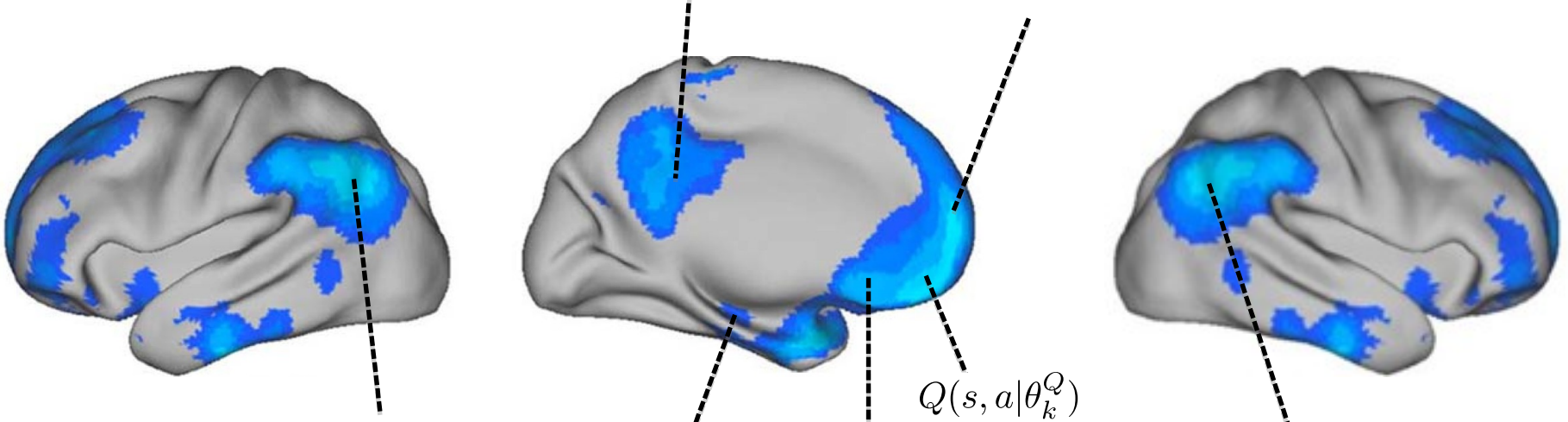
$$\nabla_{\theta^Q_{k+1}} \mathcal{L}(\theta^Q_{k+1}) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(D)} \left[\underbrace{(Q(s,a|\theta^Q_{k+1}) - y)}_{\text{regret}} \underbrace{\nabla_{\theta^Q_{k+1}} Q(s,a|\theta^Q_{k+1})}_{???} \right]$$

parameter update (backprop!)



s, a **(really ?)**

world knowledge

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

value function

r_t

Instantaneous reward

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

experience sampling

$a = \text{argmax}_{\tilde{a}\in\mathcal{A}} Q(s,\tilde{a}|\theta^Q_k) + u(s,\tilde{a})$

policy function (greedy!)