Policy Gradient mathematics

A/Prof Richard Yi Da Xu
richardxu.com

University of Technology Sydney (UTS)

February 14, 2021

Content

1. Policy Gradient Theorem
2. Mathematics on Trusted Region Optimization in RL
3. Natural Gradients on TRPO
4. Proximal Policy Optimization (PPO)
5. Conjugate Gradient Algorithm

This lecture is referenced heavily from:

- `https://lilianweng.github.io/lil-log/2018/04/08/policy-gradient-algorithms.html`. I borrowed it heavily, please check her goodies on RL and GAN
- `https://medium.com/@jonathan_hui/rl-trust-region-policy-optimization-trpo-explained-a6ee04eeeee9`, Jonathan Hui's blog. Again, lots of goodies.
- `http://www.cs.cmu.edu/~pradeepr/convexopt/Lecture_Slides/conjugate_direction_methods.pdf`

What is Policy Gradient "on the surface"

- Gradient of Expected entire Rewards $R(\tau)$ collected by taking a "trajectory" τ following π_θ:

$$\begin{aligned}\nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] &= \mathbb{E}_{\tau\sim\pi_\theta}\left[R(\tau)\cdot \nabla_\theta \log \mathbb{P}_\theta(\tau)\right] \\ &= \mathbb{E}_{\tau\sim\pi_\theta}\left[R(\tau)\cdot \nabla_\theta \left(\sum_{t=0}^{T-1} \log \pi_\theta(a_t|s_t)\right)\right]\end{aligned}$$

- Derivative of Log-likelihood of Policy Gradient is:

$$\begin{aligned}\nabla_\theta \log \mathbb{P}_\theta(\tau) &= \nabla_\theta \log\left(\mu(s_0)\prod_{t=0}^{T-1}\pi_\theta(a_t|s_t)P(s_{t+1}|s_t,a_t)\right) \\ &= \nabla_\theta\left[\underbrace{\log\mu(s_0)}_{\text{no } \theta} + \sum_{t=0}^{T-1}\left(\log\pi_\theta(a_t|s_t) + \underbrace{\log P(s_{t+1}|s_t,a_t)}_{\text{no } \theta}\right)\right] \\ &= \nabla_\theta \sum_{t=0}^{T-1}\log\pi_\theta(a_t|s_t)\end{aligned}$$

- $\log p(s_{t+1}|s_t,a_t)$ has no θ is controversial, we need to see why

Significance of Policy Gradient Theorem

- we use an alternative representation:

$$J(\theta) \equiv V^{\pi}(s_0)$$

which we can expand using recursion as needed for unknow T:

- Computing gradient $\nabla_\theta J(\theta)$ is **difficult** because it depends on both:
  1. action selection **directly** determined by π_θ, and
  2. stationary state following action selection behavior **indirectly** determined by π_θ
- difficult to estimate policy update effect on state distribution for generally unknown environment
- however, **Policy gradient theorem** states:

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} Q^{\pi}(s, a) \pi_\theta(a|s) \\ &\propto \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} Q^{\pi}(s, a) \nabla_\theta \pi_\theta(a|s)\end{aligned}$$

- **significance**: above objective function does **not** involve derivative of state distribution $d^{\pi}(.)$

Proof of Policy Gradient Theorem

- We want a policy to maximize $J(\theta) \equiv V^{\pi}(s)$:
- first step is always to write $V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi_\theta(a|s) Q^{\pi}(s, a)$:

$$\nabla_\theta V^{\pi}(s) = \nabla_\theta \Big(\sum_{a \in \mathcal{A}} \underbrace{\pi_\theta(a|s)}_{u} \underbrace{Q^{\pi}(s, a)}_{v} \Big)$$

$$= \underbrace{\sum_{a \in \mathcal{A}} \Big(\nabla_\theta \pi_\theta(a|s) Q^{\pi}(s, a) +}_{=\phi(s) \text{ which contain } \nabla_\theta} \underbrace{\sum_{a \in \mathcal{A}} \pi_\theta(a|s) \nabla_\theta Q^{\pi}(s, a)}_{\text{see how we make } \nabla_\theta \text{ disappear in this term}} \Big)$$

$$= \phi(s) + \sum_{a \in \mathcal{A}} \Big(\pi_\theta(a|s) \nabla_\theta \sum_{s'} \sum_{r} P(s', r|s, a) (\underbrace{r + V^{\pi}(s')}_{\text{immediate \& future reward}}) \Big)$$

$$= \phi(s) + \sum_{a \in \mathcal{A}} \Big(\pi_\theta(a|s) \sum_{s'} \sum_{r} P(s', r|s, a) \nabla_\theta V^{\pi}(s') \Big) \quad \text{remove part independent of } \theta$$

$$= \phi(s) + \sum_{a \in \mathcal{A}} \Big(\pi_\theta(a|s) \sum_{s'} P(s'|s, a) \nabla_\theta V^{\pi}(s') \Big) \quad \text{retain marginal by integrate out } r$$

- $V^{\pi}(s)$:

$$V^{\pi}(s_t) = \mathbb{E}_{a_t, s_{t+1}, a_{t+1}, \dots}\left[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}) \Big| s_t \right]$$

$$= \mathbb{E}_{a_t, s_{t+1}, a_{t+1}, \dots}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \right] \quad \text{let } r_{t+k} \equiv r(s_{t+k})$$

$$\text{by induction, } V^{\pi}(s_{t+1}) = \underline{\mathbb{E}_{a_{t+1}, s_{t+2}, a_{t+2}, \dots}\left[\sum_{k=0}^{\infty} \gamma^k r(s_{(t+1)+k}) \Big| s_{t+1} \right]}$$

writing $V^\pi(s)$ and $Q^\pi(s,a)$

- $Q^\pi(s,a)$:

$$
\begin{aligned}
Q^\pi(s_t, a_t) &= \mathbb{E}_{s_{t+1}, a_{t+1}, \dots}\left[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}) \Big| s_t, a_t\right] \\
&= \mathbb{E}_{s_{t+1}, a_{t+1}, \dots}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \Big| s_t, a_t\right] \quad \text{let } r_{t+k} \equiv r(s_{t+k}) \\
&= \mathbb{E}_{s_{t+1}, a_{t+1}, \dots}\left[r_t + \sum_{k=1}^{\infty} \gamma^k r_{t+k} \Big| s_t, a_t\right] \\
&= \mathbb{E}_{s_{t+1}, a_{t+1}, \dots}\left[r_t + \gamma \sum_{k=0}^{\infty} \gamma^k r_{(t+1)+k} \Big| s_t, a_t\right] \\
&= \mathbb{E}_{s_{t+1}}\Bigg[\underbrace{r_t}_{\text{only } s_{t+1} \text{ affect it}} \underline{+\gamma \mathbb{E}_{a_{t+1}, s_{t+2}, a_{t+2}, \dots}\left[\sum_{k=0}^{\infty} \gamma^k r_{(t+1)+k} \Big| s_{t+1}\right]} \Bigg| s_t, a_t\Bigg] \\
&= \mathbb{E}_{s_{t+1}}\left[r_t + \gamma V^\pi(s_{t+1})\right]
\end{aligned}
$$

writing advantage function

- if we choose Advantage function to be:

$$A^{\pi}(s, a) = Q_w^{\pi}(s, a) - V_v^{\pi}(s)$$

i.e., if we to construct two neural networks for Q and V, is very inefficient:

