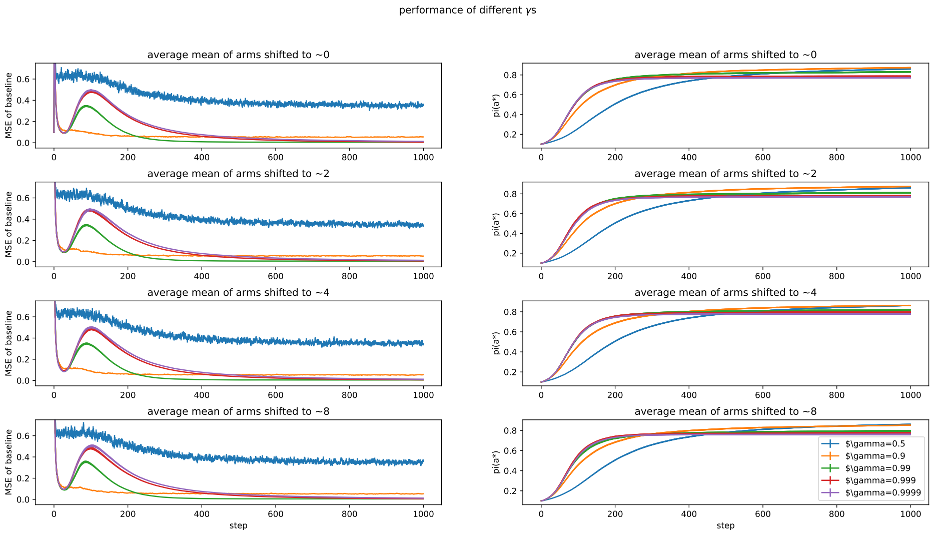

Figure 4: MSE of the baseline and probability of the optimal action under the learned policy at each step for all values of γ considered. Plots include mean and standard error bars over all environments.

References

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning : An Introduction, 2nd edition*. MIT Press, DRAFT, 2017. URL `http://incompleteideas.net/sutton/book/bookdraft2016sep.pdf`.