A. Proofs

All the provided proofs operate under the setting where $\mu(a|s)$ has full support over the action space. When this assumption is not satisfied, the provided proofs can be transferred by assuming we are operating in a new MDP M_μ as defined below.

Given the MDP $M = \langle \mathcal{S}, \mathcal{A}, r, \mathcal{P}, \gamma \rangle$ and $\mu(a|s)$, let us define the new MDP $M_\mu = \langle \mathcal{S}_\mu, \mathcal{A}_\mu, r, \mathcal{P}, \gamma \rangle$, where $\mathcal{S}_\mu$ denotes the set of reachable states by μ, and $\mathcal{A}_\mu$ is $\mathcal{A}$ restricted to the support of $\mu(a|s)$ in each state in $\mathcal{S}_\mu$.

A.1. Contraction Mapping

Theorem 3.1. *In the tabular setting, for any $N \in \mathbb{N}$, $\mathcal{T}_\mu^N$ is a contraction operator in the $\mathcal{L}_\infty$ norm. Hence, with repeated applications of the $\mathcal{T}_\mu^N$, any initial Q function converges to a unique fixed point.*

Proof. Let Q_1 and Q_2 be two arbitrary Q functions.

$$\mathcal{T}_\mu^N Q_1 - \mathcal{T}_\mu^N Q_2 \ _\infty = \tag{13}$$

$$\max_{s,a} \ \Big(r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N}[\max_{\{a_i\}^N} Q_1(s',a')]\Big) - \Big(r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N}[\max_{\{a_i\}^N} Q_2(s',a')]\Big) \ = \tag{14}$$

$$\gamma \cdot \max_{s,a} \ \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N} \Big[\max_{\{a_i\}^N} Q_1(s',a') - \max_{\{a_i\}^N} Q_2(s',a') \Big] \ \leq \tag{15}$$

$$\gamma \cdot \max_{s,a} \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N} \ \max_{\{a_i\}^N} Q_1(s',a') - \max_{\{a_i\}^N} Q_2(s',a') \ \leq \tag{16}$$

$$\gamma \cdot \max_{s,a} \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N} \left\| Q_1 - Q_2 \right\|_\infty = \tag{17}$$

$$\gamma \cdot \left\| Q_1 - Q_2 \right\|_\infty \tag{18}$$

where line 17 is due to the following: Let $\hat{a} = \arg\max_{\{a_i\}^N} Q_1(s', a_i)$,

$$\max_{\{a_i\}^N} Q_1(s',a') - \max_{\{a_i\}^N} Q_2(s',a') = Q_1(s',\hat{a}) - \max_{\{a_i\}^N} Q_2(s',a') \tag{19}$$

$$\leq Q_1(s',\hat{a}) - Q_2(s',\hat{a}) \tag{20}$$

$$\leq \left\| Q_1 - Q_2 \right\|_\infty \tag{21}$$

□

A.2. Limiting Behavior

Theorem 3.3. *Let π_μ^* denote the optimal policy from the class of policies whose actions are restricted to lie within the support of the policy $\mu(a|s)$. Let Q_μ^* denote the Q-value function corresponding to π_μ^*. Furthermore, let Q_μ denote the Q-value function of the policy $\mu(a|s)$. Let $\mu^*(s) := \int_{\text{Support}(\pi_\mu^*(a|s))} \mu(a|s)$ denote the probability of optimal actions under $\mu(a|s)$. Under the assumption that $\inf_s \mu^*(s) > 0$ and $r(s,a)$, we have that,*

$$Q_\mu^1 = Q_\mu \qquad \textit{and} \qquad \lim_{N \to \infty} Q_\mu^N = Q_\mu^*$$

Let $\mu^*(s) := \int_{\text{Support}(\pi_\mu^*(a|s))} \mu(a|s)$ denote the probability of optimal actions under $\mu(a|s)$. To show $\lim_{N\to\infty} Q_\mu^N = Q_\mu^*$, we also require the additional assumption that $\inf_s \mu^*(s) > 0$.

Proof. Given that,

$$\mathcal{T}_\mu^1 Q(s,a) := r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N \sim \mu(\cdot|s')} \left[Q(s',a') \right] \tag{22}$$

the unique fixed-point of $\mathcal{T}_\mu^1$ is the Q-value function of the policy $\mu(a|s)$. Hence $Q_\mu^1 = Q_\mu$.

The second part of this theorem will be proven as a Corollary to Theorem 3.5 □

A.3. Increasingly Better Policies

Theorem 3.4. *For all $N, M \in \mathbb{N}$, where $N > M$, we have that $\forall s \in \mathcal{S}, \forall a \in \text{Support}(\mu(\cdot|s))$, $Q_\mu^N(s,a) \geq Q_\mu^M(s,a)$. Hence, $\pi_\mu^N(a|s)$ is at least as good of a policy as $\pi_\mu^M(a|s)$.*

Proof. It is sufficient to show that $\forall s, a, Q_\mu^{N+1}(s,a) \geq Q_\mu^N(s,a)$. We will do so by induction. Let Q^i denote the resulting function after applying $\mathcal{T}_\mu^{N+1}$, i times, starting from Q_μ^N.

Base Case

By definition $Q^0 := Q_\mu^N$. Let $s \in \mathcal{S}$, $a \in \mathcal{A}$.

$$
\begin{aligned}
Q^1(s,a) &= \mathcal{T}_\mu^{N+1} Q^0(s,a) && (23)\\
&= r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^{N+1} \sim \mu(a'|s')} [\max_{\{a_i\}^{N+1}} Q^0(s',a')] && (24)\\
&\geq r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^{N} \sim \mu(a'|s')} [\max_{\{a_i\}^{N}} Q^0(s',a')] && (25)\\
&= r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^{N} \sim \mu(a'|s')} [\max_{\{a_i\}^{N}} Q_\mu^N(s',a')] && (26)\\
&= Q_\mu^N(s,a) && (27)\\
&= Q^0(s,a) && (28)
\end{aligned}
$$

Induction Step

Assume $\forall s, a, Q^i(s,a) \geq Q^{i-1}(s,a)$.

$$
\begin{aligned}
Q^{i+1}(s,a) - Q^i(s,a) &= \mathcal{T}_\mu^{N+1} Q^i(s,a) - \mathcal{T}_\mu^{N+1} Q^{i-1}(s,a) && (29)\\
&= \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^{N+1} \sim \mu(a'|s')} [\max_{\{a_i\}^{N+1}} Q^i(s',a') - \max_{\{a_i\}^{N+1}} Q^{i-1}(s',a')] && (30)\\
&\geq 0 && (31)
\end{aligned}
$$

Hence, by induction we have to $\forall i, j, i > j \implies \forall s, a, Q^i(s,a) \geq Q^j(s,a)$. Since $Q^0 = Q_\mu^N$ and $\lim_{i \to \infty} Q^i = Q_\mu^{N+1}$, we have than $\forall s, a, Q_\mu^{N+1}(s,a) \geq Q_\mu^N(s,a)$. Thus π_μ^{N+1} is a better policy than π_μ^N, and by a simple induction argument, π_μ^N is a better policy than π_μ^M when $N > M$.

□

A.4. Bounds

Theorem 3.5. *For $s \in \mathcal{S}$ let,*

$$\Delta(s) = \max_{a \in \text{Support}(\mu(\cdot|s))} Q_\mu^*(s,a) - \mathbb{E}_{\{a_i\}^N \sim \mu(\cdot|s)} [\max_{b \in \{a_i\}^N} Q_\mu^*(s,b)]$$

The suboptimality of Q_μ^N can be upperbounded as follows,

$$\| Q_\mu^N - Q_\mu^* \|_\infty \leq \frac{\gamma}{1-\gamma} \max_{s,a} \mathbb{E}_{s'} \Big[\Delta(s') \Big] \leq \frac{\gamma}{1-\gamma} \max_s \Delta(s) \tag{32}$$

The same also holds when Q_μ^* is replaced with Q_μ^N in the definition of Δ.

Proof. The two versions where $\Delta(s)$ is defined in terms of Q_μ^N and Q_μ^* have very similar proofs.