- now we can substitute $Q^{\pi}(s_t, a_t) \equiv \mathbb{E}_{s_{t+1}}\left[r_t + V^{\pi}(s_{t+1})\right]$:

$$\begin{aligned}
A^{\pi}(s_t, a_t) &= Q^{\pi}(s_t, a_t) - V^{\pi}(s_t) \\
&= \mathbb{E}_{s_{t+1}}\left[r_t + \gamma V^{\pi}(s_{t+1})\right] - V^{\pi}(s_t) \\
&= \mathbb{E}_{s_{t+1}}\left[r_t + V^{\pi}(s_{t+1}) - V^{\pi}(s_t)\right] \quad \text{put } V^{\pi}(s_t) \text{ inside integral won't affect it} \\
A^{\pi}(s, a) &= \mathbb{E}_{s' \sim P(s'|s,a)}\left[r(s) + \gamma V^{\pi}(s') - V^{\pi}(s)\right] \quad \text{drop } t \text{ and write } s' \equiv s_{t+1}
\end{aligned}$$

- given above, one may approximate $A^{\pi}(s, a)$, by $s' \sim P(s'|s, a)$ and compute:

$$A^{\pi}(s, a) = r(s) + \gamma V_v^{\pi}(s') - V_v^{\pi}(s)$$

using only one network v

Difference between two polices $J(\pi)$ and $J(\beta)$

- start with some not-so-well-known quantity $\mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi}(s_t, a_t)\right]$

$$
\begin{aligned}
&\mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi}(s_t, a_t)\right] \\
&= \mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{s'\sim P(s'|s,a)}\left[r(s) + \gamma V^{\pi}(s') - V^{\pi}(s)\right]\right] \\
&= \mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty}\gamma^t \left[r(s_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)\right]\right] \quad \text{inner sum } \mathbb{E}_{s'\sim P(s'|s,a)} \text{ absorbed by outer} \\
&= \mathbb{E}_{\tau|\beta}\left[\underbrace{\sum_{t=0}^{\infty}\gamma^t \left[\gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)\right]}_{\text{all terms } t \geq 1 \text{ cancels out except } t=0} + \sum_{t=0}^{\infty}\gamma^t\left[r(s_t)\right]\right] \\
&= \mathbb{E}_{\tau|\beta}\left[-V^{\pi}(s_0) + \sum_{t=0}^{\infty}\gamma^t r(s_t)\right] \\
&= -\mathbb{E}_{s_0|\beta}\left[V^{\pi}(s_0)\right] + \mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right] \\
&= \underbrace{-\mathbb{E}_{s_0\sim P(s_0)}\left[V^{\pi}(s_0)\right]}_{\text{independent of policy}} + \mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]
\end{aligned}
$$

Difference between two polices $J(\pi)$ and $J(\beta)$

- start with some not-so-well-known quantity $\mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi}(s_t, a_t)\right]$

$$
\begin{aligned}
\mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi}(s_t, a_t)\right] &= -\mathbb{E}_{s_0\sim P(s_0)}\left[V^{\pi}(s_0)\right] + \mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right] \\
&= -\mathbb{E}_{s_0\sim P(s_0)}\left[V^{\pi}(s_0)\right] + \mathbb{E}_{s_0\sim P(s_0)}\left[\underbrace{\mathbb{E}_{a_0,s_1,a_2,\dots|\beta}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]}_{V^{\beta}(s_0)}\right] \\
&= -J(\pi) + J(\beta)
\end{aligned}
$$

- this implies:

$$
J(\beta) = J(\pi) + \mathbb{E}_{\tau|\beta}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi}(s_t, a_t)\right]
$$

Policy Gradient Theorem (1)

- let $\rho^{\pi}(s \to s', t)$ to be the probability of transition from state $s \to s'$ in t steps.

$$
\begin{aligned}
\nabla_\theta V^\pi(s) =& \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \nabla_\theta V^\pi(s') \\
=& \phi(s) + \sum_{s'} \underline{\sum_a \pi_\theta(a|s) P(s'|s,a)} \nabla_\theta V^\pi(s') \qquad \text{switch the two summation places} \\
=& \phi(s) + \sum_{s'} \underline{\rho^\pi(s \to s', 1)} \nabla_\theta V^\pi(s') \qquad \text{expand this recursion: } s \to s' \text{ and } s' \to s'' \\
=& \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \left[\phi(s') + \sum_{s''} \rho^\pi(s' \to s'', 1) \nabla_\theta V^\pi(s'') \right] \\
=& \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \phi(s') + \underbrace{\sum_{s'} \sum_{s''} \rho^\pi(s \to s', 1) \rho^\pi(s' \to s'', 1)}\, \nabla_\theta V^\pi(s'') \\
=& \phi(s) + \sum_{s'} \rho^\pi(s \to s', 1) \phi(s') + \underbrace{\sum_{s''} \rho^\pi(s \to s'', 2)} \qquad \underbrace{\nabla_\theta V^\pi(s'')}_{\text{Repeatedly expand } \nabla_\theta V^\pi(.):} \\
=& \underbrace{\rho^\pi(s \to s, 0)}_{=1} \phi(s) + \sum_{s^{(1)} \in \mathcal{S}} \rho^\pi(s \to s^{(1)}, 1) \phi(s^{(1)}) + \sum_{s^{(2)} \in \mathcal{S}} \rho^\pi(s \to s^{(2)}, 2) \phi(s^{(2)}) + \ldots \\
=& \sum_{\{s^{(t)}\} \in \mathcal{S}} \sum_{t=0}^{\infty} \rho^\pi(s \to s^{(t)}, t) \phi(s^{(t)})
\end{aligned}
$$

- as it roll out to fullest, you see the role of $\nabla_\theta V^\pi(s^\infty)$ becomes negligible

