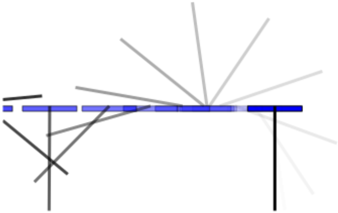
Distributional Hamilton-Jacobi-Bellman Equations for Continuous-Time Reinforcement Learning

Harley Wiltzer David Meger Marc G. Bellemare

Continuous-Time Reinforcement Learning

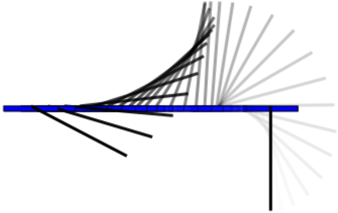
Continuous-Time Reinforcement Learning

Control frequency: ω



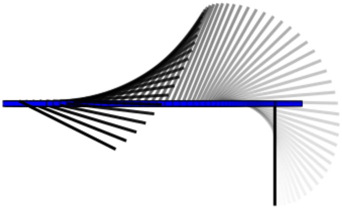
Continuous-Time Reinforcement Learning

Control frequency: 2ω



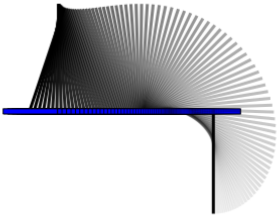
Continuous-Time Reinforcement Learning

Control frequency: 5ω



Continuous-Time Reinforcement Learning

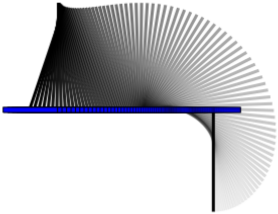
Control frequency: 10ω



- Challenge: learn a value function that converges as $\omega \uparrow \infty$.

Continuous-Time Reinforcement Learning

Control frequency: 10ω



- Challenge: learn a value function that converges as $\omega \uparrow \infty$.
- Characterized by the HJB Equation:

$$V^\pi(x)\log\gamma + r(x) + \langle \nabla V^\pi(x), \mu_\pi(x)\rangle + \frac{1}{2}\mathsf{Tr}(\sigma_\pi(x)^\top \mathsf{H}V^\pi(x)\sigma_\pi(x)) = 0$$

Contribution 1: The *Distributional* HJB Equation

- Goal: rather than learn the expected return, learn the return *distribution*.

Contribution 1: The *Distributional* HJB Equation

- Goal: rather than learn the expected return, learn the return *distribution*.

$$V^{\pi}(x)\log\gamma + r(x) + \langle \nabla V^{\pi}(x), \mu_{\pi}(x)\rangle + \frac{1}{2}\mathsf{Tr}(\sigma_{\pi}(x)^{\top}\mathsf{H}V^{\pi}(x)\sigma_{\pi}(x)) = 0 \tag{HJB}$$

Contribution 1: The *Distributional* HJB Equation

- Goal: rather than learn the expected return, learn the return *distribution*.

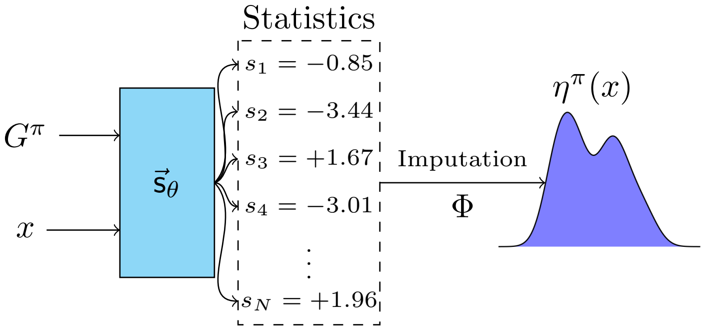
$$V^{\pi}(x)\log\gamma + r(x) + \langle \nabla V^{\pi}(x), \mu_{\pi}(x)\rangle + \frac{1}{2}\mathsf{Tr}(\sigma_{\pi}(x)^{\top}\mathsf{H}V^{\pi}(x)\sigma_{\pi}(x)) = 0 \tag{HJB}$$

