→ TRPO

Trust Region methods

Konpat Preechakul
Chulalongkorn University
November 2019

Recap policy gradient

- Policy gradient theorem

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

  - Variance problem

- Advantage

$$A(s,a) = Q(s,a) - V(s)$$

  - Variance reduction
  - A2C

$$\nabla J(\theta) = \mathbb{E}_{s,a}\left[A(s,a)\nabla\log\pi_{\theta}(a|s)\right]$$

Recap policy gradient

- Off-policy gradient
  - Off-policy critic
  - Off-policy actor

$$\nabla_\theta J(\theta) \approx \sum_s d^b(s) \sum_a Q^\pi(s,a) \nabla_\theta \pi_\theta(a|s)$$

- Deterministic policy gradient
  - DDPG

$$\nabla_\theta J(\theta) = \sum_s d^\pi(s) \nabla_a Q_\phi(s,a)|_{a=\pi(s)} \nabla_\theta \pi_\theta(s)$$

Today’s topic

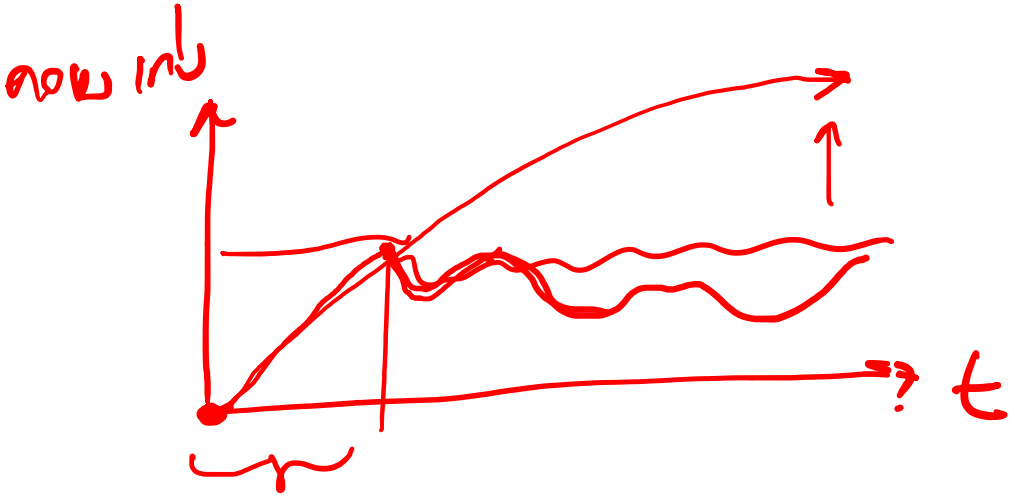
Today topics

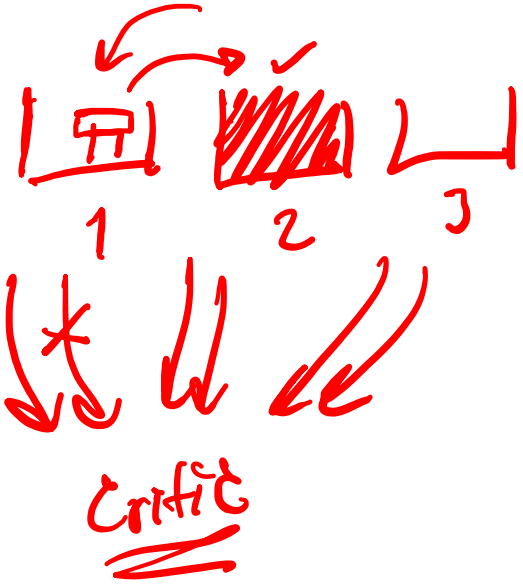
- Why policy gradient fails?
- More robust policy gradient
  - Trust region methods
- Natural gradients

Why policy gradient fails?

Bad critic

- Critic is not oracle, it has its flaws
- **What are some flaws?**





Bad critic

- Critic is not oracle, it has its flaws
- How to reduce the flaws?
- **Critic is prone to forgetting**
- When does critic forget?

Critic is prone to forgetting

- On-policy training "limits" kind of data the critic sees

- If the data is concentrated in "late game", the critic forgets "early game" states
  - It is likely to give gibberish Q to the actor

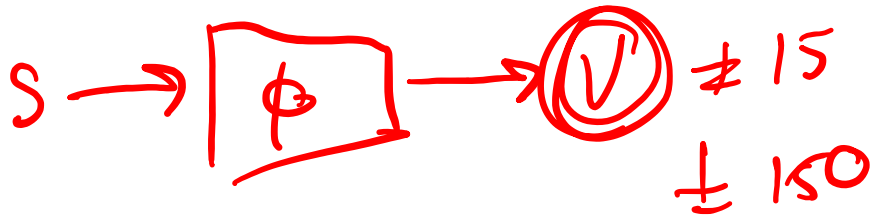
- Underlines using replay, parallel actors