Policy Gradient Theorem (2)

- starting from a state s_0:

$$
\begin{aligned}
\nabla_\theta J(\theta) &\equiv \nabla_\theta V^\pi(s_0) \\
&= \sum_s \underbrace{\sum_{t=0}^{\infty} \rho^\pi(s_0 \to s, t)}_{\eta(s)} \phi(s) \\
&= \sum_s d^\pi(s)\phi(s) \quad \text{where } d^\pi(s) \equiv \frac{\eta(s)}{\sum_s \eta(s)} \text{ is a normalized version of } \eta(s) \\
&= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a)
\end{aligned}
$$

- $d^\pi(s)$ acts like the weight of derivative concerning particular s
- in words, it says using policy π, which is the probability end up in a particular state s
- also giving enough t, one may transition from s_0 to any any state s

Policy Gradient Theorem (3)

- to write $\nabla_\theta J(\theta)$ in terms of $\mathbb{E}_\pi\left[.\right]$

$$
\begin{aligned}
\nabla_\theta J(\theta) &\propto \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} Q^\pi(s,a)\nabla_\theta \pi_\theta(a|s) \\
&= \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} \pi_\theta(a|s) Q^\pi(s,a) \underbrace{\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}} \\
&= \underbrace{\sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} \pi_\theta(a|s)}_{\mathbb{E}_\pi} \left[Q^\pi(s,a) \underbrace{\nabla_\theta \log \pi_\theta(a|s)} \right] \\
&= \mathbb{E}_\pi\left[Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s)\right]
\end{aligned}
$$

- so we have the final equation:

$$
\nabla_\theta J(\theta) \propto \mathbb{E}_\pi\left[Q^\pi(s,a)\nabla_\theta \ln \pi_\theta(a|s)\right]
$$

Variance reduction using Baseline

- subtract a baseline function $B(s)$ from policy gradient, note $B(s)$ only depends on state s, **not depends on action** a, such that:

$$\mathbb{E}_\pi \left[\underbrace{Q^\pi(s,a)}_{\text{replace with} B(s)} \nabla_\theta \ln \pi_\theta(a|s) \right]$$

$$\begin{aligned}
&\text{so we have: } \mathbb{E}_\pi \left[B(s) \nabla_\theta \ln \pi_\theta(a|s) \right] \\
&= \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(s,a) B(s) \\
&= \sum_{s \in \mathcal{S}} d^\pi(s) B(s) \nabla_\theta \sum_{a \in \mathcal{A}} \pi_\theta(s,a) \\
&= 0
\end{aligned}$$

- A good baseline is $B(s) = V^\pi(s)$:

$$\begin{aligned}
\text{without baseline} \qquad & \nabla_\theta J(\theta) = \mathbb{E}_\pi \left[Q^\pi(s,a) \nabla_\theta \ln \pi_\theta(a|s) \right] \\
\text{with baseline} \qquad & \nabla_\theta J(\theta) = \mathbb{E}_\pi \left[\nabla_\theta \ln \pi_\theta(a|s) (Q^\pi(s,a) - V^\pi(s)) \right] \\
& \qquad\quad\; = \mathbb{E}_\pi \left[\nabla_\theta \ln \pi_\theta(a|s) A^\pi(s,a) \right]
\end{aligned}$$

Off policy

- change behavioral distribution from π to β but target policy is still $\pi_\theta(a|s)$:

$$J(\theta) = \sum_{s\in\mathcal{S}} d^{\beta}(s) \sum_{a\in\mathcal{A}} Q^{\pi}(s,a)\pi_\theta(a|s) = \mathbb{E}_{s\sim d^{\beta}}\left[\sum_{a\in\mathcal{A}} Q^{\pi}(s,a)\pi_\theta(a|s)\right]$$

- adding Importance sampling

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{s\sim d^{\beta}}\left[\sum_{a\in\mathcal{A}} \underbrace{Q^{\pi}(s,a)}_{u}\,\underbrace{\pi_\theta(a|s)}_{v}\right] \\
&= \mathbb{E}_{s\sim d^{\beta}}\left[\sum_{a\in\mathcal{A}} \left(Q^{\pi}(s,a)\nabla_\theta \pi_\theta(a|s) + \pi_\theta(a|s)\nabla_\theta Q^{\pi}(s,a)\right)\right] \\
&\overset{(i)}{\approx} \mathbb{E}_{s\sim d^{\beta}}\left[\sum_{a\in\mathcal{A}} Q^{\pi}(s,a)\nabla_\theta \pi_\theta(a|s)\right] \qquad \text{big assumption: Ignore the red part:} \\
&= \mathbb{E}_{s\sim d^{\beta}}\left[\sum_{a\in\mathcal{A}} \beta(a|s)\frac{\pi_\theta(a|s)}{\beta(a|s)} Q^{\pi}(s,a)\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}\right] \\
&= \mathbb{E}_{\beta}\left[\underbrace{\frac{\pi_\theta(a|s)}{\beta(a|s)}}_{\text{importance weights}} Q^{\pi}(s,a)\nabla_\theta \ln \pi_\theta(a|s)\right]
\end{aligned}$$

- using $\beta = \pi_{k\theta_k}(a|s)$, you have on-policy, so we use β genericaly

Importance sampling interpretation

- looking at:

$$\nabla_\theta J(\theta) = \mathbb{E}_\beta \left[\frac{\pi_\theta(a|s)}{\beta(a|s)} Q^\pi(s,a) \nabla_\theta \ln \pi_\theta(a|s) \right]$$

- this can also be interpreted by:

$$\begin{aligned}
\nabla_\theta J(\theta) \ \Big|\theta_{\text{old}} &= \nabla_\theta \mathbb{E}_{\pi_{\theta_{\text{old}}}(a|s)} \Big[\log(\pi_\theta(a|s)) Q^\pi(s,a) \Big] \Big| \theta_{\text{old}} \\
&= \mathbb{E}_{\pi_{\theta_{\text{old}}}(a|s)} \left[\left(\left(\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \right) \Big| \theta_{\text{old}} \right) Q^\pi(s,a) \right] \\
&= \mathbb{E}_{\pi_{\theta_{\text{old}}}(a|s)} \left[\left(\frac{\nabla_\theta \pi_\theta(a|s) \big| \theta_{\text{old}}}{\pi_{\theta_{\text{old}}}} \right) Q^\pi(s,a) \right] \qquad \pi_\theta(a|s) \text{ is determined, but } \nabla_\theta \pi_\theta(a|s) \text{ nc} \\
&= \mathbb{E}_{\pi_{\theta_{\text{old}}}(a|s)} \left[\left(\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} \right) Q^\pi(s,a) \right] \Bigg| \theta_{\text{old}} \\
&= \mathbb{E}_{\pi_{\theta_{\text{old}}}(a|s)} \left[\nabla_\theta \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} \right) Q^\pi(s,a) \right] \Bigg| \theta_{\text{old}} \\
\implies J(\theta) &= \mathbb{E}_{\pi_{\theta_{\text{old}}}(a|s)} \left[\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} \right) Q^\pi(s,a) \right] \Bigg| \theta_{\text{old}}
\end{aligned}$$

