KL-Divergence, Trust-Region and Natural PG

Roberto Capobianco

SAPIENZA
UNIVERSITÀ DI ROMA

Recap

Policy Gradient Theorem (Infinite Setting)

The policy gradient theorem generalizes the likelihood ratio approach

$$\nabla_\theta J(\pi_\theta) = \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s,a\sim\mathbb{P}_h^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a\,|\,s) \cdot Q^{\pi_\theta}(s,a)$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a\,|\,s) \cdot \boxed{Q^{\pi_\theta}(s,a)]}\right.$$

Policy Evaluation!

Policy Gradient Theorem (Infinite Setting)

The policy gradient theorem generalizes the likelihood ratio approach

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s,a\sim\mathbb{P}_h^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a\,|\,s) \cdot Q^{\pi_\theta}(s,a) \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \left[\nabla_\theta \ln \pi_\theta(a\,|\,s) \cdot \boxed{Q^{\pi_\theta}(s,a)} \right]\end{aligned}$$

We can use the return G as an unbiased estimate of Q (MC)

REINFORCE

Initialize policy parameters θ arbitrarily

for each episode $\{s_1, a_1, r_2, \cdots, s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$ **do**

  for $t = 1$ to $T - 1$ **do**

    $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) G_t$

  endfor

endfor

return θ

VARIANCE!

Baseline

To reduce the variance we can introduce baselines (function of state)

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma} \boxed{\mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \Big[\nabla_\theta \ln \pi_\theta(a \mid s)} \cdot \Big(Q^{\pi_\theta}(s,a) - \boxed{b(s)}\Big)\Big]$$

Is this term introducing a bias? NO!

Value Function as Baseline

As baselines have to be action-independent, a common choice for a baseline is

$$b(s) = V^{\pi_\theta}(s)$$

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a\,|\,s)\boxed{\left(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)\right)}\right]$$

Called Advantage Function

$$\nabla_\theta J(\theta_t) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_{\theta_t}}}\left[\nabla_\theta \ln \pi_{\theta_t}(a\,|\,s)\boxed{\left(A^{\pi_{\theta_t}}(s,a)\right)}\right]$$

Advantage Function

Intuition: the advantage function tells us how good an action is compared to the average value of the state

Value of an action in the state

$$Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)$$

Average value of the state

REINFORCE with Baseline

Initialize policy parameter θ, baseline b

for iteration$=1, 2, \cdots$ **do**

Collect a set of trajectories by executing the current policy

At each timestep t in each trajectory τ^i, compute

Return $\boxed{G_t^i = \sum_{t'=t}^{T-1} r_{t'}^i}$, and

Advantage estimate $\hat{A}_t^i = G_t^i - b(s_t)$.

Re-fit the baseline, by minimizing $\sum_i \sum_t ||b(s_t) - G_t^i||^2$,

Update the policy, using a policy gradient estimate $\hat{g}$,

Which is a sum of terms $\nabla_\theta \log \pi(a_t|s_t, \theta)\hat{A}_t$.

(Plug $\hat{g}$ into SGD or ADAM)

endfor

We're still using the return and collecting MC samples

Credits: Emma Brunskill

Sapienza Università di Roma

Advantage Function

$$Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$$

If we can access the true value function, the TD error is an unbiased estimate of the advantage function

$$\begin{aligned}
\mathbb{E}_{\pi_\theta}\left[\delta^{\pi_\theta} | s, a\right] &= \mathbb{E}_{\pi_\theta}\left[r + \gamma V^{\pi_\theta}(s') | s, a\right] - V^{\pi_\theta}(s) \\
&= Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) \\
&= A^{\pi_\theta}(s,a)
\end{aligned}$$

So we can use the TD error to compute the policy gradient

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\ \delta^{\pi_\theta}\right]$$

$$\delta^{\pi_\theta} = r + \gamma V^{\pi_\theta}(s') - V^{\pi_\theta}(s)$$

Can be approximated!

Credits: David Silver

Reducing Variance with Critic

Motivation: Monte-Carlo policy gradient still has high variance!

We can estimate V/Q by using a *critic*

Such critic is also parameterized

$$Q_w(s, a) \approx Q^{\pi_\theta}(s, a)$$

MC vs TD Policy Gradient

- In MC policy gradient, the target is the return G

$$\Delta\theta = \alpha(G_t - V_v(s_t))\nabla_\theta \log \pi_\theta(s_t, a_t)$$

- In Actor-Critic the target is a TD target and relies on bootstrapping
  - Multiple timescales are possible (not only 1-step)
  - Also TD-lambda with forward/backward view

$$\Delta\theta = \alpha(r + \gamma V_v(s_{t+1}) - V_v(s_t))\nabla_\theta \log \pi_\theta(s_t, a_t)$$

Credits: David Silver

Actor-Critic with LFA

Critic $Q_w(s,a) = \phi(s,a)^\top w$ updates weights w by linear TD(0)
Actor updates weights by policy gradient

function QAC
  Initialise s, θ
  Sample $a \sim \pi_\theta$
  for each step **do**
    Sample reward $r = \mathcal{R}_s^a$; sample transition $s' \sim \mathcal{P}_{s,\cdot}^a$
    Sample action $a' \sim \pi_\theta(s', a')$
    $\delta = r + \gamma Q_w(s', a') - Q_w(s, a)$
    $\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(s, a) Q_w(s, a)$
    $w \leftarrow w + \beta \delta \phi(s, a)$
    $a \leftarrow a', s \leftarrow s'$
  end for
end function

Credits: David Silver

Policy Gradient Summary

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) G_t\right] && \text{REINFORCE} \\
&= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\; Q^w(s,a)\right] && \text{Q Actor-Critic} \\
&= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\; A^w(s,a)\right] && \text{Advantage Actor-Critic} \\
&= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\; \delta\right] && \text{TD Actor-Critic} \\
&= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\; \delta e\right] && \text{TD}(\lambda)\text{ Actor-Critic}
\end{aligned}
$$

Critic does policy evaluation to estimate Q, V or A using bootstrapping (*if it uses MC we do not call it a critic*)

End Recap

Policy Iteration Recall

Procedure:

