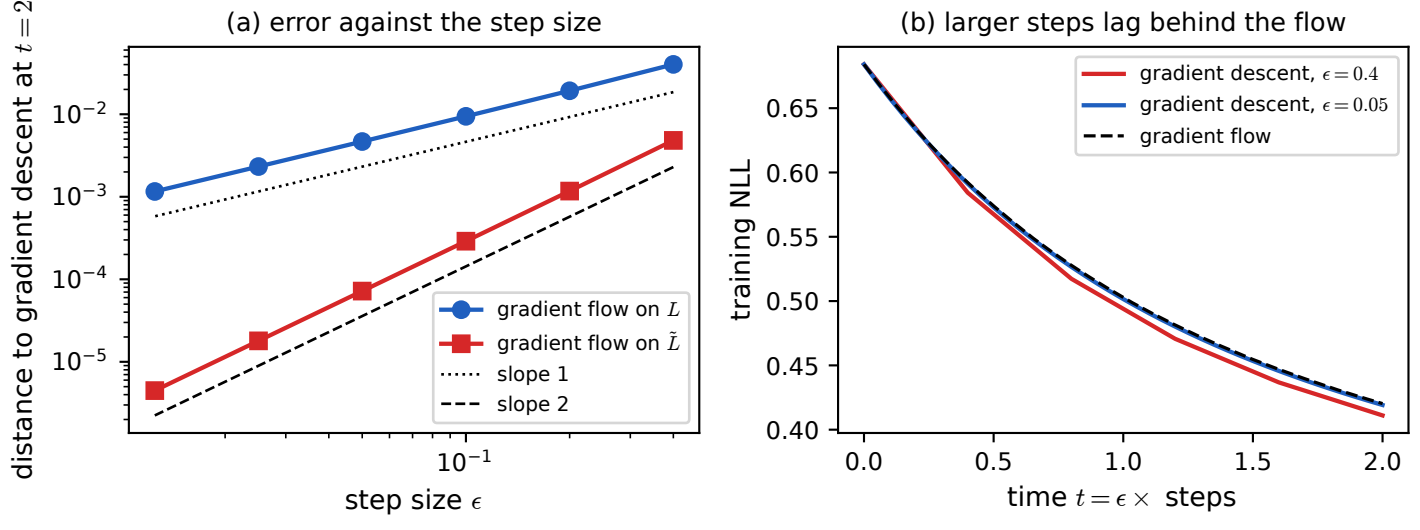

(a) error against the step size
distance to gradient descent at $t = 2$
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-1}
step size ϵ
gradient flow on L
gradient flow on $\tilde{L}$
slope 1
slope 2
(b) larger steps lag behind the flow
training NLL
0.65
0.60
0.55
0.50
0.45
0.40
0.0
0.5
1.0
1.5
2.0
time $t = \epsilon \times$ steps
gradient descent, $\epsilon = 0.4$
gradient descent, $\epsilon = 0.05$
gradient flow