Trust region policy optimization (TRPO)

- look at the equation for off-policy + baseline:

$$\nabla_\theta J(\theta) = \mathbb{E}_\beta \Big[\underbrace{\frac{\pi_\theta(a|s)}{\beta(a|s)}} \nabla_\theta \ln \pi_\theta(a|s)\big(Q^\pi(s,a) - V^\pi(s)\big)\Big]$$

- θ_k is the policy before update, as we do not need to update each time. It can be made same as β (then we have on-policy)

$$\begin{aligned} J(\theta) &= \sum_{s\in\mathcal{S}} \rho^{\pi_{\theta_{\text{old}}}} \sum_{a\in\mathcal{A}} \left(\pi_\theta(a|s)\hat{A}_{\theta_{\text{old}}}(s,a)\right) \\ &= \sum_{s\in\mathcal{S}} \rho^{\pi_{\theta_{\text{old}}}} \sum_{a\in\mathcal{A}} \left(\beta(a|s)\frac{\pi_\theta(a|s)}{\beta(a|s)}\hat{A}_{\theta_{\text{old}}}(s,a)\right) \\ &= \mathbb{E}_{s\sim\rho^{\pi_{\theta_{\text{old}}}}, a\sim\beta}\left[\frac{\pi_\theta(a|s)}{\beta(a|s)}\hat{A}_{\theta_{\text{old}}}(s,a)\right] \end{aligned}$$

- as a side note, if we were to take derivatives to compute for gradient descent:

$$\begin{aligned} \nabla_\theta J(\theta) &= \sum_{s\in\mathcal{S}} \rho^{\pi_{\theta_{\text{old}}}} \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s)\hat{A}_{\theta_{\text{old}}}(s,a) \\ &= \sum_{s\in\mathcal{S}} \rho^{\pi_{\theta_{\text{old}}}} \sum_{a\in\mathcal{A}} \beta(a|s)\frac{\pi_\theta(a|s)}{\beta(a|s)}\frac{\nabla_\theta\pi_\theta(a|s)}{\pi_\theta(a|s)}\hat{A}_{\theta_{\text{old}}}(s,a) \\ &= \mathbb{E}_{s\sim\rho^{\pi_{\theta_{\text{old}}}}, a\sim\beta}\left[\frac{\pi_\theta(a|s)}{\beta(a|s)} \log\left(\nabla_\theta\pi_\theta(a|s)\right)\hat{A}_{\theta_{\text{old}}}(s,a)\right] \end{aligned}$$

TRPO objective

- objective function, assume we let $\beta \equiv \theta_{\text{old}}$:

$$\max_{\pi} \left(J(\pi) - J(\beta)\right)$$

- basically, finding the best new policy π to improve upon the previous behavioral policy β
- however, we need it to:

$$\begin{aligned}
&\max_{\pi}(J(\pi) - J(\beta)) \\
&\qquad J(\pi) - J(\beta) \geq \mathcal{L}_{\beta}(\pi) - C\,\mathbb{E}_{s\sim d_k^{\beta}}[\mathrm{KL}(\pi\|\beta)(s)] \\
&\qquad\qquad = \underbrace{\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t)\right] - C\,\mathbb{E}_{s\sim d_k^{\beta}}[\mathrm{KL}(\pi\|\beta)[s]]}_{\text{lower bound } \mathcal{L}_{\beta}(\pi)}
\end{aligned}$$

- so we just need to maximize $\mathcal{L}_{\beta}(\pi)$ instead

Why equality occurs ($\pi = \beta$)?

$$J(\beta) - J(\beta) = \underbrace{\mathbb{E}_{\tau \sim \beta}\left[\sum_{t=0}^{\infty} \gamma^t \frac{\beta(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t)\right]}_{=\text{what?}} - \underbrace{C\, \mathbb{E}_{s \sim d_k^{\beta}}[\mathrm{KL}(\beta \| \beta)[s]]}_{=0\text{, well, that's KL}}$$

- looking at $\mathbb{E}_{\tau \sim \beta}\left[\sum_{t=0}^{\infty} \gamma^t A^{\beta}(s_t, a_t)\right]$:

$$\begin{aligned}
\mathbb{E}_{\tau \sim \beta}\left[\sum_{t=0}^{\infty} \gamma^t A^{\beta}(s_t, a_t)\right] &= \sum_{t=0}^{\infty} \gamma^t \sum_{a_t \in \mathcal{A}} A^{\beta}(s_t, a_t) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{a_t \in \mathcal{A}} \left(Q^{\beta}(s_t, a_t) - V^{\beta}(s_t) \right) \\
&= \sum_{t=0}^{\infty} \gamma^t \left(\sum_{a_t \in \mathcal{A}} Q^{\beta}(s_t, a_t) \right) - V^{\beta}(s_t) \\
&= \sum_{t=0}^{\infty} \gamma^t \left(V^{\beta}(s_t) - V^{\beta}(s_t) \right) = 0
\end{aligned}$$

- As a side note: if instead we look at $\mathbb{E}_{\tau \sim \beta}\left[\sum_{t=0}^{\infty} \gamma^t f(a_t) A^{\beta}(s_t, a_t)\right]$:

$$\begin{aligned}
\mathbb{E}_{\tau \sim \beta}\left[\sum_{t=0}^{\infty} f(a_t) \gamma^t A^{\beta}(s_t, a_t)\right] &= \sum_{t=0}^{\infty} \gamma^t \sum_{a_t \in \mathcal{A}} f(a_t) \left(Q^{\beta}(s_t, a_t) - V^{\beta}(s_t) \right) \\
&= \sum_{t=0}^{\infty} \gamma^t \underbrace{\left(\sum_{a_t \in \mathcal{A}} f(a_t) Q^{\beta}(s_t, a_t) \right)}_{\neq f(a_t) V^{\beta}(s_t)} - f(a_t) V^{\beta}(s_t)
\end{aligned}$$