1. Start with a random guess π_0 (can be deterministic or stochastic)
2. For t=0,...,T:
   a. Do **policy evaluation** and compute $Q^{\pi t}$ for all s,a
   b. Do **policy improvement** as $\pi_{t+1}=\text{argmax}_a Q^{\pi t}(s,a)$ for all s

This algorithm only makes progress, and the performance progress of the policy is monotonic

Policy Iteration Recall

Procedure:

1. Start with a random guess π_0 (can be deterministic or stochastic)
2. For t=0,...,T:
   a. Do **policy evaluation** and compute $Q^{\pi t}$ for all s,a
   b. Do **policy improvement** as π_{t+1}=argmax$_a$**$A^{\pi t}$(s,a)** for all s

$$\boxed{Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)}$$

We can also use the advantage function, it's equivalent: pick an action that has the largest advantage against π at every state s

Policy Iteration Recall

Procedure:

1. Start with a random guess π_0 (can be deterministic or stochastic)
2. For t=0,...,T:
    a. Do **policy evaluation** and compute $Q^{\pi t}$ for all s,a
    b. Do **policy improvement** as π_{t+1}=argmax$_a$ **$A^{\pi t}$(s,a)** for all s

$$\boxed{Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)}$$

$$\arg\max_a Q^\pi(s,a) = \arg\max_a A^\pi(s,a)$$

Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what’s their performance difference?

Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what’s their performance difference?

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = ??$$

Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what's their performance difference?

$$
\begin{aligned}
V^{\pi}(s_0) - V^{\pi'}(s_0) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}Q^{\pi'}(s,a) - V^{\pi'}(s)\right] \\
&:= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]
\end{aligned}
$$

Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what’s their performance difference?

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}Q^{\pi'}(s,a) - V^{\pi'}(s)\right]$$

$$:= \frac{1}{1-\gamma}\boxed{\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]}$$

Average advantage value

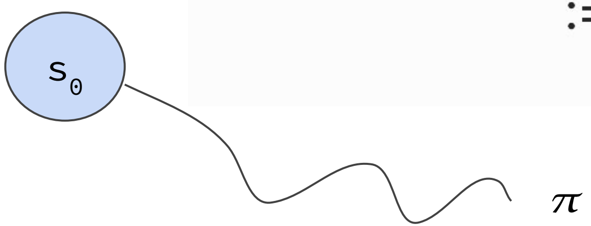
Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what's their performance difference?

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}Q^{\pi'}(s,a) - V^{\pi'}(s)\right]$$

$$:= \frac{1}{1-\gamma}\mathbb{E}_{\boxed{s\sim d^{\pi}_{s_0}}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]$$

Average advantage value



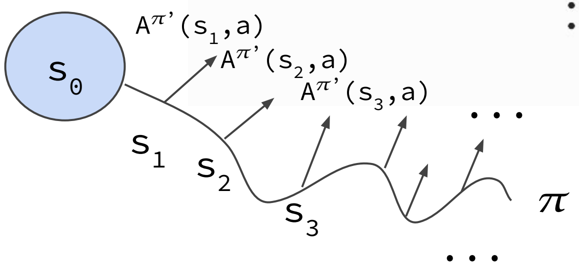
Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what's their performance difference?

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}Q^{\pi'}(s,a) - V^{\pi'}(s)\right]$$

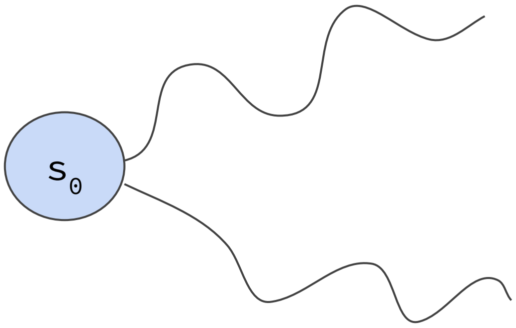
$$:= \frac{1}{1-\gamma}\mathbb{E}_{\boxed{s\sim d^{\pi}_{s_0}}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\boxed{A^{\pi'}(s,a)}\right]$$