Version with Q_μ^N

Let $\mathcal{T}^{QL}$ denote the backup operation in Q-Learning. Let $(\mathcal{T}^{QL})^m = \underbrace{\mathcal{T}^{QL} \circ \mathcal{T}^{QL} \circ ... \circ \mathcal{T}^{QL}}_{m \text{ times}}$. We know the following statements to be true:

$$Q_\mu^N = \mathcal{T}_\mu^N Q_\mu^N = r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N \sim \mu(a'|s')} [\max_{\{a_i\}^N} Q_\mu^N(s', a')] \tag{33}$$

$$\mathcal{T}^{QL} Q_\mu^N = r(s,a) + \gamma \cdot \mathbb{E}_{s'} \max_{a'} Q_\mu^N(s', a') \tag{34}$$

$$\lim_{m \to \infty} (\mathcal{T}^{QL})^m Q_\mu^N = Q^* \tag{35}$$

$$\left\| (\mathcal{T}^{QL})^{m+2} Q_\mu^N - (\mathcal{T}^{QL})^{m+1} Q_\mu^N \right\|_\infty \leq \gamma \cdot \left\| (\mathcal{T}^{QL})^{m+1} Q_\mu^N - (\mathcal{T}^{QL})^m Q_\mu^N \right\|_\infty \tag{36}$$

$$\left\| (\mathcal{T}^{QL})^{m+1} Q_\mu^N - (\mathcal{T}^{QL})^m Q_\mu^N \right\|_\infty \leq \gamma^m \cdot \left\| \mathcal{T}^{QL} Q_\mu^N - Q_\mu^N \right\|_\infty \tag{37}$$

Putting these together we have that,

$$\left\| Q_\mu^N - Q^* \right\|_\infty \leq \sum_{m=0}^{\infty} \left\| (\mathcal{T}^{QL})^{m+1} Q_\mu^N - (\mathcal{T}^{QL})^m Q_\mu^N \right\|_\infty \tag{38}$$

$$\leq \sum_{m=0}^{\infty} \gamma^m \cdot \left\| \mathcal{T}^{QL} Q_\mu^N - Q_\mu^N \right\|_\infty \tag{39}$$

$$= \frac{1}{1-\gamma} \left\| \mathcal{T}^{QL} Q_\mu^N - Q_\mu^N \right\|_\infty \tag{40}$$

$$= \frac{1}{1-\gamma} \max_{s,a} \left(r(s,a) + \gamma \cdot \mathbb{E}_{s'} \max_{a'} Q_\mu^N(s', a') \right) \tag{41}$$

$$- \left(r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N \sim \mu(a'|s')} [\max_{\{a_i\}^N} Q_\mu^N(s', a')] \right) \tag{42}$$

$$= \frac{\gamma}{1-\gamma} \max_{s,a} \mathbb{E}_{s'} \left[\max_{a'} Q_\mu^N(s', a') - \mathbb{E}_{\{a_i\}^N \sim \mu(a'|s')} [\max_{\{a_i\}^N} Q_\mu^N(s', a')] \right] \tag{43}$$

$$\leq \frac{\gamma}{1-\gamma} \max_{s'} \max_{a'} Q_\mu^N(s', a') - \mathbb{E}_{\{a_i\}^N \sim \mu(a'|s')} [\max_{\{a_i\}^N} Q_\mu^N(s', a')] \tag{44}$$

Version with Q_μ^*

Very similarly we have,

$$\left\| Q_\mu^N - Q^* \right\|_\infty \leq \sum_{m=0}^{\infty} \left\| (\mathcal{T}_\mu^N)^{m+1} Q^* - (\mathcal{T}_\mu^N)^m Q^* \right\|_\infty \tag{45}$$

$$\leq \sum_{m=0}^{\infty} \gamma^m \cdot \left\| \mathcal{T}_\mu^N Q^* - Q^* \right\|_\infty \tag{46}$$

$$= \frac{1}{1-\gamma} \left\| Q^* - \mathcal{T}_\mu^N Q^* \right\|_\infty \tag{47}$$

$$= \frac{1}{1-\gamma} \max_{s,a} \left(r(s,a) + \gamma \cdot \mathbb{E}_{s'} \max_{a'} Q^*(s', a') \right) \tag{48}$$

$$- \left(r(s,a) + \gamma \cdot \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N \sim \mu(a'|s')} [\max_{\{a_i\}^N} Q^*(s', a')] \right) \tag{49}$$

$$= \frac{\gamma}{1-\gamma} \max_{s,a} \mathbb{E}_{s'} \left[\max_{a'} Q^*(s', a') - \mathbb{E}_{\{a_i\}^N \sim \mu(a'|s')} [\max_{\{a_i\}^N} Q^*(s', a')] \right] \tag{50}$$

$$\leq \frac{\gamma}{1-\gamma} \max_{s'} \max_{a'} Q^*(s', a') - \mathbb{E}_{\{a_i\}^N \sim \mu(a'|s')} [\max_{\{a_i\}^N} Q^*(s', a')] \tag{51}$$

□

Corollary A.1. *Let V_μ, Q_μ, A_μ denote the value, Q, and advantage functions of μ respectively. When $N = 1$ we have that,*

$$\|Q_\mu - Q^*\|_\infty \leq \frac{\gamma}{1-\gamma} \max_{s'} \max_{a'} Q_\mu(s', a') - \mathbb{E}_{a' \sim \mu(a'|s')}[Q_\mu(s', a')] \tag{52}$$

$$= \frac{\gamma}{1-\gamma} \max_{s'} \max_{a'} Q_\mu(s', a') - V_\mu(s') \tag{53}$$

$$= \frac{\gamma}{1-\gamma} \max_{s', a'} A_\mu(s', a') \tag{54}$$

It is interesting how the sub-optimality can be upper-bounded in terms of a policy's own advantage function.

Corollary A.2. ***(Proof for second part of Theorem 3.3)***

Proof. We want to show $\lim_{N\to\infty} Q_\mu^N = Q^*$. More exactly, what we seek to show is the following,

$$\lim_{N\to\infty} \left\| Q_\mu^N - Q^* \right\|_\infty = 0 \tag{55}$$

or,

$$\forall \epsilon > 0, \exists N, \text{ s.t. } \forall M \geq N, \ \left\| Q_\mu^N - Q^* \right\|_\infty < \epsilon \tag{56}$$

Let $\epsilon > 0$. Recall,

$$\Delta(s) = \max_{a \in \text{Support}(\mu(\cdot|s))} Q_\mu^*(s, a) - \mathbb{E}_{\{a_i\}^N \sim \mu(\cdot|s)}[\max_{b \in \{a_i\}^N} Q_\mu^*(s, b)] \tag{57}$$

Let $\inf_s \mu^*(s) = p > 0$. Let the lower and upper bounds of rewards be ℓ and L, and let $\alpha = \frac{1}{1-\gamma}\ell$ and $\beta = \frac{1}{1-\gamma}L$. We have that,

$$\mathbb{E}_{\{a_i\}^N \sim \mu(\cdot|s)}[\max_{b \in \{a_i\}^N} Q_\mu^*(s, b)] \geq (1-p)^N \cdot \alpha + (1 - (1-p)^N) \cdot \max_{a \in \text{Support}(\mu(\cdot|s))} Q_\mu^*(s, a) \tag{58}$$

Hence $\forall s$,

$$\Delta(s) \leq (1-p)^N \cdot \max_{a \in \text{Support}(\mu(\cdot|s))} Q_\mu^*(s, a) - (1-p)^N \cdot \alpha \tag{59}$$

$$= (1-p)^N \cdot \Big(\max_{a \in \text{Support}(\mu(\cdot|s))} Q_\mu^*(s, a) - \alpha \Big) \tag{60}$$

$$\leq (1-p)^N \cdot \Big(\beta - \alpha \Big) \tag{61}$$

Thus, for large enough N we have that,

$$\left\| Q_\mu^N - Q_\mu^* \right\|_\infty \leq \frac{\gamma}{1-\gamma} \max_s \Delta(s) < \epsilon \tag{62}$$

concluding the proof. □

B. Autoregressive Generative Model

The architecture for our autoregressive generative model is inspired by the works of (Metz et al., 2017; Van de Wiele et al., 2020; Germain et al., 2015). Given a state-action pair from the dataset (s, a), first an MLP produces a d-dimensional embedding for s, which we will denote by h. Below, we use the notation a_i to denote the i^{th} index of a, and $a_{[:i]}$ to represent a slice from first up to and **not including** the i^{th} index, where indexing begins at 0. We use a discretization in each action dimension. Thus, we discretize the range of each action dimension into N uniformly sized bins, and represent a by the labels of the bins. Let ℓ_i denote the label of the i^{th} action index.

Training We use separate MLPs per action dimension. Each MLP takes in the d-dimensional state embedding and ground-truth actions before that index, and outputs N logits for the choice over bins. The probability of a given index's label is given by,

$$p(\ell_i|s, a[:i]) = \text{SoftMax}\Big(\text{MLP}_i(d, a[:i])\Big)[\ell_i] \tag{63}$$

We use standard maximum-likelihood training (i.e. cross-entropy loss).

