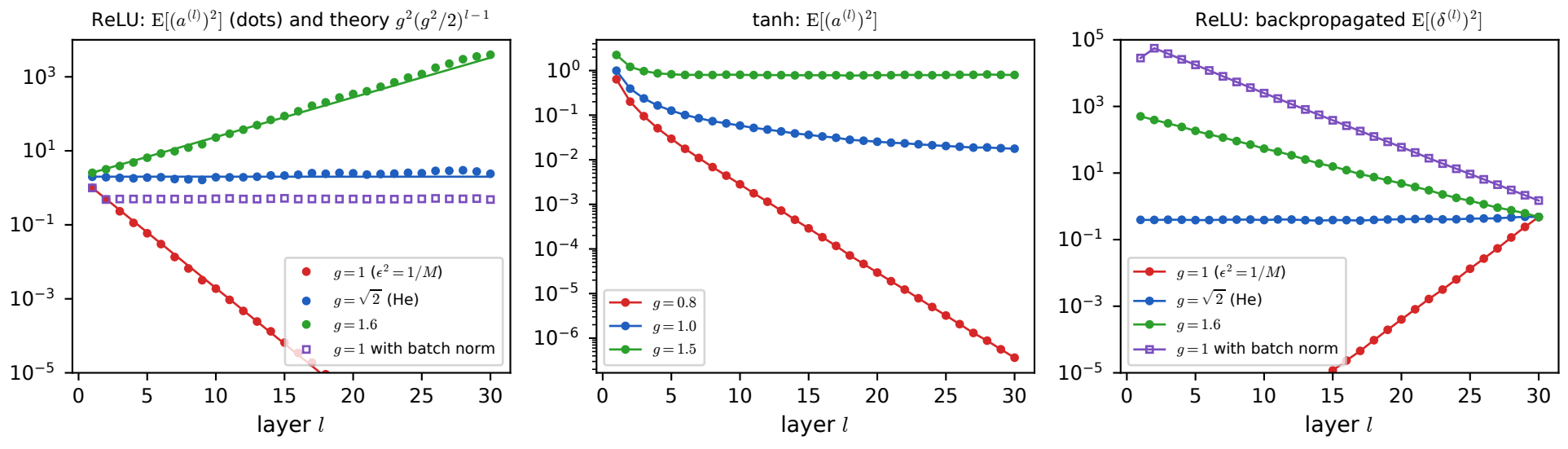

ReLU: $\mathrm{E}[(a^{(l)})^2]$ (dots) and theory $g^2(g^2/2)^{l-1}$
10^3
10^1
10^{-1}
10^{-3}
10^{-5}
0
5
10
15
20
25
30
layer l
$g=1$ ($\epsilon^2=1/M$)
$g=\sqrt{2}$ (He)
$g=1.6$
$g=1$ with batch norm
tanh: $\mathrm{E}[(a^{(l)})^2]$
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
0
5
10
15
20
25
30
layer l
$g=0.8$
$g=1.0$
$g=1.5$
ReLU: backpropagated $\mathrm{E}[(\delta^{(l)})^2]$
10^5
10^3
10^1
10^{-1}
10^{-3}
10^{-5}
0
5
10
15
20
25
30
layer l
$g=1$ ($\epsilon^2=1/M$)
$g=\sqrt{2}$ (He)
$g=1.6$
$g=1$ with batch norm