Bad critic

- Critic is not oracle, it has its flaws
- Critic is prone to forgetting
- **How to know when to trust?**

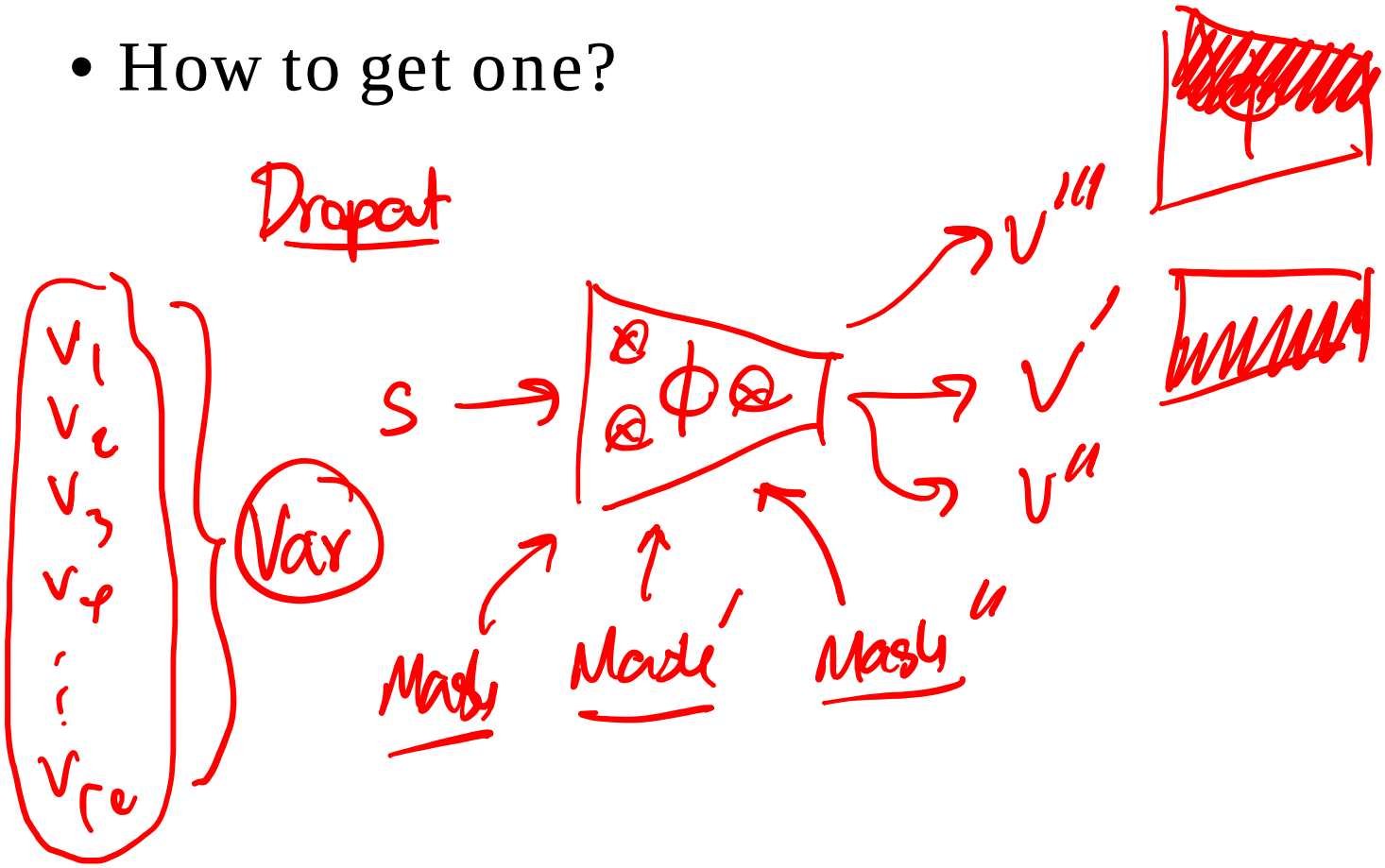
When to trust critic?

- Do critic have “confidence”?



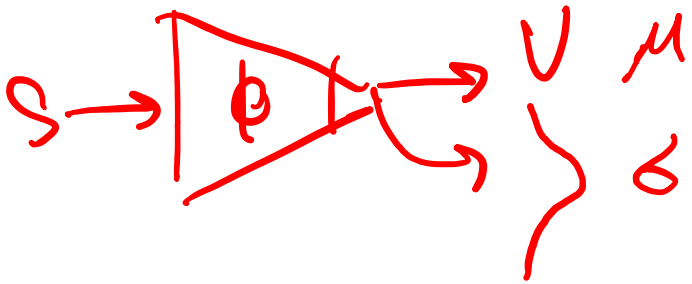
When to trust critic?