Sampling Given a state s, to sample an action we again embed the state, and sample the action indices one-by-one.

$$p(\ell_0|s) = \text{SoftMax}\Big(\text{MLP}_i(d)\Big)[\ell_0] \tag{64}$$

$$\ell_0 \sim p(\ell_0|s), a_0 \sim \text{Uniform}(\text{Bin corresponding to } \ell_0) \tag{65}$$

$$p(\ell_i|s) = \text{SoftMax}\Big(\text{MLP}_i(d, a[:i])\Big)[\ell_i] \tag{66}$$

$$\ell_i \sim p(\ell_i|s, a[:i]), a_i \sim \text{Uniform}(\text{Bin corresponding to } \ell_i) \tag{67}$$

C. Algorithm Box

Algorithm 1: Full EMaQ Training Algorithm

Offline dataset $\mathcal{D}$, Pretrain $\mu(a|s)$ on $\mathcal{D}$
Initialize K Q functions with parameters θ_i, and K target Q functions with parameters θ_i^{target}
`Ensemble` parameter λ, Exponential moving average parameter α

Function `Ensemble` (*values*) **:**
  return $\lambda \cdot \min(values) + (1-\lambda) \cdot \max(values)$

Function y_{target} (s, a, s', r, t) **:**
  $\{a'_i\}^N \sim \mu(a'|s')$
  $Qvalues \leftarrow [\ \ ]$
  for $k \leftarrow 1$ **to** N **do**
    `/* Estimate the value of action` a'_k `*/`
    $Qvalues$`.append(Ensemble(`$[Q_i^{target}(s', a'_k)$ for all $i]$`))`
  return $r + (1-t) \cdot \gamma \max(Qvalues)$

while *not converged* **do**
  Sample a batch $\{(s_m, a_m, s'_m, r_m, t_m)\}^M \sim \mathcal{D}$
  for $i = 1, ..., K$ **do**
    $\mathcal{L}(\theta_i) = \sum_m \Big(Q_i(s_m, a_m) - y_{\text{target}}(s_m, a_m, s'_m, r_m, t_m)\Big)^2$
    $\theta_i \leftarrow \theta_i -$ `AdamUpdate`$\Big(\mathcal{L}(\theta_i), \theta_i\Big)$
    $\theta_i^{\text{target}} \leftarrow \alpha \cdot \theta_i^{\text{target}} + (1-\alpha) \cdot \theta_i$

D. Inconclusive Experiments

D.1. Updating the Proposal Distribution

Akin to the work of (Van de Wiele et al., 2020), we considered maintaining a second proposal distribution $\tilde{\mu}$ that is updated to distill $\arg\max_{\{a_i\}^N} Q(s, a)$, and sampling from the mixture of μ and $\tilde{\mu}$. In our experiments however, we did not observe noticeabel gains. This may potentially be due to the relative simplicity of the Mujoco benchmark domains, and may become more important in more challenging domains with more uniformly distributed $\mu(a|s)$.

Algorithm 2: Test-Time Policy π_{test}

Function `TestEnsemble(`*values*`)`**:**
  return $\lambda \cdot \min(values) + (1-\lambda) \cdot \max(values)$

Function $\pi_{\text{test}}(s)$**:**
  $\{a_i\}^N \sim \mu(a|s)$
  return $\arg\max_{\{a_i\}^N}$ `TestEnsemble`$\big([Q_i(s,a) \text{ for all } i]\big)$

E. Laundry List

- Autoregressive models are slow to generate samples from and EMaQ needs to take many samples, so it was slower to train than the alternative methods. However, this may be addressed by better generative models and engineering effort.

F. Online RL

EMaQ is also applicable to online RL setting. Combining strong offline RL methods with good exploration policies has the potential for producing highly sample-efficient online RL algorithms. Concretely, we refer to online RL as the setting where iteratively, a batch of M environment steps with an exploration policy are interleaved with M RL updates (Levine et al., 2020; Matsushima et al., 2020).

EMaQ is designed to remain within the support of the provided training distribution. This however, is problematic for online RL which requires good exploration interleaved with RL updates. To this end, first, we modify our autoregressive proposal distribution $\mu(a|s)$ by dividing the logits of all softmaxes by $\tau > 1$. This has the effect of smoothing the $\mu(a|s)$ distribution, and increasing the probability of sampling actions from the low-density regions and the boundaries of the support. Given this online proposal distribution, a criteria is required by which to choose amongst sampled actions. While there exists a rich literature on how to design effective RL exploration policies (Weng, 2020), in this work we used a simple UCB-style exploration criterion (Chen et al., 2017) as follows:

$$Q^{\text{explore}}(s,a) = \text{mean}\Big(\{Q_i(s,a)\}_K\Big) + \beta \cdot \text{std}\Big(\{Q_i(s,a)\}_K\Big) \tag{68}$$

Given N sampled actions from the modified proposal distribution, we take the action with highest Q^{explore}.

We compare the online variant of EMaQ with entropy-constrained Soft Actor Critic (SAC) with automatic tuning of the temperature parameter (Haarnoja et al., 2018). For EMaQ we swept the temperatures and used a fixed bin size of 40, 8 Q-function ensembles and $N = 200$. For fairness of comparisons, we also ran SAC with similar sweeps over different collection batch sizes and number of Q-function ensembles. In the fully online setting (trajectory batch size 1, Figure 3a), EMaQ is already competitive with SAC, and more excitingly, in the deployment-efficient setting[5] (trajectory batch size 50K, Figure 3b), EMaQ can outperform SAC[6]. Figures 4 and 5 present the results for all hyperparameter settings, for SAC and EMaQ, in the batch size 1 and batch size $50K$ settings respectively. **In the fully online setting, EMaQ is already competitive with SAC, and more excitingly, in the deployment-efficient setting, EMaQ can outperform SAC.**

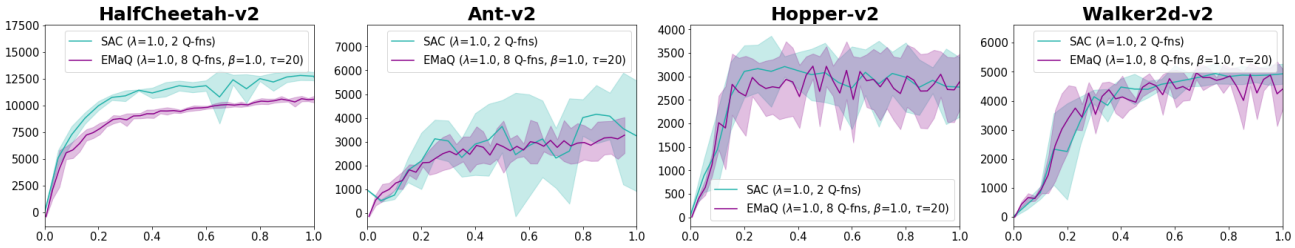
G. Offline RL Experimental Details

For each environment and data setting, we train an autoregressive model – as described above – on the provided data with 2 random seeds. These generative models are then frozen, and used by the downstream algorithms (EMaQ, BEAR, and BCQ) as the base behavior policy ($\mu(a|s)$ in EMaQ)[7].

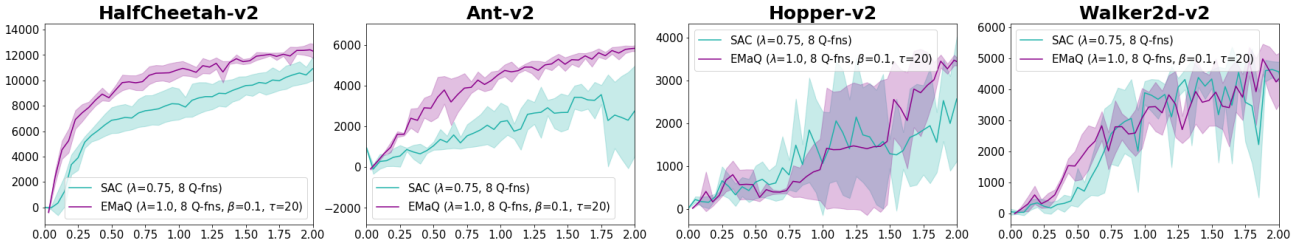
[5] By deployment-efficient we mean that less number of different policies need to be executed in the environment, which may have substantial benefits for safety and otherwise constrained domains (Matsushima et al., 2020).

[6] It must be noted that the online variant of EMaQ has more hyperparameters to tune, and the relative performance is dependent on these hyperparameters, while SAC with ensembles has the one extra ensemble mixing parameter λ to tune.