Average advantage value



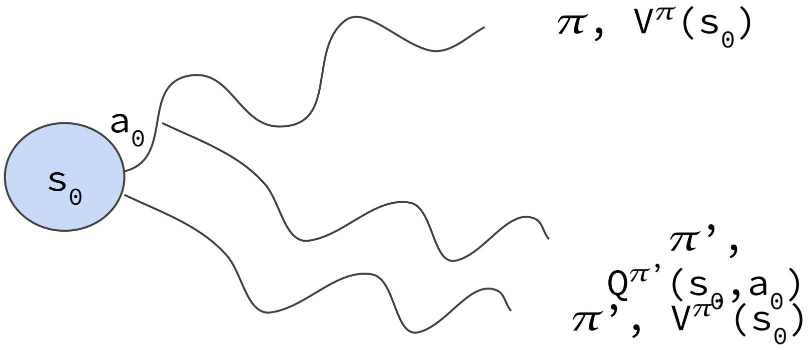
Performance Difference Lemma

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]$$



Performance Difference Lemma

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]$$



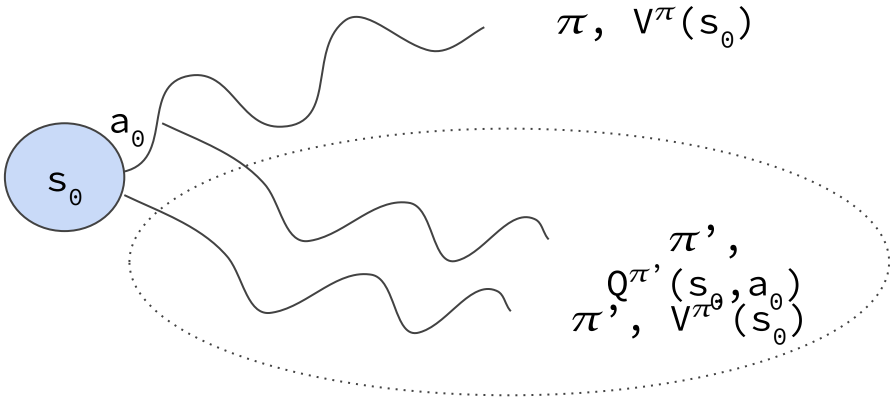
SAPIENZA
UNIVERSITÀ DI ROMA

Performance Difference Lemma

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]$$

This difference is exactly the definition of advantage

$$Q^{\pi'}(s,a) - V^{\pi'}(s)$$

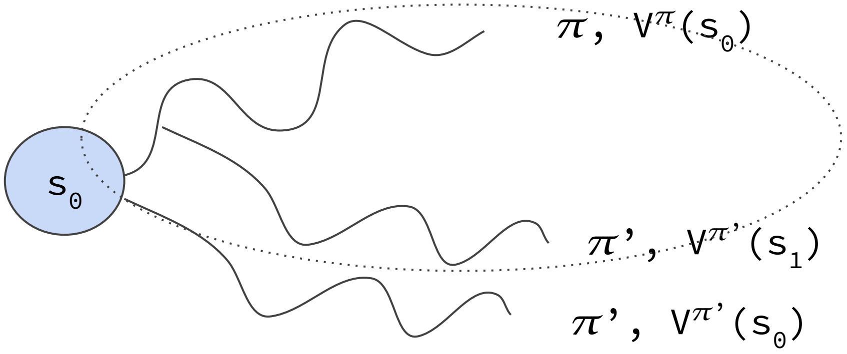


SAPIENZA
UNIVERSITÀ DI ROMA

Performance Difference Lemma

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi'}(s,a)\right]$$

We can do recursion and follow the same reasoning again



Performance Difference Lemma Proof Sketch

$$V^{\pi}(s_0) - V^{\pi'}(s_0)$$

Performance Difference Lemma Proof Sketch

$$V^{\pi}(s_0) - V^{\pi'}(s_0)$$

$$= V^{\pi}(s_0) - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s')\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

Performance Difference Lemma Proof Sketch

$$V^{\pi}(s_0) - V^{\pi'}(s_0)$$

$$= \boxed{V^{\pi}(s_0)} - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[\cancel{r(s_0, a_0)} + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s')\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

$$\mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[\cancel{r(s_0, a_0)} + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi}(s')\right]$$

Performance Difference Lemma Proof Sketch

$$V^{\pi}(s_0) - V^{\pi'}(s_0)$$

$$= V^{\pi}(s_0) - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(\cancel{s_0}, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} \boxed{V^{\pi'}(s')}\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0)$$

$$\mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(\cancel{s_0}, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} \boxed{V^{\pi}(s')}\right]$$

Performance Difference Lemma Proof Sketch

$$
\begin{aligned}
& V^{\pi}(s_0) - V^{\pi'}(s_0) \\
& = V^{\pi}(s_0) - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0) \\
& = \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \mathbb{E}_{s_1 \sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0)
\end{aligned}
$$

Performance Difference Lemma Proof Sketch

$$
\begin{aligned}
& V^{\pi}(s_0) - V^{\pi'}(s_0) \\
& = V^{\pi}(s_0) - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s') \right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s') \right] - V^{\pi'}(s_0) \\
& = \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \mathbb{E}_{s_1 \sim P(s_0, a_0)} \left[V^{\pi}(s_1) - V^{\pi'}(s_1) \right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \left[\boxed{r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0, a_0)} V^{\pi'}(s')} \right] - V^{\pi'}(s_0) \\
& = \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \mathbb{E}_{s_1 \sim P(s_0, a_0)} \left[V^{\pi}(s_1) - V^{\pi'}(s_1) \right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \left[\boxed{Q^{\pi'}(s_0, a_0)} - V^{\pi'}(s_0) \right]
\end{aligned}
$$

Apply definition

Performance Difference Lemma Proof Sketch

$$
\begin{aligned}
&V^{\pi}(s_0) - V^{\pi'}(s_0) \\
&= V^{\pi}(s_0) - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0) \\
&= \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \mathbb{E}_{s_1 \sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0) \\
&= \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \mathbb{E}_{s_1 \sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \boxed{\left[Q^{\pi'}(s_0, a_0) - V^{\pi'}(s_0)\right]} \\
&= \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \mathbb{E}_{s_1 \sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \boxed{\left[A^{\pi'}(s_0, a_0)\right]}
\end{aligned}
$$

Apply definition

SAPIENZA
UNIVERSITÀ DI ROMA

Performance Difference Lemma Proof Sketch

$$
\begin{aligned}
&V^{\pi}(s_0) - V^{\pi'}(s_0) \\
&= V^{\pi}(s_0) - \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] - V^{\pi'}(s_0) \\
&= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0,a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)}V^{\pi'}(s')\right] - V^{\pi'}(s_0) \\
&= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[Q^{\pi'}(s_0,a_0) - V^{\pi'}(s_0)\right] \\
&= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\boxed{\left[V^{\pi}(s_1) - V^{\pi'}(s_1)\right]} + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[A^{\pi'}(s_0,a_0)\right]
\end{aligned}
$$

Recursion

SAPIENZA
UNIVERSITÀ DI ROMA

Performance Difference Lemma

We know that the new policy from PI is better than the old one, but what's their performance difference?

$$V^{\pi_{new}}(s_0) - V^{\pi_{old}}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_{s_0}^{\pi_{new}}}\left[A^{\pi_{old}}(s,a)\right]$$

Advantage against old policy averaged over the new policy induced distribution

Approximate Policy Iteration Recall

Procedure:

1. Start with a random guess π_0 (can be deterministic or stochastic)
2. For t=0,...,T:
   a. Do **policy evaluation** and compute **$\hat{A}^{\pi_t}$**
   b. Do **policy improvement** as $\hat{\pi}_{t+1}=\mathrm{argmax}_a \hat{A}^{\pi_t}(s,a)$ for all s

$$\boxed{Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)}$$

For example, estimate $\hat{A}$ directly through regression

Approximate Policy Iteration Recall

Procedure:

