Carnegie Mellon
School of Computer Science

Deep Reinforcement Learning and Control

Natural Policy Gradients

Spring 2020, CMU 10-403

Katerina Fragkiadaki

Policy Gradients

Likelihood ratio gradient estimator

$$\max_{\theta} . \quad U(\theta) = \mathbb{E}_{x\sim P_\theta(x)} f(x)$$

$$\nabla U(\theta) = \mathbb{E}_{x\sim P_\theta(x)} \nabla_\theta \log P_\theta(x) f(x)$$

$$\max_{\theta} . \; U(\theta) = \mathbb{E}_{\tau\sim P_\theta(\tau)} \left[R(\tau)\right]$$

$$\nabla U(\theta) = \mathbb{E}_{\tau\sim P_\theta(\tau)} \left[\nabla_\theta \log P_\theta(\tau) R(\tau)\right]$$

Chain rule of derivatives

$$y = P_\theta(x)$$

$$\max_{\theta} . \; U(\theta) = f(P_\theta(x))$$

$$\nabla U(\theta) = \frac{df(P_\theta(x))}{d\theta} = \frac{df(y)}{dy}\frac{dy}{d\theta}$$

$$a = \pi_\theta(s)$$

$$\max_{\theta} . \quad U(\theta) = \mathbb{E} \sum_t Q(S_t, \pi_\theta(S_t))$$

$$\nabla U(\theta) = \frac{d\mathbb{E}\sum_t Q(S_t, \pi_\theta(S_t))}{d\theta} = \mathbb{E}\sum_t \frac{dQ(S_t, a)}{da}\frac{d\pi_\theta(S_t)}{d\theta}$$

Re-parametrization for Gaussian policies

$$\max_{\theta} . \quad U(\theta) = \mathbb{E}_{x\sim\mathcal{N}(\mu_\theta, \Sigma_\theta)} f(x)$$

$$\max_{\theta} . \quad U(\theta) = \mathbb{E}_{z\sim\mathcal{N}(0,I)} f(\mu_\theta + z * \sigma_\theta)$$

$$\nabla U(\theta) = \mathbb{E}_{z\sim\mathcal{N}(0,I)} \frac{df}{dx}\frac{d(\mu_\theta + z * \sigma_\theta)}{d\theta}$$

$$\max_{\theta} . \quad U(\theta) = \mathbb{E}_{A_t\sim\mathcal{N}(\mu_\theta(S_t), \sigma_\theta(S_t))} \sum_t Q(S_t, A_t)$$

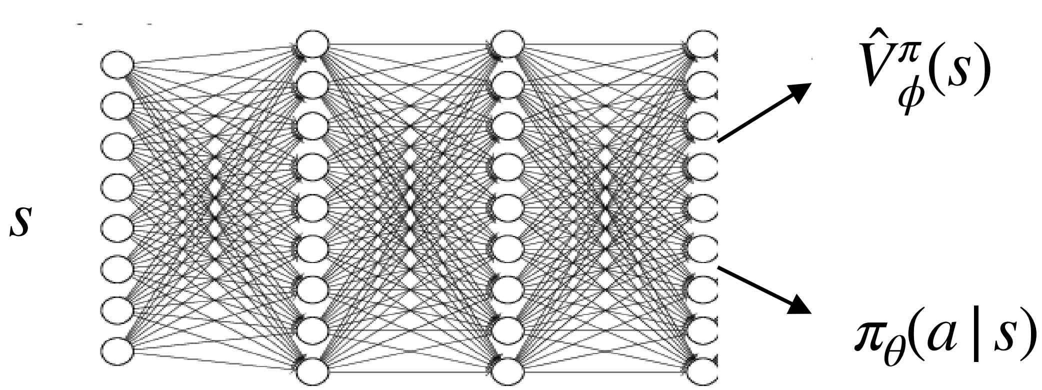
$$\max_{\theta} . \quad U(\theta) = \mathbb{E}_{z\sim\mathcal{N}(0,I)} \sum_t Q\left(S_t, \mu_\theta(S_t) + z * \sigma_\theta(S_t)\right)$$

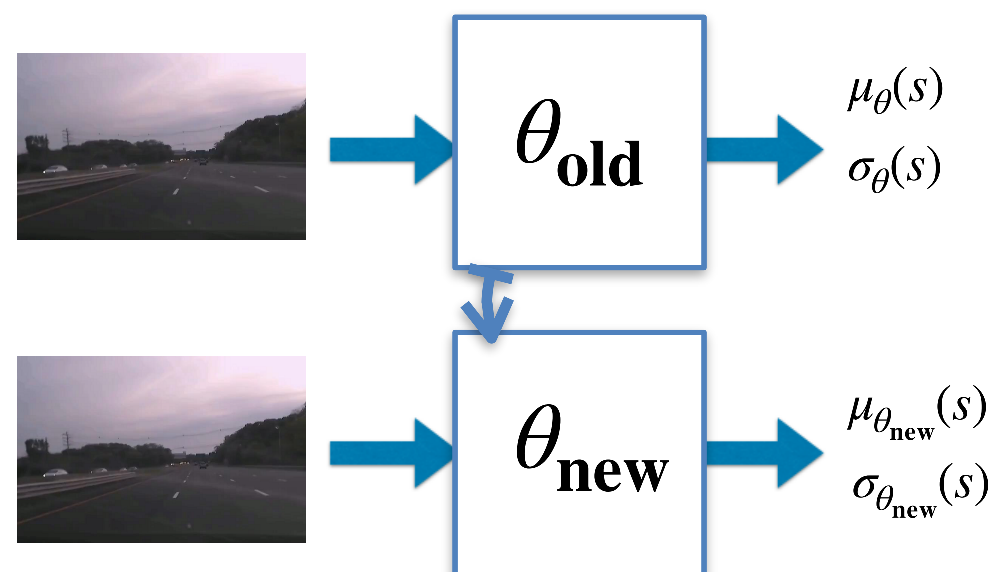
$$\nabla U(\theta) = \mathbb{E}_{z\sim\mathcal{N}(0,I)} \sum_t \frac{\partial Q(S_t, a)}{\partial a}\frac{\partial\left(\mu_\theta(S_t) + z * \sigma_\theta(S_t)\right)}{\partial\theta}$$

