Lecture 8: Exploration in tabular MDPs

Akshay Krishnamurthy
akshay@cs.umass.edu

March 23, 2022

1 Recap

So far, we have been discussing the three capabilities/challenges of generalization, credit assignment, and exploration, and we have seen techniques for addressing each challenge in isolation: empirical risk minimization for generalization, optimism for exploration, and dynamic programming for credit assignment. In terms of the pairs, contextual bandits addresses exploration and generalization but not credit assignment, while policy gradient methods address credit assignment and generalization (via linear function approximation) but not exploration. Today we will address exploration and credit assignment simultaneously. The methods that can do this fall into the "tabular reinforcement learning" paradigm.

2 Why is exploration hard?

It is not hard to construct MDP instances where random exploration has an exponentially small probability of visiting every state. One such MDP is displayed in Figure 1, and MDPs with this structure are broadly referred to as combination locks. Here the agent starts at g_0 deterministically, so if we execute actions uniformly at random, it takes 2^H episodes before we see the reward.

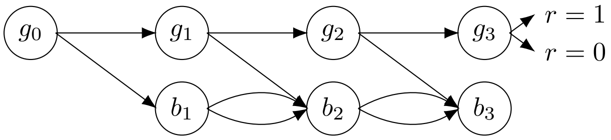

Figure 1: A combination lock MDP

It is also worthwhile to think about why policy gradient fails to address the exploration issue in the context of this combination lock. Last time we also saw that NPG's sample complexity depends on the distribution shift coefficient $\|d^{\tilde{\pi}}/\mu\|_\infty$ where $\tilde{\pi}$ is the comparator policy and μ is the starting distribution. Taking $\tilde{\pi} = \pi^\star$, this quantity is infinite. If we know some other distribution ρ and use it for estimation of the Q-functions, ρ can replace μ above, but how do we find some distribution ρ that makes the distribution shift coefficient small? Indeed, finding such a ρ requires us to learn how to visit the end of the chain, which amounts to solving the exploration problem! This is another way to see that policy gradient methods do not address the exploration challenge.

3 UCB-VI

Let us first describe the precise learning protocol and objective. We consider an episodic MDP $M = (\mathcal{S}, \mathcal{A}, P, R)$ with horizon H. For notational simplicity we assume the starting state is fixed and known, and that the reward function R is deterministic and known. (Both of these can be easily generalized, but it requires us to track and estimate a few more objects. However it does not affect the final guarantees at all.) We consider an agent interacting with this environment for T episodes, where at the onset P is unknown. In episode t, a trajectory $\tau_t :=$

$(s_0^t, a_0^t, r_0^t, s_1^t, a_1^t, r_1^t, \ldots, s_{H-1}^t, a_{H-1}^t, r_{H-1}^t)$ is generated where all actions are chosen by the agent and states/rewards are sampled according to the MDP. We measure the performance of the algorithm via regret:

$$\mathrm{Reg}(T) := TJ(\pi^\star) - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1} r_h^t\right]$$

If the algorithm deploys a policy π^t in the t^{th} episode, then we can re-write the regret as $\mathrm{Reg}(T) = \sum_t J(\pi^\star) - J(\pi^t)$.

A natural idea. A natural idea is to (1) use the data that we have collected to estimate a transition operator $\hat{P}$, (2) plan in the estimated MDP $\hat{M} = (\mathcal{S}, \mathcal{A}, \hat{P}, R)$ to find a policy $\hat{\pi}$, and (3) deploy $\hat{\pi}$ in the next episode. This doesn't work unfortunately, but it is helpful to understand why, both from a conceptual and a technical level.

We can understand this issue in the simpler stochastic MAB setting. Here this plug-in/greedy approach amounts to choosing the arm $a_t = \mathrm{argmax}_a\, \hat{\mu}_{t-1}(a)$, and it fails because it may starve the optimal arm $a^\star$. If we start by playing every arm once, there is a good chance that our empirical mean for $a^\star$ is lower than the true mean for some other arm. More specifically, we may have $\hat{\mu}(a^\star) \le \mu(\tilde{a}) \le \hat{\mu}(\tilde{a})$. From this point, we will play $\tilde{a}$, and our estimate $\hat{\mu}(\tilde{a})$ will converge to $\mu(\tilde{a})$, but this will continue to dominate our estimate for $a^\star$, so the latter will never improve.

