A. Appendices

A.1. Hyperparameter Settings

Our implementation is based on rllab (Duan et al., 2016) and as such most of the hyperparameter settings were kept to their default settings. The details are provided in Table 1.

Table 1. Experimental details

	Cliff Walker	HalfCheetah	Ant
POLOPT	TRPO	TRPO	TRPO
KL constraint	0.01	0.01	0.01
Discount rate	0.99	0.99	0.99
GAE λ	1.0	1.0	1.0
Batch size	10,000	12,500	12,500
Policy layers	(5,5)	(100,100)	(100,100)
Policy units	Tanh	ReLU	ReLU
FPO-UCB κ	2	2	2

A.2. Detailed Experimental Results

In Table 2c we present the quartiles of the expected return of the final learnt policy for each method across the 10 random starts.

A.3. Further Examination of the Learnt Policies

As explained in the Experiments section, the SREs for the HalfCheetah and Ant experiments are based on the velocities achieved by the agent; for HalfCheetah the SRE is defined as the velocity target being 4 (and carrying a large bonus reward) instead of 2 with a 2% probability of occurrence, while for Ant velocities greater than 2 has a 5% probability of incurring a large cost. Here we compare the performance of FPO-UCB(S) against the next best baseline (Enum for HalfCheetah and Random for Ant) and the Naïve baseline by visualising the velocity profiles of the final learnt policy. For each random start for each method we sampled 10 trajectories (for a total of 100 trajectories per method) and plot the histogram of the velocity at each timestep. This is presented in Figure 6.

For the HalfCheetah task, from Figure 6a we can see that the velocity profile for the Naïve approach is highly concentrated around 2. This goes to show that the Naïve approach learns a policy that does not take into account the SRE at all. On the other hand, both Enum and FPO-UCB(S) have velocity profiles with much higher variance with a lot of mass spread between 2 and 4. This goes to show that both of them take into account the effect of SREs. However, FPO-UCB(S) manages to better balance the SRE/non-SRE rewards and has slightly higher mass concentrated on 4,

Table 2. Quartiles of expected return across 10 random starts

(a) Cliff Walker

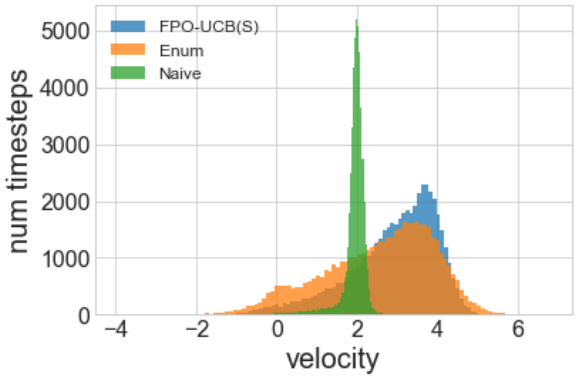
	Q1	Median	Q2
FPO-UCB(S)	427.1	441.5	450.0
FPO-UCB(A)	335.2	432.6	440.4
FPO-FITBO(S)	428.1	443.6	453.1
FPO-FITBO(A)	372.2	438.2	451.5
Naïve	-1478.7	-135.5	243
EPOpt	-44.4	282.1	354.4
ALOQ	33.5	57.2	77.2
Random	345.8	358.9	373.4

(b) HalfCheetah

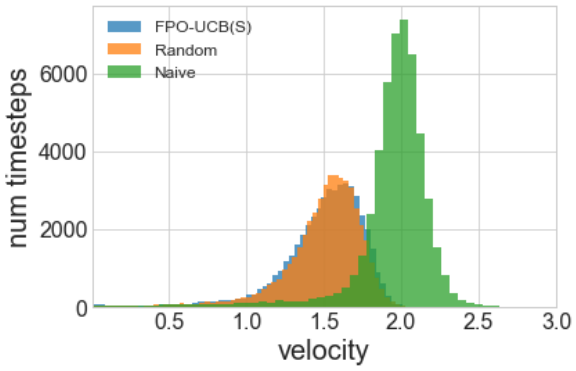
	Q1	Median	Q2
FPO-UCB(S)	3913.7	5464.0	5905.5
FPO-UCB(A)	4435.6	5231.8	5897.6
FPO-FITBO(S)	2973.9	3187.2	3923.7
FPO-FITBO(A)	3686.2	4091.1	7247.3
Naïve	1059.9	1071.1	1086.0
EPOpt	803.6	4066.0	4421.0
OFFER	1093.9	1097.4	1111.2
Random	1722.3	2132.6	2645.5
Enum	2442.8	2796.0	3428.4

(c) Ant

	Q1	Median	Q2
FPO-UCB(S)	490.6	674.2	713.3
FPO-UCB(A)	408.1	519.4	629.7
FPO-FITBO(S)	626.8	704.2	770.0
FPO-FITBO(A)	455.7	533.1	707.0
Naïve	-1746.9	-1669.4	-1585.2
EPOpt	-1732.3	-1606.6	-1454.5
Random	460.3	575.6	640.4
Enum	255.7	273.4	285.6



(a) HalfCheetah



(b) Ant

Figure 6. Histogram of the velocity profile of the final learnt policies for each method.

which in turn leads to it significantly outperforming Enum.

For the Ant task, unsurprisingly once again the Naïve approach completely ignores the SREs, and exhibits a velocity profile that is greater than 2 roughly 50% of the time. The velocity profiles of the Random baseline and FPO-UCB(S) are almost exactly the same. This is unexpected as there is no significant difference between the expected return of the final policies learnt by these two methods, as shown in Section 5. As noted earlier, the good performance of Random is not unsurprising since the schedule for ψ chosen by FPO-UCB(S) is close to 0.5, which is also the mean of ψ under the random baseline as $\psi \sim U(0, 1)$.

A.4. Performance in settings without SREs

FPO considers the setting where environments are characterised by SREs. A natural question to ask is how does its performance compare to the naïve method in settings where there are no SREs. To investigate this we applied FPO-UCB(S), and the naïve baseline, to the cliff walker problem presented in Section 5.1, with the modification that falling off the cliff now carries 0 reward instead of -5000. This removes the SRE, but the environment variable (the location of the cliff) is still relevant since it has a significant effect on the dynamics. The results are presented in Figure 7. Note that the performance of the Naïve method is far more stable than in the setting with SRE. However, while it is able to learn a good policy, FPO-UCB(S) still performs better since it takes into account the effect of the environment variable.

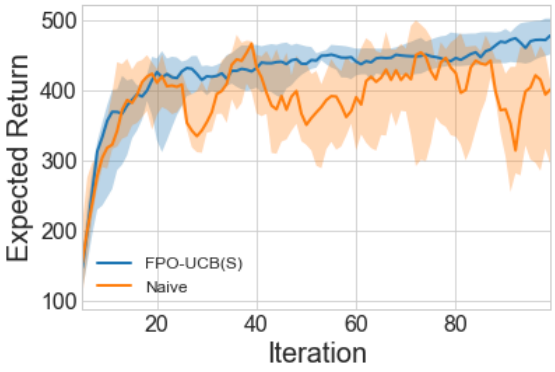


Figure 7. Results for the Cliff Walker environment without any SRE. Solid line shows the median and the shaded region the quartiles across 10 random starts.