Always improve the result

- we know,

$$
\begin{aligned}
& J(\beta) - J(\beta) = 0, \text{ and, } J(\pi) - J(\beta) \geq \mathcal{L}_{\beta}(\pi) \\
\implies & J(\pi) - J(\beta) \geq 0 \text{ after we optimized } \mathcal{L}_{\beta}(\pi)
\end{aligned}
$$

- meaning the new policy is always as good as the previous one

KL penalized vs KL constrained

Two different constraints for $\text{KL}(\pi \| \beta)$

- $\text{KL}(\pi \| \beta) = C$:

$$\max_{\pi} \Bigg[\underbrace{\mathbb{E}_{\tau \sim \beta} \Bigg[\sum_{t=0}^{\infty} \gamma^t \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t) \Bigg]}_{\mathcal{L}_{\theta_k}(\theta)} - C \, \text{KL}(\pi \| \beta) \Bigg]$$

- $\text{KL}(\pi \| \beta) \leq \delta$:

$$\max_{\pi} \Bigg[\underbrace{\mathbb{E}_{\tau \sim \beta} \Bigg[\sum_{t=0}^{\infty} \gamma^t \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t) \Bigg]}_{\mathcal{L}_{\theta_k}(\theta)} \Bigg]$$

$$\text{s.t.} \quad \text{KL}(\pi \| \beta) \leq \delta$$

- solving the above is hard, we approx both $\mathcal{L}_{\theta_k}(\theta)$ and $\text{KL}(\pi \| \beta)$ part
- $\text{KL}(\pi \| \beta)$ part need concept of **natural gradient**

Natural Gradient manifold: KL$(\pi\|\beta) = C$

- Taylor (order 1) expansion of $\mathcal{L}(\theta)$:

$$
\begin{aligned}
&\mathcal{L}(\theta + h) \approx \mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta)^\top h \\
\implies &\arg\min_h \{\mathcal{L}(\theta + h)\} \approx \arg\min_h \{\nabla_\theta \mathcal{L}(\theta)^\top h\}
\end{aligned}
$$

- look at steepest gradient descent: we minimize at an **equiv-euclidean-distance** hyper-sphere:

$$
\begin{aligned}
h^* &= \arg\min_h \{\mathcal{L}(\theta + h) : \|h\| = 1\} \\
&\approx \arg\min_h \{\nabla_\theta \mathcal{L}(\theta)^\top h : \|h\| = 1\} \\
&= -\nabla_\theta \mathcal{L}(\theta)
\end{aligned}
$$

- now instead, we minimize at an **equiv-KL-distance** manifold:

$$
\begin{aligned}
h^* &= \arg\min_h \left\{\mathcal{L}(\theta + h) : h \in \left(\mathrm{KL}[p_\theta \| p_{\theta+h}] = c\right)\right\} \\
&\approx \arg\min_h \left\{\nabla_\theta \mathcal{L}(\theta)^\top h : h \in \left(\mathrm{KL}[p_\theta \| p_{\theta+h}] = c\right)\right\}
\end{aligned}
$$

Natural Gradient manifold: $\text{KL}(\pi \| \beta) = C$

- solving

$$h^* \approx \arg\min_h \left\{ \nabla_\theta \mathcal{L}(\theta)^\top h : h \in \left(\text{KL}[p_\theta \| p_{\theta+h}] = c \right) \right\}$$

- solve using Lagrange Multiplier:

$$= \arg\min_h \left(\nabla_\theta \mathcal{L}(\theta)^\top h + \lambda (\text{KL}[p_\theta \| p_{\theta+h}] - c) \right)$$

- **if** we can prove second degree Taylor approximation:

$$\text{KL}[p_\theta \| p_{\theta+h}] \equiv \text{KL}[p(x|\theta) \| p(x|\theta+h)] \approx \frac{1}{2} h^\top \text{F} h \qquad \text{A}$$

- then,

$$\begin{aligned} h^* &\approx \arg\min_h \left(\nabla_\theta \mathcal{L}(\theta)^\top h + \lambda \big(\frac{1}{2} h^\top \text{F} h - c \big) \right) \\ \implies & \frac{\partial}{\partial h} \left(\nabla_\theta \mathcal{L}(\theta)^\top h + \frac{1}{2} \lambda\, h^\top \text{F} h - \lambda c \right) = 0 \\ & \nabla_\theta \mathcal{L}(\theta) + \lambda\, \text{F} h = 0 \\ & h = -\frac{1}{\lambda} \text{F}^{-1} \nabla_\theta \mathcal{L}(\theta) \end{aligned}$$

Ⓐ why second order Taylor Expansion $\text{KL}[p_\theta \| p_{\theta+h}] \approx \frac{1}{2} h^\top \text{F} h$

- look at taylor expansion:

$$f(x_0 + h) \approx f(\mathbf{x}) + f'(\mathbf{x})h + \frac{1}{2} h^\top f''(\mathbf{x}) h \quad | \ \mathbf{x} = x_0$$

- to avoide confusion: $x_0 \to \theta_0$ is constant, and $\theta' \to \theta$ is variable

$$\begin{aligned}
\text{KL}[p_{\theta_0} \parallel p_{\theta+h}] &\approx \text{KL}[p_{\theta_0} \parallel p_\theta] + \left((\nabla_\theta \text{KL}[p_{\theta_0} \parallel p_\theta])^\top h + \frac{1}{2} h^\top (\nabla^2_\theta \text{KL}[p_{\theta_0} \parallel p_\theta]) h \right) \Big|_{\theta=\theta_0} \\
&= \text{KL}[p_{\theta_0} \parallel p_{\theta_0}] + \underbrace{\left(\nabla_\theta \text{KL}[p_{\theta_0} \parallel p_\theta] \Big|_{\theta=\theta_0} \right)^\top}_{1} h + \frac{1}{2} h^\top \underbrace{\left(\nabla^2_\theta \text{KL}[p_{\theta_0} \parallel p_\theta] \Big|_{\theta=\theta_0} \right)}_{2} h \\
&= 0 + 0 + \frac{1}{2} h^\top \text{F} h \\
&= \frac{1}{2} h^\top \text{F} h
\end{aligned}$$

- note the ordering when computing $\nabla_\theta f(\theta, \theta_0)\Big|_{\theta=\theta_0}$: take derivative first, then substitute.
- look at KL between $p(x|\theta)$ and $p(x|\theta')$:

$$\mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] = \mathbb{E}_{p(x|\theta)}\left[\log \frac{p(x|\theta)}{p(x|\theta')}\right] = \mathbb{E}_{p(x|\theta)}[\log p(x|\theta)] - \mathbb{E}_{p(x|\theta)}[\log p(x|\theta')]$$

- taking first derivative with respect to θ':

$$\begin{aligned}
\nabla_{\theta'} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] &= \nabla_{\theta'} \left[\mathbb{E}_{p(x|\theta)}[\log p(x|\theta)] - \mathbb{E}_{p(x|\theta)}[\log p(x|\theta')]\right] \\
&= -\mathbb{E}_{p(x|\theta)}\left[\nabla_{\theta'}[\log p(x|\theta')]\right] \\
&= -\int p(x|\theta)\nabla_{\theta'}[\log p(x|\theta')]\,dx
\end{aligned}$$

① what's $\nabla_{\theta' \to \theta} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')]$?