- Do critic have “confidence”?
- How to get one?



When to trust critic?

- Do critic have “confidence”?
- How to get one?
- Confidence is challenging:
  - Critic outputs variance?



When to trust critic?

- Do critic have “confidence”?
- How to get one?
- Confidence is challenging:
  - Critic outputs variance?
  - Multiple critics?

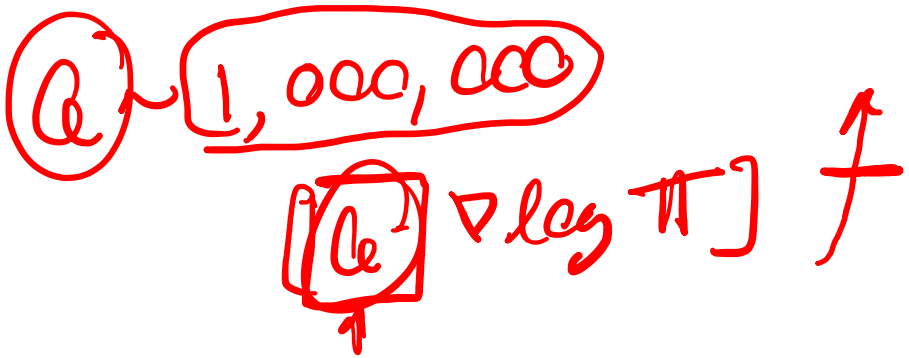
When to trust critic?

- Do critic have “confidence”?
- How to get one?
- Confidence is challenging:
  - Critic outputs variance?
  - Multiple critics?
  - Critic dropout?

Gal, Yarin, and Zoubin Ghahramani. 2015. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1506.02142.

Bad critic

- Critic is not oracle, it has its flaws
- Critic is prone to forgetting
- How to know when to trust?
- **How to limit the trust?**
- Not updating too much ...



Not updating too much

Not updating too much

- How much is too much?
- The right LR?
- What is the right LR?
- If the critic is “abruptly” large, no small LR is small enough

We need to be serious about update!

Quantifying "much"

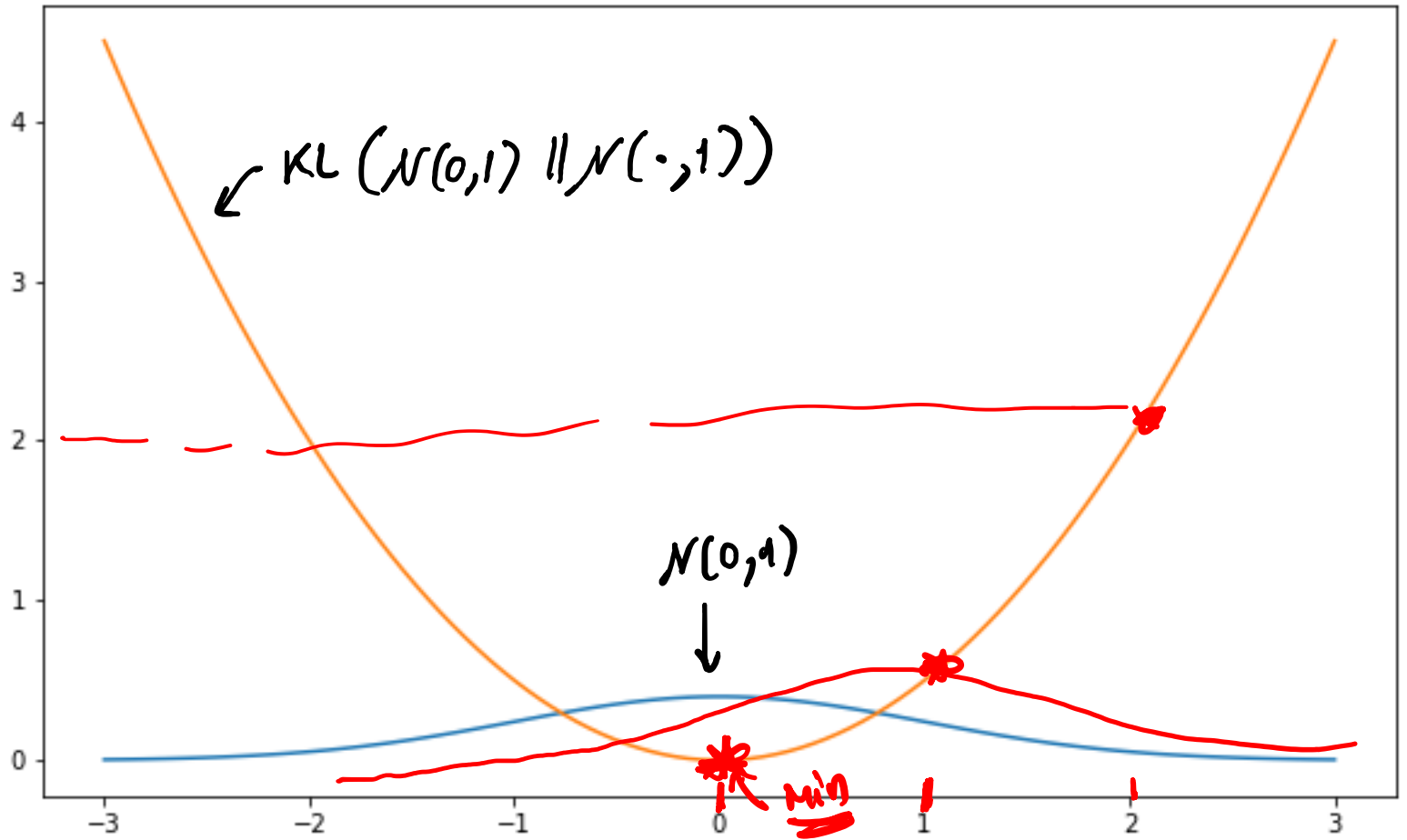
- Policies are probability functions
- We find a "distance" measure on the probabilities
- KL divergence:

$$\mathrm{KL}(P\|Q) = \sum_x P(x)[\log P(x) - \log Q(x)]$$

$\times\ d(x,y) \neq d(y,x)$

KL Divergence in picture

$$\mathrm{KL}(P\|Q) = \sum_x P(x)[\log P(x) - \log Q(x)]$$



Quantify "much" in RL

- Context of RL, for a state S:

$$\mathrm{KL}_s(\theta||\theta') = \sum_a \pi_\theta(a|s)[\log \pi_\theta(a|s) - \log \pi_{\theta'}(a|s)]$$

- For all states:

$$\mathrm{KL}(\theta||\theta') = \sum_s d^\pi(s)\mathrm{KL}_s(\theta||\theta')$$

Do policy gradient while making sure not going too far in terms of KL

$KL(\pi \| \pi') = c$

New objective for policy update

- From the policy gradient objective

$$J(\theta) = \sum_s P_{s_0} V^{\pi}(s)$$

$$\theta \leftarrow \theta + d \qquad d = \alpha \nabla_\theta J(\theta)$$

$E_{s,a}[Q(s,a) \nabla \log \pi(\,)]$

- A new update direction should be: don't worry too much

$$d^* = \underset{d}{\text{argmax}}\ J(\theta + d) \quad \text{s.t. } \mathrm{KL}(\theta \| \theta + d) = c$$

New update direction

$$d^* = \underset{d}{\operatorname{argmax}}\ J(\theta + d) \quad \text{s.t. } \mathrm{KL}(\theta \| \theta + d) = c$$

- A constrained optimization
- We relax it using **Lagrangian:**

$$\mathcal{L}(d, \lambda) = J(\theta + d) + \lambda(\mathrm{KL}(\theta \| \theta + d) - c)$$

- Optimal d is at the critical point

$$\nabla_{d,\lambda} \mathcal{L} = 0$$

Solving for the update direction

$$\mathcal{L}(d, \lambda) = J(\theta + d) + \lambda \left(\mathrm{KL}(\theta \| \theta + d) - c \right)$$

$$\nabla_{d,\lambda} \mathcal{L} = 0$$

$$\nabla_d \mathcal{L}(d, \lambda) = \nabla_d J(\theta + d) + \lambda \nabla_d \mathrm{KL}(\theta \| \theta + d) = 0$$

$$\nabla_\lambda \mathcal{L}(d, \lambda) = \mathrm{KL}(\theta \| \theta + d) - c = 0$$

