The Importance of Pessimism in Fixed-Dataset Policy Optimization

Jacob Buckman (Mila; McGill University)*
Carles Gelada (OpenAI)
Marc G. Bellemare (Google Research; Mila; McGill University; CIFAR Fellow)

This work **provides formal justification** for the properties of **every "Offline RL" algorithm** in the literature, including:

BCQ, CRR, SPIBB, BEAR, CQL, KLC, BRAC, MBS-QI, MoREL, MOPO, and more.

Our goal is an algo with minimal worst-case suboptimality,

$$\textsc{SubOpt}(\mathcal{O}(D)) = \mathbb{E}_\rho[\mathbf{v}_{\mathcal{M}}^{\pi^*_{\mathcal{M}}}] - \mathbb{E}_\rho[\mathbf{v}_{\mathcal{M}}^{\mathcal{O}(D)}].$$

We consider "value-based" algorithms,

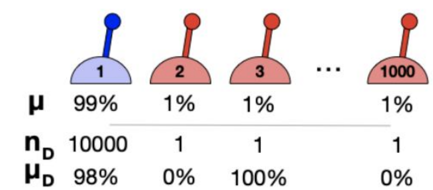
$$\mathcal{O}^{\text{VB}}_{\text{sub}}(D) := \arg\max_\pi \mathbb{E}_\rho[\mathcal{E}_{\text{sub}}(D,\pi)].$$

These algorithms can be characterized by the choice of fixed point of E_{sub}.
Suboptimality of these algos permits an "over/under decomposition",

$$\textit{SubOpt}(\mathcal{O}^{VB}(D)) \leq \inf_\pi \left(\mathbb{E}_\rho[\mathbf{v}_{\mathcal{M}}^{\pi^*_{\mathcal{M}}} - \mathbf{v}_{\mathcal{M}}^{\pi}] + \mathbb{E}_\rho[\mathbf{v}_{\mathcal{M}}^{\pi} - \mathbf{v}_{D}^{\pi}]\right) + \sup_\pi \left(\mathbb{E}_\rho[\mathbf{v}_{D}^{\pi} - \mathbf{v}_{\mathcal{M}}^{\pi}]\right)$$

The actual suboptimality depends choice of E_{sub}.
One important type of algo is "naive":

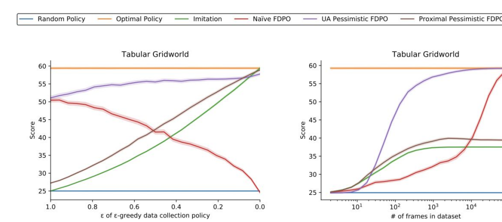
$$f_{\textit{naïve}}(\mathbf{v}^\pi) := A^\pi(\mathbf{r}_D + \gamma P_D \mathbf{v}^\pi). \qquad \textit{SubOpt}(\mathcal{O}^{VB}_{\textit{naïve}}(D)) \leq \inf_\pi \left(\mathbb{E}_\rho[\mathbf{v}_{\mathcal{M}}^{\pi^*_{\mathcal{M}}} - \mathbf{v}_{\mathcal{M}}^{\pi}] + \mathbb{E}_\rho[\boldsymbol{\mu}^\pi_{D,\delta}]\right) + \sup_\pi \mathbb{E}_\rho[\boldsymbol{\mu}^\pi_{D,\delta}]$$



This often leads to a large "sup" term. We can fix this by finding pessimistic fixed points, which let us choose the relative size of the two terms:

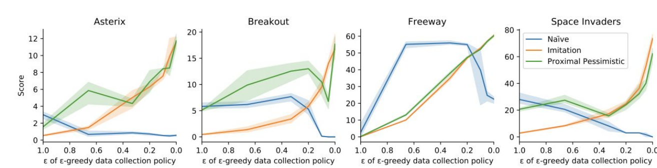
$$f_{ua}(\mathbf{v}^\pi) = A^\pi(\mathbf{r}_D + \gamma P_D \mathbf{v}^\pi) - \alpha \mathbf{u}^\pi_{D,\delta}$$

$$\textit{SubOpt}(\underline{\mathcal{O}}^{VB}_{ua}(D)) \leq \inf_\pi \left(\mathbb{E}_\rho[\mathbf{v}_{\mathcal{M}}^{\pi_{\mathcal{M}}} - \mathbf{v}_{\mathcal{M}}^{\pi}] + (1+\alpha)\cdot\mathbb{E}_\rho[\boldsymbol{\mu}^\pi_{D,\delta}]\right) + (1-\alpha)\cdot\left(\sup_\pi \mathbb{E}_\rho[\boldsymbol{\mu}^\pi_{D,\delta}]\right)$$



Implementing this algorithm requires implementing a valid uncertainty measure, which we don't know how to do right now with NNs. If we take a "trivial uncertainty" of V_{max}, we get proximal algorithms:

$$f_{\textit{proximal}}(\mathbf{v}^\pi) = A^\pi(\mathbf{r}_D + \gamma P_D \mathbf{v}^\pi) - \alpha\left(\frac{TV_S(\pi, \hat{\pi}_D)}{(1-\gamma)^2}\right)$$



The trivial uncertainty is the "worst" uncertainty, leading to a much looser bound; but it is, at least, implementable.