On a technical level, this algorithm admits the per-round regret decomposition:

$$\mu(a^\star) - \mu(a_t) \le |\mu(a^\star) - \hat{\mu}_{t-1}(a^\star)| + |\mu(a_t) - \hat{\mu}_{t-1}(a_t)| + \hat{\mu}_{t-1}(a^\star) - \hat{\mu}_{t-1}(a_t) \le \mathrm{conf}_{t-1}(a^\star) + \mathrm{conf}_{t-1}(a_t).$$

The issue is that we cannot guarantee that $\mathrm{conf}(a^\star)$ is decreasing with t, although we did show that $\mathrm{conf}_{t-1}(a_t)$ decreases with t. This is the technical reason why the greedy algorithm suffers linear regret.

Optimistic regret decomposition. In the stochastic MAB problem, optimism addresses the above issue on a technical level because it instead admits the regret decomposition

$$\mu(a^\star) - \mu(a_t) \le 2\mathrm{conf}_{t-1}(a_t). \tag{1}$$

Now, our potential function argument guarantees that $\mathrm{conf}_{t-1}(a_t)$ cannot remain large forever. Can we instantiate this idea in the episodic MDP setting? Indeed it is possible, and this is the main idea behind the algorithm UCB-VI. However, the method is quite a bit more complicated than the UCB algorithm for multi-armed bandits, so let us break it down into several components.

Perhaps the core lemma is an optimistic regret decomposition, which is quite general. To state the lemma, recall the Bellman backup operator $\mathcal{T}$ defined as

$$\mathcal{T}f : (s,a) \to R(s,a) + \mathbb{E}_{s' \sim P(s,a)}\left[\max_{a'} f(s',a')\right].$$

Additionally recall that we use $d_h^\pi \in \Delta(\mathcal{S}\times\mathcal{A})$ to denote the state-action occupancy measure of policy π at time step h. The next lemma states that if we have a Q-function that is (a) optimistic and (b) nearly "Bellman consistent," then we can obtain a regret decomposition that is analogous to (1).

Lemma 1 (Optimistic regret decomposition). *Suppose that we have $\bar{Q}$ and a confidence bound* conf *satisfying:*

$$\forall s,a,h : Q_h^\star(s,a) \le \bar{Q}_h(s,a) \le (\mathcal{T}\bar{Q}_{h+1})(s,a) + \mathrm{conf}_h(s,a), \tag{2}$$

then if $\bar{\pi}_h : s \mapsto \mathrm{argmax}_a\, \bar{Q}_h(s,a)$ is the greedy policy w.r.t. $\bar{Q}$, we have

$$J(\pi^\star) - J(\bar{\pi}) \le \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d_h^{\bar{\pi}}}\left[\mathrm{conf}_h(s,a)\right]$$

The key property of this lemma is that there is no dependence on $\pi^\star$ on the right hand side. Instead everything is measured on states and actions that we will visit if we execute $\bar{\pi}$. Another important remark is that the upper bound on $\bar{Q}_h$ in (2) is actually more favorable than asking $\bar{Q}_h$ to be not-much-larger than $Q_h^\star$ (which may seem like the more natural pre-condition). This holds because $\bar{Q}_{h+1}$ is also optimistic!

Proof. The key step is to translate from $Q^\star$ to $\bar{Q}$ and then use the greedy property of $\bar{\pi}$:

$$\begin{aligned} J(\pi^\star) - J(\bar{\pi}) &= Q_0^\star(s_0, \pi^\star(s_0)) - Q_0^{\bar{\pi}}(s_0, \bar{\pi}(s_0)) \\ &\le \bar{Q}_0(s_0, \pi^\star(s_0)) - Q_0^{\bar{\pi}}(s_0, \bar{\pi}(s_0)) \\ &\le \bar{Q}_0(s_0, \bar{\pi}(s_0)) - Q_0^{\bar{\pi}}(s_0, \bar{\pi}(s_0)) \\ &\le (\mathcal{T}\bar{Q}_1)(s_0, \bar{\pi}(s_0)) + \mathrm{conf}_0(s_0, \bar{\pi}(s_0)) - \mathbb{E}_{s_1 \sim P(s_0, \bar{\pi}(s_0))}\left[Q_1^{\bar{\pi}}(s_1, \bar{\pi}(s_1))\right] \\ &= \mathbb{E}_{(s_1,a_1)\sim d_1^{\bar{\pi}}}\left[\bar{Q}_1(s_1, a_1) - Q_1^{\bar{\pi}}(s_1, a_1)\right] + \mathbb{E}_{(s_0,a_0)\sim d_0^{\bar{\pi}}}\left[\mathrm{conf}_0(s_0, a_0)\right] \end{aligned}$$