Actor-critic

1. Sample trajectories $\{s_t^i, a_t^i\}_{i=0}^T$ by running the current policy $a \sim \pi_\theta(s)$
2. Fit value function $V_\phi^\pi(s)$ by MC or TD estimation (update ϕ)
3. Compute advantages $A^\pi(s_t^i, a_t^i) = R(s_t^i, a_t^i) + \gamma V_\phi^\pi(s_{t+1}^i) - V_\phi^\pi(s_t^i)$
4. $\nabla_\theta U(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\alpha_t^i \mid s_t^i) A^\pi(s_t^i, a_t^i)$
5. $\theta' = \theta + \boxed{\alpha} \nabla_\theta U(\theta)$

This lecture is about this stepsize





Choosing a stepsize

Reinforcement learning and policy gradients:

$$\hat{g}^{PG} \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\alpha_t^{(i)} \mid s_t^{(i)}) A^\pi(s_t^{(i)}, a_t^{(i)}), \quad \tau_i \sim \pi_\theta$$

Supervised learning using expert actions $\tilde{a} \sim \pi^*$:

$$U^{SL}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \log \pi_\theta(\tilde{\alpha}_t^{(i)} \mid s_t^{(i)}), \quad \tau_i \sim \pi^* \qquad \text{(+regularization)}$$

with gradient:

$$\hat{g}^{SL} \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\tilde{\alpha}_t^{(i)} \mid s_t^{(i)}), \quad \tau_i \sim \pi^*$$

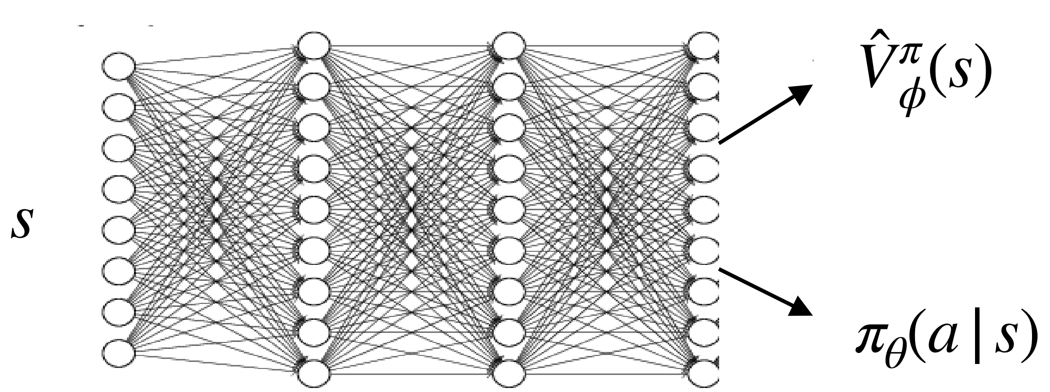
We want to optimize both objectives using gradient descent

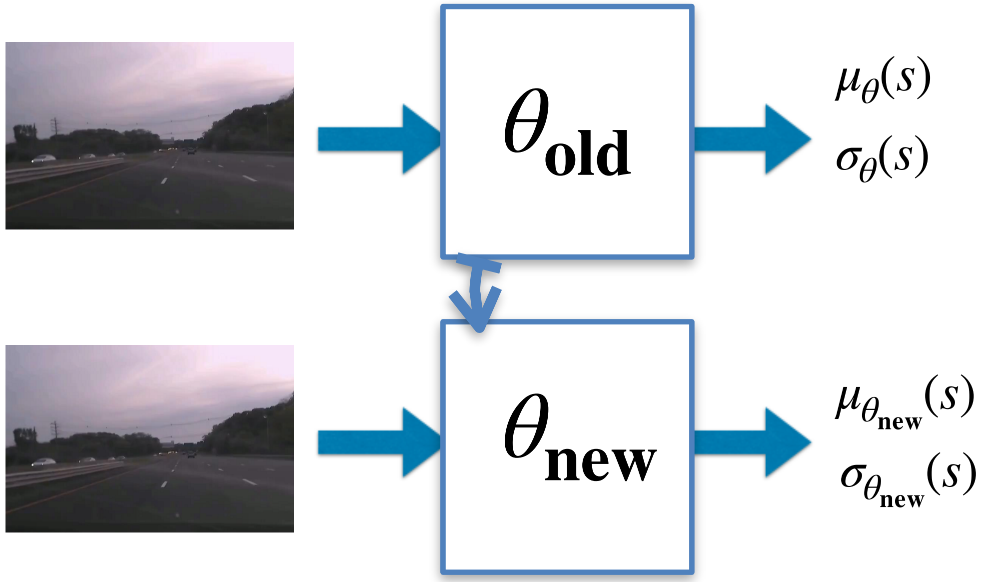
$$\theta' = \theta + \alpha \nabla_\theta U(\theta)$$

Choosing the right stepsize is more critical for RL than for SL.

Choosing a stepsize

- Step too big:Bad policy->data collected under bad policy-> we cannot recover. In Supervised Learning, data does not depend on neural network weights.
- Step too small: Not efficient use of experience. In Supervised Learning, data can be trivially re-used