- let $\theta' \to \theta$:

$$
\begin{aligned}
&\nabla_{\theta'} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] \; |\theta' \to \theta \\
&= -\int p(x|\theta) \nabla_{\theta} [\log p(x|\theta)] \, \mathrm{d}x \\
&= -\int p(x|\theta) \frac{\nabla_{\theta} [p(x|\theta)]}{p(x|\theta)} \, \mathrm{d}x = -\int \nabla_{\theta} [p(x|\theta)] \mathrm{d}x \\
&= -\nabla_{\theta} \left[\int p(x|\theta) \mathrm{d}x \right] \\
&= 0
\end{aligned}
$$

② what's $\nabla^2_{\theta' \to \theta} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')]$?

$$
\begin{aligned}
\nabla_{\theta'} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] &= -\int p(x|\theta) \nabla_{\theta'} \log p(x|\theta') \, \mathrm{d}x \\
\Longrightarrow \quad \nabla^2_{\theta'} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] &= \nabla_{\theta'} \left[-\int p(x|\theta) \nabla_{\theta'} \log p(x|\theta') \, \mathrm{d}x \right] \\
\Longrightarrow \quad \nabla^2_{\theta' \to \theta} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] &= \nabla_{\theta'} \left. \left[-\int p(x|\theta) \nabla_{\theta'} \log p(x|\theta') \, \mathrm{d}x \right] \right|_{\theta' = \theta} \\
&= -\int p(x|\theta) \, \nabla_\theta \left[\nabla_\theta \left[\log p(x|\theta) \right] \right] \mathrm{d}x
\end{aligned}
$$

$$
\begin{aligned}
&\nabla^2_{\theta' \to \theta} \text{KL}[p(x|\theta) \,\|\, p(x|\theta')] \\
&= -\int p(x|\theta)\, \nabla_\theta \left[\nabla_\theta \left[\log p(x|\theta)\right] \right] \text{d}x = -\int p(x|\theta)\, \nabla_\theta \left[\frac{\nabla_\theta [p(x|\theta)]}{p(x|\theta)} \right] \text{d}x \\
&= -\int p(x|\theta)\, \nabla_\theta \left[\underbrace{\nabla_\theta [p(x|\theta)]}_{u}\, \underbrace{p(x|\theta)^{-1}}_{v} \right] \text{d}x \\
&= -\int p(x|\theta) \left[\underbrace{-\nabla_\theta [p(x|\theta)] p(x|\theta)^{-2} \nabla_\theta [p(x|\theta)]}_{uv'} + \underbrace{\nabla^2_\theta [p(x|\theta)] p(x|\theta)^{-1}}_{u'v} \right] \text{d}x \quad \text{scalar form} \\
&= -\int p(x|\theta) \left[\nabla^2_\theta [p(x|\theta)] p(x|\theta)^{-1} - \nabla_\theta [p(x|\theta)]^2 p(x|\theta)^{-2} \right] \text{d}x \\
&= -\int p(x|\theta) \left[\frac{\nabla^2_\theta [p(x|\theta)]}{p(x|\theta)} \right] \text{d}x + \int p(x|\theta) \left[\left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right) \left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right)^\top \right] \text{d}x \quad \text{vector-matrix form} \\
&= -\int \nabla^2_\theta [p(x|\theta)] \text{d}x + \mathbb{E}_{p(x|\theta)} \left[\left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right) \left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right)^\top \right] \\
&= -\nabla^2_\theta \left[\int p(x|\theta) \text{d}x \right] + \mathbb{E}_{p(x|\theta)} \left[\nabla \log p(x|\theta)\, \nabla \log p(x|\theta)^\top \right] \\
&= 0 + \text{F} \\
&= \text{F}
\end{aligned}
$$

Something about Fisher Information (1)

- now, let's have a look at the second derivative:

$$
\begin{aligned}
\nabla^2_{\theta_i,\theta_j}[\log p_\theta(x)] &= \nabla^2_{\theta_i,\theta_j}\left(\frac{\nabla_{\theta_j} p_\theta(x)}{p_\theta(x)}\right) = \nabla_{\theta_i}\left(\frac{\nabla_{\theta_j} p_\theta(x)}{p_\theta(x)}\right) \\
&= \nabla_{\theta_i}\left(\underbrace{\nabla_{\theta_j} p_\theta(x)}_{u}\,\underbrace{p_\theta(x)^{-1}}_{v}\right) \\
&= \underbrace{\frac{\nabla^2_{\theta_i,\theta_j} p_\theta(x)}{p_\theta(x)}}_{u'v} - \underbrace{\frac{\nabla_{\theta_i} p_\theta(x)}{p_\theta(x)}\frac{\nabla_{\theta_j} p_\theta(x)}{p_\theta(x)}}_{uv'}
\end{aligned}
$$

$$
\begin{aligned}
\implies \mathbb{E}_{p(x|\theta)}\left[\nabla^2_{\theta_i,\theta_j}[\log p_\theta(x)]\right] &= \mathbb{E}_{p(x|\theta)}\left[\frac{\nabla^2_{\theta_i,\theta_j} p_\theta(x)}{p_\theta(x)}\right] - \mathbb{E}_{p(x|\theta)}\left[\frac{\nabla_{\theta_i} p_\theta(x)}{p_\theta(x)}\frac{\nabla_{\theta_j} p_\theta(x)}{p_\theta(x)}\right] \\
&= 0 - \mathbb{E}_{p(x|\theta)}\left[\nabla_{\theta_i}[\log(p_\theta(x))]\nabla_{\theta_j}[\log(p_\theta(x))]\right] \\
&= 0 - \mathrm{F}_{i,j}
\end{aligned}
$$

Something about Fisher Information (2)

