Carnegie Mellon

School of Computer Science

Deep Reinforcement Learning and Control

# Natural Policy Gradients

Fall 2020, CMU 10-703

Katerina Fragkiadaki

# Stepsize for Actor-Critic?

0. Initialize policy parameters $\theta$ and critic parameters $\phi$ .

1. Sample trajectories $\{\tau_i = \{s_t^i, a_t^i\}_{i=0}^T\}$ by deploying the current policy $\pi_\theta(a_t \mid s_t)$ .

2. Fit value function $V_\phi^\pi(s)$ by MC or TD estimation (update $\phi$)

3. Compute action 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. $\nabla_\theta U(\theta) \approx \hat{g} = \dfrac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\alpha_t^i \mid s_t^i) A^\pi(s_t^i, a_t^i)$

5. $\theta \leftarrow \theta + \alpha \nabla_\theta U(\theta)$

What should be the step size?

# Choosing a stepsize in RL VS SL

- Reinforcement learning objective

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

with gradient:

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

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

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

with gradient:

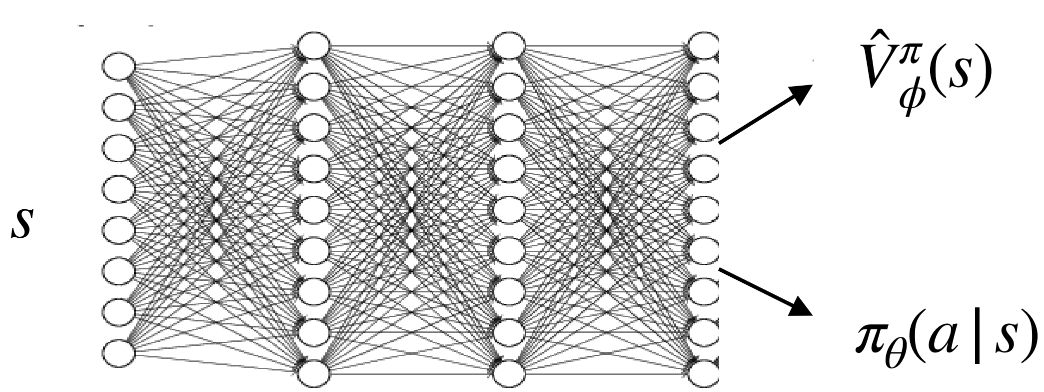
$$\hat{g}^{SL} \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\tilde{\alpha}_t^{(i)} \,|\, s_t^{(i)}), \quad \tau_i \sim \pi^*$$

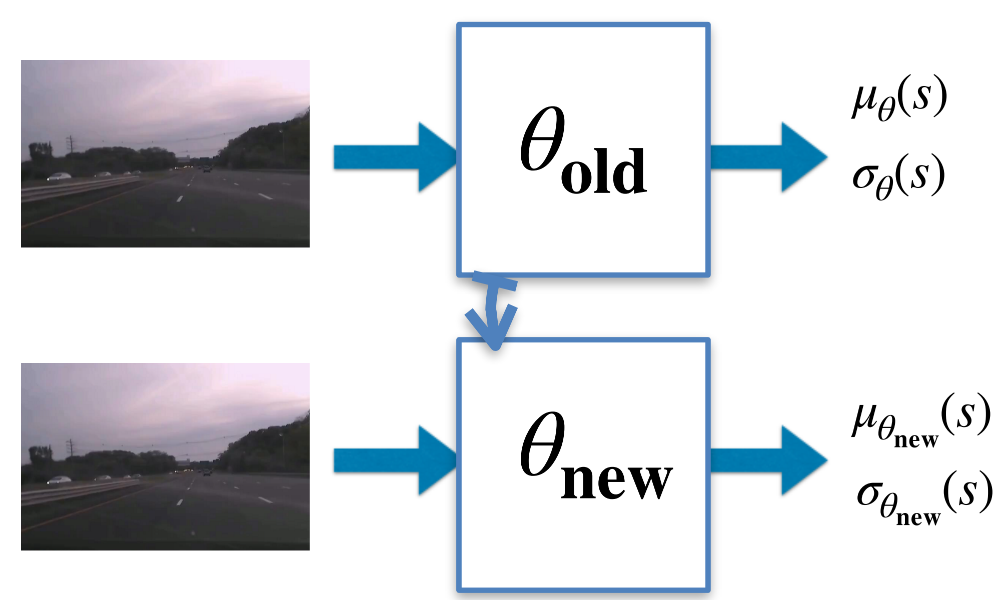
We want to take a gradient step:

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

# 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.

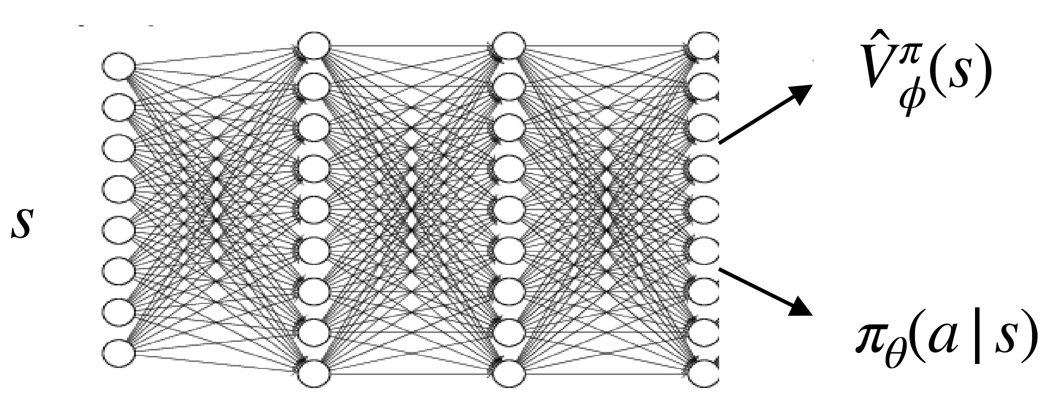


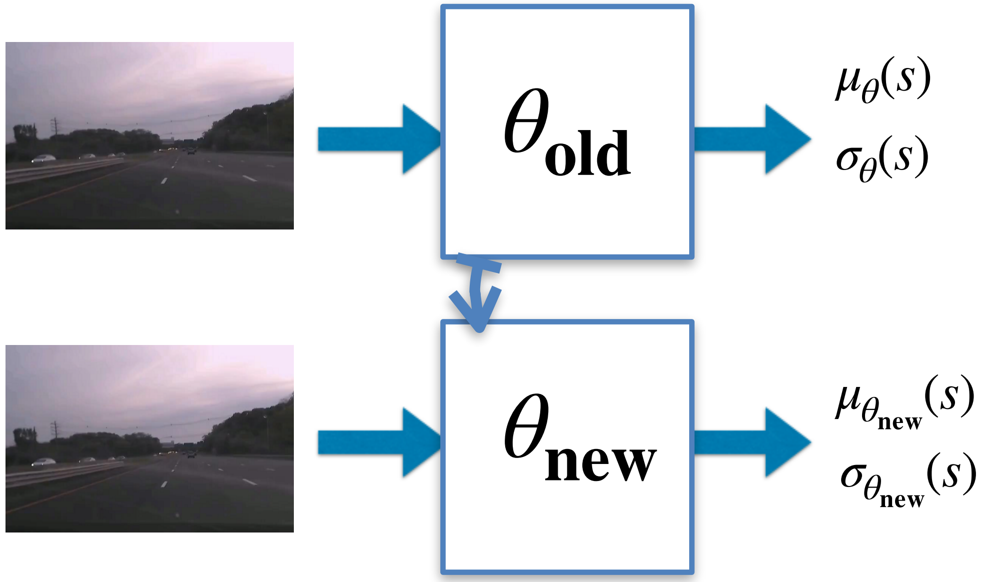



# Choosing a stepsize

- 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 $\pi_{\theta_{\mathbf{old}}}(s)$ and $\pi_{\theta_{\mathbf{new}}}(s)$