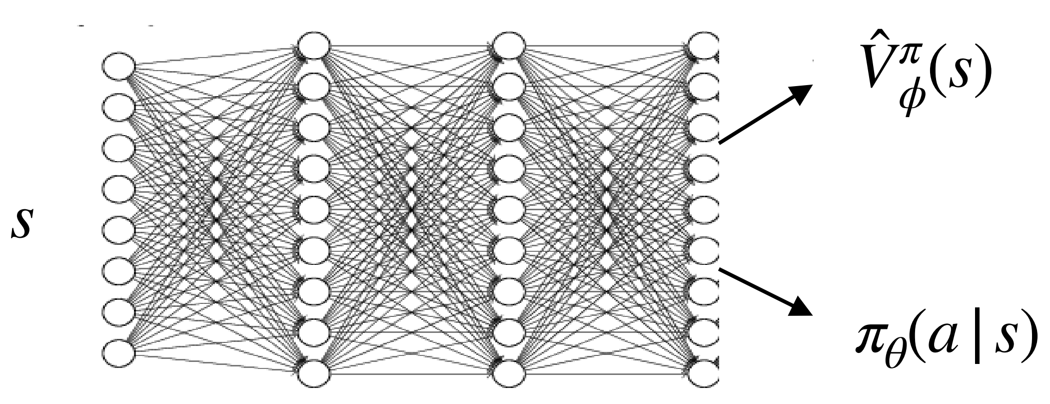


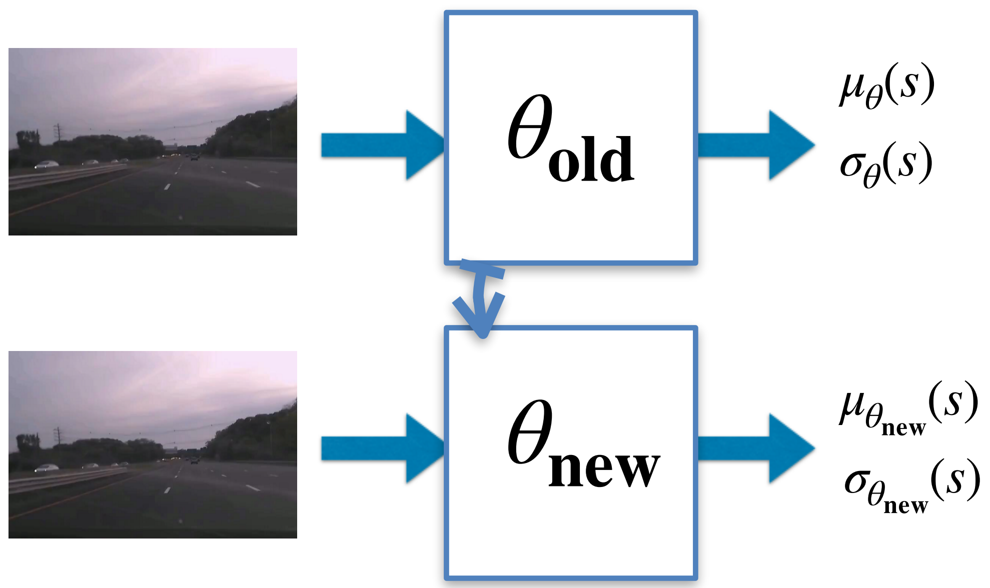


Choosing a stepsize

- Step too big:Bad policy->data collected under bad policy-> we cannot recover. In Supervised Learning, data does not depend on neural network weights.
- Step too small: Not efficient use of experience. In Supervised Learning, data can be trivially re-used

Gradient descent in parameter space does not take into account the resulting distance in the (output) policy space between $\pi_{\theta_{\text{old}}}(s)$ and $\pi_{\theta_{\text{new}}}(s)$

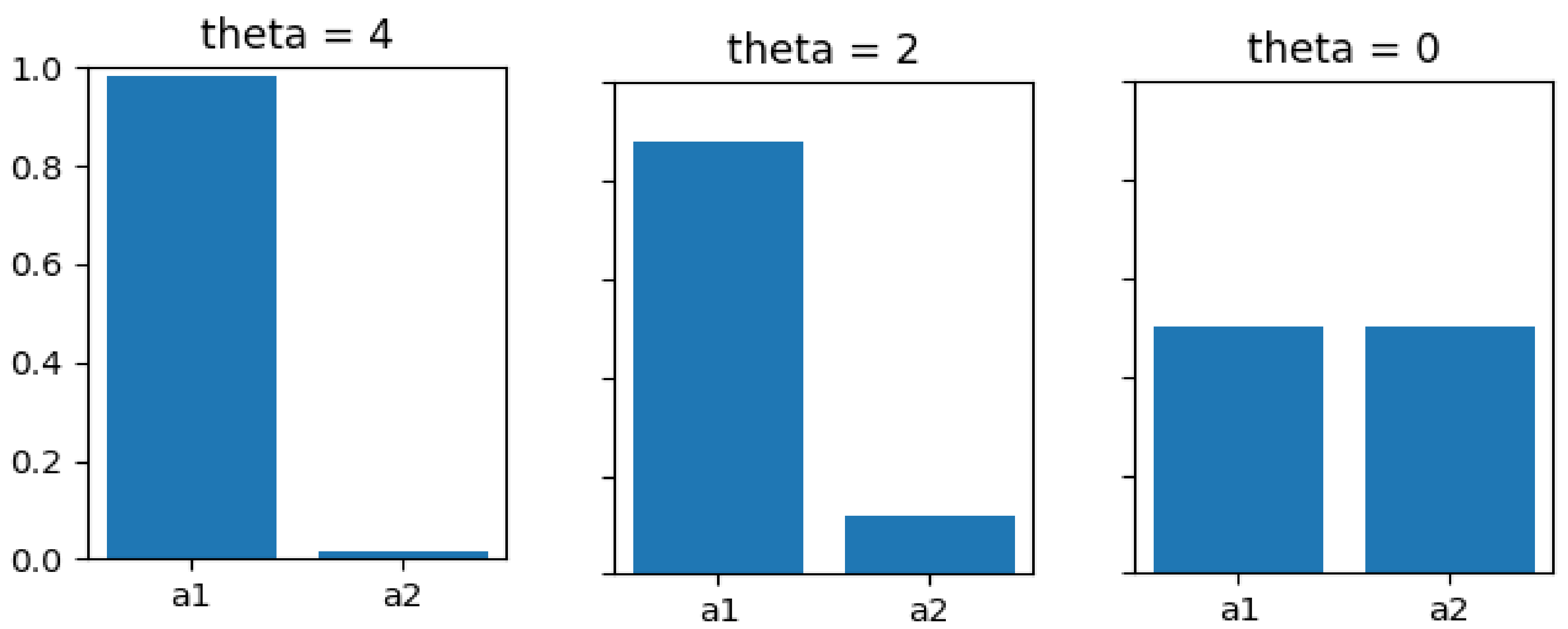




Hard to choose stepsizes

Consider a family of policies with parametrization:

$$\pi_\theta(a) = \begin{cases} \sigma(\theta) & a = 1 \\ 1 - \sigma(\theta) & a = 2 \end{cases}$$



The same parameter step $\Delta\theta = -2$ changes the policy distribution more or less dramatically depending on where in the parameter space we are.