- as a consequence, one may compute:

$$\mathrm{F}_{i,j} = \mathbb{E}_{p(x|\theta)}\left[\nabla_{\theta_i}[\log(p_\theta(x))]\nabla_{\theta_j}[\log(p_\theta(x))]\right]$$

or,

$$\mathrm{F}_{i,j} = -\mathbb{E}_{p(x|\theta)}\left[\nabla^2_{\theta_i,\theta_j}[\log p_\theta(x)]\right]$$

- of course, we pick the easier of the two!
- now we just proved that,

$$\mathrm{F} = \left(\nabla^2_{\theta}\mathrm{KL}[p_{\theta_0} \parallel p_{\theta}]\Big|_{\theta=\theta_0}\right)$$

Final equation: $\text{KL}(\pi \| \beta) = C$

repeat the steps until convergence:

1. feed-forward
2. compute $\nabla_\theta J(\theta_n)$
3. Compute: $\text{F} = \mathbb{E}_{p(x|\theta_n)}\left[\nabla_\theta [J(\theta_n)] \nabla_\theta [J(\theta_n)]^\top\right]$
4. $\theta_{n+1} = \theta_n - \alpha \, \text{F}^{-1} \nabla_{\theta_n} J(\theta_n)$

Then, for policy gradient, we just need to have:

$$\begin{aligned} \theta_{n+1} &= \theta_n - \alpha \, \text{F}^{-1} \nabla_\theta \left(\sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \pi_{\theta_n}(a|s) Q^\pi(s,a) \right) \\ &= \theta_n - \alpha \, \text{F}^{-1} \left(\sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \nabla_\theta \log \pi_{\theta_n}(a|s) Q^\pi(s,a) \right) \end{aligned}$$

Compatible Function Approximation (1): about $\mathsf{F}^{-1}\nabla_{\theta_n} J(\theta_n)$

- $J(\pi_\theta) \equiv J(\theta) = \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} \pi_\theta(a|s) Q^\pi(s,a)$
- if $\tilde{w} = \mathsf{F}^{-1}\nabla_\theta J(\theta)$ is a single natural policy gradient step, then:
- If we can prove $\tilde{w}$ also minimize sqaured error:

$$\tilde{w} = \arg\min_w \left(\sum_s d^\pi(s) \sum_a \pi_\theta(a|s) \left(w^\top \nabla_\theta \log \pi(a|s,\theta) - Q^\pi(s,a) \right)^2 \right)$$

- **interpretation**: Good actions, i.e., those with large $Q^\pi(s,a)$ value should have feature vectors $\nabla_\theta \log \pi(a|s,\theta)$ that have a large inner product with the natural gradient $\tilde{w}$.

Compatible Function Approximation (2)

- We start the reverse: let $\tilde{w}$ minimize sqaured error:

$$\tilde{w} = \arg\min_{w} \left(\sum_{s} d^{\pi}(s) \sum_{a} \pi_{\theta}(a|s) \left(w^{\top} \nabla_{\theta} \log \pi(a|s,\theta) - Q^{\pi}(s,a) \right)^{2} \right)$$

- then,

$$\begin{aligned}
\nabla_{w} \epsilon(\tilde{w}) &= 0 \\
\nabla_{w} \epsilon(w) &= \nabla_{w} \left(\sum_{s} d^{\pi}(s) \sum_{a} \pi_{\theta}(a|s) \left(\nabla_{\theta} \log \pi_{\theta}(a|s)^{\top} w - Q^{\pi}(s,a) \right)^{2} \right) \\
&\Longrightarrow \sum_{s} d^{\pi}(s) \sum_{a} \pi_{\theta}(a|s) \nabla_{\theta} \log \pi_{\theta}(a|s) \left(\nabla_{\theta} \log \pi_{\theta}(a|s)^{\top} \tilde{w} - Q^{\pi}(s,a) \right) = 0 \\
&\Longrightarrow \underbrace{\sum_{s} d^{\pi}(s) \sum_{a} \pi_{\theta}(a|s) \nabla_{\theta} \log \pi_{\theta}(a|s) \nabla_{\theta} \log \pi_{\theta}(a|s)^{\top}}_{F(\theta)} \tilde{w} \\
&\qquad = \sum_{s} d^{\pi}(s) \sum_{a} \underline{\pi_{\theta}(a|s) \nabla_{\theta} \log \pi_{\theta}(a|s)} Q^{\pi}(s,a) \\
&\qquad = \underbrace{\sum_{s} d^{\pi}(s) \sum_{a} \underline{\nabla_{\theta} \pi_{\theta}(a|s)} Q^{\pi}(s,a)}_{\nabla_{\theta} J(\theta)} \\
&\Longrightarrow F\tilde{w} = \nabla_{\theta} J(\theta) \\
&\Longrightarrow \tilde{w} = F^{-1} \nabla_{\theta} J(\theta)
\end{aligned}$$

Solve TRPO $\text{KL}(\pi \| \beta) \leq \delta$

- elements of the objective equation:

$$
\begin{aligned}
\mathcal{L}_{\theta_k}(\theta) &\approx \underbrace{\mathcal{L}_{\theta_k}(\theta_k)}_{0} + g^\top(\theta - \theta_k) \\
&= g^\top(\theta - \theta_k) \qquad \text{where } g = \nabla_\theta \mathcal{L}_{\theta_k}(\theta) \mid \theta_k \\
\bar{\text{KL}}(\theta \| \theta_k) &\approx \underbrace{\bar{\text{KL}}(\theta_k \| \theta_k)}_{0} + \underbrace{\nabla_\theta \bar{\text{KL}}(\theta_k \| \theta_k)}_{0} + \frac{1}{2}(\theta - \theta_k)^\top \text{F}(\theta - \theta_k) \\
&= \frac{1}{2}(\theta - \theta_k)^\top \text{F}(\theta - \theta_k) \qquad \text{where F} = \nabla_\theta{}^2 \bar{\text{KL}}(\theta \| \theta_k) \mid \theta_k
\end{aligned}
$$

Objective function

- objective function of:

$$\max_{\pi} \underbrace{\left[\mathbb{E}_{\tau \sim \beta} \left[\sum_{t=0}^{\infty} \gamma^t \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t) \right] \right]}_{\mathcal{L}_{\theta_k}(\theta)}$$

$$\text{s.t.} \quad \text{KL}(\pi \| \beta) \leq \delta$$

can be re-formulated as:

$$\theta_{k+1} = \arg\max_{\theta} \left[g^{\top} (\theta - \theta_k) \right]$$

$$\text{s.t.} = \frac{1}{2} (\theta - \theta_k)^{\top} \text{F} (\theta - \theta_k) \leq \delta$$

- answer:

$$\theta_{k+1} = \theta_k + \frac{1}{\sqrt{g^{\top} \text{F}^{-1} g}} \text{F}^{-1} g$$