[7] While in the original presentation of BCQ and BEAR the behvior policy is learned online, there is technically no reason for this to be the case, and in theory both methods should benefit from this pretraining



(a) SAC vs. EMaQ, Trajectory Batch Size 1: For easier visual interpretration we plot a single hyperparameter setting of EMaQ that tended to perform well across the 4 domains considered. The hyperparameters considered were $N = 200$, $\lambda = 1.0$, $\beta = 1.0$, $\tau \in \{1, 5, 10, 20\}$. SAC performed worse when using 8 Q-functions as in EMaQ. x-axis unit is 1 million environment steps.



(b) SAC vs. EMaQ, Trajectory Batch Size 50K: For easier visual interpretration we plot a single hyperparameter setting of EMaQ that tended to perform well across the 4 domains considered. The hyperparameters considered were $N = 200$, $\lambda \in \{0.75, 1.0\}$, $\beta \in \{0.1, 1.0\}$, $\tau \in \{1, 5, 10, 20\}$. x-axis unit is 1 million environment steps.

Figure 3. Online RL results under different trajectory batch sizes.

G.1. Comparing Offline RL Methods

Following the bechmarking efforts of (Wu et al., 2019), the range of clipping factor considered for BCQ was $\Phi \in \{0.005, 0.015, 0.05, 0.15, 0.5\}$, and the range of target divergence value considered for BEAR was $\epsilon \in \{0.015, 0.05, 0.15, 0.5, 1.5\}$. For both methods, the larger the value of the hyperparameter is, the more the learned policy is allowed to deviate from the $\mu(a|s)$.

The rest of the hyperparameters use can be found in Table 2. The autoregressive models have the following architecture sizes (refer to Appendix B for description of the models used). The state embedding MLP consists of 2 hidden layers of dimension 750 with relu activations, followed by a linear embedding into a 750 dimensional state representation. The individual MLP for each action dimension consist of 3 hidden layers of dimension 256 with relu activations. Each action dimension is discretized into 40 equally sized bins.

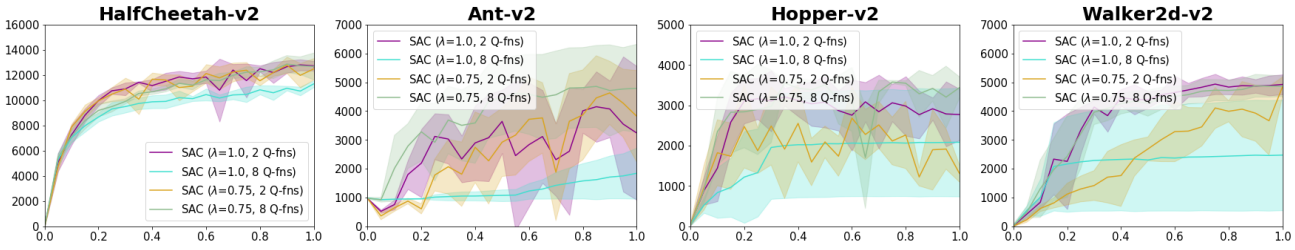
G.2. EMaQ Ablation Experiment

Hyperparameters are identical to those in Table 2, except batch size is 100 and number of updates is 500K.

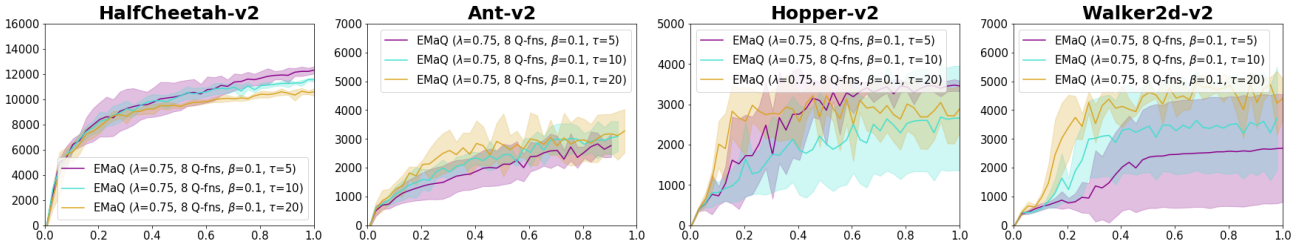
G.3. Details for Table 1 Experiments

Generative Model The generative models used are almost identical to the description in Appendix B, with a slight modification that $\text{MLP}_i(d, a[:i])$ is replace with $\text{MLP}_i(d, \text{Lin}_i(a[:i]))$ where Lin_i is a linear transformation. This change was not necessary for good performance; it was as architectural detail that we experimented with and did not revert prior generating Table **??**. The model dimensions for each domain are shown in 3 in the following format (state embedding MLP hidden size, state embedding MLP number of layers, action MLP hidden size, action MLP number of layers, Ouput size of Lin_i, number of bins for action discretization). Increasing the number of discretization bins from 40 (value for standard Mujoco experiments) to 80 was the most important change. Output dimension of state-embedding MLP is the same as the hidden size.

Hyperparameters Table 3 shows the hyperparameters used for the experiments in Table 1.



(a) SAC batch 1 results



(b) EMaQ batch 1 results

Figure 4. All results for batch size 1

H. VAE Results

H.1. Implementation

We also ran experiments with VAE parameterizations for $\mu(a|s)$. To be approximately matched in parameter count with our autoregressive models, the encoder and decoder both have 3 hidden layers of size 1024 with relu activations. The dimension of the latent space was twice the number of action dimensions. The decoder outputs a vector v which, and the decoder action distribution is defined to be $\mathcal{N}(\text{Tanh}(v), I)$. When sampling from the VAE, following prior work, samples from the VAE prior (spherical normal distribution) were clipped to the range $[-0.5, 0.5]$ and mean of the decoder distibution was used (i.e. the decoder distribution was not sampled from). The KL divergence loss term was weighted by 0.5. This VAE implementation was the one used in the benchmarking codebase of (Wu et al., 2019), so we did not modify it.

H.2. Results

As can be seen in Figure 6, EMaQ has a harder time improving upon $\mu(a|s)$ when using the VAE architecture described above. However, as can be seen in Figure 7, BCQ and BEAR do show some variability as well when switching to the VAEs. Since as an algorithm EMaQ is much more reliant on $\mu(a|s)$, our hypothesis is that if it is true that the autoregressive models better captured the action distribution, letting EMaQ not make poor generalizations to out-of-distribution actions. Figures 8 and 9 show autoregressive and VAE results side-by-side for easier comparison.

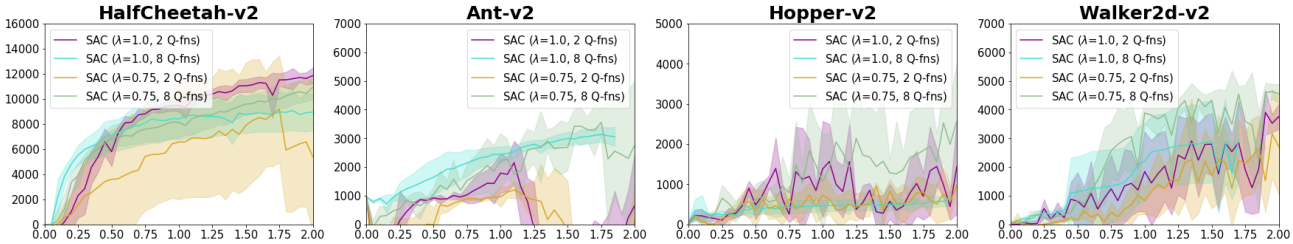
I. EMaQ Medium-Expert Setting Results

In HalfCheetah, increasing N significantly slows down the convergence rate of the training curves; while large Ns continue to improve, we were unable to train them long enough for convergence. In Walker, for EMaQ, BCQ, and most hyperparameter settings of BEAR, training curves have a prototypical shape of a hump, where performance improves up to a certain high value, and then continues to fall very low. In Hopper, for higher values of N in EMaQ we observed that increasing batch size from 100 to 256 largely resolved the poor performance, but for consistency we did not alter Figure 1 with these values.