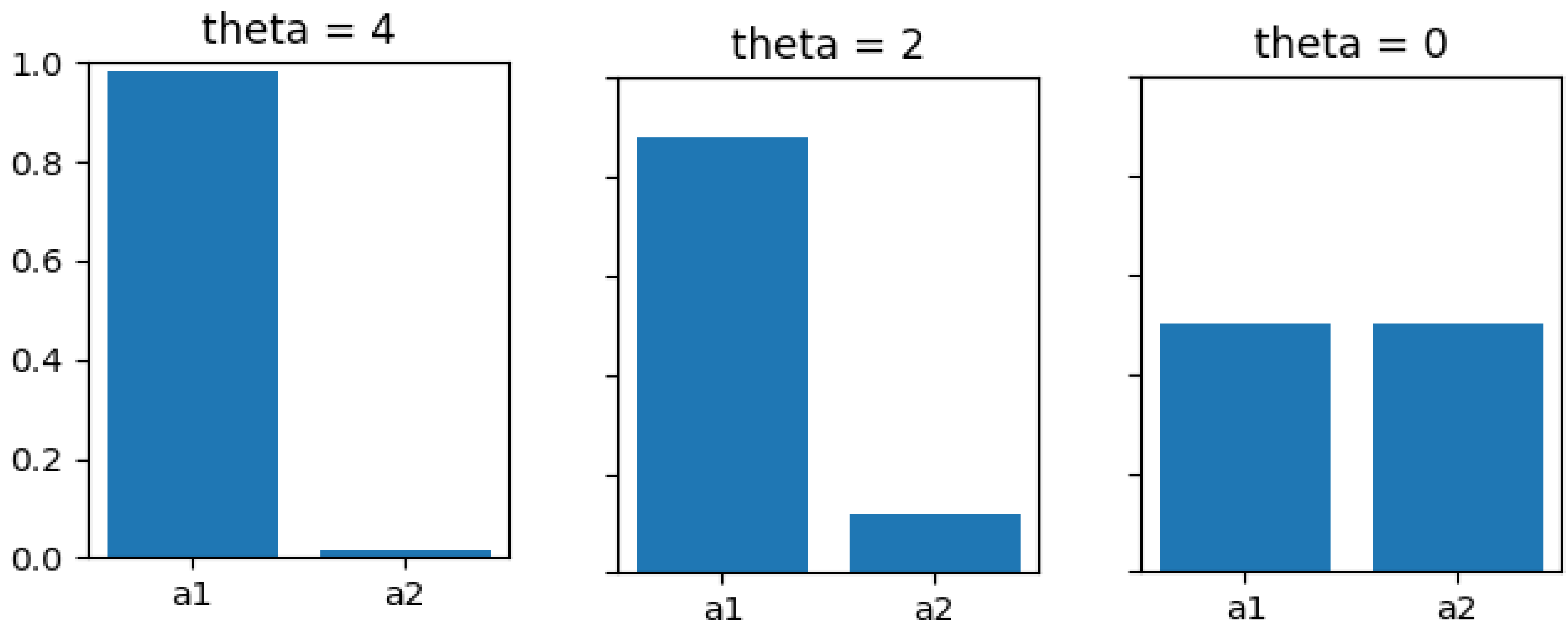


# Choosing a stepsize

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 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 Distribution Space

Consider a parameterized distribution $\pi_\theta$ and an objective $U(\theta)$ that depends on $\theta$ through $\pi_\theta$ and for which we want to take a gradient step.

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

# Gradient Descent in Distribution Space

Consider a parameterized distribution $\pi_\theta$ and an objective $U(\theta)$ that depends on $\theta$ through $\pi_\theta$ and for which we want to take a gradient step.

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

- Gradient descent: the step in parameter space is determined by considering the Euclidean distance of the parameter vectors before and after the update:

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

Euclidean distance in parameter space

It is hard to predict how different is $\pi_{\theta_{new}}$ from $\pi_{\theta_{old}}$. It is hard to pick the threshold epsilon.

# Gradient Descent in Distribution Space

Consider a parameterized distribution $\pi_\theta$ and an objective $U(\theta)$ that depends on $\theta$ through $\pi_\theta$ and for which we want to take a gradient step.

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

- Gradient descent: the step in parameter space is determined by considering the Euclidean distance of the parameter vectors before and after the update:

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

Euclidean distance in parameter space

It is hard to predict how different is $\pi_{\theta_{new}}$ from $\pi_{\theta_{old}}$. It is hard to pick the threshold epsilon.