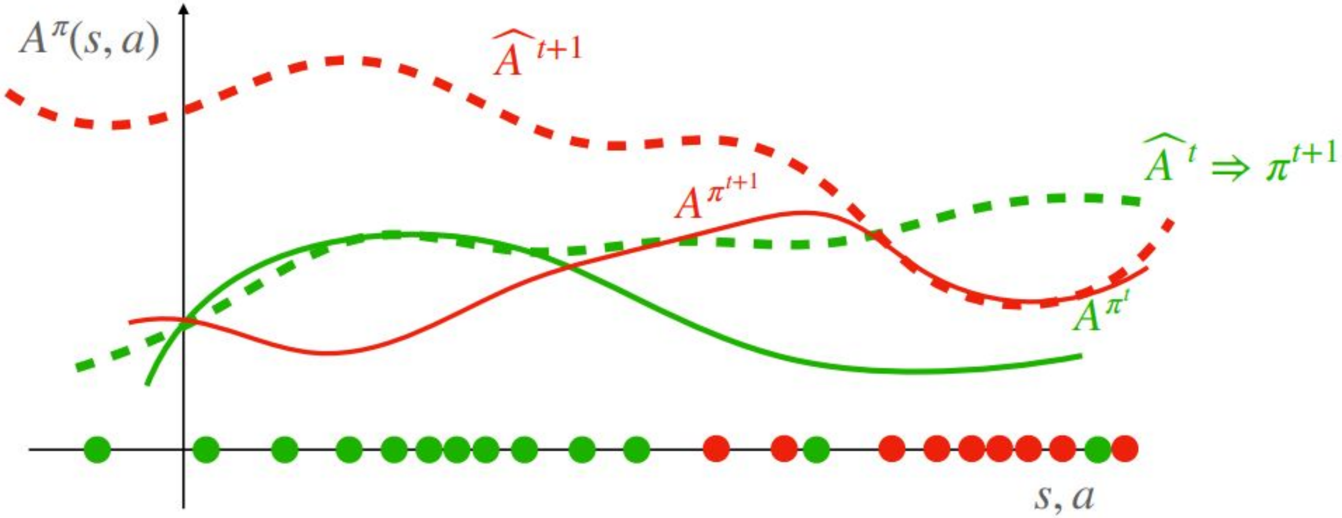
1. Start with a random guess π_0 (can be deterministic or stochastic)
2. For t=0,...,T:
    a. Do **policy evaluation** and compute **$\hat{A}^{\pi t}$**
    b. Do **policy improvement** as $\hat{\pi}_{t+1}=\text{argmax}_a$ **$\hat{A}^{\pi t}(s,a)$** for all s

$$\boxed{Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)}$$

π^ is an approximate greedy policy

$$\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\widehat{\pi}(s))\right] \approx \max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$$

Approximate Policy Iteration Recall



No monotonic improvement

Credits: Wen Sun

Conservative Policy Iteration

Oscillations are due to the distribution change induced by the policy

Can we design an update rule that does not change the distribution so much?

$$d^{\pi^t} \approx d^{\pi^{t+1}}$$

$$\mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right] \approx \mathbb{E}_{s \sim d_\mu^{\pi^{t+1}}} \left[A^{\pi^t}(s, \pi^{t+1}(s)) \right]$$

Incremental Update of CPI

Oscillations are due to the distribution change induced by the policy

Can we design an update rule that does not change the distribution so much?

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^{t}(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

$$\|\pi^{t+1}(\cdot \mid s) - \pi^{t}(\cdot \mid s)\|_1 \leq 2\alpha \longrightarrow \|d_\mu^{\pi^{t+1}}(\cdot) - d_\mu^{\pi^{t}}(\cdot)\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

Incremental Update of CPI

If we set alpha appropriately we can get back monotonic improvement until termination

$$\boxed{\begin{array}{l}\text{If } \max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}}[A^{\pi^t}(s, \pi(s))] \leq \epsilon \\ \quad \textbf{Return } \pi^t\end{array}}$$

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^t(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

$$\|\pi^{t+1}(\cdot \mid s) - \pi^t(\cdot \mid s)\|_1 \leq 2\alpha \longrightarrow \|d_\mu^{\pi^{t+1}}(\cdot) - d_\mu^{\pi^t}(\cdot)\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

Problem of CPI

I now need to retain all the old policies in memory: what if they are all large neural networks?

$$\pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^{t}(\cdot \mid s) + \alpha\pi'(\cdot \mid s), \forall s$$

Problem of CPI

I now need to retain all the old policies in memory: what if they are all large neural networks?

Let's use KL-Divergence

KL-Divergence

Given two distributions Q and P, KL-Divergence is defined as

$$KL(P|Q) = \mathbb{E}_{x\sim P}\left[\ln \frac{P(x)}{Q(x)}\right]$$

$$\boxed{KL(P|Q) \geq 0}$$

$\boxed{Q = P}$ $\quad KL(P|Q) = KL(Q|P) = 0$

$\boxed{P = \mathcal{N}(\mu_1, \sigma^2 I), Q = \mathcal{N}(\mu_2, \sigma^2 I)}$ $\quad KL(P|Q) = \|\mu_1 - \mu_2\|_2^2/\sigma^2$

Trust-Region Formulation for Policy Update

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$

$$\text{s.t.,}\ KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \le \delta$$

$$\boxed{\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots}$$

KL-Divergence of State Distribution

$$\boxed{KL(P\,|\,Q) = \mathbb{E}_{x\sim P}\left[\ln \frac{P(x)}{Q(x)}\right]}$$

$$\boxed{\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0\,|\,s_0)P(s_1\,|\,s_0, a_0)\pi_\theta(a_1\,|\,s_1)\ldots}$$

$$KL\left(\rho_{\pi_{\theta_t}}\,|\,\rho_{\pi_\theta}\right) = \mathbb{E}_{\tau\sim\rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)}$$

KL-Divergence of State Distribution

$$KL(P \mid Q) = \mathbb{E}_{x \sim P}\left[\ln \frac{P(x)}{Q(x)}\right]$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$KL\left(\rho_{\pi_{\theta_t}} \mid \rho_{\pi_\theta}\right) = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)}$$

Initial state distribution, as well as next state distribution simplify, because they are the same. We are only left with the different policies.