$\nabla_\theta J(\theta)$

Problematic terms

Cannot calculate easily

① $\nabla_d J(\theta + d)$ ∇J_θ ✓

$J \approx \bar{J}(\theta+d) = J(\theta) + \nabla_\theta J(\theta) \cdot d$

$\nabla_d \bar{J}(\theta+d) = \nabla_\theta J(\theta)$

② $\nabla_d \mathrm{KL}(\theta \| \theta + d)$

$\approx$

$\overline{KL}(\theta \| \theta+d) = \cancel{KL(\theta\|\theta)}\ 0$

$+ \cancel{\nabla_\theta KL(\theta\|\theta) d}\ 0$

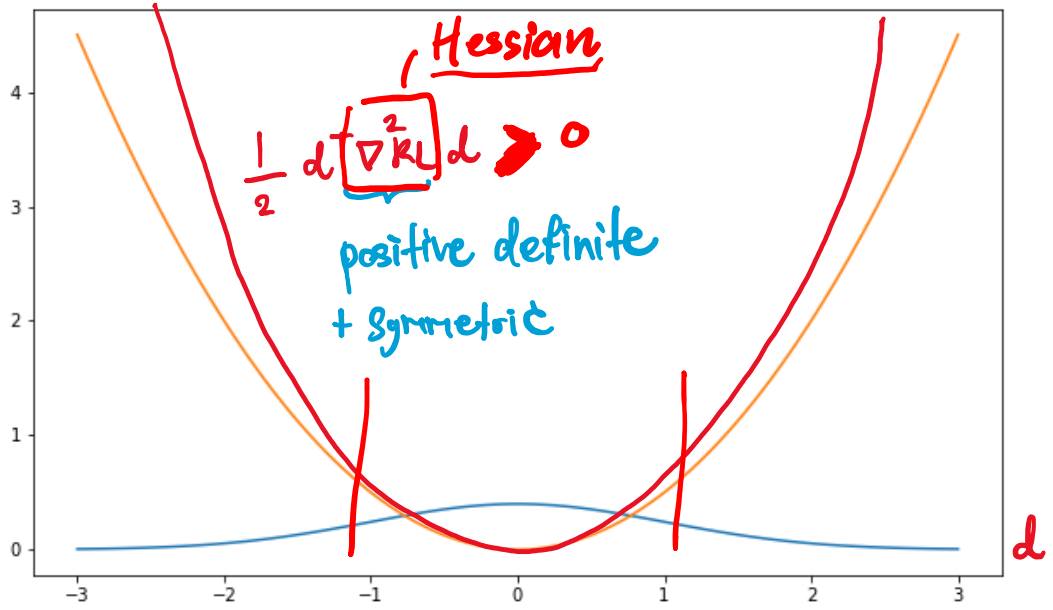
$+ \frac{1}{2} d^T \nabla_\theta^2 KL\, d$

Problematic term

Taylor's to the second order:

$$\mathrm{KL}(\theta \| \theta + d) \approx \overline{\mathrm{KL}}(\theta \| \theta + d)$$

$$= \mathrm{KL}(\theta \| \theta) + \nabla_{\theta'} \mathrm{KL}(\theta \| \theta')|_{\theta'=\theta} d + \frac{1}{2} d^T \nabla^2_{\theta'} \mathrm{KL}(\theta \| \theta')|_{\theta'=\theta} d$$



$\pi \xrightarrow{c} \pi'$

Solving for the update direction

$$0 = \nabla_d J(\theta + d) + \lambda \nabla_d KL(\theta \| \theta + d)$$

$$\nabla_d \frac{1}{2} d^T \nabla^2 KL\, d$$

$$\frac{1}{2} \nabla^2 KL\, d^2$$

$$0 \approx \nabla_\theta J(\theta) + \lambda \nabla_\theta^2 KL(\theta \| \theta) \cdot d$$

$$-\nabla_\theta J(\theta) = \lambda \nabla_\theta^2 KL \cdot d$$

$$-\underbrace{(\nabla_\theta^2 KL)^{-1} \nabla_\theta J}_{s} = d$$

$$s = -(\nabla_\theta^2 KL)^{-1} \nabla_\theta J$$

$$d = +\frac{s}{\lambda}$$

$$KL(\theta \| \theta + d) - c = 0$$

$$\frac{1}{2} d^T \nabla^2 KL\, d - c = 0$$

$$\frac{1}{2}\left(\frac{s}{\lambda}\right)^T \nabla^2 K \left(\frac{s}{\lambda}\right) = 2c\lambda^2$$

$$\frac{s^T \nabla^2 KL\, s}{2c} = \lambda^2$$

$$\sqrt{\frac{s^T \nabla^2 KL\, s}{2c}} = \lambda$$

New update direction

$$d = \frac{s}{\lambda}$$

lr

What does it really mean

$$s = -(\nabla^2_\theta \mathrm{KL})^{-1} \nabla_\theta J(\theta)$$

$$\lambda = \sqrt{\frac{s^T \nabla^2 \mathrm{KL} s}{2c}}$$

previously

$\nabla_\theta L(\theta)$

$$\theta \leftarrow \theta + d$$

Inverse of KL?