- Natural gradient descent: the step 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!

# Gradient Descent in Distribution Space

Consider a parameterized distribution $\pi_\theta$ and an objective $U(\theta)$ that depends on $\theta$ through $\pi_\theta$ and for which we want to take a gradient step.

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

- Gradient descent: the step in parameter space is determined by considering the Euclidean distance of the parameter vectors before and after the update:

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

Euclidean distance in parameter space

It is hard to predict how different is $\pi_{\theta_{new}}$ from $\pi_{\theta_{old}}$. It is hard to pick the threshold epsilon.

- Natural gradient descent: the step 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

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

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 objective and second order for the KL!

$$\approx \arg\max_d U(\theta_{old}) + \nabla_\theta U(\theta)|_{\theta=\theta_{old}} \cdot d - \lambda(\mathrm{D}_{\mathrm{KL}}(\pi_{\theta_{old}} | \pi_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(\pi_{\theta_{old}} | \pi_\theta)|_{\theta=\theta_{old}}$$

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

# Solving the KL Constrained Problem

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

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 objective and second order for the KL!

$$\approx \arg \max_d U(\theta_{old}) + \nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d - \lambda(\mathrm{D}_{\mathrm{KL}}(\pi_{\theta_{old}} \,|\, \pi_{\theta_{old}}) + d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(\pi_{\theta_{old}} \,|\, \pi_\theta)\,|_{\theta=\theta_{old}}$$

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

what is this?

The police gradient: $\nabla_\theta \log \pi_\theta(a \,|\, s) A(a \,|\, s)\,|_{\theta=\theta_{old}}$

# Solving the KL Constrained Problem

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

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 objective and second order for the KL!

$$\approx \arg \max_d U(\theta_{old}) + \nabla_\theta U(\theta)\,|_{\theta=\theta_{old}} \cdot d - \lambda(\cancel{\mathrm{D}_{\mathrm{KL}}(\pi_{\theta_{old}} | \pi_{\theta_{old}})} + \cancel{d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(\pi_{\theta_{old}} | \pi_\theta)\,|_{\theta=\theta_{old}}}$$

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

# Solving the KL Constrained Problem

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

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 objective and second order for the KL!

$$\approx \arg\max_d U(\theta_{old}) + \nabla_\theta U(\theta)|_{\theta=\theta_{old}} \cdot d - \lambda(\cancel{\mathrm{D}_{\mathrm{KL}}(\pi_{\theta_{old}} | \pi_{\theta_{old}})} + \cancel{d^\top \nabla_\theta \mathrm{D}_{\mathrm{KL}}(\pi_{\theta_{old}} | \pi_\theta)|_{\theta=\theta_{old}}}$$

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

# Solving the KL Constrained Problem

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

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 objective and second order for the KL!

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

# Taylor expansion of KL

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

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

# Taylor expansion of KL

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

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

# Taylor expansion of KL

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

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

The Fisher information matrix

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

Can be approximated by sampling:

$$\mathbf{F}(\theta_{old}) \approx \sum_{i=1, x^{(i)} \sim p_{\theta_{old}}}^{N} \left[ \nabla_\theta \log p_\theta(x^{(i)}) |_{\theta=\theta_{old}} \nabla_\theta \log p_\theta(x^{(i)}) |_{\theta=\theta_{old}}^\top \right]$$

# Solving the KL Constrained Problem

Unconstrained penalized objective:

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

First order Taylor expansion for the objective and second order for the KL!

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

Substitute for the information matrix:

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

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

# 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:

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

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

# Solving the KL Constrained Problem

Setting the gradient to zero:

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

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

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

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

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

# Solving the KL Constrained Problem

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

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

# 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: $g_N = \mathbf{F}^{-1}(\theta_{old})\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}}$ 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: $g_N = \mathbf{F}^{-1}(\theta_{old})\,\nabla_\theta U(\theta)\,|_{\theta=\theta_{old}}$

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

How shall we choose stepsize along the natural gradient direction

# Stepsize along the Natural Gradient direction

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

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

By the 2nd order Taylor expansion of KL:

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

I want the KL between old and new policies to be at most $\epsilon$.

