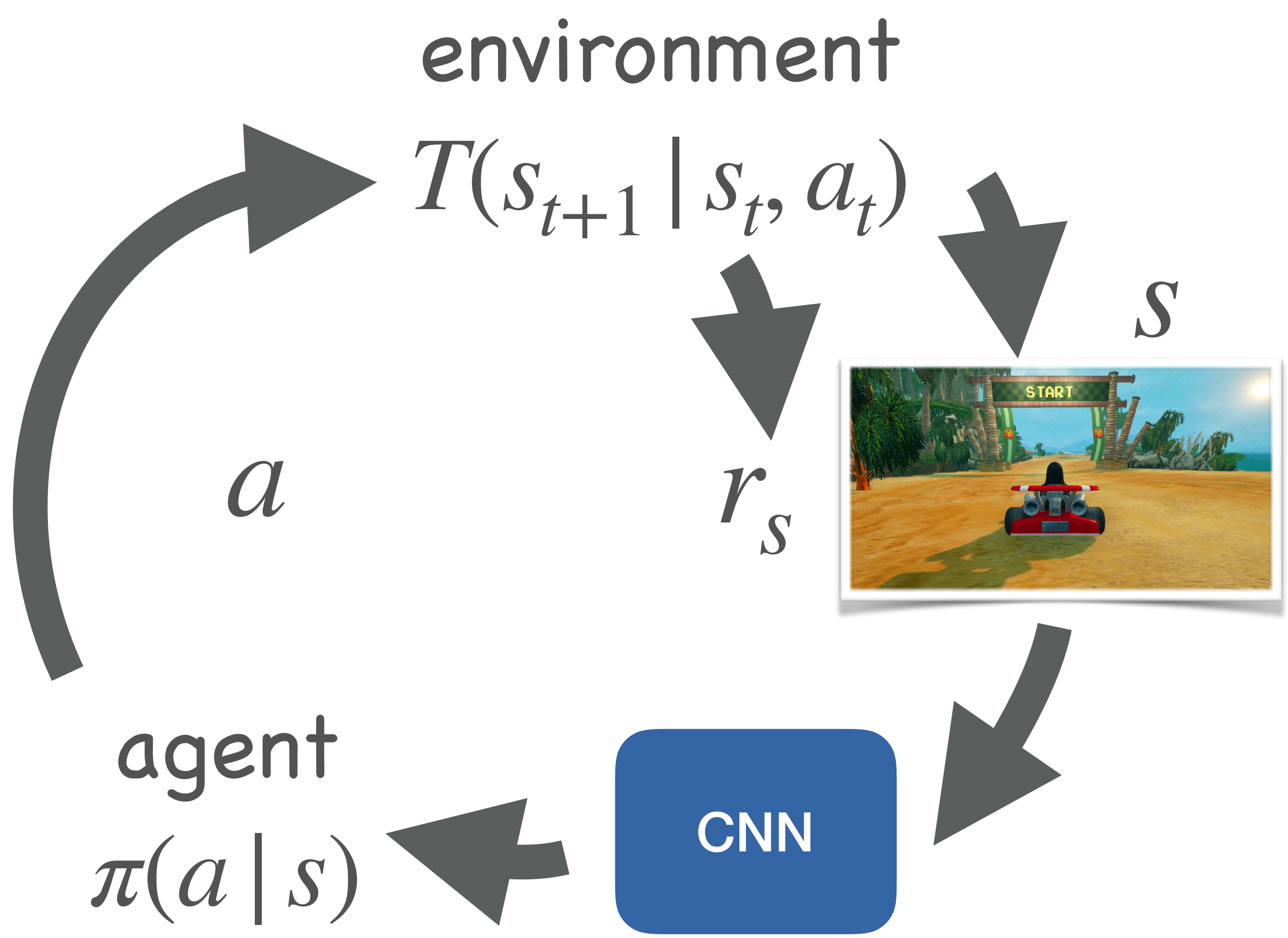
Non-differentiability

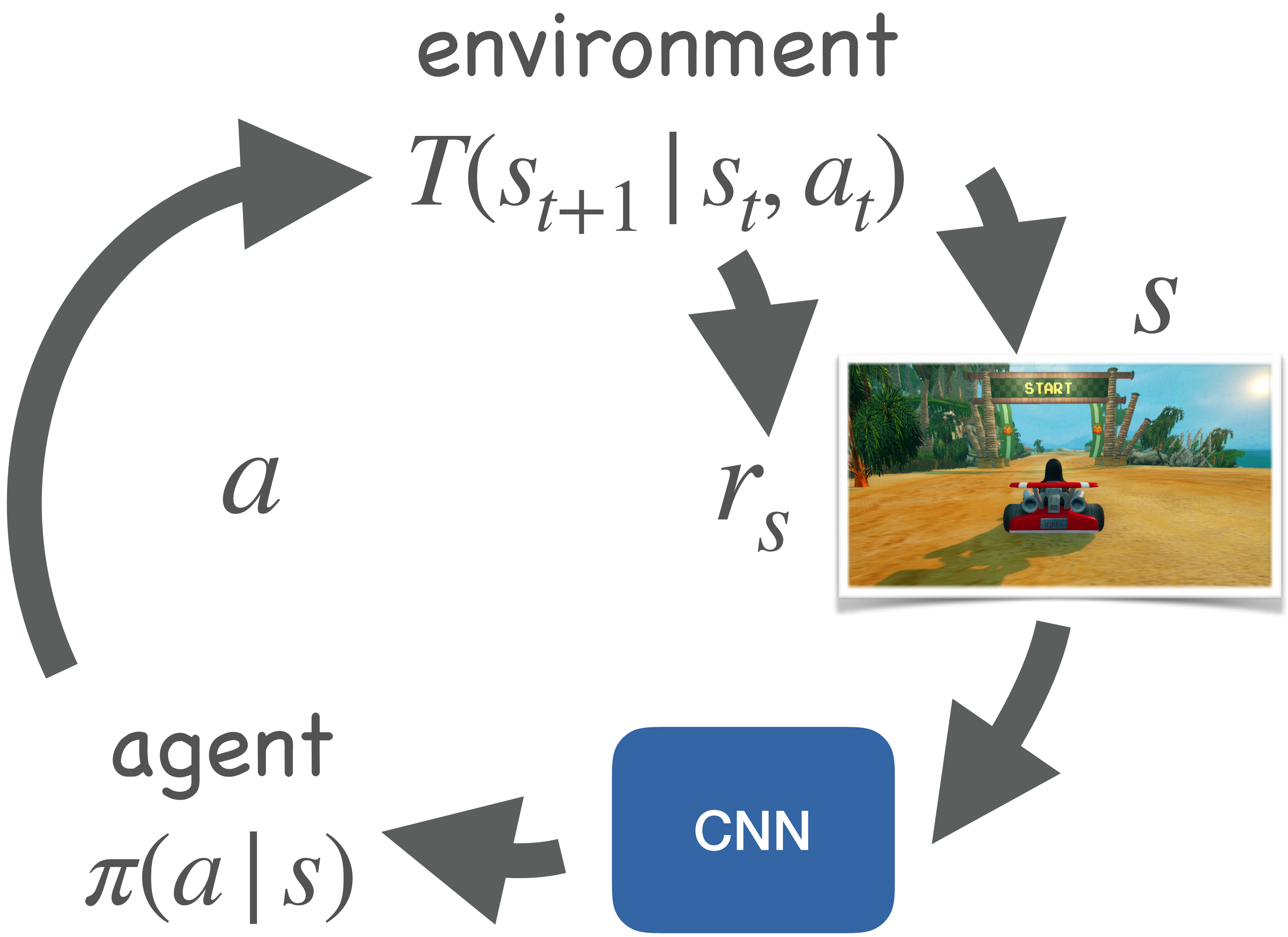
Deep learning for action

- Why not just learn a policy that maximizes reward?
  - Hard to optimize!



