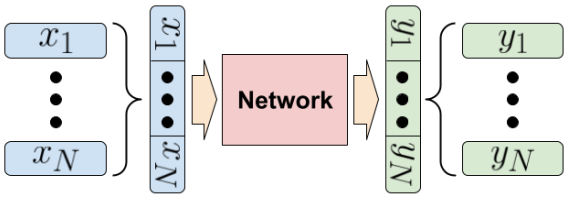
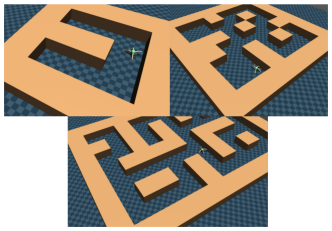


(a) Visual depiction of MIMO Ensemble



(b) D4RL antmaze tasks. Figure taken from Fu et al. (2020).

A Efficient Ensembles & Ensemble Ablations Table

Domain	CQL	Deep Ens.	MIMO-0.5	MIMO-1	Multi-Head	Batch Ens	$N = 1$	Shared LCB	Shared Min
halfcheetah-medium	50.1	50.6	40.0	44.4	49.8	40.2	48.6	47.9	36.0
hopper-medium	90.2	88.2	60.2	65.1	94.8	41.6	50.2	30.0	43.6
walker2d-medium	82.8	83.6	80.7	83.6	84.5	13.2	84.9	85.5	0.0
halfcheetah-mixed	47.6	52.1	0.0	6.7	49.4	43.0	49.6	51.0	2.3
hopper-mixed	61.1	82.1	22.7	32.5	45.5	35.0	30.5	32.5	6.7
walker2d-mixed	20.5	24.0	14.5	16.8	15.9	12.0	11.5	26.3	0.0
antmaze-large-diverse	15.8	64.8	11.0	26.0	33.5	0.0	noisy 30.5	0.0	0.0
antmaze-large-play	14.9	68.8	7.5	20.5	27.5	0.0	noisy 11.0	0.0	0.0

Table 3: Results for efficient ensembles and ensemble ablations.

B D4RL Gym Details

All policies and Q-functions are a 3 layer neural network with relu activations and hidden layer size 256. The policy output is a normal distribution that is squashed to $[-1, 1]$ using the tanh function. All methods were trained for 1M steps. CQL and MSG are trained with behavioral cloning (BC) for the first 50K steps. F-BRC pretrains with 1M steps of BC.

B.1 Hyperparameter Search

For all methods we performed hyperparameter search using 2 seeds, and for the final choice of hyperparameter ran experiments with 5 new random seed.

MSG $\beta \in \{0., -1., -2., -4. - 8.\}, \alpha \in \{0., 0.1, 0.5, 1., 2.\}$

CQL $\alpha \in \{0., 0.1\} + np.exp(np.linspace(np.log(0.1), np.log(10.), steps = 23))$

F-BRC $\alpha \in \{0.\} + np.exp(np.linspace(np.log(0.01), np.log(10.0), steps = 24))$

C Antmaze Details

We use the same hyperparameter search procedure as Gym results, with same architectures. The only difference is that models are trained for 2M steps and at evaluation time they are rolled out for 100 episodes instead of 10.

In prior work, the rewards in the offline dataset are converted using the formula $4(r - 0.5)$. We also use the same reward transformation.

D THEORY

For our notation to match Lee et al. (2019), throughout this section we use the f to denote the network Q, and we use x instead of (s, a).

In Lee et al. (2019) it is shown that when training an infinitely wide neural network to perform regression using mean squared error, subject to technical conditions on the learning rate used, the predictions of the trained network are equivalent to if we had linearized (Taylor expanded) the network at its initialization, and performed trained the linearized network instead. This means that after t iterations of our policy evaluation procedure, $\forall x, t, \quad f_t^{\text{lin}}(x) = f_t(x)$. Hence we only need to study the evolution of the linearized network f^{lin}.

The theorems in the main manuscript are direct corollaries of the following two derivations.

D.1 EACH HAVE THEIR OWN TARGET

For a single – infinitely wide – ensemble member, using the equations in Lee et al. (2019) (section 2.2, equations 9-10-11) we can write the following recursive updates for our policy evaluation procedure