KL-Divergence of State Distribution

$$KL(P \mid Q) = \mathbb{E}_{x \sim P}\left[\ln \frac{P(x)}{Q(x)}\right]$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$KL\left(\rho_{\pi_{\theta_t}} \mid \rho_{\pi_\theta}\right) = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)} = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_t}(a_h \mid s_h)}{\pi_\theta(a_h \mid s_h)}$$

SAPIENZA
UNIVERSITÀ DI ROMA

KL-Divergence of State Distribution

$$KL(P\,|\,Q) = \mathbb{E}_{x\sim P}\left[\ln \frac{P(x)}{Q(x)}\right]$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0\,|\,s_0)P(s_1\,|\,s_0, a_0)\pi_\theta(a_1\,|\,s_1)\ldots$$

$$KL\left(\rho_{\pi_{\theta_t}}\,|\,\rho_{\pi_\theta}\right) = \mathbb{E}_{\tau\sim\rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)} = \mathbb{E}_{\tau\sim\rho_{\pi_{\theta_t}}} \sum_{h=0}^{\infty} \ln \frac{\pi_{\theta_t}(a_h\,|\,s_h)}{\pi_\theta(a_h\,|\,s_h)}$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s_h,a_h\sim d_\mu^{\pi_{\theta_t}}}\left[\ln \frac{\pi_{\theta_t}(a_h\,|\,s_h)}{\pi_\theta(a_h\,|\,s_h)}\right] := \ell(\theta)$$

Trust-Region Formulation for Policy Update

$$\max_{\pi_\theta} \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(s)} A^{\pi_{\theta_t}}(s, a) \right]$$

$$\text{s.t.,} \boxed{KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta}$$

This is our trust-region, that maintains the distributions not so far

$$\boxed{\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\dots}$$

Trust-Region Formulation for Policy Update

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$

$$\text{s.t.,} \boxed{KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta}$$

How do we optimize this?

Trust-Region Formulation for Policy Update

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$

$$\text{s.t.,} \boxed{KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta}$$

How do we optimize this?

Remember: the trajectory distribution is actually unknown and we do not know the transition function!

Trust-Region Formulation for Policy Update

$$\max_{\pi_\theta} \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(s)} A^{\pi_{\theta_t}}(s, a) \right]$$

$$\text{s.t.,} \boxed{KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta}$$

How do we optimize this?

1st or 2nd order Taylor expansion

Trust-Region Optimization: Objective Function

Let’s first simplify and linearize the objective function using a 1st order Taylor Expansion

$$\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)}A^{\pi_{\theta_t}}(s,a)\right] \approx \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)}A^{\pi_{\theta_t}}(s,a)\right] + \underbrace{\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)}\nabla_\theta \ln \pi_{\theta_t}(a\,|\,s)A^{\pi_{\theta_t}}(s,a)\right]}_{\nabla_\theta J(\pi_{\theta_t})} \cdot (\theta-\theta_t)$$

Trust-Region Optimization: Objective Function

Let's first simplify and linearize the objective function using a 1st order Taylor Expansion

$$\mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a \sim \pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right] \approx \boxed{\mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a \sim \pi_{\theta_t}(s)} A^{\pi_{\theta_t}}(s,a)\right]} + \underbrace{\mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a \sim \pi_{\theta_t}(s)} \nabla_\theta \ln \pi_{\theta_t}(a \mid s) A^{\pi_{\theta_t}}(s,a)\right]}_{\nabla_\theta J(\pi_{\theta_t})} \boxed{\cdot} (\theta - \theta_t)$$

Advantage of the policy against itself is 0

Inner product

Trust-Region Optimization: Objective Function

Let's first simplify and linearize the objective function using a 1st order Taylor Expansion

$$\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)}A^{\pi_{\theta_t}}(s,a)\right] \approx \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)}A^{\pi_{\theta_t}}(s,a)\right] + \underbrace{\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)}\nabla_\theta \ln \pi_{\theta_t}(a\,|\,s)A^{\pi_{\theta_t}}(s,a)\right]}_{\nabla_\theta J(\pi_{\theta_t})} \cdot (\theta-\theta_t)$$

$$= \boxed{\nabla_\theta J(\pi_{\theta_t})^\top(\theta-\theta_t)}$$

Trust-Region Optimization: Constraint

Let’s then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta)$$

Trust-Region Optimization: Constraint

Let’s then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta)$$

$$\ell(\theta) \approx \ell(\theta_t) + \nabla \ell(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

Trust-Region Optimization: Constraint

Let's then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta)$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \nabla \ell(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

$$\ell(\theta_t) = KL(\rho_{\theta_t} | \rho_{\theta_t}) = 0$$

Trust-Region Optimization: Constraint

Let's then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

$\boxed{\ln(a/b) = \ln a - \ln b}$

Trust-Region Optimization: Constraint

Let’s then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}}\left[\ln \frac{\boxed{\pi_{\theta_t}(a_h | s_h)}}{\pi_\theta(a_h | s_h)}\right]$$

Does not depend on the variation of theta

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top(\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

Trust-Region Optimization: Constraint

Let’s then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\boxed{\pi_{\theta_t}(a_h | s_h)}}{\pi_\theta(a_h | s_h)} \right]$$

Does not depend on the variation of theta

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

$$\nabla_\theta \ell(\theta) |_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(-\nabla_\theta \ln \pi_\theta(a_h | s_h) |_{\theta=\theta_t} \right)$$

$\boxed{\ln(a/b) = \ln a - \ln b}$

Expectation has nothing to do with gradient, so we bring gradient inside

Trust-Region Optimization: Constraint

Let’s then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

Bring sum inside:
this sums to 1

$$\nabla_\theta \ell(\theta)|_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a|s)\left(- \nabla_\theta \ln \pi_\theta(a_h | s_h)|_{\theta=\theta_t} \right) = - \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \cancel{\pi_{\theta_t}(a|s)} \frac{\boxed{\nabla_\theta \pi_{\theta_t}(a|s)}}{\cancel{\pi_{\theta_t}(a|s)}} = 0$$

Trust-Region Optimization: Constraint

Let's then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

Trust-Region Optimization: Constraint

Let's then simplify and linearize the constraint using a 2nd order Taylor Expansion

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)} (\theta - \theta_t)$$

Trust-Region Optimization: Constraint

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\boxed{\pi_{\theta_t}(a_h | s_h)}}{\pi_\theta(a_h | s_h)} \right]$$