Deep learning for action

- Two sources of non-differentiability
  - Sampling
  - Environment



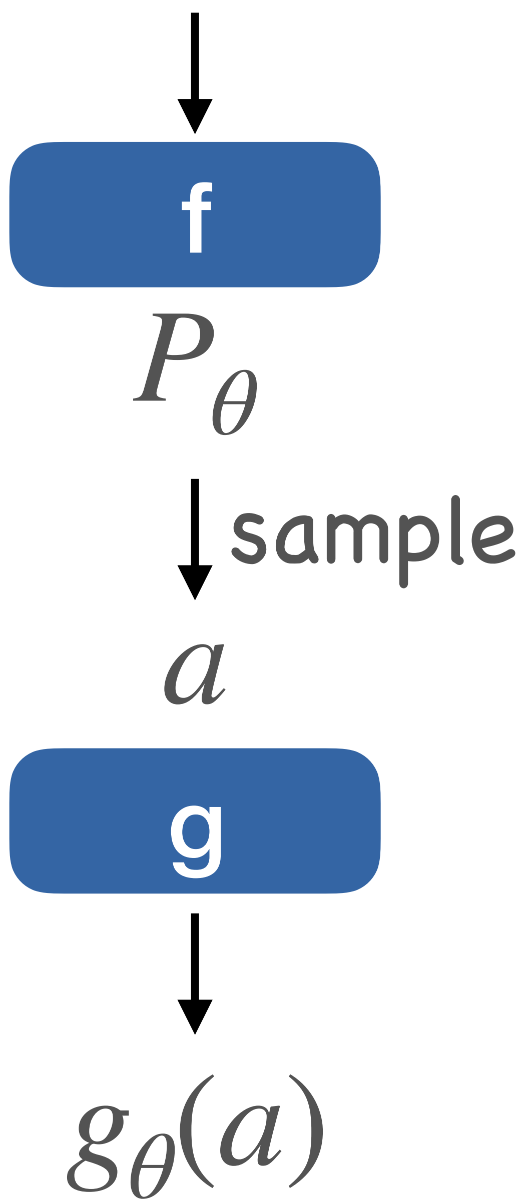
Differentiating sampling

- Compute gradient of

$$\mathbb{E}_{a \sim P_\theta}[g_\theta(a)] = \sum_a P_\theta(a) g_\theta(a)$$

$$\frac{\partial}{\partial\theta}\mathbb{E}_{a \sim P_\theta}[g_\theta(a)] = \sum_a g_\theta(a) \frac{\partial}{\partial\theta} P_\theta(a)$$

- $$+ \sum_a P_\theta(a) \frac{\partial}{\partial\theta} g_\theta(a)$$



Differentiating sampling - Issues

$$\frac{\partial}{\partial \theta}\mathbb{E}_{a \sim P_\theta}[g_\theta(a)] = \sum_a g_\theta(a)\frac{\partial}{\partial \theta}P_\theta(a) + \sum_a P_\theta(a)\frac{\partial}{\partial \theta}g_\theta(a)$$

- Large sum over all samples / action
  - Generally intractable

Reparametrization trick

- For continuous distributions
    - Rewrite
    $$P_\theta(a) = \frac{1}{\sigma_\theta} P\left(\frac{a - \mu_\theta}{\sigma_\theta}\right)$$
    - e.g. standard normal
    - $\mathbb{E}_{a \sim P_\theta}[g_\theta(a)] = \int_{\tilde{\Omega}} P\left(b\right) g_\theta(b\sigma_\theta + \mu_\theta) db$