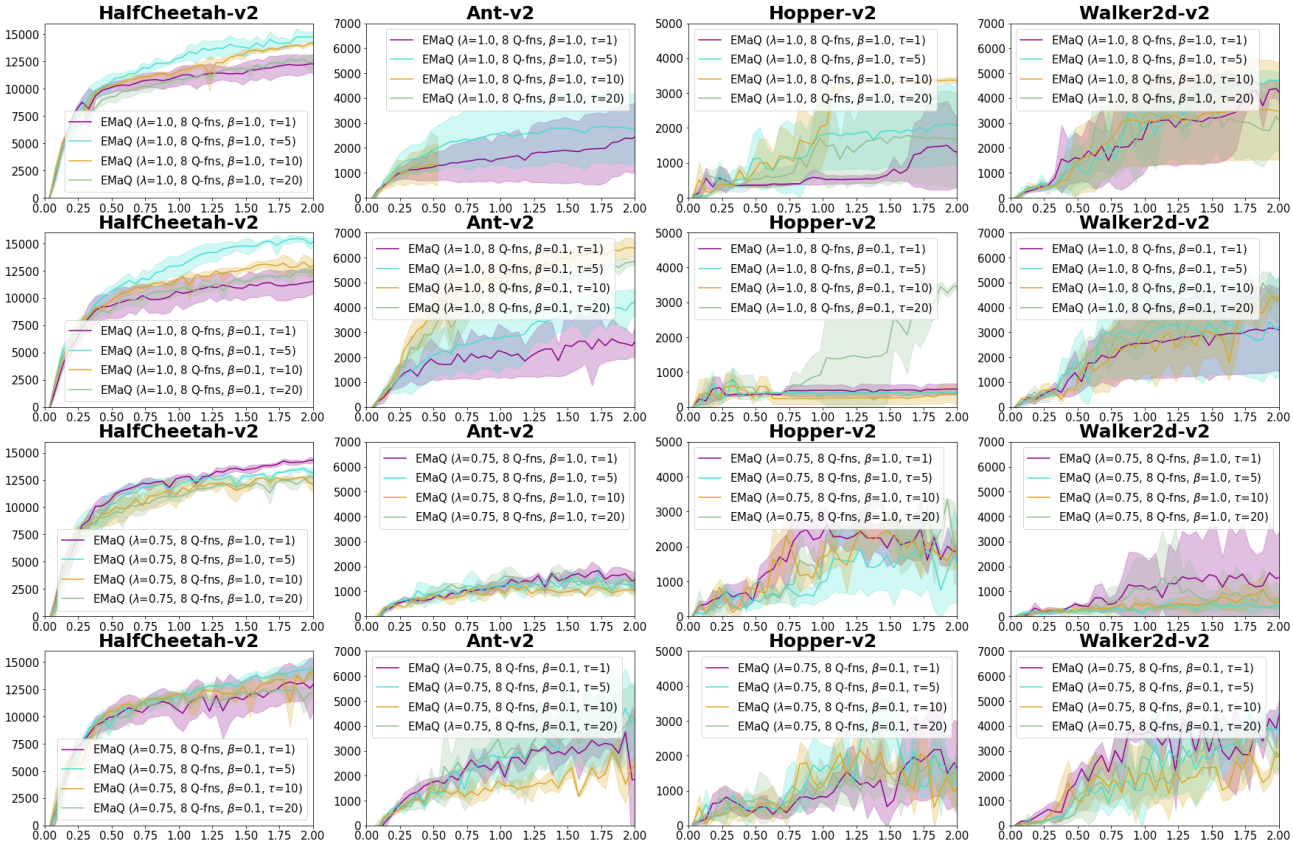
J. Comparison with Softmax Backup Operators

For clarity of writing, we will write the forms for deterministic dynamics and remove the expectations over the next state.

An interesting connection to our proposed backup operators would be the following Softmax backup operator with similarities



(a) SAC batch 50K results



(b) EMaQ batch 50K results

Figure 5. All results for batch size 50K

Shared Hyperparameters	
λ	1.0
Batch Size	256
Num Updates	1e6
Num Q Functions	8
Q Architecture	MLP, 3 layers, 750 hid dim, relu
μ lr	5e-4
α	0.995
EMaQ Hyperparameters	
Q lr	1e-4
BEAR Hyperparameters	
π Architecture	MLP, 3 layers, 750 hid dim, relu
Q lr	1e-3
π lr	3e-5
BCQ Hyperparameters	
π Architecture	MLP, 3 layers, 750 hid dim, relu
Q lr	1e-4
π lr	5e-4

Table 2. Hyperparameters for Mujoco Experiments

to EMaQ,

$$\mathcal{T}_\mu^\alpha Q(s,a) := r(s,a) + \mathbb{E}_{soft(a'|s')}[Q(s',a')] \tag{69}$$

$$soft(a|s) \propto \mu(a|s) \cdot \exp(\alpha \cdot Q(s,a)) \tag{70}$$

The policy corresponding to $soft(a|s)$ is a policy that aims to maximize Q-values, subject to a KL-constraint between itself and the policy $\mu(a|s)$. The looser the constraint, the larger the effective α and the farther the policy will be from μ. One approach to Monte Carlo estimation of the expectation on the right hand side could be to take samples using methods from the energy-based generative modelling literature.

An alternative approach which will more closely resembles EMaQ is to use self-normalized importance sampling,

$$\mathcal{T}_\mu^\alpha Q(s,a) := r(s,a) + \mathbb{E}_{soft(a'|s')}[Q(s',a')] \tag{71}$$

$$= r(s,a) + \sum_{\{a_i\}^N \sim \mu(a'|s')} w_i \cdot Q(s',a') \tag{72}$$

$$\tilde{w}_i = \frac{\mu(a_i'|s') \cdot \exp(\alpha \cdot Q(s',a_i'))}{\mu(a_i'|s')} \tag{73}$$

$$w_i = \frac{\tilde{w}_i}{\sum \tilde{w}_i} \tag{74}$$

$$= softmax(\alpha \cdot Q(s',a_i'))[i] \tag{75}$$

In this form, the soft backup is similar to EMaQ, where instead of taking ths max Q-value over the N samples, we take an average over the N Q-values, weighted by the softmax probabilities in equation 75. For a given N, the $\alpha = 0$ would be equivalent to Q-evaluation of the policy $\mu(a|s)$, and as $\alpha \to \infty$, the soft backups approach EMaQ backups.

In Figure 10 we present empirical results with the soft backup operators, under a large range $\alpha \in \{1, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$, in the Halfcheetah settings. The EMaQ and soft-EMaQ were run with the same architectures, but were smaller than the ones used for the results in the main text. We used the same checkpoints of the generative models as for the results in the main text. The test-time policy for both approaches is the same, sampling N actions and taking the argmax action under the ensemble Q-value. The only difference between the EMaQ and soft-EMaQ implementations was a one-line change to replace `max` with a softmax average of the Q-values.

Some interesting observations are the following: As anticipated, the soft EMaQ backups approach EMaQ as the value of α is increased. However, the necessary value of α to match the performance of EMaQ can be quite large. In the medium-expert

Shared Hyperparameters	
λ	1.0
Batch Size	128
Num Updates	1e6
Num Q Functions	16
Q Architecture	MLP, 4 layers, 256 hid dim, relu
α	0.995
μ lr	5e-4
Kitchen μ Arch Params	$(256, 4, 128, 1, 128, 80)$
Antmaze μ Arch Params	$(256, 4, 128, 1, 128, 80)$
Adroit μ Arch Params	$(256, 4, 128, 1, 128, 80)$
EMaQ Hyperparameters	
Q lr	1e-4
Kitchen N's Searched	$\{4, 8, 16, 32, 64\}$
Antmaze N's Searched	$\{50, 100, 150, 200\}$
Adroit N's Searched	$\{16, 32, 64, 128\}$
BEAR Hyperparameters	
π Architecture	MLP, 4 layers, 256 hid dim, relu
Q lr	1e-4
π lr	5e-4
BCQ Hyperparameters	
π Architecture	MLP, 4 layers, 256 hid dim, relu
Q lr	1e-4
π lr	5e-4

Table 3. Hyperparameters for Table 1 Experiments

setting, where figure 1 suggests challenges arising from the combination of large Ns and function-approximators, we did not gain much advantage from soft backups, and only $\alpha \in \{8, 16, 32\}$ seem to have provided some mitigation of the problem for $N = 25$. Since the soft backup introduces an additional hyperparameter that cannot be determined ahead of time, and does not seem to provide an advantage (at least in the limited Halfcheetah settings considered), from a practical perspective, we would prefer to use the regular EMaQ backup.

K. Examining Sample-Max Regularization in a Simplified Setting

K.1. Exact TD Backups, Infinite Batch Data, Inexact Behavior Estimate

For simplification of notation and equations, in this section we assume that dynamics are deterministic.