The final equality holds because $\bar{\pi}$ is the greedy policy w.r.t. $\bar{Q}$, so that $\max_a \bar{Q}_1(s,a) = \bar{Q}_1(s, \bar{\pi}_1(s))$. The first term in last line has the same for as the third line, except that we are one time step ahead. So we can recursively apply this argument to prove the lemma. □

Optimistic planning via bonuses. The next issue we need to address is the construction of a Q function that satisfies (2). For this, UCB-VI leverages a form of *optimistic value iteration*. We need to specify the estimator for the transition operator and the confidence bonus.

Suppose we are in episode t, and let $N_{t-1}(s,a,s'), N_{t-1}(s,a)$ denote the number of times we have seen (s,a,s') or (s,a) in the past $t-1$ trajectories, that is:

$$N^{t-1}(s,a) = \sum_{i=1}^{t-1}\sum_{h=0}^{H-2} \mathbf{1}\{s_h^i = s, a_h^i = a\}, \qquad N^{t-1}(s,a,s') = \sum_{i=1}^{t-1}\sum_{h=0}^{H-2} \mathbf{1}\{s_h^i = s, a_h^i = a, s_{h+1}^i = s'\}$$

(We throw away the last step, since we do not see s_H, we cannot estimate the transitions from these samples.)

Then, we can estimate the transition model via $P^{t-1}(s' \mid s,a) = \frac{N^{t-1}(s,a,s')}{N^{t-1}(s,a)}$, which is just the empirical estimate. As discussed above, if we perform value iteration using P^{t-1} we may not obtain an optimistic Q-function. However, if we instead consider the optimistic value iteration procedure defined as:

$$Q_{H-1}^{t-1}(s,a) = R(s,a), \qquad Q_h^{t-1}(s,a) = \min\{H, \underbrace{R(s,a) + b^{t-1}(s,a)}_{\text{reward bonus}} + \sum_{s'} P^{t-1}(s' \mid s,a) \max_{a'} Q_{h+1}^{t-1}(s',a')\} \tag{3}$$

and set $b^{t-1}(s,a) = H\min\{1, \sqrt{\frac{S\Delta}{N^{t-1}(s,a)}}\}$ with $\Delta = c\log(SAHT/\delta)$ for some universal constant c, then we can show that Q^{t-1} is optimistic with high probability. This is the contents of the next lemma.

Before we state and prove the lemma it is helpful to provide some intutition. The procedure is performing standard value iteration in a different MDP $\hat{M} = (\mathcal{S}, \mathcal{A}, P^{t-1}, R + b^{t-1}, \mu)$, which uses the estimated transition operator and a reward function that includes an *exploration bonus*. The bonus is high for the (s,a) pairs with low visitation, which incentivizes the agent to visit those states. Further, because the bonus is tied to the estimation error of the transition model, a policy that gets large bonus in $\hat{M}$ will definitely visit some unknown region when deployed in M. To see this, observe that up until the first time the bonus is "large" the two MDPs behave very similarly. Thus, the exploration bonus intuitively takes care of the exploration challenge.

Lemma 2. *Set $b^{t-1}(s,a) = H\min\{1, \sqrt{\frac{S\Delta}{N^{t-1}(s,a)}}\}$ and perform the update* (3)*. Then with probability at least $1-\delta$:*

$$\forall t,h,s,a : Q_h^\star(s,a) \le Q_h^{t-1}(s,a) \le (\mathcal{T}Q_{h+1}^{t-1})(s,a) + 2b^{t-1}(s,a).$$

Proof. By a martingale version of Bernstein's inequality, we can show that for all (s,a,t) we have

$$|P^{t-1}(s' \mid s,a) - P(s' \mid s,a)| \lesssim \sqrt{\frac{P(s' \mid s,a)\log(SAHT/\delta)}{N^{t-1}(s,a)}} + \frac{\log(SAHT/\delta)}{N^{t-1}(s,a)},$$

with probability at least $1-\delta$. Recall that the total variation distance $\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\sum_s |P(s) - Q(s)|$ is 1/2 of the ℓ_1 distance. By an application of Cauchy-Schwarz, this implies that we have a total-variation approximation to the the transition operator in the sense that

$$\forall t,s,a : \|P^{t-1}(s,a) - P(s,a)\|_{\mathrm{TV}} \le \sqrt{\frac{S\Delta}{N^{t-1}(s,a)}}.$$

