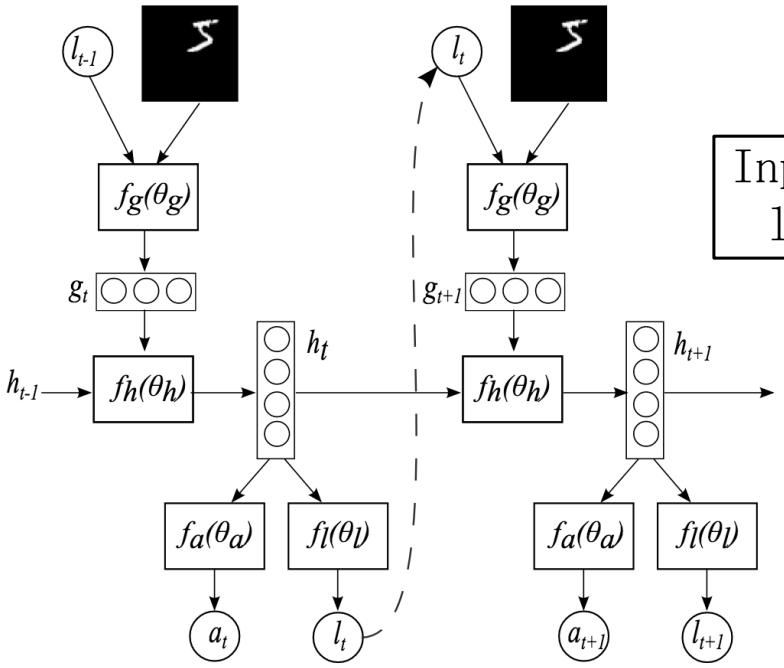

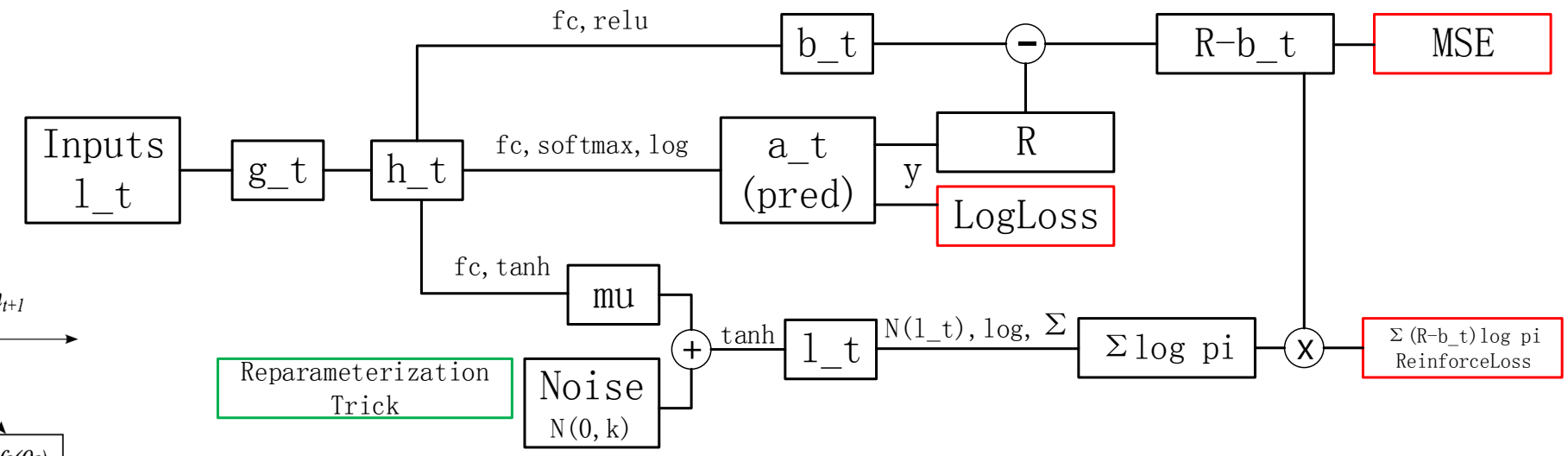

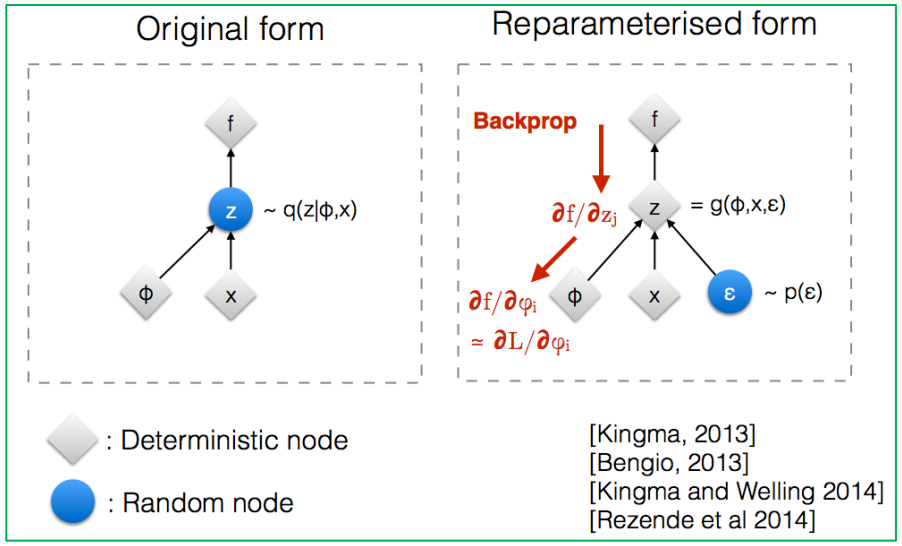

$$J_\theta = E_{\tau\sim p(\tau;\theta)}[r(\tau)] = \int r(\tau)p(\tau;\theta)d\tau = \int d\tau r(\tau)\int \nabla_\theta p(\tau;\theta)d\theta$$

$$= \int d\tau r(\tau)\int p(\tau;\theta)\nabla_\theta \log p(\tau;\theta)d\theta = \int d\theta \int p(\tau;\theta)r(\tau)\nabla_\theta \log p(\tau;\theta)d\tau$$

$$= \int d\theta E_{\tau\sim p(\tau;\theta)}[r(\tau)\nabla_\theta \log p(\tau;\theta)] = E_{\tau\sim p(\tau;\theta)}[r(\tau)\log p(\tau;\theta)]$$

$$\approx \frac{1}{|batch|}\sum_{batch} r(\tau)\log p(\tau;\theta) \triangleq \frac{1}{|batch|}\sum_{batch}\sum_{t\geq 0}[r(\tau)-b_t]\log p(\tau_t;\theta)$$

重设参数技巧：本来通过N(mu, k)采样得到l_t，但是采样会阻断梯度传播，因此引入noise作为输入，达到采样的随机性效果，不影响梯度反传。