Does not depend on the variation of theta

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)} (\theta - \theta_t)$$

$$\nabla_\theta^2 \ell(\theta)|_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(-\nabla_\theta^2 \ln \pi_\theta(a | s)|_{\theta=\theta_t} \right)$$

Expectation has nothing to do with gradient, so we bring gradient inside

$\boxed{\text{ln(a/b) = ln a - ln b}}$

Trust-Region Optimization: Constraint

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2} (\theta - \theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)} (\theta - \theta_t)$$

$$\nabla_\theta^2 \ell(\theta) |_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(-\nabla_\theta^2 \ln \pi_\theta(a | s) |_{\theta=\theta_t} \right)$$

$$\nabla_\theta \ln \pi_\theta(a_h | s_h) |_{\theta=\theta_t} = \frac{\nabla_\theta \pi_{\theta_t}(a | s)}{\pi_{\theta_t}(a | s)}$$

$$\nabla_\theta^2 \ln \pi_\theta(a | s) |_{\theta=\theta_t}$$?

Sapienza Università di Roma

Trust-Region Optimization: Constraint

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}}\left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)}\right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)}(\theta - \theta_t)$$

$$\nabla_\theta^2 \ell(\theta)|_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s)\left(-\nabla_\theta^2 \ln \pi_\theta(a | s)|_{\theta=\theta_t}\right)$$

$$\nabla_\theta \ln \pi_\theta(a_h | s_h)|_{\theta=\theta_t} = \frac{\nabla_\theta \pi_{\theta_t}(a | s)}{\pi_{\theta_t}(a | s)}$$

$$\nabla_\theta^2 \ln \pi_\theta(a | s)|_{\theta=\theta_t} \quad ?$$

We just have to compute the gradient of this now

SAPIENZA
UNIVERSITÀ DI ROMA

Trust-Region Optimization: Constraint

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2} (\theta - \theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)} (\theta - \theta_t)$$

$$\nabla_\theta^2 \ell(\theta) |_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(- \nabla_\theta^2 \ln \pi_\theta(a | s) |_{\theta=\theta_t} \right)$$

$$\nabla_\theta \ln \pi_\theta(a_h | s_h) |_{\theta=\theta_t} = \frac{\nabla_\theta \pi_{\theta_t}(a | s)}{\pi_{\theta_t}(a | s)}$$

$$\nabla_\theta^2 \ln \pi_\theta(a | s) |_{\theta=\theta_t}$$?

(f/g)’ = f’/g – fg’/g^2

Trust-Region Optimization: Constraint

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2} (\theta - \theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)} (\theta - \theta_t)$$

$$\nabla_\theta^2 \ell(\theta) |_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(-\nabla_\theta^2 \ln \pi_\theta(a | s) |_{\theta=\theta_t} \right)$$

$$= -\mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(\frac{\nabla_\theta^2 \pi_{\theta_t}(a | s)}{\pi_{\theta_t}(a | s)} - \frac{\nabla_\theta \pi_{\theta_t}(a | s) \nabla_\theta \pi_{\theta_t}(a | s)^\top}{\pi_{\theta_t}^2(a | s)} \right)$$

(f/g)’ = f’/g - fg’/g^2

Trust-Region Optimization: Constraint

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)} (\theta - \theta_t)$$

$$\nabla_\theta^2 \ell(\theta) |_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(-\nabla_\theta^2 \ln \pi_\theta(a | s) |_{\theta=\theta_t} \right)$$

$$= -\mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \cancel{\pi_{\theta_t}(a | s)} \left(\frac{\boxed{\nabla_\theta^2 \pi_{\theta_t}(a | s)}}{\cancel{\pi_{\theta_t}(a | s)}} - \frac{\nabla_\theta \pi_{\theta_t}(a | s) \nabla_\theta \pi_{\theta_t}(a | s)^\top}{\pi_{\theta_t}^2(a | s)} \right)$$

Bring sum inside:
this sums to 1

(f/g)' = f'/g - fg'/g^2

SAPIENZA UNIVERSITÀ DI ROMA

Trust-Region Optimization: Constraint

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s_h,a_h\sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top (\theta-\theta_t) + \frac{1}{2} (\theta-\theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)} (\theta-\theta_t)$$

$$\nabla_\theta^2 \ell(\theta)|_{\theta=\theta_t} = \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s)\left(-\nabla_\theta^2 \ln \pi_\theta(a | s)|_{\theta=\theta_t}\right) = -\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(\frac{\nabla_\theta^2 \pi_{\theta_t}(a | s)}{\pi_{\theta_t}(a | s)} - \frac{\nabla_\theta \pi_{\theta_t}(a | s)\,\nabla_\theta\pi_{\theta_t}(a | s)^\top}{\pi_{\theta_t}^2(a | s)}\right)$$

$$= \mathbb{E}_{s,a\sim d_\mu^{\pi_{\theta_t}}} \left[\boxed{\nabla_\theta \ln \pi_{\theta_t}(a | s) \left(\nabla_\theta \ln\pi_{\theta_t}(a | s)\right)^\top} \right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

This is the **Fisher Information Matrix**

Trust-Region Optimization: Constraint

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}}\left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)}\right]$$

$$\ell(\theta) \approx \boxed{\ell(\theta_t)} + \boxed{\nabla \ell(\theta_t)}^\top(\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \boxed{\nabla_\theta^2 \ell(\theta_t)}(\theta - \theta_t)$$

Easy to compute, as we know how to compute the gradient of the log likelihood of the policy

$$= \mathbb{E}_{s,a \sim d_\mu^{\pi_{\theta_t}}}\left[\boxed{\nabla_\theta \ln \pi_{\theta_t}(a | s)\left(\nabla_\theta \ln \pi_{\theta_t}(a | s)\right)^\top}\right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

This is the **Fisher Information Matrix**

Trust-Region Optimization: Simplified Constraint

$$KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \approx \frac{1}{2}(\theta - \theta_t)^\top F_{\theta_t}(\theta - \theta_t)$$

$$F_{\theta_t} := \mathbb{E}_{s,a \sim d_\mu^{\pi_{\theta_t}}}\left[\nabla_\theta \ln \pi_{\theta_t}(a|s)\left(\nabla_\theta \ln \pi_{\theta_t}(a|s)\right)^\top\right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

F is always positive semi-definite

Simplified Trust-Region Formulation

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \le \delta$$

$$\boxed{\begin{aligned}&\max_\theta \nabla_\theta J(\pi_{\theta_t})^\top(\theta-\theta_t)\\ &\text{s.t. } (\theta-\theta_t)^\top F_{\theta_t}(\theta-\theta_t) \le \delta\end{aligned}}$$

Simplified Trust-Region Formulation

$$\boxed{\begin{aligned} &\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t) \\ &\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta \end{aligned}}$$

This looks very easy and we can compute the solution in closed form!