Let π_β denote the true behavior policy. Let $d_\beta(s, a)$ denote the state-action distribution of π_β. Let μ denote our estimate of the behavior policy which may have inaccuracies, namely, it can sample actions outside the support of π_β, and hence reach unobserved states. Assuming that we can perform exact backups in each iteration (no function approximation assumption), but taking into account that $\pi_\beta \neq \mu$, as each iteration we have the following EMaQ backup:

$$\forall s, a \in \texttt{Support}(d_\beta(s, a)), \mathcal{T}_\mu^N Q(s, a) := r(s, a) + \gamma \cdot \mathbb{E}_{\{a'_i\}^N \sim \mu(\cdot|s')} \left[\max_{a' \in \{a'_i\}^N} Q(s', a') \right] \tag{76}$$

The main difference between this and Equation 5 is that in 5, due to the assumption that μ was the behavior policy π_β, we implicitly had $\forall s, a \in \texttt{Support}(d_\mu(s, a))$. Now that $\pi_\beta \neq \mu$, in this scenario of exact backups without function approximation, the Q-values for $s, a \notin \texttt{Support}(d_\beta(s, a))$ will never be updated from their original values at initialization.

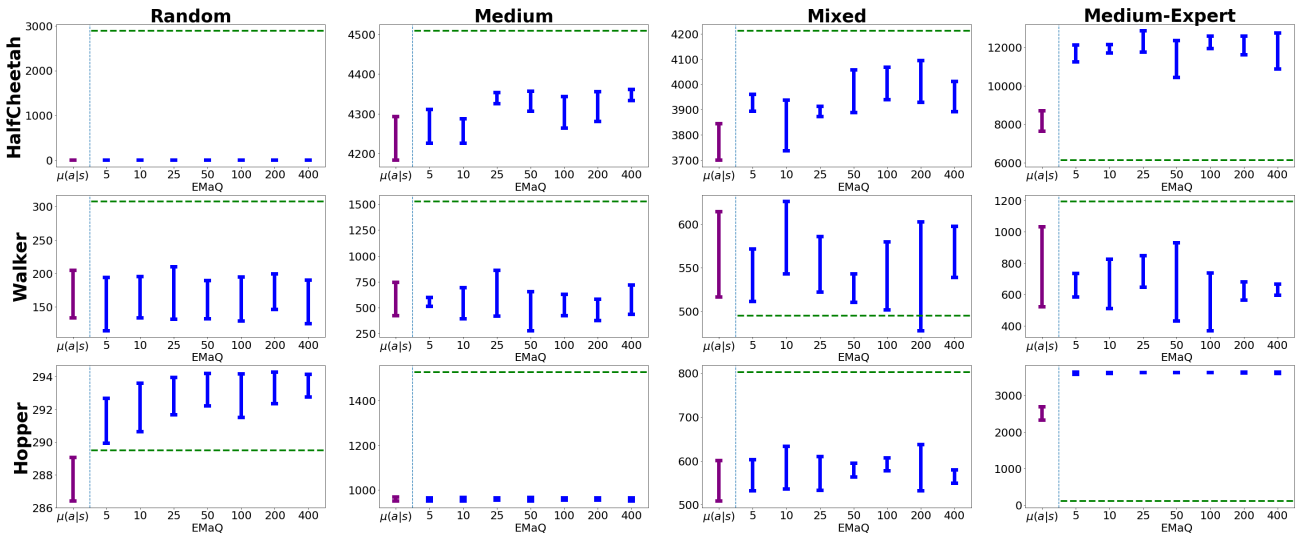


Figure 6. Results for evaluating EMaQ on D4RL (Fu et al., 2020b) benchmark domains when using the described VAE implementation, with $N \in \{5, 10, 25, 50, 100, 200, 400\}$. Values above $\mu(a|s)$ represent the result of evaluating the base behavior policies. Horizontal green lines represent the reported performance of BEAR in the D4RL benchmark (apples to apples comparisons in Figure 7).

For $s \in \texttt{Support}(d_\beta(s,a))$, let $p(s)$ denote the probability that $\mu(a|s)$ is inside the support of $\pi_\beta(a|s)$. We now have:

$$\mathcal{T}_\mu^N Q(s,a) := r(s,a) + p(s')^N \cdot \gamma \cdot \mathbb{E}_{\{a'_i\}^N \in \text{ support}} \left[\max_{a' \in \{a'_i\}^N} Q(s',a') \right] + (1 - p(s')^N) \cdot \gamma \cdot \mathbb{E}_{\text{else}} \left[\max_{a' \in \{a'_i\}^N} Q(s',a') \right] \tag{77}$$

The interaction between the out-of-distribution Q-values and the backups is a very complex one, which depends on the Q-values for in vs. out of distribution actions, at any given iteration during training. We can however try to consider hypothetical adversarial scenarios to gain a better sense of the behavior of the backups.

As discussed above, the Q-values for out-of-distribution (OOD) actions do not change from their initialization value – in the current setting of exact backups without function approximation. Let us imagine a particularly unfortunate scenario, where for all OOD actions we have initialized $Q(s,a) = \frac{1}{1-\gamma} R_{max}$ while in "reality", all such actions end in a terminal state with return $\frac{1}{1-\gamma} R_{min}$. Hence:

$$\mathcal{T}_\mu^N Q(s,a) := r(s,a) + p(s')^N \cdot \gamma \cdot \mathbb{E}_{\{a'_i\}^N \in \text{ support}} \left[\max_{a' \in \{a'_i\}^N} Q(s',a') \right] + (1 - p(s')^N) \cdot \frac{\gamma}{1-\gamma} R_{max} \tag{78}$$

This results in a state-dependent bonus reward and a state-dependent discount $p(s')^N \cdot \gamma$. If $p(s')$ is very unbalanced across states, some states with higher likelihood of OOD actions would receive a larger bonus than others which would be problematic. On the other hand, if $p(s')$ is relatively constant across states, the bonus reward would be relatively similar across states; **however the discount would still be $p(s')^N \cdot \gamma$ which is more myopic than γ.** This suggests for the existence of a myopic bias, and the experimentally one should increase the value of γ used. We have not done experiments with this adjustment, however investigating the effect of this matter on the Antmaze medium and large domains – which explicitly test for extremely long-range sparse rewards (reward of 1 at goal position and 0 everywhere else) – may lead to non-trivial gains in performance.

K.2. Moving Closer to Function Approximation Setting

To move closer to the function approximation setup, we can continue to make the assumptions from the previous section, except that the Q-values for out-of-distribution (OOD) actions are somehow related to the values for actions inside the

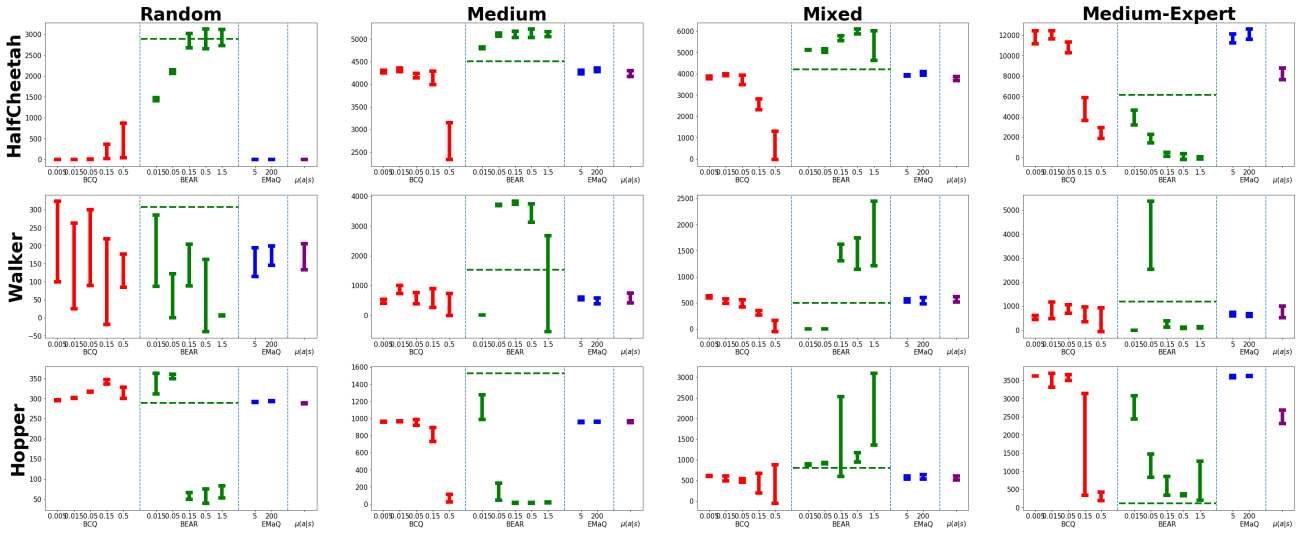


Figure 7. Comparison of EMaQ, BCQ, and BEAR on D4RL (Fu et al., 2020b) benchmark domains when using when using the described VAE implementation for $\mu(a|s)$. For both BCQ and BEAR, from left to right the allowed deviation from $\mu(a|s)$ increases. Horizontal green lines represent the reported performance of BEAR in the D4RL benchmark.