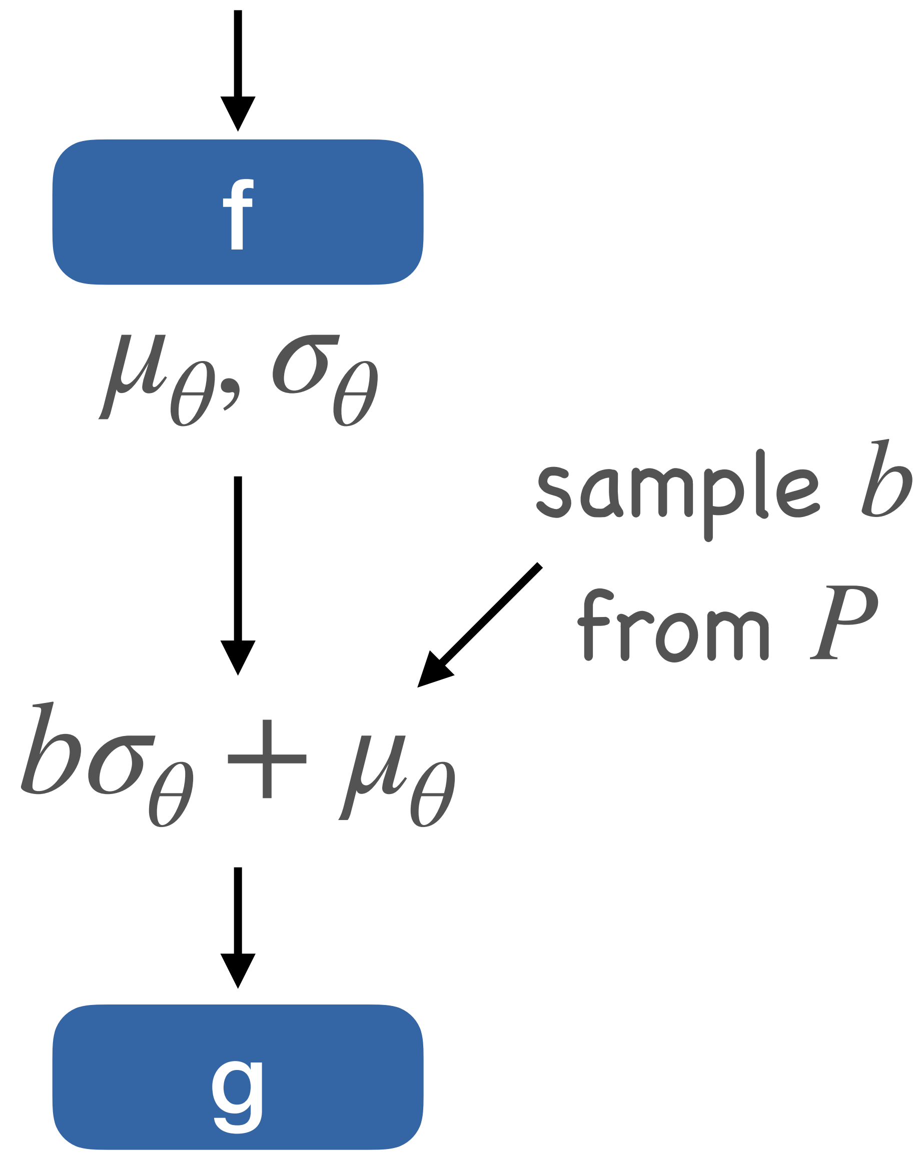
Auto-Encoding Variational Bayes, Kingma and Welling, ICLR 2014