Simplified Trust-Region Solution

$$\max_{\theta} \boxed{\nabla_{\theta} J(\pi_{\theta_t})}^{\top} \boxed{(\theta - \theta_t)}$$

$$\text{s.t. } \boxed{(\theta - \theta_t)}^{\top} F_{\theta_t} \boxed{(\theta - \theta_t)} \leq \delta$$

Let’s first simplify the notation:

$\theta - \theta_t = \Delta$

$\nabla_{\theta} J(\pi_{\theta t}) = \nabla$

SAPIENZA UNIVERSITÀ DI ROMA

Credits: Wen Sun

Simplified Trust-Region Solution

$$\max_{\Delta} \nabla^{\top} \Delta,$$
$$\text{s.t. } \Delta^{\top} F \Delta \leq \delta$$

Let's first simplify the notation:

$\theta - \theta_t = \Delta$

$\nabla_\theta J(\pi_{\theta t}) = \nabla$

Credits: Wen Sun

Simplified Trust-Region Solution

$$\max_{\Delta} \nabla^{\top}\Delta, \quad \text{s.t. } \Delta^{\top} F \Delta \leq \delta$$

Let's then introduce $F^{1/2}$

For a positive definite matrix this can be obtained from the Eigen Decomposition: $F = U\Sigma U^{\top}$, $F^{1/2} = U\sqrt{\Sigma}U^{\top}$

Credits: Wen Sun

Simplified Trust-Region Solution

$$\max_{\Delta} \nabla^{\top} \Delta,$$
$$\text{s.t. } \Delta^{\top} F \Delta \leq \delta$$

$(F^{1/2})^2 = F$

$F^{1/2}F^{-1/2} = I$

$\longrightarrow$

$\max_{\Delta} \nabla^{\top} F^{1/2} F^{-1/2} \Delta$

$\text{s.t.} (F^{1/2}\Delta)^{\top}(F^{1/2}\Delta) \leq \delta$

Credits: Wen Sun

Simplified Trust-Region Solution

$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^{\mathsf{T}} \widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^{\mathsf{T}}\widetilde{\Delta} \leq \delta$$

$$\widetilde{\Delta} := F^{1/2}\Delta$$

$(F^{1/2})^2 = F$

$F^{1/2}F^{-1/2} = I$

$\longrightarrow$

$\max_{\Delta} \nabla^{\mathsf{T}} F^{1/2} F^{-1/2} \Delta$

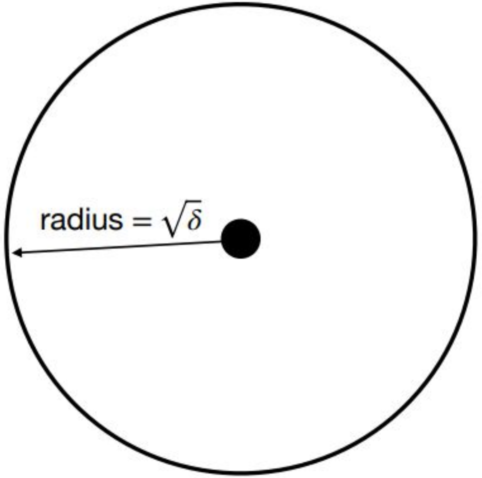
$\text{s.t.} (F^{1/2}\Delta)^{\mathsf{T}} (F^{1/2}\Delta) \leq \delta$

Simplified Trust-Region Solution

$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^{\top}\widetilde{\Delta},$$

$$\widetilde{\Delta} := F^{1/2}\Delta$$

$$\text{s.t. } \boxed{\widetilde{\Delta}^{\top}\widetilde{\Delta} \leq \delta}$$

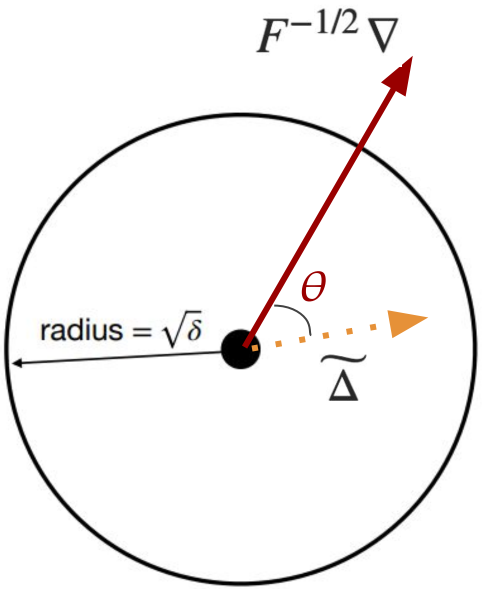


This is my (ball) constraint: the norm of $\Delta^{\sim}$ has to be ≤ δ (so, any vector $\Delta^{\sim}$ falls in this ball)

Credits: Wen Sun

Simplified Trust-Region Solution

This is the direction of the gradient



$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^{\top}\widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^{\top}\widetilde{\Delta} \leq \delta$$