support, and not fixed at the initialization value. To try to study this setup in a more pessimistic setting, assume that for a given state, the values of all OOD actions is the maximum Q-value inside the π_β support, $Q(s)_{max}$, plus a positive state-dependent term $f(s)$. Similar to above, after rearranging, we have:

$$\mathcal{T}_\mu^N Q(s,a) := [r(s,a) + (1-p(s')^N)\cdot\gamma\cdot f(s')] + p(s')^N\cdot\gamma\cdot\mathbb{E}_{\{a'_i\}^N\in\text{ support}}\left[\max_{a'\in\{a'_i\}^N} Q(s',a')\right] + (1-p(s')^N)\cdot\gamma\cdot Q(s')_{max} \tag{79}$$

K.3. Qualitative Difference in Training Curves

In Appendix section L we present a short discussion on qualitative observations comparing the training curves of EMaQ vs. BCQ, demonstrating stability in EMaQ training.

L. Qualitative Differences in Training Curves

We have sometimes observed that the curves representing agent performance throughout training can be significantly more stable under EMaQ in comparison to BEAR and BCQ. A domain where the differences are particularly striking are the `antmaze-umaze` and `antmaze-umaze-diverse` domains. In Figure 4 we have included plots of agent performance during training under the variety of considered hyperparameters and random seeds. It can be seen that in these two domains, initially the BCQ agents improves in performance close to the performance of EMaQ, and the drastically degrades with more training. In constrast, EMaQ agents remain stable even after twice as many training iterations as BCQ, which may indicate the downside of the heuristic perturbation model for constraining actions.

M. Larger Plots for Visibility

Due to larger size of plots, each plot is shown on a separate page below. For ablation results, see Figure 11. For MuJoCo results, see Figure 12.

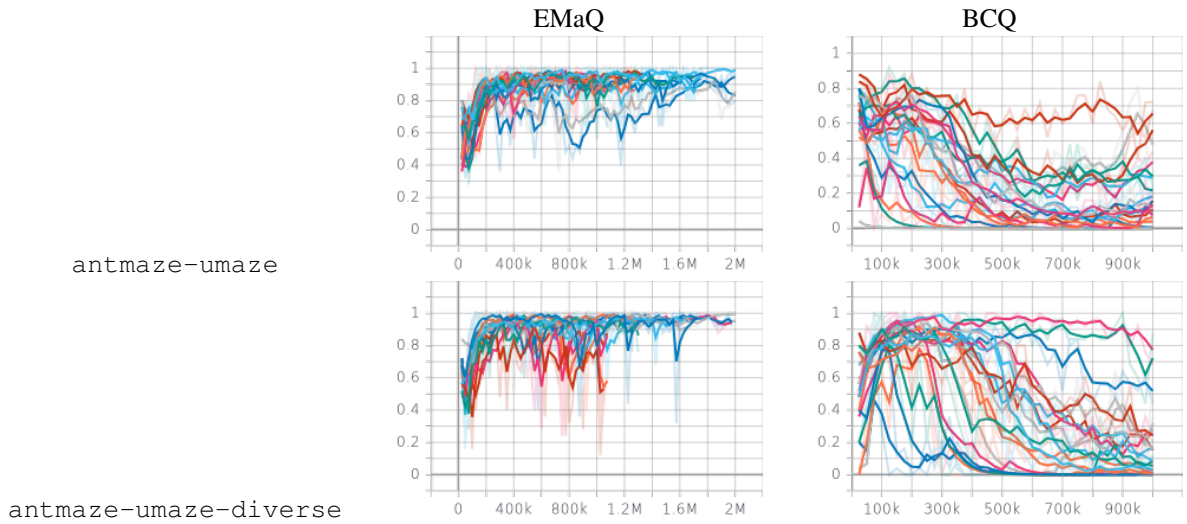


Table 4. Comparison of agent returns throughout training, under the variety of hyperparameters and and random seeds, in the small ant domains. We observe that EMaQ is significantly more stable than BCQ in these domains, even though the values of N in EMaQ were fairly large for these plots $N \in \{50, 100, 150, 200\}$.

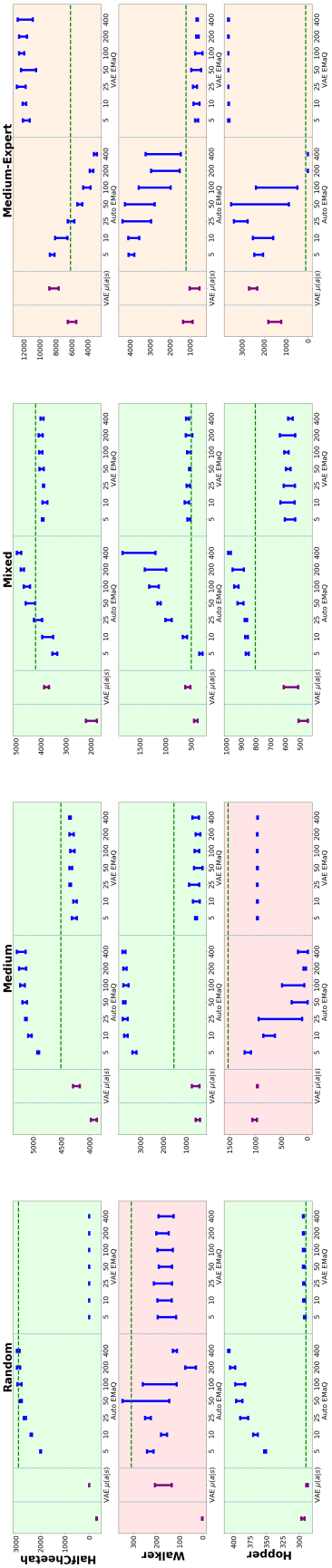


Figure 8. Results with both autoregressive and VAE models in one plot for easier comparison.

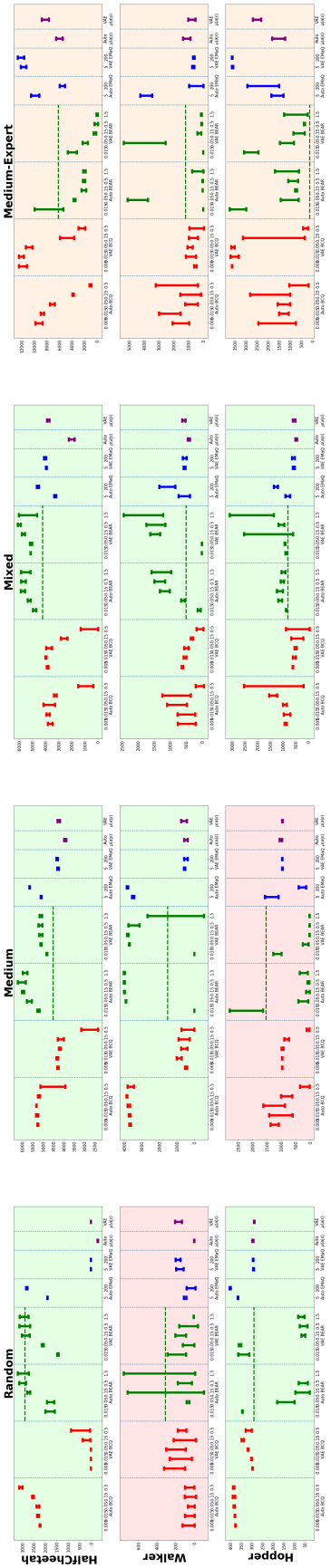


Figure 9. Results with both autoregressive and VAE models in one plot for easier comparison.

