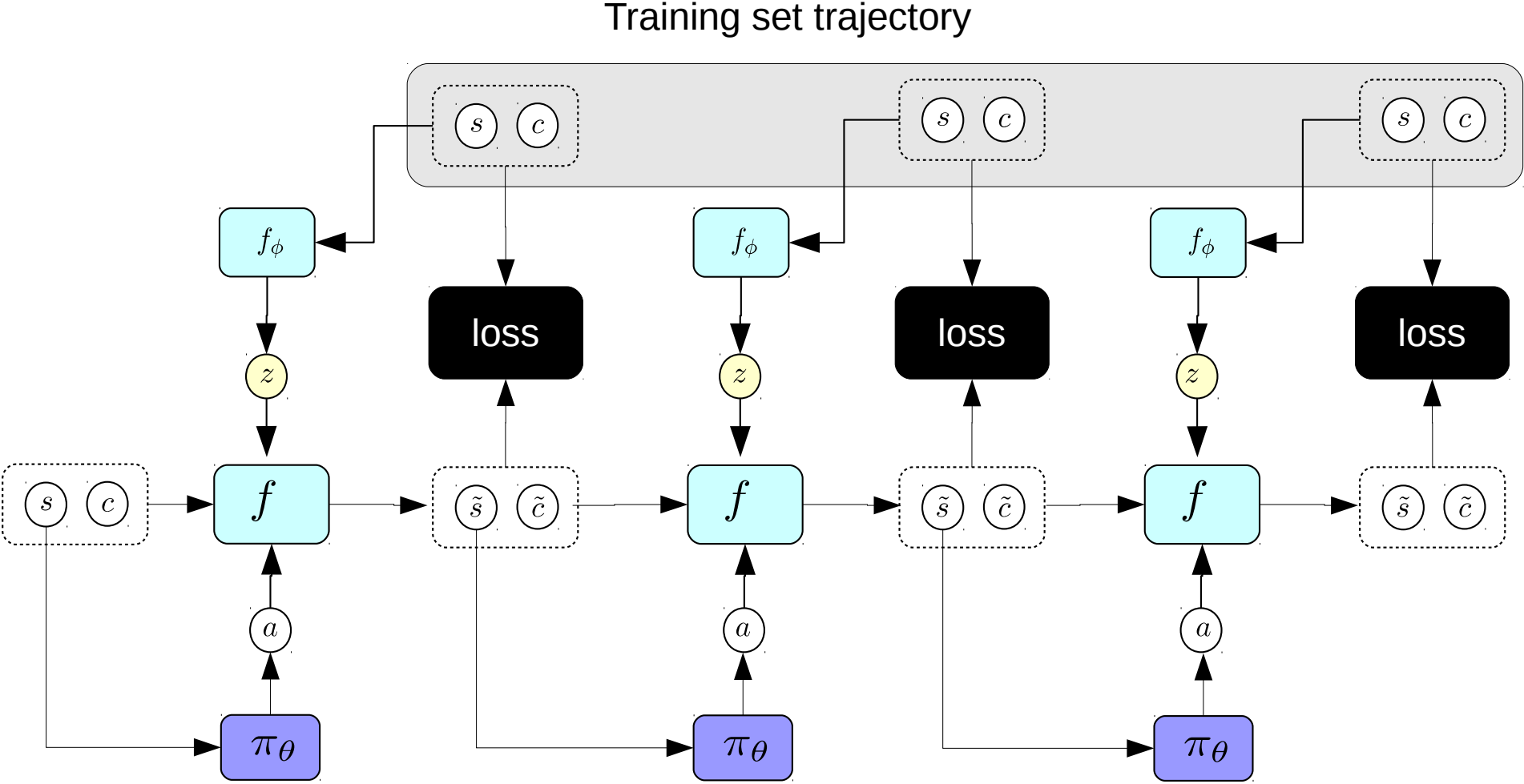

Stochastic policy network (optimized)

loss $= \|s - \tilde{s}\|_2 + \lambda\tilde{c}$