Notation

We will use the following to denote values of parameters and corresponding policies before and after an update:

$$\theta_{old} \rightarrow \theta_{new}$$

$$\pi_{old} \rightarrow \pi_{new}$$

$$\theta \rightarrow \theta'$$

$$\pi \rightarrow \pi'$$

Gradient Descent in Parameter Space

Consider a parameterized distribution π_θ and an objective $U(\theta)$ that depends on θ through π_θ.
The stepwise in gradient descent results from solving the following optimization problem:

$$d* = \arg\max_{\|d\| \leq \epsilon} U(\theta + d)$$

Euclidean distance in parameter space

SGD: $\theta_{new} = \theta_{old} + d*$

It is hard to predict the result of the parameter update $\theta_{new} = \theta_{old} + d*$ on the parameterized distribution $\pi(\theta)$. It is hard to pick the threshold epsilon.

Gradient Descent in Distribution Space

Consider a parameterized distribution π_θ and an objective $U(\theta)$ that depends on θ through π_θ.
The stepwise in gradient descent results from solving the following optimization problem:

$$d* = \arg\max_{\|d\| \leq \epsilon} U(\theta + d)$$

SGD: $\theta_{new} = \theta_{old} + d*$

Euclidean distance in parameter space

It is hard to predict the result of the parameter update $\theta_{new} = \theta_{old} + d*$ on the parameterized distribution $\pi(\theta)$. It is hard to pick the threshold epsilon.

Natural gradient descent: the stepwise in parameter space is determined by considering the KL divergence in the distributions before and after the update:

$$d* = \arg\max_{\mathrm{KL}(\pi_\theta \| \pi_{\theta+d}) \leq \epsilon} U(\theta + d)$$

KL divergence in distribution space

Easier to pick the distance threshold!

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

Solving the KL Constrained Problem

Unconstrained penalized objective:

$$d* = \arg\max_d U(\theta + d) - \lambda(\mathrm{D}_{\mathrm{KL}}\left[\pi_\theta \| \pi_{\theta+d}\right] - \epsilon)$$

$$\approx \arg\max_d U(\theta_{old}) + \nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d - \frac{1}{2}\lambda(d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}\left[\pi_{\theta_{old}} \| \pi_\theta\right]|_{\theta=\theta_{old}} d) + \lambda\epsilon$$

(First order Taylor expansion for the loss and second order for the KL)

Taylor expansion of KL

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) \approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta)|_{\theta=\theta_{old}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta)|_{\theta=\theta_{old}} d$$

$$\boxed{\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) = \mathbb{E}_{x \sim p_{\theta_{old}}} \log \left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)} \right)}$$

Taylor expansion of KL

$$D_{KL}(p_{\theta_{old}}|p_\theta) \approx D_{KL}(p_{\theta_{old}}|p_{\theta_{old}}) + d^\top \boxed{\nabla_\theta D_{KL}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}}} + \frac{1}{2}d^\top \nabla_\theta^2 D_{KL}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} d$$

$$\boxed{D_{KL}(p_{\theta_{old}}|p_\theta) = \mathbb{E}_{x\sim p_{\theta_{old}}} \log\left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)}\right)}$$

Taylor expansion of KL

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) \approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_{\theta_{old}}) + d^\top \boxed{\nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} d$$

$$\boxed{\nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}}} = -\nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_\theta(x)|_{\theta=\theta_{old}} + \nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_{\theta_{old}}(x)|_{\theta=\theta_{old}}$$

$$\boxed{\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) = \mathbb{E}_{x\sim p_{\theta_{old}}} \log\left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)}\right)}$$

Taylor expansion of KL

$$\mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) \approx \mathrm{D_{KL}}(p_{\theta_{old}} | p_{\theta_{old}}) + d^\top \boxed{\nabla_\theta \mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}}} + \frac{1}{2} d^\top \nabla^2_\theta \mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}} d$$

$$\begin{aligned} \boxed{\nabla_\theta \mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}}} &= -\nabla_\theta \mathbb{E}_{x \sim p_{\theta_{old}}} \log P_\theta(x) |_{\theta=\theta_{old}} + \nabla_\theta \mathbb{E}_{x \sim p_{\theta_{old}}} \log P_{\theta_{old}}(x) |_{\theta=\theta_{old}} \\ &= -\mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta \log P_\theta(x) |_{\theta=\theta_{old}} \end{aligned}$$

$$\boxed{\mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) = \mathbb{E}_{x \sim p_{\theta_{old}}} \log \left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)} \right)}$$

Taylor expansion of KL

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) \approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} d$$

$$\begin{aligned}
\nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} &= -\nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_\theta(x)|_{\theta=\theta_{old}} + \nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_{\theta_{old}}(x)|_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x\sim p_{\theta_{old}}} \nabla_\theta \log P_\theta(x)|_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x\sim p_{\theta_{old}}} \frac{1}{P_{\theta_{old}}(x)} \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}}
\end{aligned}$$

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) = \mathbb{E}_{x\sim p_{\theta_{old}}} \log\left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)}\right)$$

Taylor expansion of KL

$$D_{KL}(p_{\theta_{old}}|p_\theta) \approx D_{KL}(p_{\theta_{old}}|p_{\theta_{old}}) + d^\top \boxed{\nabla_\theta D_{KL}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}}} + \frac{1}{2} d^\top \nabla_\theta^2 D_{KL}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} d$$

$$\boxed{\nabla_\theta D_{KL}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}}} = -\nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_\theta(x)|_{\theta=\theta_{old}} + \nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_{\theta_{old}}(x)|_{\theta=\theta_{old}}$$

$$= -\mathbb{E}_{x\sim p_{\theta_{old}}} \nabla_\theta \log P_\theta(x)|_{\theta=\theta_{old}}$$

$$= -\mathbb{E}_{x\sim p_{\theta_{old}}} \frac{1}{P_{\theta_{old}}(x)} \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}}$$

$$= \int_x P_{\theta_{old}}(x) \frac{1}{P_{\theta_{old}}(x)} \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}}$$

$$\boxed{D_{KL}(p_{\theta_{old}}|p_\theta) = \mathbb{E}_{x\sim p_{\theta_{old}}} \log\left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)}\right)}$$