$$\hat{\Theta}_0^{-1} := \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \tag{5}$$
$$C := \hat{\Theta}_0(\mathcal{X}', \mathcal{X})\hat{\Theta}_0^{-1} \tag{6}$$
$$f_{t+1}^{\text{lin}}(\mathcal{X}) = \mathcal{Y}_t \tag{7}$$
$$= \mathcal{R} + \gamma f_t^{\text{lin}}(\mathcal{X}') \tag{8}$$
$$\forall x, f_{t+1}^{\text{lin}}(x) = f_0(x) + \hat{\Theta}_0(x, \mathcal{X})\hat{\Theta}_0^{-1}(\mathcal{Y}_t - f_0(\mathcal{X})) \tag{9}$$
$$\mathcal{Y}_{t+1} = \mathcal{R} + \gamma f_t^{\text{lin}}(\mathcal{X}') \tag{10}$$
$$f_{t+1}^{\text{lin}}(\mathcal{X}') = f_0(\mathcal{X}') + \hat{\Theta}_0(\mathcal{X}', \mathcal{X})\hat{\Theta}_0^{-1}(\mathcal{Y}_t - f_0(\mathcal{X})) \tag{11}$$
$$= f_0(\mathcal{X}') + \hat{\Theta}_0(\mathcal{X}', \mathcal{X})\hat{\Theta}_0^{-1}(\mathcal{R} + \gamma f_t^{\text{lin}}(\mathcal{X}') - f_0(\mathcal{X})) \tag{12}$$
$$= f_0(\mathcal{X}') + C\mathcal{R} + \gamma C f_t^{\text{lin}}(\mathcal{X}') - C f_0(\mathcal{X}) \tag{13}$$
$$= f_0(\mathcal{X}') + C\mathcal{R} - C f_0(\mathcal{X}) + \gamma C f_t^{\text{lin}}(\mathcal{X}') \tag{14}$$
$$= \dots \tag{15}$$
$$= (1 + \ldots + \gamma^t C^t)\Big(f_0(\mathcal{X}') + C\mathcal{R} - C f_0(\mathcal{X})\Big) + (\gamma C)^{t+1} f_0(\mathcal{X}') \tag{16}$$

In the infinite-width regime, the above equations imply the following distributions,

$$\mathop{\mathbb{E}}_{ensemble}[\mathcal{Y}_t] = (1 + \ldots + \gamma^t C^t)\mathcal{R} \tag{17}$$
$$\forall x, \mathop{\mathbb{E}}_{ensemble}[f_{t+1}^{\text{lin}}(x)] = (1 + \ldots + \gamma^t C^t)C\mathcal{R} \tag{18}$$

In addition to the formulae for the expectations above, a key observation from equation 16 to note is that the variance/covariance of terms of interest is modulated by this expression:

$$(1 + \ldots + \gamma^t C^t) \tag{19}$$

This means that the uncertainty being "backed-up/accumulated" through dynamic programming.

D.2 THEY ALL SHARE THE TARGET

Let us consider the setting where all ensemble memebers use their mean as the target.

$$\hat{\Theta}_0^{-1} := \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \tag{20}$$

$$C := \hat{\Theta}_0(\mathcal{X}', \mathcal{X})\hat{\Theta}_0^{-1} \tag{21}$$

$$\forall x, f_{t+1}^{\text{lin}}(x) = f_0(x) + \hat{\Theta}_0(x, \mathcal{X})\hat{\Theta}_0^{-1}(\mathcal{Y}_t - f_0(\mathcal{X})) \tag{22}$$

$$\mathcal{Y}_{t+1} = \mathcal{R} + \gamma \mathop{\mathbb{E}}_{ensemble}[f_t^{\text{lin}}(\mathcal{X}')] \tag{23}$$

$$= \mathcal{R} + \gamma C \mathcal{Y}_t \tag{24}$$

$$= \dots \tag{25}$$

$$= (1 + \ldots + \gamma^{t+1} C^{t+1})\mathcal{R} + \mathbb{E}[\gamma^{t+2} C^{t+1} f_0(\mathcal{X}')] \tag{26}$$

$$= (1 + \ldots + \gamma^{t+1} C^{t+1})\mathcal{R} \tag{27}$$

$$f_{t+1}^{\text{lin}}(\mathcal{X}) = \mathcal{Y}_t \tag{28}$$

$$\forall x, f_{t+1}^{\text{lin}}(x) = f_0(x) + C\Big((1 + \ldots + \gamma^t C^t)\mathcal{R} - f_0(\mathcal{X})\Big) \tag{29}$$

In the infinite-width regime, the above equations imply the following distributions,

$$\mathop{\mathbb{E}}_{ensemble}[\mathcal{Y}_t] = (1 + \ldots + \gamma^t C^t)\mathcal{R} \tag{30}$$

$$\forall x, \mathop{\mathbb{E}}_{ensemble}[f_{t+1}^{\text{lin}}(x)] = (1 + \ldots + \gamma^t C^t)C\mathcal{R} \tag{31}$$

As we can see, **the expectation of the learned values is identical** to the previous case where targets were independent. **However, the variance/covariance is completely different.** This key difference arises from the difference between equations 16 and 29. In the setting where the target was shared, the covariance was modulated by the term $(1 + \ldots + \gamma^t C^t)$. However, in the case where the target is the mean, there is no such modulation of the covariance. **The uncertainties obtained with mean targets identical to regressing the uncertainties obtained by regressing $\mathcal{Y}_t$ using an ensemble.**