to the repeated application of the chain-rule, the computation of the gradient itself is often mathematically more involved than a sampling-based estimate.

Let us consider an example where the immediate reward r only depends on the state (generalizations to control-dependent rewards are straightforward) and the system dynamics are deterministic, such that $\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{u}_t) = f(\boldsymbol{x}_t, \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{\theta}))$, where f is a (nonlinear) transition function, $\pi_{\boldsymbol{\theta}}$ is the (deterministic) policy, and $\boldsymbol{\theta}$ are the policy parameters. The gradient of the long-term reward $J_{\boldsymbol{\theta}} = \sum_t \gamma^t r(\boldsymbol{x}_t)$ with respect to the policy parameters is obtained by applying the chain-rule repeatedly:

$$\frac{\mathrm{d}J_{\boldsymbol{\theta}}}{\mathrm{d}\boldsymbol{\theta}} = \sum_t \gamma^t \frac{\mathrm{d}r(\boldsymbol{x}_t)}{\mathrm{d}\boldsymbol{\theta}} = \sum_t \gamma^t \frac{\partial r(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t} \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}\boldsymbol{\theta}} \tag{3.20}$$

$$= \sum_t \gamma^t \frac{\partial r(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t} \left(\frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{x}_{t-1}} \frac{\mathrm{d}\boldsymbol{x}_{t-1}}{\mathrm{d}\boldsymbol{\theta}} + \frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{u}_{t-1}} \frac{\mathrm{d}\boldsymbol{u}_{t-1}}{\mathrm{d}\boldsymbol{\theta}} \right) . \tag{3.21}$$

From these equations we observe that the total derivative $\mathrm{d}\boldsymbol{x}_t / \mathrm{d}\boldsymbol{\theta}$ depends on the total derivative $\mathrm{d}\boldsymbol{x}_{t-1} / \mathrm{d}\boldsymbol{\theta}$ at the previous time step. Therefore, the derivative $\mathrm{d}J_{\boldsymbol{\theta}} / \mathrm{d}\boldsymbol{\theta}$ can be computed iteratively.

Extension to Probabilistic Models and Stochastic MDPs. For the extension to derivatives in stochastic MDPs and/or probabilistic models, we have to make a few adaptations to the gradients in Equation (3.20)–(3.21): When the state $\boldsymbol{x}_t$ is represented by a probability distribution $p(\boldsymbol{x}_t)$, we have to compute the *expected* reward $\mathbb{E}[r(\boldsymbol{x}_t)] = \int r(\boldsymbol{x}_t) p(\boldsymbol{x}_t) \, \mathrm{d}\boldsymbol{x}_t$. Moreover, we need to compute the derivatives with respect to the parameters of the state distribution, assuming that $p(\boldsymbol{x}_t)$ has a parametric representation.

For example, if $p(\boldsymbol{x}_t) = \mathcal{N}\big(\boldsymbol{x}_t \,|\, \boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}_t^x\big)$, we compute the derivatives of $\mathbb{E}[r(\boldsymbol{x}_t)]$ with respect to the mean $\boldsymbol{\mu}_t^x$ and covariance $\boldsymbol{\Sigma}_t^x$ of the state distribution and continue applying the chain-rule similarly to Equation (3.20)–(3.21): With the definition $\mathcal{E}_t \coloneqq \mathbb{E}_{\boldsymbol{x}_t}[r(\boldsymbol{x}_t)]$, we obtain the