- Naïve inverse is not possible to calculate online

$$\nabla_\theta^2 \mathrm{KL}^{-1} = \left(\mathbb{E}_s\left[\nabla_\theta^2 \mathrm{KL}_s\right]\right)^{-1}$$

$\nabla_{\theta'}^2 \mathrm{KL}(\pi_\theta(\cdot|s) \| \pi_{\theta'}(\cdot|s))\big|_{\theta'=\theta}$

- We use:

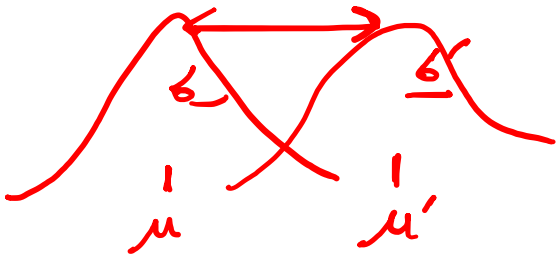
$$\nabla_\theta^2 \mathrm{KL}^{-1} \approx \mathbb{E}_s\left[\left(\nabla_\theta^2 \mathrm{KL}_s\right)^{-1}\right]$$

$D^2L(x)$

Finally, Limited trust PG is

$$\theta \leftarrow \theta + d \qquad d = \frac{s}{\lambda} \qquad \lambda = \sqrt{\frac{s^T \nabla^2 \mathrm{KL} s}{2c}}$$

Trust region



$$s = (\nabla_\theta^2 \mathrm{KL})^{-1} \nabla_\theta J(\theta)$$

$$s \approx \mathbb{E}_s \left[\left(\nabla_\theta^2 \mathrm{KL}_s \right)^{-1} \mathbb{E}_a \left[Q^\pi(s,a) \nabla_\theta \log \pi_\theta(a|s) \right] \right]$$

PG

* Calculation of inverse is expensive

There are tricks to improve this even further

TRPO

Schulman, John, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2015. "Trust Region Policy Optimization." *arXiv [cs.LG]*. arXiv. https://doi.org/10.1063/1.4927398.

Bad critic

- Critic is not oracle, it has its flaws
- Critic is prone to forgetting
- How to know when to trust?
- **How to limit the trust?**

Bad critic

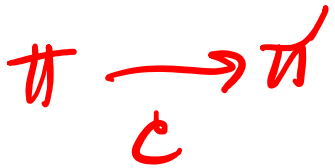
- Critic is not oracle, it has its flaws
- Critic is prone to forgetting
- How to know when to trust?
- How to limit the trust? ←

- **Still doesn’t guarantee policy improvement**

Policy improvement guarantee

Making, Decision, and Control Sergey Levine. n.d. “Advanced Policy Gradients.”

Motivation



- We now have better update
- But, still need to know how large or small the "C" (trust parameter) is
- C could be "varying"
- Too large C could degrade policy
- Too small C is too conservative
- **We want to find C that is just right**

Forming the problem

- We want to "guarantee" policy improvement

$$J(\theta) = \mathbb{E}_{s_0 \sim P(s_0)}[V^{\pi}(s_0)]$$

$$J(\theta') \geq J(\theta)$$

- Objective becomes:

$$\underset{\theta'}{\operatorname{argmax}}\, J(\theta') - J(\theta)$$

- How to estimate?

The problem of estimation

- We want $J(\theta')$

- We need:
  1. Create a new policy
  2. Evaluate the policy

$J(\theta')$

$\checkmark\ a, s \sim \pi$

$\times\ a, s \sim \pi'$

- Aim:
  - Estimate the new policy from **what we have**

Write it in another form

$$J(\theta') - J(\theta) = \underbrace{E_{s_0}[V^{\pi'}(s_0)]}_{} - \underbrace{E_{s_0}[V^{\pi}(s_0)]}_{}$$

(a) ~~V_1 V_2 V_3~~ ~~V_1 V_2 V_3~~

① $E_{s_0}[V^{\pi}(s_0)] = E_{s_0}\left[E_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t V(s_t) - \sum_{t=1}^{\infty}\gamma^t V(s_t)\right]\right]$

$V^{\pi}(s_0)$

$$\uparrow \quad E_{\tau\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t V^{\pi}(s_t) - \gamma^{t+1}V^{\pi}(s_{t+1})\right]$$

$Q \approx r + V'$ $\quad Q - V = r + V' - V$

$V = V$ $\quad A =$

② $E_{s_0}[V^{\pi'}(s_0)] = E_{s_0}\left[E_{\tau\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]\right]$

$$J(\theta') - J(\theta) = E_{\tau\sim\pi'}\left[\sum_{t=0}^{\infty}\gamma^t r_t - \gamma^t V^{\pi}(s_t) + \gamma^{t+1}V^{\pi}(s_{t+1})\right]$$

$$= E_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi}(s_t, a_t)\right]$$

π' → π

$$J(\theta') - J(\theta) = \mathop{E}_{\tau \sim \pi'}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi}(s_t, a_t)\right] \qquad s_0\ s_1\ s_2 \dots s_\infty$$

$$= \sum_{t=0}^{\infty} \mathop{E}_{s_t \sim \pi'}\left[\mathop{E}_{a_t \sim \pi'}\left[\gamma^t A^{\pi}(s_t, a_t)\right]\right]$$

$$= \sum_{t=0}^{\infty} \mathop{E}_{s_t \sim \pi'}\left[\mathop{E}_{a_t \sim \pi}\left[\frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)} \cdot \gamma^t A^{\pi}(s_t, a_t)\right]\right]$$