Taylor expansion of KL

$$D_{KL}(p_{\theta_{old}}|p_\theta) \approx D_{KL}(p_{\theta_{old}}|p_{\theta_{old}}) + d^\top \boxed{\nabla_\theta D_{KL}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}}} + \frac{1}{2} d^\top \nabla_\theta^2 D_{KL}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} d$$

$$\begin{aligned}
\boxed{\nabla_\theta D_{KL}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}}} &= -\nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_\theta(x)|_{\theta=\theta_{old}} + \nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_{\theta_{old}}(x)|_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x\sim p_{\theta_{old}}} \nabla_\theta \log P_\theta(x)|_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x\sim p_{\theta_{old}}} \frac{1}{P_{\theta_{old}}(x)} \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}} \\
&= \int_x P_{\theta_{old}}(x) \frac{1}{P_{\theta_{old}}(x)} \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}} \\
&= \int_x \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}}
\end{aligned}$$

$$\boxed{D_{KL}(p_{\theta_{old}}|p_\theta) = \mathbb{E}_{x\sim p_{\theta_{old}}} \log\left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)}\right)}$$

Taylor expansion of KL

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) \approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} d$$

$$\begin{aligned}
\nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} &= -\nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_\theta(x)|_{\theta=\theta_{old}} + \nabla_\theta \mathbb{E}_{x\sim p_{\theta_{old}}} \log P_{\theta_{old}}(x)|_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x\sim p_{\theta_{old}}} \nabla_\theta \log P_\theta(x)|_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x\sim p_{\theta_{old}}} \frac{1}{P_{\theta_{old}}(x)} \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}} \\
&= \int_x P_{\theta_{old}}(x) \frac{1}{P_{\theta_{old}}(x)} \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}} \\
&= \int_x \nabla_\theta P_\theta(x)|_{\theta=\theta_{old}} \\
&= \nabla_\theta \int_x P_\theta(x)|_{\theta=\theta_{old}} \\
&= 0
\end{aligned}$$

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) = \mathbb{E}_{x\sim p_{\theta_{old}}} \log\left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)}\right)$$

Taylor expansion of KL

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) \approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_{\theta_{old}}) + d^\top \boxed{\nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}} d$$

$$\boxed{\nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}}} = -\mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta^2 \log P_\theta(x) |_{\theta=\theta_{old}}$$

$$\boxed{\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) = \mathbb{E}_{x \sim p_{\theta_{old}}} \log \left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)} \right)}$$

Taylor expansion of KL

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) \approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} d$$

$$\begin{aligned}\nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} &= -\mathbb{E}_{x\sim p_{\theta_{old}}} \nabla_\theta^2 \log P_\theta(x)|_{\theta=\theta_{old}} \\ &= -\mathbb{E}_{x\sim p_{\theta_{old}}} \nabla_\theta \left(\frac{\nabla_\theta P_\theta(x)}{P_\theta(x)} \right)|_{\theta=\theta_{old}}\end{aligned}$$

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) = \mathbb{E}_{x\sim p_{\theta_{old}}} \log \left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)} \right)$$

Taylor expansion of KL

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) \approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} d$$

$$\begin{aligned}
\nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta)|_{\theta=\theta_{old}} &= -\mathbb{E}_{x\sim p_{\theta_{old}}} \nabla_\theta^2 \log P_\theta(x)|_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x\sim p_{\theta_{old}}} \nabla_\theta \left(\frac{\nabla_\theta P_\theta(x)}{P_\theta(x)} \right)|_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x\sim p_{\theta_{old}}} \left(\frac{\nabla_\theta^2 P_\theta(x) P_\theta(x) - \nabla_\theta P_\theta(x) \nabla_\theta P_\theta(x)^\top}{P_\theta(x)^2} \right)|_{\theta=\theta_{old}}
\end{aligned}$$

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}}|p_\theta) = \mathbb{E}_{x\sim p_{\theta_{old}}} \log \left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)} \right)$$

Taylor expansion of KL

$$\mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) \approx \mathrm{D_{KL}}(p_{\theta_{old}} | p_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}} d$$

$$\begin{aligned}
\nabla_\theta^2 \mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}} &= -\mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta^2 \log P_\theta(x) |_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta \left(\frac{\nabla_\theta P_\theta(x)}{P_\theta(x)} \right) |_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x \sim p_{\theta_{old}}} \left(\frac{\nabla_\theta^2 P_\theta(x) P_\theta(x) - \nabla_\theta P_\theta(x) \nabla_\theta P_\theta(x)^\top}{P_\theta(x)^2} \right) |_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x \sim p_{\theta_{old}}} \frac{\nabla_\theta^2 P_\theta(x) |_{\theta=\theta_{old}}}{P_{\theta_{old}}(x)} + \mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta \log P_\theta(x) \nabla_\theta \log P_\theta(x)^\top |_{\theta=\theta_{o}}
\end{aligned}$$

$$\mathrm{D_{KL}}(p_{\theta_{old}} | p_\theta) = \mathbb{E}_{x \sim p_{\theta_{old}}} \log \left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)} \right)$$

Taylor expansion of KL

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) \approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}} + \frac{1}{2} d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}} d$$

$$\begin{aligned}
\nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) |_{\theta=\theta_{old}} &= -\mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta^2 \log P_\theta(x) |_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta \left(\frac{\nabla_\theta P_\theta(x)}{P_\theta(x)} \right) |_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x \sim p_{\theta_{old}}} \left(\frac{\nabla_\theta^2 P_\theta(x) P_\theta(x) - \nabla_\theta P_\theta(x) \nabla_\theta P_\theta(x)^\top}{P_\theta(x)^2} \right) |_{\theta=\theta_{old}} \\
&= -\mathbb{E}_{x \sim p_{\theta_{old}}} \frac{\nabla_\theta^2 P_\theta(x) |_{\theta=\theta_{old}}}{P_{\theta_{old}}(x)} + \mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta \log P_\theta(x) \nabla_\theta \log P_\theta(x)^\top |_{\theta=\theta_{old}} \\
&= \mathbb{E}_{x \sim p_{\theta_{old}}} \nabla_\theta \log P_\theta(x) \nabla_\theta \log P_\theta(x)^\top |_{\theta=\theta_{old}}
\end{aligned}$$

$$\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) = \mathbb{E}_{x \sim p_{\theta_{old}}} \log \left(\frac{P_{\theta_{old}}(x)}{P_\theta(x)} \right)$$

