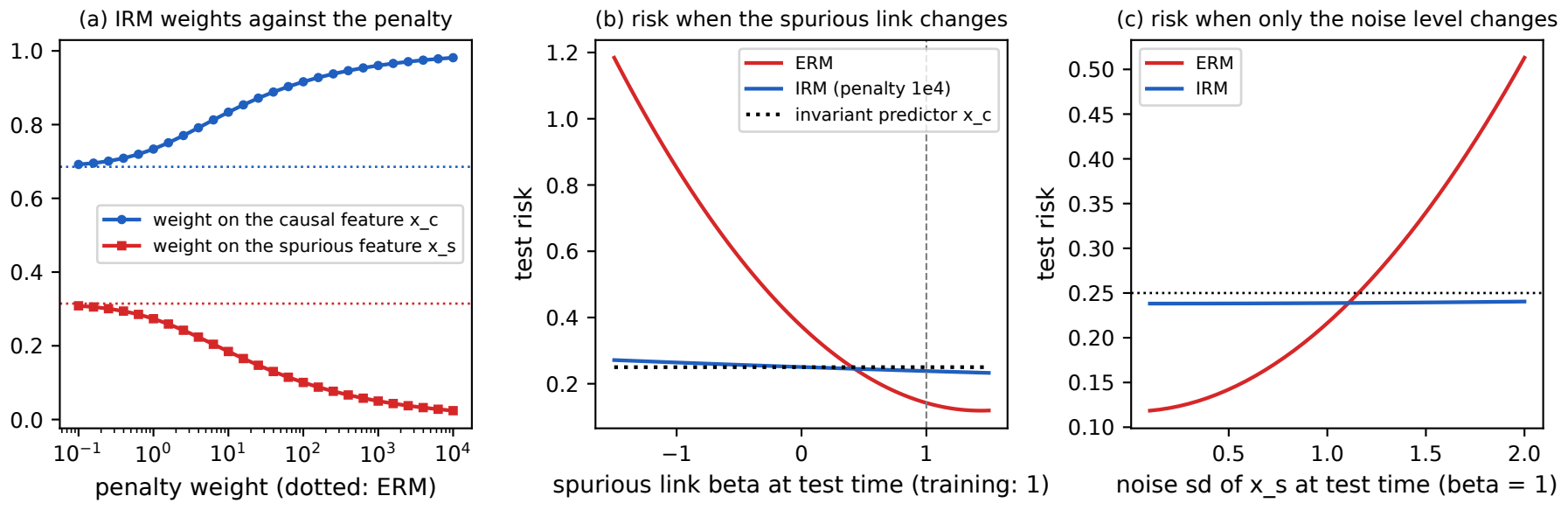

(a) IRM weights against the penalty
1.0
0.8
0.6
0.4
0.2
0.0
weight on the causal feature x_c
weight on the spurious feature x_s
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
penalty weight (dotted: ERM)
(b) risk when the spurious link changes
test risk
1.2
1.0
0.8
0.6
0.4
0.2
ERM
IRM (penalty 1e4)
invariant predictor x_c
−1
0
1
spurious link beta at test time (training: 1)
(c) risk when only the noise level changes
test risk
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
ERM
IRM
0.5
1.0
1.5
2.0
noise sd of x_s at test time (beta = 1)