IS

Write it in another form

$$J(\theta') - J(\theta) = \sum_t \mathbb{E}_{s_t \sim P'(s_t)} \left[\mathbb{E}_{a_t \sim \pi} \left[\frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)} \gamma^t A^{\pi}(s_t, a_t) \right] \right]$$

Can we **lower bound** it while using **only what we have?**

$$J(\theta') - J(\theta) \geq \sum_t \mathbb{E}_{s_t \sim P(s_t)} [\mathbb{E}_{a_t \sim \pi} [\dots]]$$

We need to bound: $|P(s_t) - P'(s_t)| \leq$ [redacted]

Bound the state probability

Intuition: $\pi' \approx \pi \rightarrow P' \approx P$

We will show only a "glimpse"

* Assume π is deterministic $a_t = \pi(s_t)$

Let:

$$\pi'(a_t \neq \pi(s_t)|s_t) \leq \epsilon$$

not taking a_t ← closeness

Then:

τ: $s_0 \rightarrow s_1 \rightarrow s_2 \ldots \rightarrow s_t$

$$P'(s_t) = (1-\epsilon)^t P(s_t) + (1-(1-\epsilon)^t) P_{\text{oth}}(s_t)$$

π'

taking action like π — taking at least 1 wrong action

Bound the probability

$t \to b$

$$\left| P(s_t) - P'(s_t) \right|$$

$$= \left| p(s_t) - (1-\varepsilon)^t p(s_t) - \left(1-(1-\varepsilon)^t\right) p_{oth}(s_t) \right|$$

$$p(s_t)\left(1-(1-\varepsilon)^t\right)$$

$$= \left(1-(1-\varepsilon)^t\right) \left| p(s_t) - p_{oth}(s_t) \right|$$

$\left| p(s_t) - p_{oth}(s_t) \right| \leq 1$

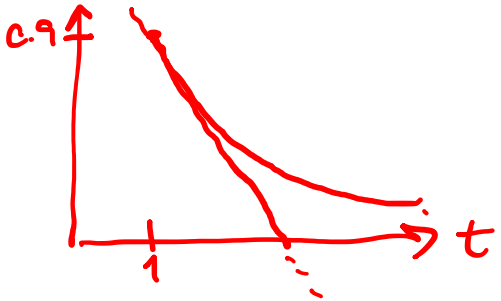
$$\leq \left(1-(1-\varepsilon)^t\right) 1$$

$$\leq \varepsilon t$$

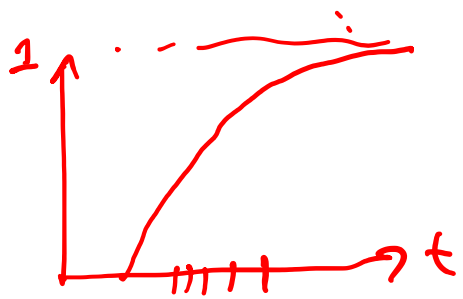
$(1-\varepsilon)^t$

$= 0.9^t$

$\varepsilon = 0.1$



$1-(1-\varepsilon)^t$



General policy case

- It is also possible to show (not here):

$$|\pi'(a|s) - \pi(a|s)| \leq \epsilon$$

$$|P(s_t) - P'(s_t)| \leq \epsilon t$$

- In optimization, we don't have π'
- We need to estimate it, if so:
- Using KL instead would ease estimation

Using KL

- It is possible to show (not here; Pollard, 2000):

$$|\pi'(a|s) - \pi(a|s)| \leq \sqrt{\frac{1}{2}\mathrm{KL}^{\max}(\pi\|\pi')} = \epsilon'$$

$$|P(s_t) - P'(s_t)| \leq \epsilon' t$$

Schulman, John, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2015. "Trust Region Policy Optimization." *arXiv [cs.LG]*. arXiv. https://doi.org/10.1063/1.4927398.

Bound some function

$$\mathbb{E}_{s_t \sim P'(s_t)}[f(s_t)] \geq \mathbb{E}_{s \sim \pi}$$

$$|P(s_t) - P'(s_t)| \leq \epsilon t$$

$$= \sum_{s_t} P'(s_t)[f(s_t)]$$

$$\geq \sum_{s_t} P(s_t) f(s_t) - |P(s_t) - P'(s_t)| \cdot \max_s f(s)$$

$$\geq \sum_{s_t} P(s_t) f(s_t) - \epsilon t \cdot \max_s f(s)$$

Returning to our objective

$$\mathcal{A}_\pi(\pi') = \mathbb{E}_{a_t \sim \pi}\left[\frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)}\gamma^t A^\pi(s_t, a_t)\right]$$