Fisher Information Matrix

Exactly equivalent to the Hessian of KL divergence!

$$\mathbf{F}(\theta) = \mathbb{E}_{x \sim p_\theta}\left[\nabla_\theta \log p_\theta(x)\, \nabla_\theta \log p_\theta(x)^\top\right]$$

$$\mathbf{F}(\theta_{old}) = \nabla^2_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta)\,|_{\theta=\theta_{old}}$$

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta) &\approx \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta)\,|_{\theta=\theta_{old}} + \frac{1}{2} d^\top \nabla^2_\theta \mathrm{D}_{\mathrm{KL}}(p_{\theta_{old}} | p_\theta)\,|_{\theta=\theta_{old}}\, d \\ &= \frac{1}{2} d^\top \mathbf{F}(\theta_{old}) d \\ &= \frac{1}{2}(\theta - \theta_{old})^\top \mathbf{F}(\theta_{old})(\theta - \theta_{old})\end{aligned}$$

Since KL divergence is roughly analogous to a distance measure between distributions, Fisher information serves as a local distance metric between distributions: how much you change the distribution if you move the parameters a little bit in a given direction.

Solving the KL Constrained Problem

Unconstrained penalized objective:

$$d^* = \arg\max_d U(\theta + d) - \lambda(\mathrm{D}_{\mathrm{KL}}\left[\pi_\theta \| \pi_{\theta+d}\right] - \epsilon)$$

First order Taylor expansion for the loss and second order for the KL:

$$\approx \arg\max_d U(\theta_{old}) + \nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d - \frac{1}{2}\lambda(d^\top \nabla_\theta^2 \mathrm{D}_{\mathrm{KL}}\left[\pi_{\theta_{old}} \| \pi_\theta\right]|_{\theta=\theta_{old}} d) + \lambda\epsilon$$

Substitute for the information matrix:

$$= \arg\max_d \nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d - \frac{1}{2}\lambda(d^\top \mathbf{F}(\theta_{old}) d)$$

$$= \arg\min_d - \nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d + \frac{1}{2}\lambda(d^\top \mathbf{F}(\theta_{old}) d)$$

Natural Gradient Descent

Setting the gradient to zero:

$$0 = \frac{\partial}{\partial d}\left(-\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d + \frac{1}{2}\lambda(d^\top \mathbf{F}(\theta_{old})d)\right)$$

$$= -\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} + \frac{1}{2}\lambda(\mathbf{F}(\theta_{old}))d$$

$$d = \frac{2}{\lambda}\mathbf{F}^{-1}(\theta_{old})\,\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}}$$

The natural gradient: $g_N = \mathbf{F}^{-1}(\theta_{old})\,\nabla_\theta U(\theta)$

$$\theta_{new} = \theta_{old} + \alpha \cdot g_N$$

Natural Gradient Descent

Setting the gradient to zero:

$$0 = \frac{\partial}{\partial d}\left(-\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d + \frac{1}{2}\lambda(d^\top \mathbf{F}(\theta_{old})d)\right)$$

$$= -\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} + \frac{1}{2}\lambda(\mathbf{F}(\theta_{old}))d$$

$$d = \frac{2}{\lambda}\mathbf{F}^{-1}(\theta_{old})\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}}$$

The natural gradient: $g_N = \mathbf{F}^{-1}(\theta_{old})\nabla_\theta U(\theta)$ what is this?

$$\theta_{new} = \theta_{old} + \alpha \cdot g_N$$

The police gradient:
$\nabla_\theta \log \pi_\theta(a\,|\,s)A(a\,|\,s)$

Natural Gradient Descent

Setting the gradient to zero:

$$0 = \frac{\partial}{\partial d}\left(-\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d + \frac{1}{2}\lambda(d^\top \mathbf{F}(\theta_{old})d) \right)$$

$$= -\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} + \frac{1}{2}\lambda(\mathbf{F}(\theta_{old}))d$$

$$d = \frac{2}{\lambda}\mathbf{F}^{-1}(\theta_{old})\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}}$$

The natural gradient: $g_N = \mathbf{F}^{-1}(\theta_{old})\nabla_\theta U(\theta)$ what is this?

$$\theta_{new} = \theta_{old} + \alpha \cdot g_N$$

Stepsize along the natural gradient direction

The police gradient:
$\nabla_\theta \log \pi_\theta(a\,|\,s)A(a\,|\,s)$

Stepsize along the Natural Gradient direction

The natural gradient: $g_N = \mathbf{F}^{-1}(\theta_{old})\,\nabla_\theta U(\theta)$

$$\theta_{new} = \theta_{old} + \alpha \cdot g_N$$

Let's solve for the stepzise along the natural gradient direction!

$$\mathrm{D_{KL}}(\pi_{\theta_{old}} | \pi_\theta) \approx \frac{1}{2}(\theta - \theta_{old})^\top \mathbf{F}(\theta_{old})(\theta - \theta_{old}) = \frac{1}{2}(\alpha g_N)^\top \mathbf{F}(\alpha g_N)$$

I want the KL between old and new policies to be at most ϵ: $\frac{1}{2}(\alpha g_N)^\top \mathbf{F}(\alpha g_N) = \epsilon$

$$\alpha = \sqrt{\frac{2\epsilon}{(g_N^\top \mathbf{F}^{-1} g_N)}}$$

Natural Gradient Descent

Algorithm 1 Natural Policy Gradient

Input: initial policy parameters θ_0
for $k = 0, 1, 2, \ldots$ **do**
  Collect set of trajectories $\mathcal{D}_k$ on policy $\pi_k = \pi(\theta_k)$
  Estimate advantages $\hat{A}_t^{\pi_k}$ using any advantage estimation algorithm
  Form sample estimates for

- policy gradient $\hat{g}_k$ (using advantage estimates)
- and KL-divergence Hessian / Fisher Information Matrix $\hat{F}_k^{-1}$

