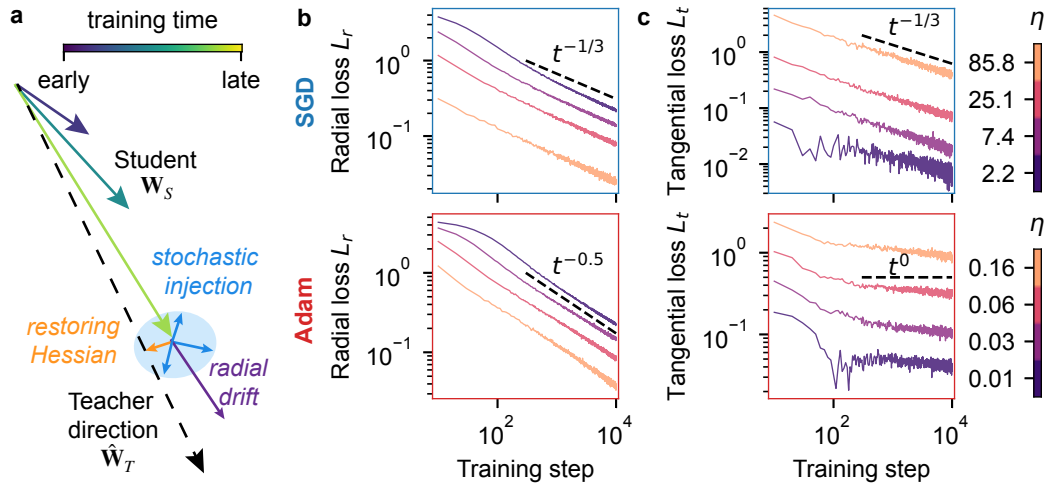

a
training time
early
late
Student
$\mathbf{W}_S$
stochastic
injection
restoring
Hessian
radial
drift
Teacher
direction
$\hat{\mathbf{W}}_T$
b
SGD
Radial loss L_r
$t^{-1/3}$
Adam
Radial loss L_r
$t^{-0.5}$
Training step
c
Tangential loss L_t
$t^{-1/3}$
η
85.8
25.1
7.4
2.2
Tangential loss L_t
t^{0}
η
0.16
0.06
0.03
0.01
Training step