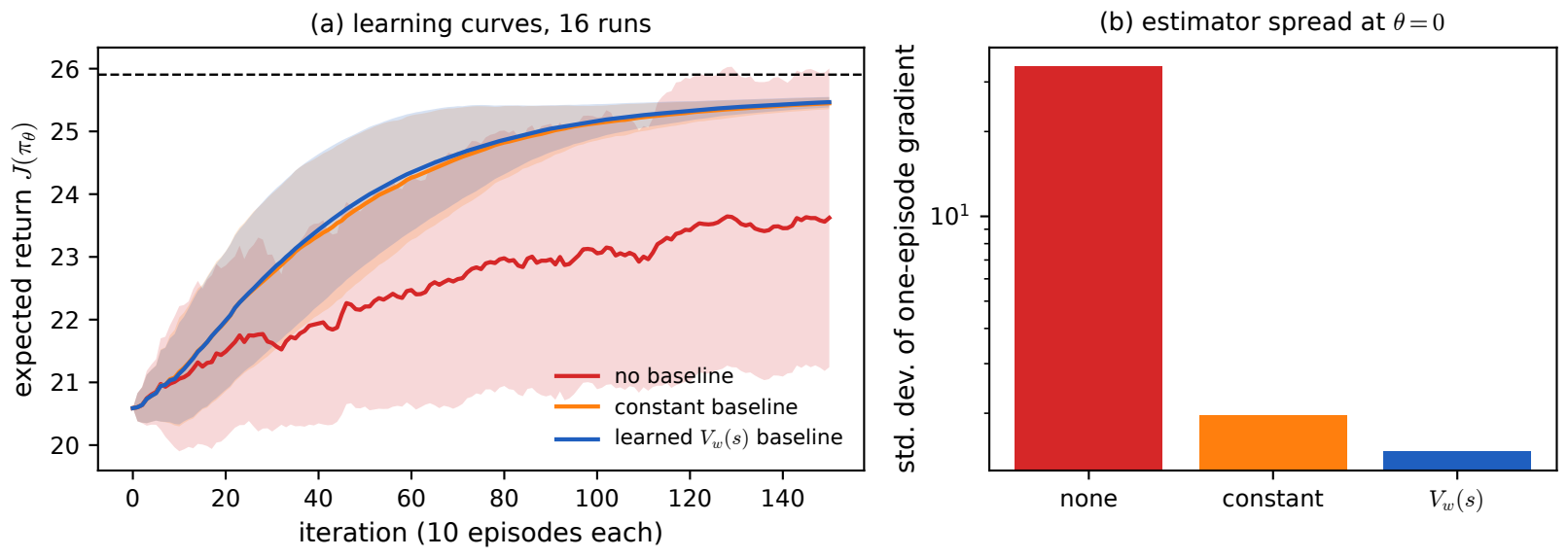

(a) learning curves, 16 runs
expected return $J(\pi_\theta)$
iteration (10 episodes each)
no baseline
constant baseline
learned $V_w(s)$ baseline
(b) estimator spread at $\theta = 0$
std. dev. of one-episode gradient
none
constant
$V_w(s)$