  Compute Natural Policy Gradient update:

$$\theta_{k+1} = \theta_k + \sqrt{\frac{2\epsilon}{\hat{g}_k^T \hat{F}_k^{-1} \hat{g}_k}} \hat{F}_k^{-1} \hat{g}_k$$

end for

Both use samples from the current policy $\pi_k = \pi(\theta_k)$

Off-policy learning with Importance Sampling

$$
\begin{aligned}
U(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[R(\tau) \right] \\
&= \sum_{\tau} \pi_\theta(\tau) R(\tau)
\end{aligned}
$$

Off-policy learning with Importance Sampling

$$\begin{aligned} U(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[R(\tau) \right] \\ &= \sum_\tau \pi_\theta(\tau) R(\tau) \\ &= \sum_\tau \pi_{\theta_{old}}(\tau) \frac{\pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau) \end{aligned}$$

Off-policy learning with Importance Sampling

$$
\begin{aligned}
U(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta(\tau)}\left[R(\tau)\right] \\
&= \sum_{\tau} \pi_\theta(\tau) R(\tau) \\
&= \sum_{\tau} \pi_{\theta_{old}}(\tau) \frac{\pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau) \\
&= \mathbb{E}_{\tau \sim \pi_{\theta_{old}}} \frac{\pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau)
\end{aligned}
$$

Off-policy learning with Importance Sampling

$$
\begin{aligned}
U(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta(\tau)}\left[R(\tau)\right] \\
&= \sum_\tau \pi_\theta(\tau) R(\tau) \\
&= \sum_\tau \pi_{\theta_{old}}(\tau) \frac{\pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau) \\
&= \mathbb{E}_{\tau \sim \pi_{\theta_{old}}} \frac{\pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau)
\end{aligned}
$$

$$
\nabla_\theta U(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{old}}} \frac{\nabla_\theta \pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau)
$$

Off-policy learning with Importance Sampling

$$
\begin{aligned}
U(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta(\tau)}\left[R(\tau)\right] \\
&= \sum_\tau \pi_\theta(\tau) R(\tau) \\
&= \sum_\tau \pi_{\theta_{old}}(\tau) \frac{\pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau) \\
&= \mathbb{E}_{\tau \sim \pi_{\theta_{old}}} \frac{\pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau)
\end{aligned}
$$

$$
\nabla_\theta U(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{old}}} \frac{\nabla_\theta \pi_\theta(\tau)}{\pi_{\theta_{old}}(\tau)} R(\tau)
$$

$$
\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} = \mathbb{E}_{\tau \sim \pi_{\theta_{old}}} \nabla_\theta \log \pi_\theta(\tau)\,|_{\theta=\theta_{old}} R(\tau)
$$

Gradient evaluated at θ_{old} is unchanged.

Trust region Policy Optimization

Due to the quadratic approximation, the KL constraint may be violated! What if we just do a line search to find the best stepsize, making sure:

- I am improving my objective $U(\theta)$
- The KL constraint is not violated.

Algorithm 2 Line Search for TRPO

Compute proposed policy step $\Delta_k = \sqrt{\frac{2\delta}{\hat{g}_k^T \hat{H}_k^{-1} \hat{g}_k}} \hat{H}_k^{-1} \hat{g}_k$
for $j = 0, 1, 2, ..., L$ **do**
  Compute proposed update $\theta = \theta_k + \alpha^j \Delta_k$
  if $\mathcal{L}_{\theta_k}(\theta) \geq 0$ and $\bar{D}_{KL}(\theta || \theta_k) \leq \delta$ **then**
    accept the update and set $\theta_{k+1} = \theta_k + \alpha^j \Delta_k$
    break
  end if
end for

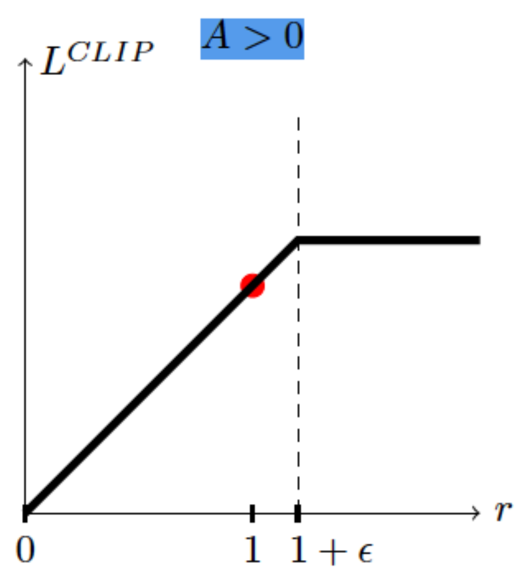
Trust Region Policy Optimization, Schulman et al. 2015

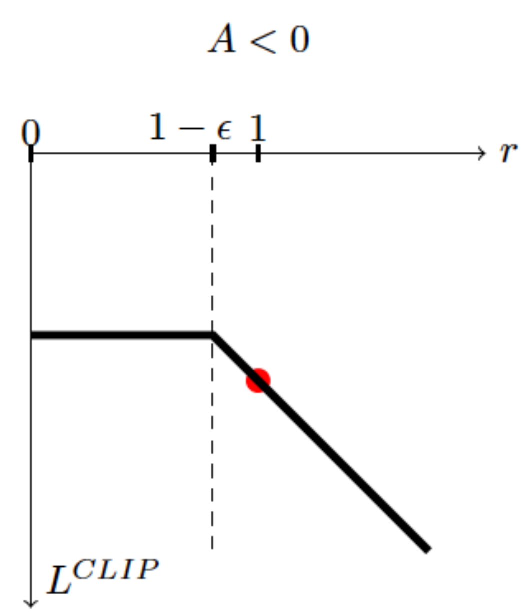
Proximal Policy Optimization

Can I achieve similar performance without second order information (no Fisher matrix!)

