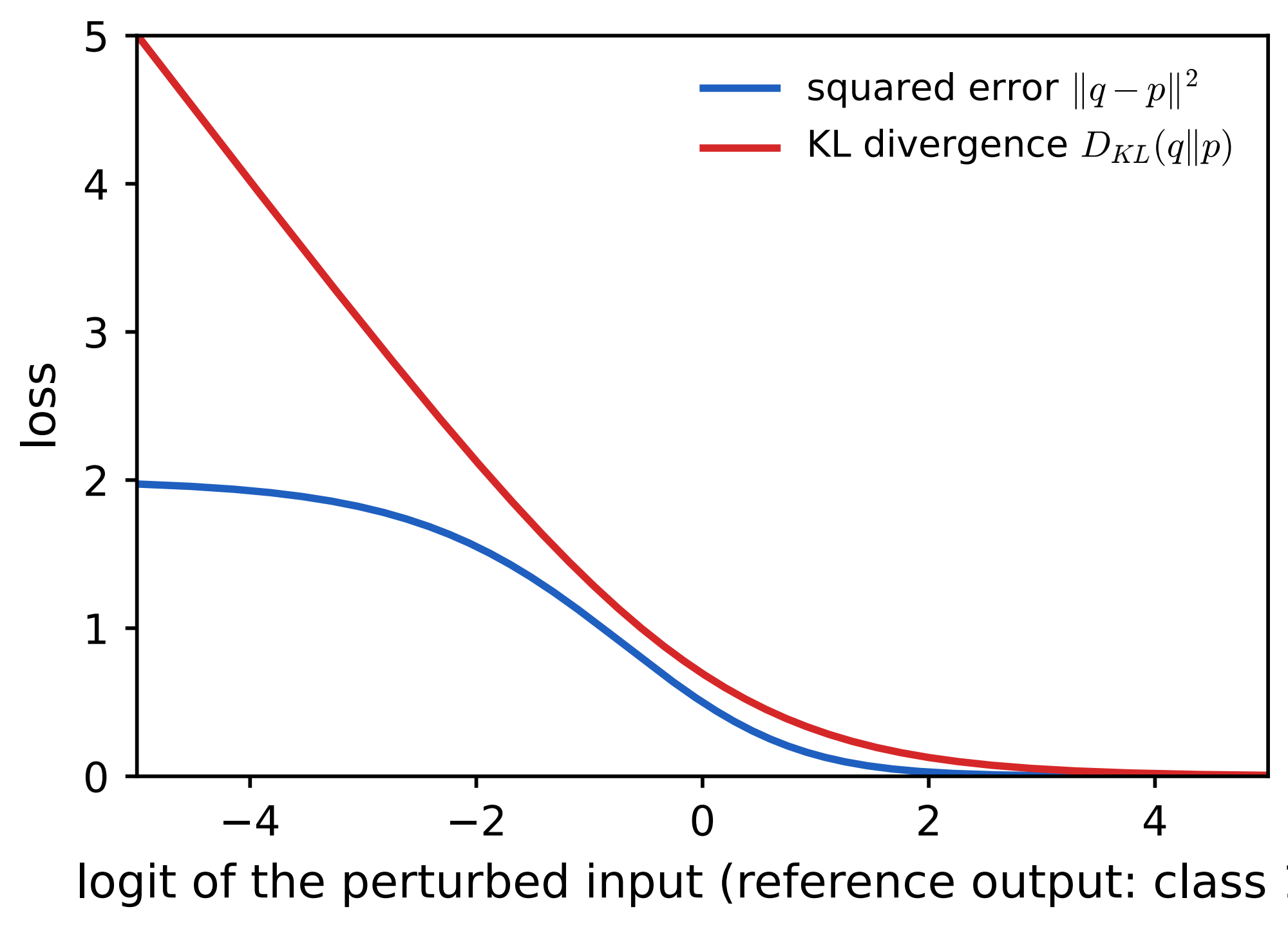

squared error $\|q-p\|^2$
KL divergence $D_{KL}(q\|p)$
loss
5
4
3
2
1
0
−4
−2
0
2
4
logit of the perturbed input (reference output: class