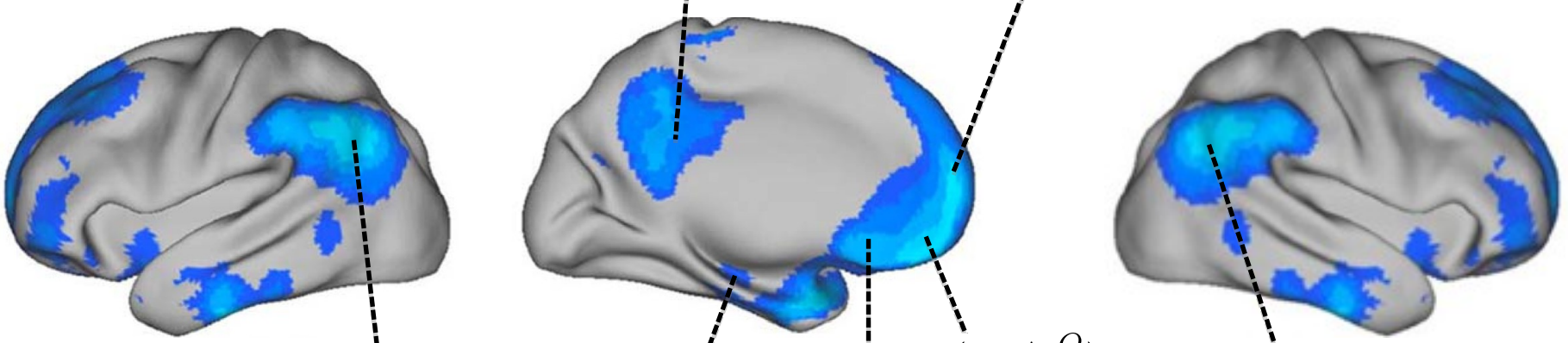
$$\nabla_{\theta^Q_{k+1}} \mathcal{L}(\theta^Q_{k+1}) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(D)} \left[\underbrace{(Q(s,a|\theta^Q_{k+1}) - y)}_{\text{regret}} \underbrace{\nabla_{\theta^Q_{k+1}} Q(s,a|\theta^Q_{k+1})}_{\text{aversiveness}} \right]$$

parameter update

$$Q^\pi(s,a) = \mathbb{E}_{s'\sim p(s'|s,a)}[r(s,a) + \gamma \max_{a'\in\mathcal{A}} Q^\pi(s',a')]$$

Bellman equation



s, a

world knowledge

$Q(s,a|\theta^Q_k)$

value function

r_t

reward

$(s,a,r,s') \sim U(\mathcal{D})$

experience sampling

$Q(s,a|\theta^Q_{k+1}) - y$

Predition error signaling