Reparametrization trick

- Compute gradient

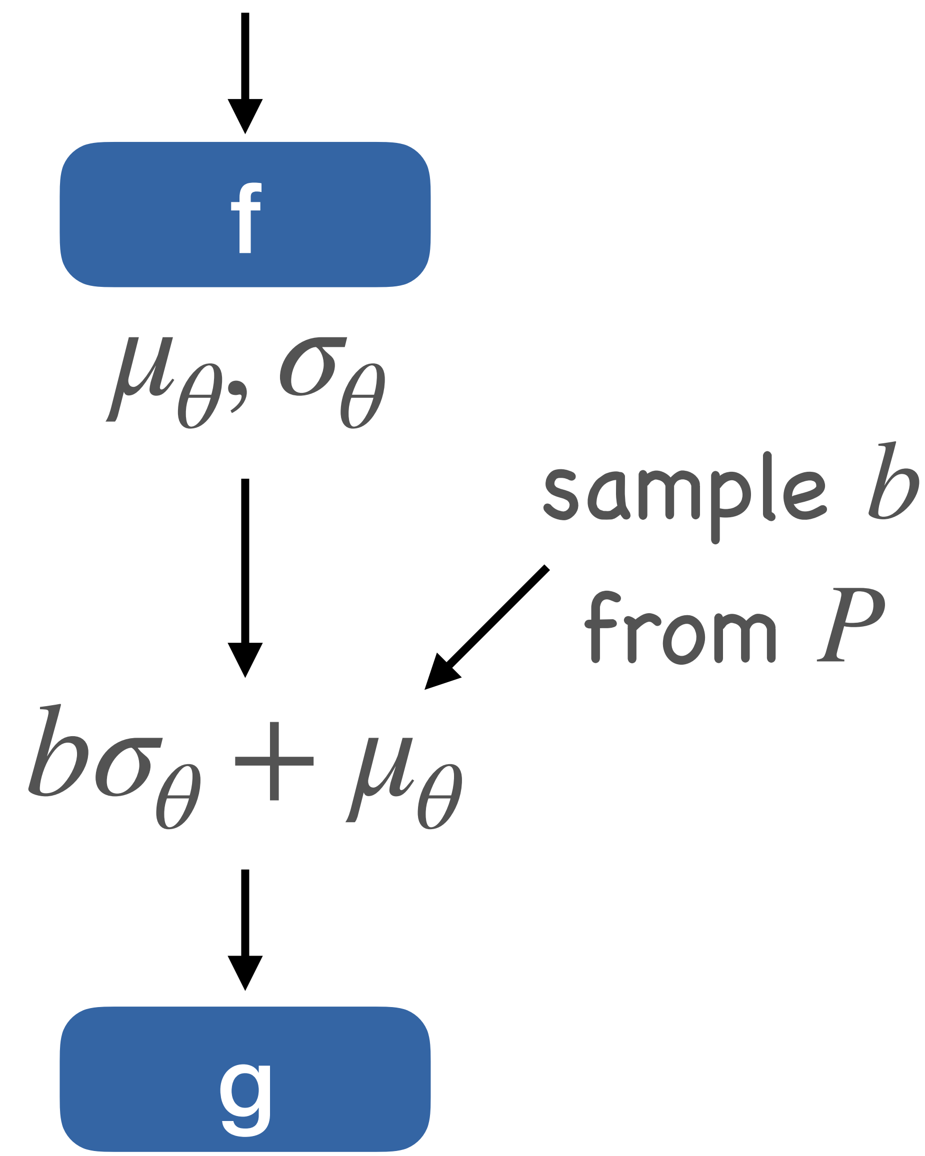
$$\frac{\partial}{\partial\theta}\mathbb{E}_{b\sim P}\left[g_\theta(b\sigma_\theta+\mu_\theta)\right]=\mathbb{E}_{b\sim P}\left[\frac{\partial}{\partial\theta}g_\theta(b\sigma_\theta+\mu_\theta)\right]$$

- Gradient computation by sampling



Reparametrization trick – discrete variables

- $\mathbb{E}_{a \sim P_\theta}[g_\theta(a)] = \sum_a P_\theta(a) g_\theta(a)$
- No change of variables
  - No differentiable function that maps to discrete distribution
- Continuous relaxation of one-hot vectors
  - Gumbel softmax



- The Concrete Distribution: a Continuous Relaxation of Discrete Random Variables, Maddison et al., ICLR 2017
- Categorical Reparameterization with Gumbel-Softmax, Jang et al, ICLR 2017

Differentiating the environment

- Quite hard
  - Up next