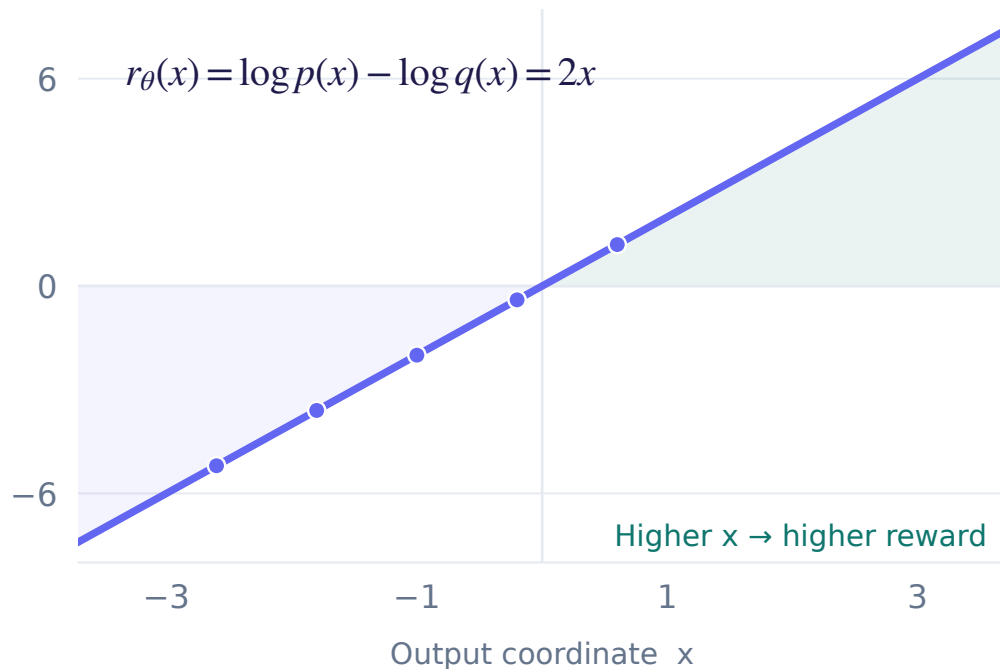


One mismatch. Two kinds of feedback.

Student $q = \mathcal{N}(-1, 1)$ · Target $p = \mathcal{N}(1, 1)$

— Target p
- - Student q

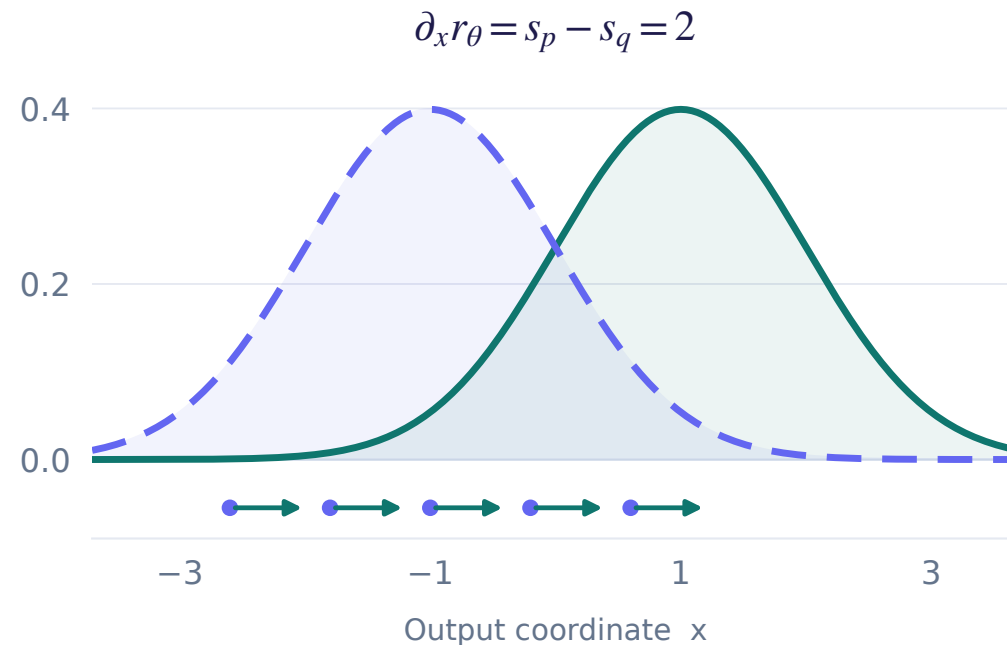
A Weight sampled log probabilities



SCALAR FEEDBACK

Reward $\times$ parameter log-probability gradient

B Move differentiable outputs



DIRECTIONAL FEEDBACK

Score difference $\times$ generator Jacobian