KKT condition

Let $x \equiv (\theta - \theta_k)$:

- primal:

$$f = \max \left[g^\top x \mid \frac{1}{2} x^\top \mathrm{F} x \leq \delta, \quad x, c \in \mathbb{R}^n, \quad \mathrm{F} \in \mathbb{R}^{n \times n} \right]$$

- Lagrangian

$$\begin{aligned} \mathcal{L}(x, \lambda) &= -g^\top x + \lambda \frac{1}{2} (x^\top \mathrm{F} x - 2\delta) \\ \implies \nabla_x \mathcal{L}(x, \lambda) &= -g + \lambda \mathrm{F} x \end{aligned}$$

- **KKT conditions**:

$$-g + \lambda \mathrm{F} x = 0, \quad \lambda \geq 0, \quad \lambda (x^\top \mathrm{F} x - 2\delta) = 0, \quad x^\top \mathrm{F} x \leq 2\delta$$

find x^*

- condition $\lambda(x^\top \mathrm{F}x - 2\delta) = 0$ states two cases: if $x^\top \mathrm{F}x < 2\delta \implies \lambda = 0$, and from condition $-g + \lambda \mathrm{F}x = 0 \implies g = 0$, which can **not** be the max
  Hence we take another case: $\lambda > 0$, $x^\top Hx = 2\delta$
- find expression of λ without having x

$$
\begin{aligned}
-g + \lambda \mathrm{F}x = 0 \quad &\implies x = \frac{1}{\lambda}\mathrm{F}^{-1}g \\
x^\top \mathrm{F}x &= \left(\frac{1}{\lambda}\mathrm{F}^{-1}g\right)^\top \mathrm{F}\left(\frac{1}{\lambda}\mathrm{F}^{-1}g\right) \\
&= \frac{1}{\lambda^2}g^\top \underbrace{\mathrm{F}^{-1}}_{\text{symmetric}} \mathrm{F}\mathrm{F}^{-1}g = \frac{1}{\lambda^2}g^\top \mathrm{F}^{-1}g = 2\delta \\
&\implies \lambda^2 = \frac{g^\top \mathrm{F}^{-1}g}{2\delta} \\
&\implies \lambda = \sqrt{\frac{g^T \mathrm{F}^{-1}g}{2\delta}} \qquad \text{since} \lambda \geq 0
\end{aligned}
$$

- substitute λ in the expression of x:

$$
x^* = \frac{1}{\lambda}\mathrm{F}^{-1}g = \sqrt{\frac{2\delta}{g^T \mathrm{F}^{-1}g}}\mathrm{F}^{-1}g
$$

Gradient descend via finding maximum first

- solving it using:

$$
\begin{aligned}
x \equiv (\theta - \theta_k) \implies x^* \equiv (\theta_{k+1} - \theta_k) \\
\implies \theta_{k+1} = \theta_k + \sqrt{\frac{2\delta}{\hat{g}_k \hat{\mathrm{F}}_k^{-1} \hat{g}_k}} \hat{\mathrm{F}}_k^{-1} \hat{g}_k
\end{aligned}
$$

- $\hat{\mathrm{F}}_k^{-1}$ is too computational! but we don't need to compute it, however, we can compute $\hat{\mathrm{F}}_k^{-1} \hat{g}_k$ together!

How to find $\bar{\mathsf{F}}\bar{g}$?

- how does it translate to our problem, i.e.,

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

- if matrix $Q \in \mathbb{R}^{n \times n}$ is positive definite, then minimal value $\mathbf{x}^*$ is:

$$Q\mathbf{x}^* = b \implies \quad x^* = Q^{-1} b$$

- as per Conjugate Gradient Ascend algorithm, which requires computation of Qd_k, or $\bar{\mathsf{F}}\bar{g}_k$ (note, not $\bar{\mathsf{F}}^{-1}\bar{g}$)
- **Direct method** can help with it:

$$\mathsf{F}_{ij} = \frac{\partial}{\partial \theta_j} \frac{\partial \mathsf{KL}}{\partial \theta_i}$$

$$\begin{aligned} f_k &= \sum_j \mathsf{F}_{kj} g_j = \sum_j \frac{\partial}{\partial \theta_j} \frac{\partial \mathsf{KL}}{\partial \theta_k} g_j = \left(\frac{\partial}{\partial \theta} \frac{\partial \mathsf{KL}}{\partial \theta_k} \right)^\top g \\ &= \frac{\partial}{\partial \theta_k} \underbrace{\sum_j \frac{\partial \mathsf{KL}}{\partial \theta_j} g_j}_{\text{scalar}} = \frac{\partial}{\partial \theta_k} \underbrace{\left(\frac{\partial \mathsf{KL}}{\partial \theta} \right)^\top g}_{\text{scalar}} \end{aligned}$$

- please refer to my notes on **Conjugate Gradient Descend** `https://github.com/roboticcam/machine-learning-notes/blob/master/files/conjugate.pdf`

Proximal Policy Optimization (PPO)

- the **penalised** version is expressed as:

$$\max_{\pi}\left[\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t\frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}A^{\beta}(s_t,a_t)\right]-C\sqrt{\mathbb{E}_{s\sim d_k^{\beta}}[\mathrm{KL}(\pi\|\beta)[s]]}\right]$$

- PPO is expressed as, using $r_t(\theta)=\frac{\pi_\theta(a|s)}{\beta(a_t|s_t)}$:

$$\max_{\pi}\left[\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t\min\left(\underbrace{r_t(\theta)A^{\beta}(s_t,a_t)},\underbrace{\mathrm{clip}\big(r_t(\theta),1-\epsilon,1+\epsilon\big)A^{\beta}(s_t,a_t)}\right)\right]\right]$$

- if $r_t(\theta)$ falls outside $(1-\epsilon)$ and $(1+\epsilon)$, $A^{\beta}(s_t,a_t)$ will be clipped
- sign of $A^{\beta}(s_t,a_t)$ plays a part:
  1. if $A^{\beta}(s_t,a_t)>0$, PPO clips at $r_t(\theta)=1+\epsilon$
  2. if $A^{\beta}(s_t,a_t)<0$, PPO clips at $r_t(\theta)=1-\epsilon$
- Therefore PPO is **not** the same as:

$$\max_{\pi}\left[\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t\min\left(\underbrace{r_t(\theta)},\underbrace{\mathrm{clip}\big(r_t(\theta),1-\epsilon,1+\epsilon\big)}\right)A^{\beta}(s_t,a_t)\right]\right]$$