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}$$

$$\max_{\theta}. \quad L^{CLIP} = \mathbb{E}_t \left[\min\left(r_t(\theta) A(s_t, a_t), \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon \right) A(s_t, a_t) \right) \right]$$





Proximal Policy Optimization Algorithms. *J. Schulman, F. Wolski, P. Dhariwal, A. Radfor and O. Klimov*

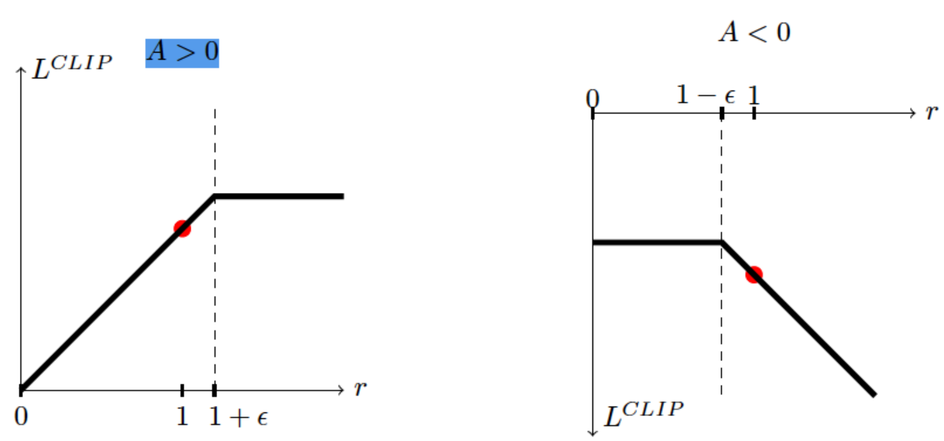
PPO: Clipped Objective

- Recall the surrogate objective:

$$L^{IS}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[r_t(\theta) \hat{A}_t \right]$$

- Form a lower bound via clipped importance ratio:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right) \hat{A}_t \right) \right]$$



Proximal Policy Optimization Algorithms. *J. Schulman, F. Wolski, P. Dhariwal, A. Radfor and O. Klimov*

PPO: Clipped Objective

Input: initial policy parameters θ_0, clipping threshold ϵ
for $k = 0, 1, 2, \ldots$ **do**
Collect set of partial trajectories $\mathcal{D}_k$ on policy $\pi_k = \pi(\theta_k)$
Estimate advantages $\hat{A}_t^{\pi_k}$ using any advantage estimation algorithm
Compute policy update

$$\theta_{k+1} = \arg\max_{\theta} \mathcal{L}_{\theta_k}^{CLIP}(\theta)$$

by taking K steps of minibatch SGD (via Adam), where

$$\mathcal{L}_{\theta_k}^{CLIP}(\theta) = \mathop{\mathrm{E}}_{\tau \sim \pi_k} \left[\sum_{t=0}^{T} \left[\min(r_t(\theta) \hat{A}_t^{\pi_k}, \mathrm{clip}\left(r_t(\theta), 1 - \epsilon, 1 + \epsilon\right) \hat{A}_t^{\pi_k}) \right] \right]$$

end for

- Clipping prevents policy from having incentive to go far away from θ_{k+1}
- Clipping seems to work at least as well as PPO with KL penalty, but is simpler to implement

PPO: Clipped Objective

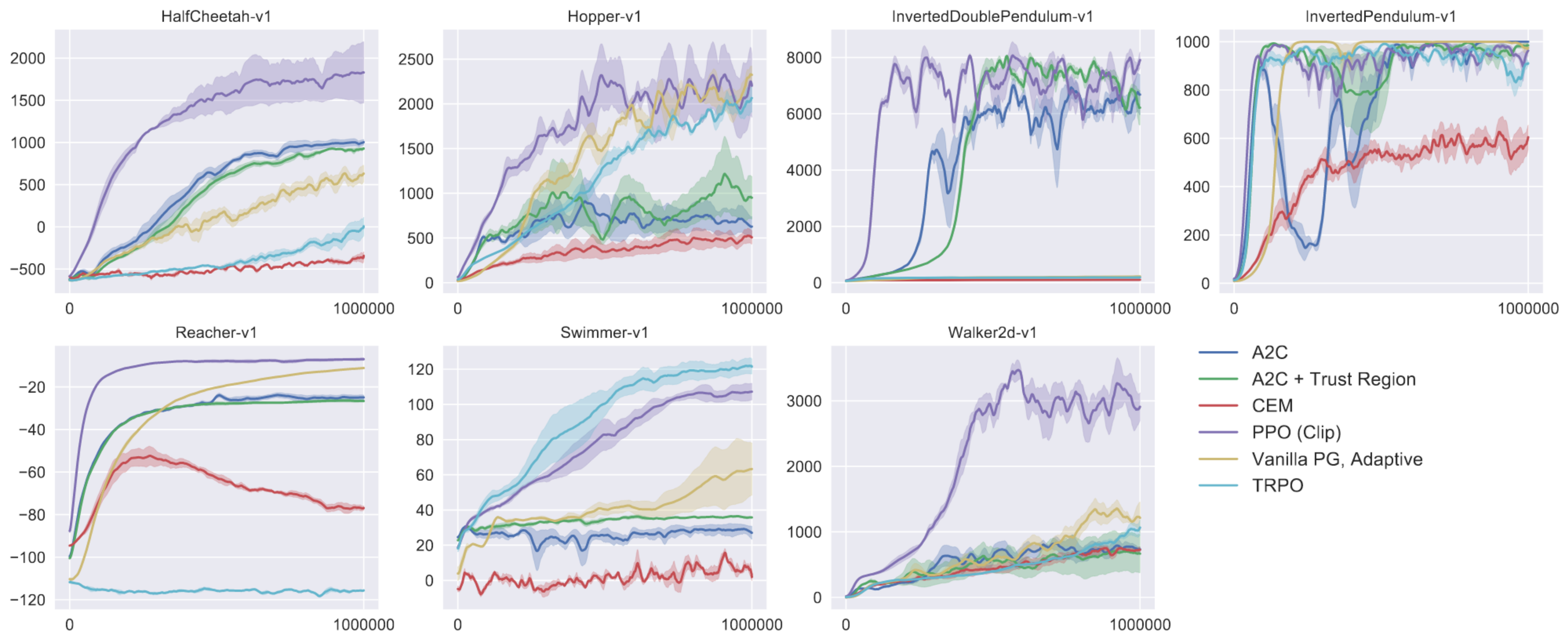


Figure: Performance comparison between PPO with clipped objective and various other deep RL methods on a slate of MuJoCo tasks. [10]

Summary

- Gradient Descent in Parameter VS distribution space
- Natural gradients: we need to keep track of how the KL changes from iteration to iteration
- Natural policy gradients
- Clipped objective works well