Using this inequality, let us prove the lemma. To establish optimism consider step h and inductively assume that Q_{h+1}^{t-1} is optimistic, that is $Q_{h+1}^{t-1}(s,a) \geq Q_{h+1}^\star(s,a)$ for all s,a. If the update for $Q_h^{t-1}(s,a)$ takes the $\min(H,\cdot)$ then, since $Q_h^\star(s,a) \leq H$, we have already satisfied optimism. So we can assume that $Q_h^{t-1}(s,a)$ is defined by the second part of the min. Note that in this case, $b^{t-1}(s,a) < H$ and in fact we can see that $b^{t-1}(s,a)$ is $1/H$ times the right hand side of the TV bound above. Therefore, in this case,

$$\begin{aligned}
Q_h^{t-1}(s,a) &= R(s,a) + b^{t-1}(s,a) + \sum_{s'} P^{t-1}(s' \mid s,a) \max_{a'} Q_{h+1}^{t-1}(s',a') \\
&\geq R(s,a) + b^{t-1}(s,a) + \sum_{s'} P^{t-1}(s' \mid s,a) V_{h+1}^\star(s') \\
&= R(s,a) + b^{t-1}(s,a) + \sum_{s'} P(s' \mid s,a) V_{h+1}^\star(s') + \sum_{s'} \left(P^{t-1}(s' \mid s,a) - P(s' \mid s,a)\right) V_{h+1}^\star(s') \\
&\geq R(s,a) + b^{t-1}(s,a) + \sum_{s'} P(s' \mid s,a) V_{h+1}^\star(s') - \|P^{t-1}(s,a) - P(s,a)\|_{\mathrm{TV}} \cdot H \\
&\geq R(s,a) + \sum_{s'} P(s' \mid s,a) V_{h+1}^\star(s') = Q_h^\star(s,a).
\end{aligned}$$

Here, the first inequality uses the inductive property and the second is precisely Holder's inequality, since $0 \leq Q_h^\star \leq H$. Finally we use that the bonus is set to dominate the error term, and the definition of $Q_h^\star$.

The upper bound is similar:

$$\begin{aligned}
Q_h^{t-1}(s,a) &\leq R(s,a) + b^{t-1}(s,a) + \sum_{s'} P^{t-1}(s' \mid s,a) \max_{a'} Q_{h+1}^{t-1}(s',a') \\
&\leq b^{t-1} + (\mathcal{T} Q_{h+1}^{t-1})(s,a) + \sum_{s'} \left(P^{t-1}(s' \mid s,a) - P(s' \mid s,a)\right) \left(\max_{a'} Q_{h+1}^{t-1}(s',a')\right) \\
&\leq 2b^{t-1} + (\mathcal{T} Q_{h+1}^{t-1})(s,a)
\end{aligned}$$

□

Potential function. Based on the above two lemmas, we can prove the following corollary.

Corollary 3. *If in every iteration we perform optimistic value iteration* (3) *and set π^t to be the greedy policy w.r.t., Q^{t-1} then with probability at least $1-2\delta$ we have*

$$\sum_{t=1}^T J(\pi^\star) - J(\pi^t) \leq 2 \sum_{t=1}^T \sum_{h=0}^{H-2} b^{t-1}(s_h^t, a_h^t) + O(H\sqrt{T\log(1/\delta)})$$

Proof. By Lemma 2 we can instantiate Lemma 1 with the confidence term equal to $b^{t-1}(s,a)$. This reveals that with probability $1-\delta$

$$\sum_{t=1}^T J(\pi^\star) - J(\pi^t) \leq 2 \sum_{t=1}^T \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d_h^{\pi^t}}[b^{t-1}(s,a)].$$

We obtain the result by observing that $(s_h^t, a_h^t) \sim d_h^{\pi^t}$, since we deploy policy π^t in the t^{th} episode. So we can relate the "empirical" state visitation to the expected state visitation via a martingale concentration argument. □

To finish the proof, we need to show that the bonuses shrink over time. This leverages a potential function argument similar to the one for the standard UCB algorithm.