$$\Downarrow$$

$$\begin{aligned}\langle \nabla_x F_{\eta^{\pi}}(x,z), \mu_{\pi}(x)\rangle - (r(x) + z\log\gamma)\frac{\partial}{\partial z}F_{\eta^{\pi}}(x,z)\\ + \frac{1}{2}\mathsf{Tr}(\sigma_{\pi}(x)^{\top}\mathsf{H}_x F_{\eta^{\pi}}(x,z)\sigma_{\pi}(x)) = 0\end{aligned} \tag{DHJB}$$

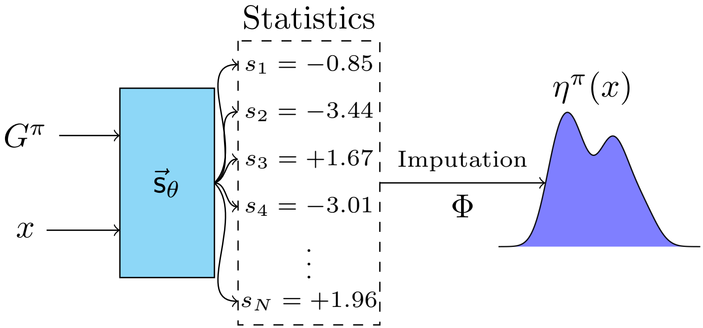
Contribution 2: The Statistical HJB Loss

- Goal: Represent the DHJB equation in finite space.



Contribution 2: The Statistical HJB Loss

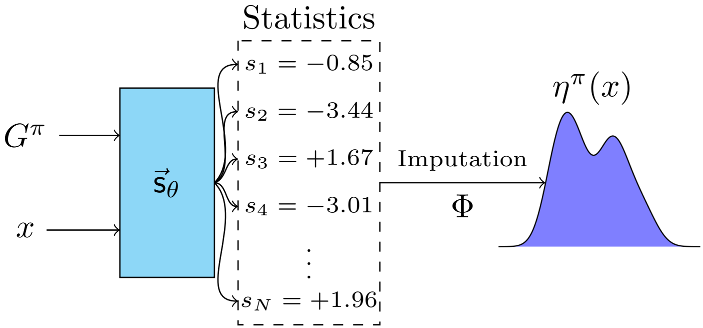
- Goal: Represent the DHJB equation in finite space.



$$\nabla_{\vec{\mathsf{s}}(x)}\Phi(\vec{\mathsf{s}}(x), z)^\top \mathsf{J}\vec{\mathsf{s}}(x)\mu_\pi(x) - (r(x) + z\log\gamma)\frac{\partial}{\partial z}\Phi(\vec{\mathsf{s}}(x), z) \qquad \text{(SHJB)}$$
$$+\frac{1}{2}\mathsf{Tr}\left[\sigma_\pi(x)^\top\left(\mathbf{K}^x_\Phi(x,z) + \mathbf{K}^s_\Phi(x,z)\right)\sigma_\pi(x)\right] \xrightarrow{N\uparrow\infty} 0$$

Contribution 2: The Statistical HJB Loss

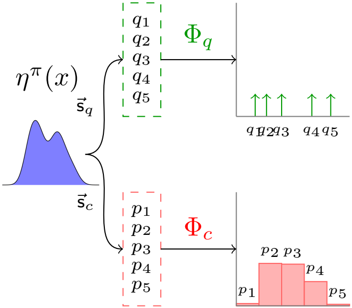
- Goal: Represent the DHJB equation in finite space.