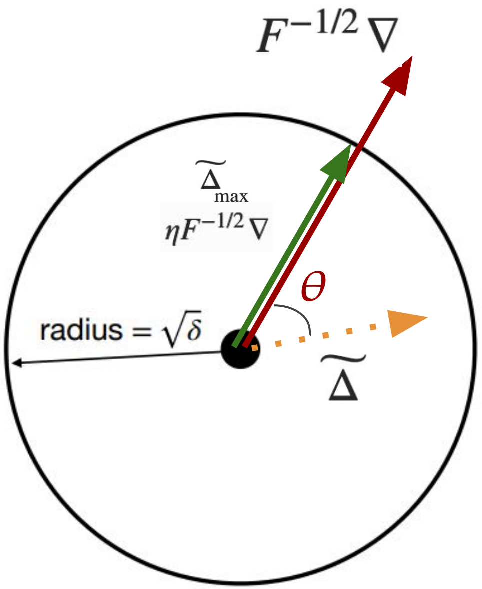
$$\widetilde{\Delta} := F^{1/2}\Delta$$

What I do care about now is the inner product between the vector $F^{-1/2}\Delta$ and any vector $\Delta^{\sim}$ in this ball

Credits: Wen Sun

Simplified Trust-Region Solution

This is the direction of the gradient



$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^{\top}\widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^{\top}\widetilde{\Delta} \leq \delta$$

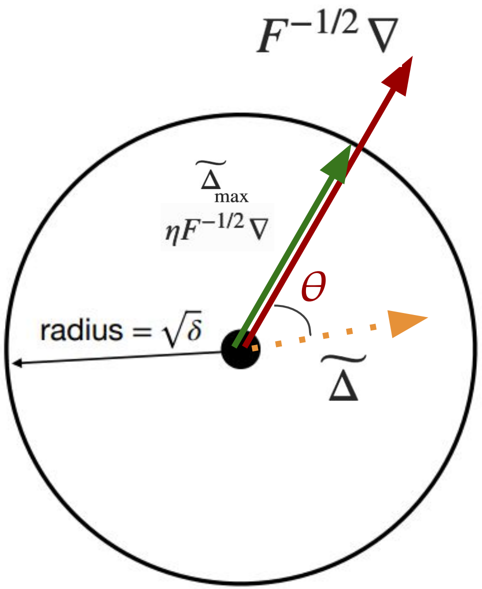
$$\widetilde{\Delta} := F^{1/2}\Delta$$

Which vector does maximize this inner product? The green one: minimum angle (same direction $F^{-1/2}\Delta$), maximum length (scaled by η)

Credits: Wen Sun

Simplified Trust-Region Solution

This is the direction of the gradient



$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^{\top}\widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^{\top}\widetilde{\Delta} \leq \delta$$

$$\widetilde{\Delta} := F^{1/2}\Delta$$

$$\|\eta F^{-1/2}\nabla\|_2 = \sqrt{\delta} \quad \Rightarrow \eta = \sqrt{\frac{\delta}{\nabla^{\top}F^{-1}\nabla}}$$

$$\widetilde{\Delta}_{max} := \sqrt{\frac{\delta}{\nabla^{\top}F^{-1}\nabla}}F^{-1/2}\nabla$$

$$\boxed{\Delta_{max} := \sqrt{\frac{\delta}{\nabla^{\top}F^{-1}\nabla}}F^{-1}\nabla}$$

Credits: Wen Sun

Natural Policy Gradient

$$\boxed{\begin{aligned}&\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top \left(\theta - \theta_t\right)\\ &\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t}(\theta - \theta_t) \le \delta\end{aligned}}$$

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t}) \qquad \eta = \sqrt{\frac{\delta}{\nabla_\theta J(\pi_{\theta_t})^\top F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})}}$$

The same solution can be obtained by applying Lagrange multipliers

$$\min_{\lambda \le 0} \max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t) + \lambda \left((\theta - \theta_t)^\top F_{\theta_t}(\theta - \theta_t) - \delta \right)$$

Natural Policy Gradient

$$\boxed{\begin{aligned} &\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t) \\ &\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta \end{aligned}}$$

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t}) \qquad \eta = \sqrt{\frac{\delta}{\nabla_\theta J(\pi_{\theta_t})^\top F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})}}$$

F is generally invertible, but in case it is not you can use pseudo-inverse or add regularization ($F = F + \lambda I$ with λ very small)

Natural Policy Gradient

$$\boxed{\begin{aligned} &\max_{\theta} \nabla_{\theta} J(\pi_{\theta_t})^{\top}(\theta - \theta_t) \\ &\text{s.t. } (\theta - \theta_t)^{\top} F_{\theta_t}(\theta - \theta_t) \leq \delta \end{aligned}}$$

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_{\theta} J(\pi_{\theta_t}) \qquad \eta = \sqrt{\frac{\delta}{\nabla_{\theta} J(\pi_{\theta_t})^{\top} F_{\theta_t}^{-1} \nabla_{\theta} J(\pi_{\theta_t})}}$$

Step size (η) depends on the allowed trust region (δ is a hyper-parameter that we typically set to a small number like 1e-2 or 1e-3)

Natural Policy Gradient

$$\boxed{\begin{aligned} &\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t) \\ &\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \le \delta \end{aligned}}$$

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t}) \qquad \eta = \sqrt{\frac{\delta}{\nabla_\theta J(\pi_{\theta_t})^\top F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})}}$$

F is pre-conditioning our gradient, instead of just fully going for it

TRPO: Line Search

Due to the quadratic approximation, the KL constraint might be violated: we solve this by doing a simple line search

for $j = 0, 1, 2, \ldots, L$ **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

Natural Policy Gradient: Additional Comments

We want to keep two distributions close, but parameters can change a lot: learning rate (η) is very high if eigen-values of F are very small (as the matrix is inverted)

Generally, Natural PG moves faster than standard/plain PG

If we have many parameters, computing & inverting F is too heavy!

Extending TRPO: Proximal Policy Optimization

If we have many params, we can impose KL divergence as a regularization term and optimize (simply through SG Ascent)

$$\max_{\theta} \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta}(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right] \underbrace{-\lambda \mathbb{E}_{s\sim d_{\mu}^{\pi^t}}\left[\mathrm{KL}\left(\pi_{\theta_t}(a|s) \,|\, \pi_{\theta}(a|s)\right)\right]}_{\text{regularization}}$$

using importance weighting and expanding KL divergence through expectation

$$\ell(\theta) := \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \frac{\pi_{\theta}(a|s)}{\pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a)\right] - \lambda \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}} \mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)}\left[-\ln \pi_{\theta}(a|s)\right]$$