Lemma 4 (Potential function)**.** *Consider any sequence of T trajectories $\tau^t = \{s_h^t, a_h^t\}_{h=0}^{H-1}$ for $t = 1, \ldots, T$. Then*

$$\sum_{t=1}^T \sum_{h=0}^{H-1} b_{t-1}(s_h^t, a_h^t) \leq O(H^2 S\sqrt{AT\Delta})$$

Proof. This proof is very similar to the MAB potential function argument, except now we have SA counters that are being incremented. There is also one wrinkle which is that we may visit the same (s, a) pair multiple times within an episode, meaning that we can suffer $1/\sqrt{N_t(s,a)}$ up to H times. On the other hand, if we do then we increment the counter by H.

To address this wrinkle it is best to break apart each counter into H counters, one per time step. Focusing on one (s, a) pair we need to bound $\sum_t z_t \sqrt{1/N^{t-1}(s,a)}$ with the update $N^t(s,a) \leftarrow N^{t-1}(s,a) + z_t$ where $z_t \in \{0, \ldots, H\}$ is the number of times we visit (s, a) in the t^{th} episode. The easiest way to control this sum is to instead create a counter for each time step h, so we define $N_h^{t-1}(s,a) = \sum_{i=1}^{t-1} \mathbf{1}\{s_h^i = s, a_h^i = a\}$ and $z_{t,h} = \mathbf{1}\{s_h^t = s, a_h^t = a\}$. Then

$$\sum_t \frac{z_t}{\sqrt{N^{t-1}(s,a)}} = \sum_t \sum_h \frac{z_{t,h}}{\sqrt{N^{t-1}(s,a)}} \leq \sum_t \sum_h \frac{z_{t,h}}{\sqrt{N_h^{t-1}(s,a)}} \leq O(\sum_h \sqrt{N_h^T(s,a)}).$$

Now let us express the sum in terms of each (s, a) pair:

$$\sum_{t=1}^{T} \sum_{h=0}^{H-1} b_{t-1}(s_h^t, a_h^t) = O(H\sqrt{S\Delta}) \sum_{s,a} \sum_t \frac{z_t(s,a)}{\sqrt{N^{t-1}(s,a)}} \leq O(H\sqrt{S\Delta}) \sum_{s,a,h} \sqrt{N_h^T(s,a)} \leq O(H\sqrt{S\Delta}) \cdot O(H\sqrt{SAT})$$

Here the final step follows from Cauchy-Schwarz and the fact that $\sum_{s,a,h} N_h^T(s,a) = HT$ since we make a total of HT updates over the course of the learning process. So in the final bound, we pay $H\sqrt{S}$ due to the size of the bonuses, we pay $\sqrt{HSA}$ for the number of counters we track, and we pay $\sqrt{HT}$ for the number of updates. □

Putting together the pieces of the above argument, we arrive at the main theorem for today. Before stating the theorem, we summarize the algorithm. Before each episode $t \in [T]$ the algorithm constructs a estimate P^{t-1} of the transition operator of the MDP and performs an optimistic version of value iteration using reward bonuses $b^{t-1}(s,a)$ that are derived from a concentration argument. Optimistic value iteration computes a Q-function Q^{t-1}, and the algorithm simply deploys the greedy policy with respect to this Q-function for the next episode. The guarantee is:

Theorem 5 (UCB-VI regret bound)**.** *The UCB-VI algorithm guarantees that with probability at least* $1 - \delta$

$$\mathrm{Reg}(T) \leq O(H^2 S \sqrt{AT \log(SAHT/\delta)})$$

3.1 Refinements

The above analysis for UCB-VI is correct, but it is quite loose. In fact the same algorithm, with a sharper definition of the bonuses $b^{t-1}(s,a)$ achieves $\tilde{O}(H^2\sqrt{SAT} + H^3 S^2 A \log(T))$ regret. The key improvement is that the leading $\sqrt{T}$ term only scales with $\sqrt{S}$ rather than with S. In other words if we want to find an ϵ-suboptimal policy and ϵ is small, then we need $T = \Omega(SAH^4/\epsilon^2)$ episodes which is *linear* in S. This is perhaps surprising since the transition operator has S^2A unknown parameters, and the algorithm is "model-based" in that it tries to estimate the transition operator. However the optimal Q-function only has SA parameters, and in fact we just need to estimate the transition operator well-enough to approximate $Q^\star$.