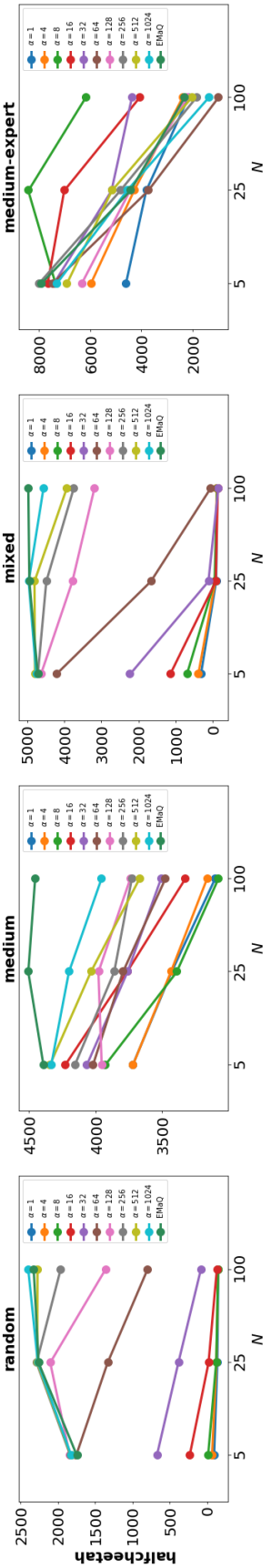


Figure 10. Comparison of Soft-EMaQ with EMaQ under a large range of hyperparameters in the Halfcheetah domain.

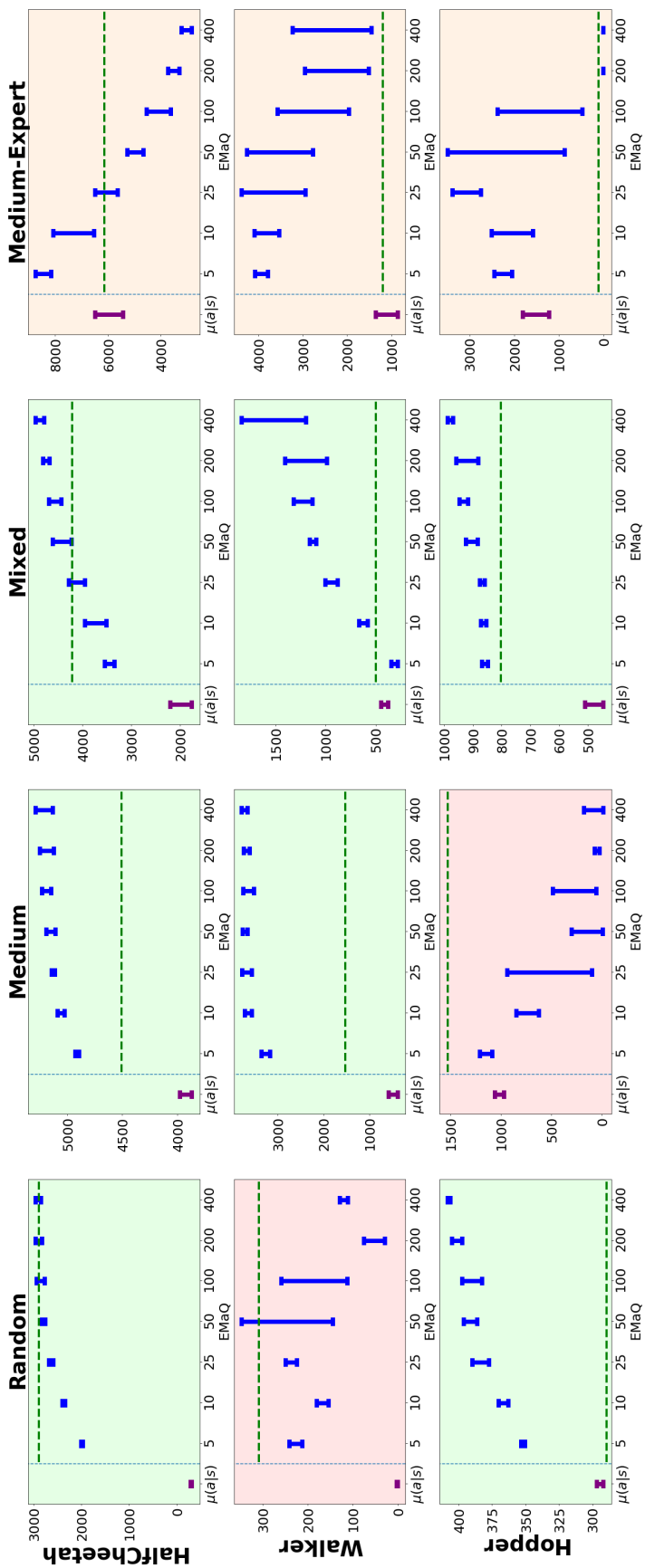


Figure 11. Results for evaluating EMaQ on D4RL (Fu et al., 2020b) benchmark domains, with $N \in \{5, 10, 25, 50, 100, 200, 400\}$. Values above $\mu(a|s)$ represent the result of evaluating the base behavior policies. Horizontal green lines represent the reported performance of BEAR in the D4RL benchmark (apples to apples comparisons in Figure 2). Refer to main text (Section 5.1) for description of color-coding.

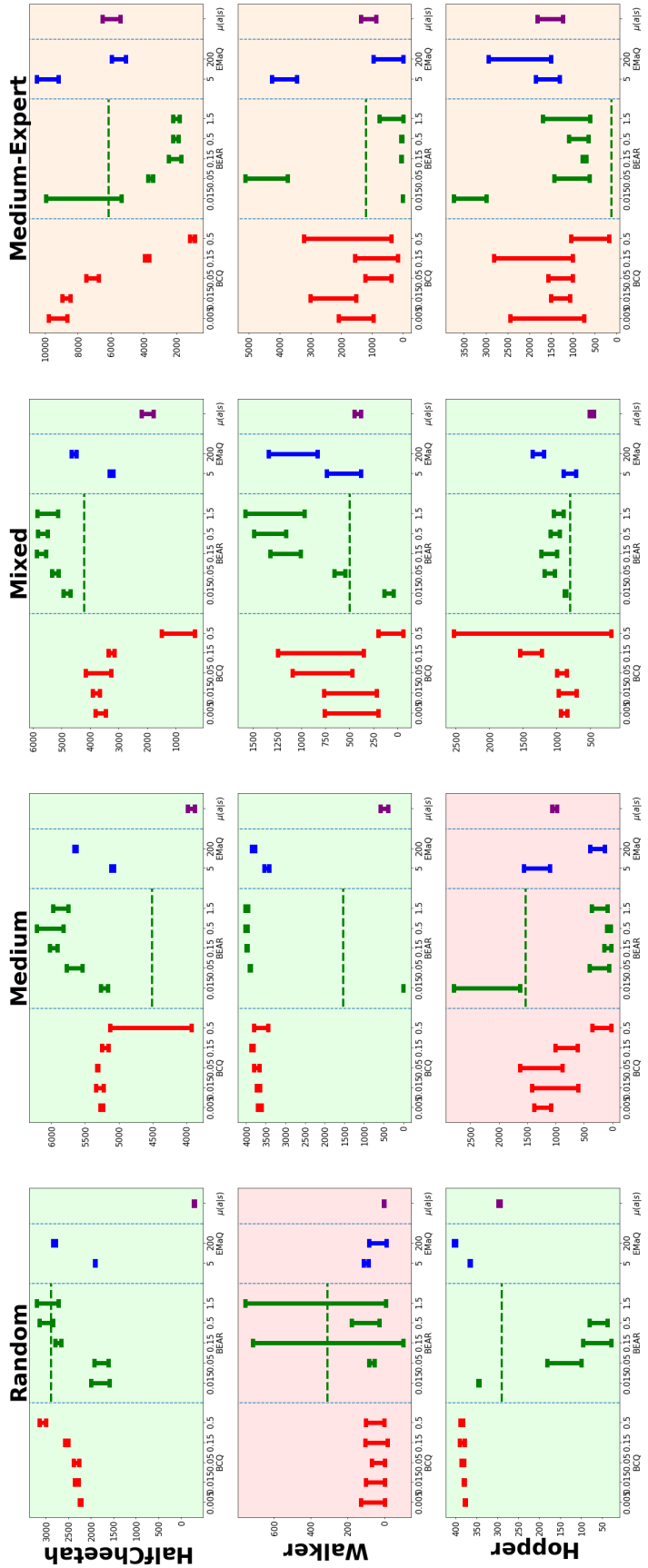


Figure 12. Comparison of EMaQ, BCQ, and BEAR on D4RL (Fu et al., 2020b) benchmark domains when using our proposed autoregressive $\mu(a|s)$. For both BCQ and BEAR, from left to right the allowed deviation from $\mu(a|s)$ increases. Horizontal green lines represent the reported performance of BEAR in the D4RL benchmark. Color-coding follows Figure 1.