$$\nabla_{\vec{\mathsf{s}}(x)} \Phi(\vec{\mathsf{s}}(x), z)^\top \mathsf{J}\vec{\mathsf{s}}(x) \mu_\pi(x) - (r(x) + z \log \gamma) \frac{\partial}{\partial z} \Phi(\vec{\mathsf{s}}(x), z) \qquad \text{(SHJB)}$$

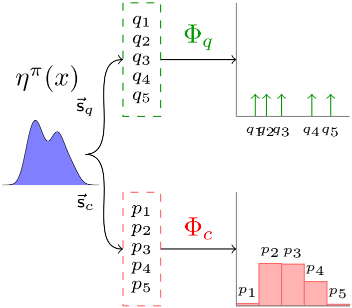
$$+ \frac{1}{2} \mathsf{Tr} \left[\sigma_\pi(x)^\top \underbrace{\left(\mathbf{K}^x_\Phi(x,z) + \mathbf{K}^s_\Phi(x,z)\right)}_{\text{"Spatial \& Statistical Diffusivity"}} \sigma_\pi(x) \right] \xrightarrow{N \uparrow \infty} 0$$

Diffusivities of some common imputation strategies



- When Φ_q is the *quantile imputation strategy*, the statistical diffusivity vanishes.

Diffusivities of some common imputation strategies



- When Φ_q is the *quantile imputation strategy*, the statistical diffusivity vanishes.
- And when Φ_c is the *categorical imputation strategy*, the SHJB is very complex.

The Quantile Case

We show that, when return distributions are represented as empirical distributions,

$$\begin{cases} \langle \nabla_x \vec{\mathsf{s}}_k(x), \mu_\pi(x) \rangle + r(x) + \vec{\mathsf{s}}_k(x) \log \gamma + \frac{1}{2} \mathsf{Tr} \left(\sigma_\pi(x)^\top \mathsf{H}_x \vec{\mathsf{s}}_k(x) \sigma_\pi(x) \right) = 0 \\ \vec{\mathsf{s}}_k(x) = F^{-1}_{\eta^\pi}(\hat{\tau}_k) \\ \hat{\tau}_k = \frac{k - \frac{1}{2}}{N} \\ k \in [N] \end{cases}$$

The Quantile Case

We show that, when return distributions are represented as empirical distributions,

$$\begin{cases} \overbrace{\langle \nabla_x \vec{\mathsf{s}}_k(x), \mu_\pi(x) \rangle + r(x) + \vec{\mathsf{s}}_k(x) \log \gamma + \frac{1}{2} \mathsf{Tr}\left(\sigma_\pi(x)^\top \mathsf{H}_x \vec{\mathsf{s}}_k(x) \sigma_\pi(x)\right) = 0}^{\text{The HJB Equation!}} \\ \vec{\mathsf{s}}_k(x) = F^{-1}_{\eta^\pi}(\hat{\tau}_k) \\ \hat{\tau}_k = \frac{k - \frac{1}{2}}{N} \\ k \in \{1, 2, \ldots, N\} \end{cases}$$

The Quantile Case

We show that, when return distributions are represented as empirical distributions,

$$\begin{cases} \overbrace{\langle \nabla_x \vec{\mathsf{s}}_k(x), \mu_\pi(x) \rangle + r(x) + \vec{\mathsf{s}}_k(x) \log \gamma + \frac{1}{2} \mathsf{Tr}\left(\sigma_\pi(x)^\top \mathsf{H}_x \vec{\mathsf{s}}_k(x) \sigma_\pi(x) \right) = 0}^{\text{The HJB Equation!}} \\ \vec{\mathsf{s}}_k(x) = F^{-1}_{\eta^\pi}(\hat{\tau}_k) \\ \hat{\tau}_k = \frac{k - \frac{1}{2}}{N} \\ k \in \{1, 2, \dots, N\} \end{cases}$$

- Notably, distributional dynamic programming reduces to dynamic programming.

Thanks!

Check out our poster, #4711, to learn more.