To establish this improved bound, there are two observations. The first is that our application of Holder's inequality to establish optimism in Lemma 2 is quite loose. We only care that P^{t-1} is accurate in the sense that

$$\sum_{s'} \left(P^{t-1}(s' \mid s,a) - P(s' \mid s,a)\right) V_h^\star(s'),$$

is small. Since $V_h^\star$ is just a single function, passing to total-variation distance is quite loose. This reasoning shows that we can set the bonus a factor of $\sqrt{S}$ smaller and still satisfy optimism.

Unfortunately, we cannot use this for the upper bound in Lemma 2 since we care to control the action of $P^{t-1} - P$ on the estimated value function V^{t-1}, which is *data-dependent*. The fact that this function is data-dependent means we cannot avoid the $\sqrt{S}$ factor in the concentration argument here, since it could correlated with the deviation in $P^{t-1} - P$ (essentially we are doing uniform convergence over all functions). However there is a trick we can use here, which is to add and subtract $V^\star_{h+1}$:

$$\begin{aligned}
&\sum_{s'} \left(P^{t-1}(s' \mid s,a) - P(s' \mid s,a)\right) V^{t-1}_{h+1}(s') \\
&\quad = \sum_{s'} \left(P^{t-1}(s' \mid s,a) - P(s' \mid s,a)\right) \left(V^{t-1}_{h+1}(s') - V^\star_{h+1}(s')\right) + \sum_{s'} \left(P^{t-1}(s' \mid s,a) - P(s' \mid s,a)\right) V^\star_{h+1}(s') \\
&\quad \lesssim \sum_{s'} \left(P^{t-1}(s' \mid s,a) - P(s' \mid s,a)\right) \left(V^{t-1}_{h+1}(s') - V^\star_{h+1}(s')\right) + H\sqrt{\frac{\Delta}{N^{t-1}(s,a)}}
\end{aligned}$$

From here, we exploit the optimistic property of V^{t-1}_{h+1} to relate the first term to the error at the next time step:

$$\begin{aligned}
&\sum_{s'} \left(P^{t-1}(s' \mid s,a) - P(s' \mid s,a)\right) \left(V^{t-1}_{h+1}(s') - V^\star_{h+1}(s')\right) \\
&\quad \lesssim \sqrt{\frac{S\sum_{s'} P(s' \mid s,a)(V^{t-1}_{h+1}(s') - V^\star_{h+1}(s'))^2\Delta}{N^{t-1}(s,a)}} + \frac{HS\Delta}{N^{t-1}(s,a)} \\
&\quad \lesssim \sqrt{\frac{HS\sum_{s'} P(s' \mid s,a)(V^{t-1}_{h+1}(s') - V^\star_{h+1}(s'))\Delta}{N^{t-1}(s,a)}} + \frac{HS\Delta}{N^{t-1}(s,a)} \\
&\quad \lesssim \frac{\mathbb{E}_{s'\sim P(s,a)}[V^{t-1}_{h+1}(s') - V^\star_{h+1}(s')]}{H} + \frac{H^2S\Delta}{N^{t-1}(s,a)} + \frac{HS\Delta}{N^{t-1}(s,a)}.
\end{aligned}$$

Here we use a Bernstein-type inequality to control the "product of differences" $(P^{t-1} - P^\star)(V^{t-1} - V^\star)$ in terms of the "variance," and then we use the optimistic property of V^{t-1} to relate the variance to the "downstream error." The last step uses the AM-GM inequality to obtain a linear term in the overestimation at the next time.

Already we can see that the S factors appear in the "lower order" terms that have a $1/N^{t-1}(s,a)$ rather than the square root. On the other hand we have a $1/H$ factor of the downstream error. This ends up exponentiating through the recursion:

$$\mathbb{E}_{d^{\pi^t}_h}[V^{t-1}_h(s) - V^\star_h(s)] \lesssim (1 + 1/H)\mathbb{E}_{s'\sim d^{\pi^t}_{h+1}}\left[V^{t-1}_{h+1}(s') - V^\star_{h+1}(s')\right] + \text{error terms},$$

but fortunately $(1 + 1/H)^H \leq e$ for all H. So this does not cause any significant issues.