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

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

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

**Algorithm 1** Natural Policy Gradient

Input: initial policy parameters $\theta_0$

**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**

TRPO= NPG +Linesearch+monotonic improvement theorem

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

**Algorithm 3** Trust Region Policy Optimization

Input: initial policy parameters $\theta_0$

**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-vector product function $f(v) = \hat{H}_k v$

Use CG with $n_{cg}$ iterations to obtain $x_k \approx \hat{H}_k^{-1} \hat{g}_k$

Estimate proposed step $\Delta_k \approx \sqrt{\frac{2\epsilon}{x_k^T \hat{H}_k x_k}} x_k$

Perform backtracking line search with exponential decay to obtain final update

$$\theta_{k+1} = \theta_k + \alpha^j \Delta_k$$

**end for**

# 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 $\bar{\mathbb{A}}_{\pi_{old}}(\pi)$
- The KL constraint is not violated.

**Algorithm 2** Line Search for TRPO

Compute proposed policy step $\Delta_k = \sqrt{\frac{2\epsilon}{\hat{g}_k^T \hat{H}_k^{-1} \hat{g}_k}} \hat{H}_k^{-1} \hat{g}_k$
**for** $j = 0, 1, 2, \ldots, L$ **do**
  Compute proposed update $\theta = \theta_k + \alpha^j \Delta_k$
  **if** $\bar{\mathbb{A}}_{\pi_{old}}(\pi) \geq 0$ 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!)

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

# Proximal Policy Optimization

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

- Adaptive KL Penalty
  - Policy update solves unconstrained optimization problem

$$\theta_{k+1} = \arg\max_{\theta} \bar{\mathbb{A}}_{\theta_k}(\theta) - \beta_k \bar{D}_{KL}(\theta||\theta_k)$$

  - Penalty coefficient $\beta_k$ changes between iterations to approximately enforce KL-divergence constraint

# Proximal Policy Optimization

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

- Adaptive KL Penalty
  - Policy update solves unconstrained optimization problem

$$\theta_{k+1} = \arg\max_{\theta} \bar{\mathbb{A}}_{\theta_k}(\theta) \, - \beta_k \bar{D}_{KL}(\theta || \theta_k)$$

  - Penalty coefficient $\beta_k$ changes between iterations to approximately enforce KL-divergence constraint
- Clipped Objective
  - New objective function: let $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_k}(a_t|s_t)$. Then

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

    where $\epsilon$ is a hyperparameter (maybe $\epsilon = 0.2$)
  - Policy update is $\theta_{k+1} = \arg\max_\theta \mathcal{L}_{\theta_k}^{CLIP}(\theta)$


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

# PPO: Adaptive KL Penalty

- Using several epochs of minibatch SGD, optimize the KL-penalized objective

$$L^{KLPEN}(\theta) = \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t - \beta\,\mathrm{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]\right]$$

- Compute $d = \hat{\mathbb{E}}_t[\mathrm{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]]$
  - If $d < d_{\text{targ}}/1.5$, $\beta \leftarrow \beta/2$
  - If $d > d_{\text{targ}} \times 1.5$, $\beta \leftarrow \beta \times 2$

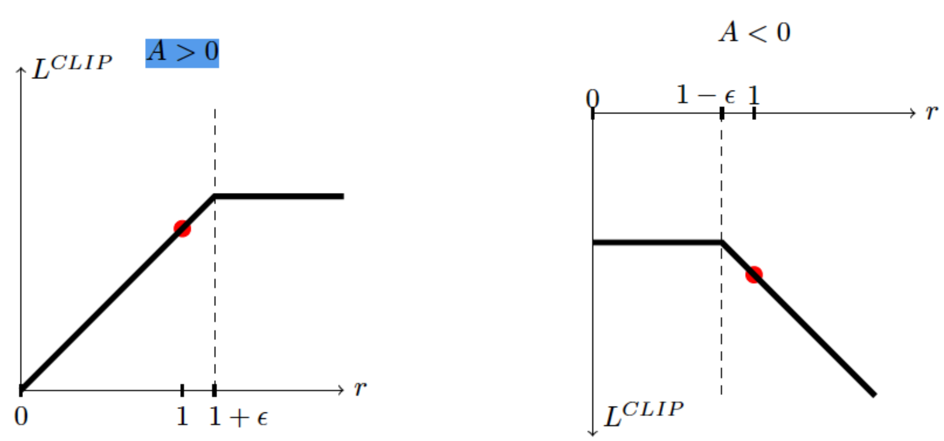
# PPO: Clipped Objective

- Recall the surrogate objective:

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

- Form a lower bound via clipped importance ratio:

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




Slide from Joshua Achiam

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

# PPO: Clipped Objective

Input: initial policy parameters $\theta_0$, clipping threshold $\epsilon$

**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 $\theta_{k+1}$
- Clipping seems to work at least as well as PPO with KL penalty, but is simpler to implement

Slide from Joshua Achiam

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

# PPO: Clipped Objective

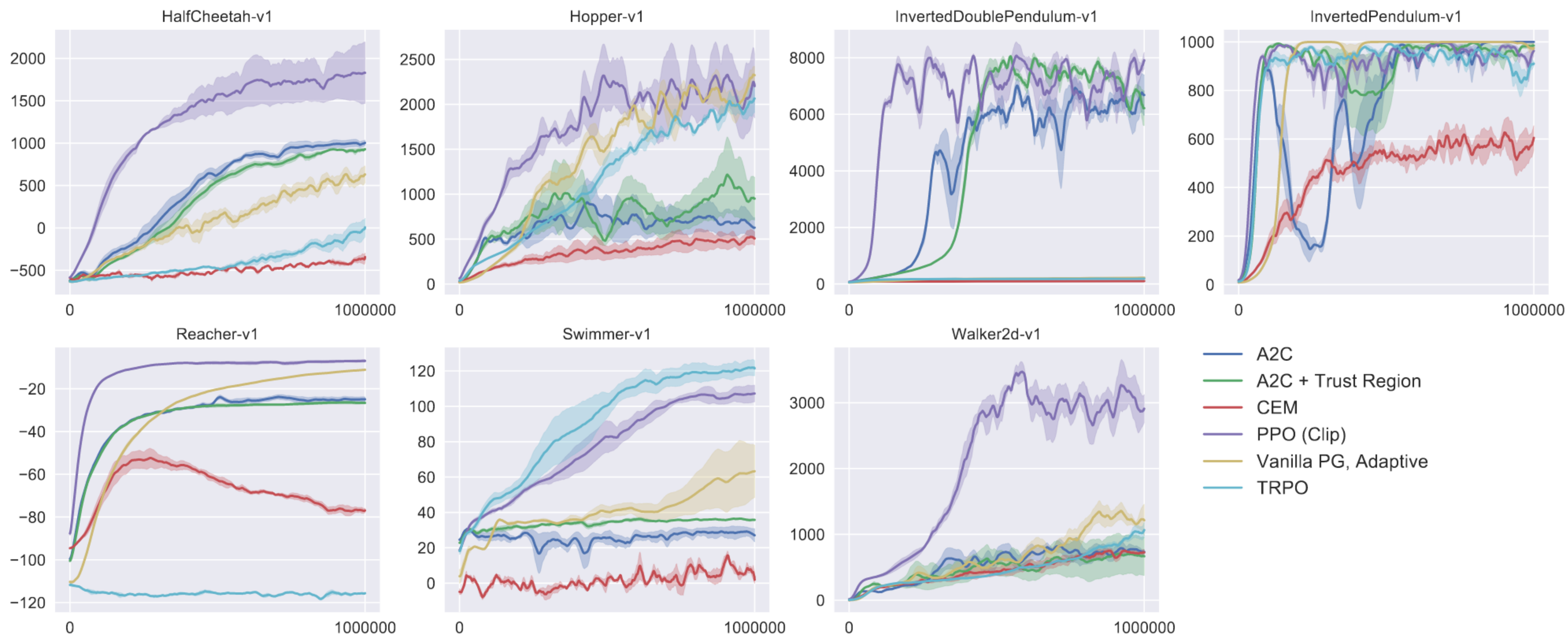



Figure: Performance comparison between PPO with clipped objective and various other deep RL methods on a slate of MuJoCo tasks. [10]