(1) $$J(\theta') - J(\theta) = \sum_t \mathbb{E}_{s_t \sim P'(s_t)}[\mathcal{A}_\pi(\pi')]$$

Lower bound:

$$J(\theta') - J(\theta) = \sum_t \mathbb{E}_{s_t \sim P(s_t)}[\mathcal{A}_\pi(\pi')] - \sum_t \epsilon t \mathcal{O}\left(\frac{r_{\max}}{1-\gamma}\right)$$

small ϵ → policy improvement s.t. $\sqrt{\frac{1}{2}\mathrm{KL}^{\max}(\pi\|\pi')} = \epsilon$

New objective

$$J(\theta') - J(\theta) = \sum_t \mathbb{E}_{s_t \sim P(s_t)} \left[\mathcal{A}_\pi(\pi') \right] - \sum_t \epsilon t \mathcal{O}\left(\frac{r_{\max}}{1-\gamma} \right)$$

$$\text{s.t. } \sqrt{\frac{1}{2} \mathrm{KL}^{\max}(\pi \| \pi')} = \epsilon$$

- Bound is very loose
- We should interpret as:
- **Keep KL small, policy improves!**

Approaching optimization

$$J(\theta') - J(\theta) = \sum_t \mathbb{E}_{s_t \sim P(s_t)} \left[\mathcal{A}_\pi(\pi') \right] - \sum_t \epsilon t \mathcal{O}\left(\frac{r_{\max}}{1-\gamma} \right)$$

$$\text{s.t. } \sqrt{\frac{1}{2} \mathrm{KL}^{\max}(\pi \| \pi')} = \epsilon$$

Becomes:

$$\underset{\theta'}{\operatorname{argmax}} \mathcal{J}(\theta') \quad \text{s.t. } \mathrm{KL}^{\max}(\pi \| \pi_{\theta'}) = \epsilon$$

Approaching optimization

$$\underset{\theta'}{\operatorname{argmax}}\ \mathcal{J}(\theta') \quad \text{s.t. } \mathrm{KL}^{\max}(\pi \| \pi_{\theta'}) = \epsilon$$

$\mathrm{KL}^{\max}(\pi \| \pi_{\theta'})$ is impractical to estimate

We need to go for all S to get the max

Approximate as:

$$\underset{\theta'}{\operatorname{argmax}}\ \mathcal{J}(\theta') \quad \text{s.t. } \mathbb{E}_s[\mathrm{KL}_s(\pi \| \pi_{\theta'})] = \epsilon$$

Limited trust PG

Seem familiar?

① Policy improvement guarantee:

$$\uparrow \underset{\theta'}{\operatorname{argmax}} \mathcal{J}(\theta') \quad \text{s.t. } \mathbb{E}_s\left[\mathrm{KL}_s(\pi \| \pi_{\theta'})\right] = \epsilon$$

How to calculate $\nabla_{\theta'} \mathcal{J}(\theta')$

② Limited trust PG:

$$\uparrow \underset{d}{\operatorname{argmax}} J(\theta + d) \quad \text{s.t. } \mathrm{KL}(\theta \| \theta + d) = c$$

Estimating the gradient

$$\nabla_{\theta'} J(\theta') = \sum_t \mathbb{E}_{s_t, a_t \sim \pi} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi(a_t|s_t)} \gamma^t A^\pi(s_t, a_t) \right]$$

$Q^\pi \nabla \log \pi$

Same old Taylor trick:

$$\nabla_{\theta'} J(\theta') \approx \nabla_{\theta'} J(\theta')|_{\theta'=\theta}$$

$$\nabla_{\theta'} J(\theta') = \sum_t E_\pi \left[\frac{\nabla_{\theta'} \pi_{\theta'}}{\pi_\theta} \gamma^t A^\pi \right] \Big|_{\theta'=\theta}$$

$$= \sum_t E_\pi \left[\frac{\nabla_\theta \pi_\theta}{\pi_\theta} \gamma^t A^\pi \right] = \sum_t E_\pi \left[\gamma^t A^\pi \nabla_\theta \log \pi_\theta \right]$$

$\nabla \log \pi$ ← $\frac{\nabla \pi}{\pi}$; $\sum_t E_\pi$ = $E_{a,s \sim \pi}$; $\gamma^t A^\pi$ = Q^π

① Limited trust PG

② Policy improvement guarantee

$$s \approx \mathbb{E}_s \left[\left(\nabla^2_\theta \mathrm{KL}_s \right)^{-1} \mathbb{E}_a \left[Q^\pi(s,a) \nabla_\theta \log \pi_\theta(a|s) \right] \right]$$

- Turns out to be the same as "Limited trust PG"
- Limit trust => Policy improvement with high chance
- We still cannot "guarantee" we don't know epsilon

- There is another interpretation of $\nabla^2_\theta \mathrm{KL}_s$

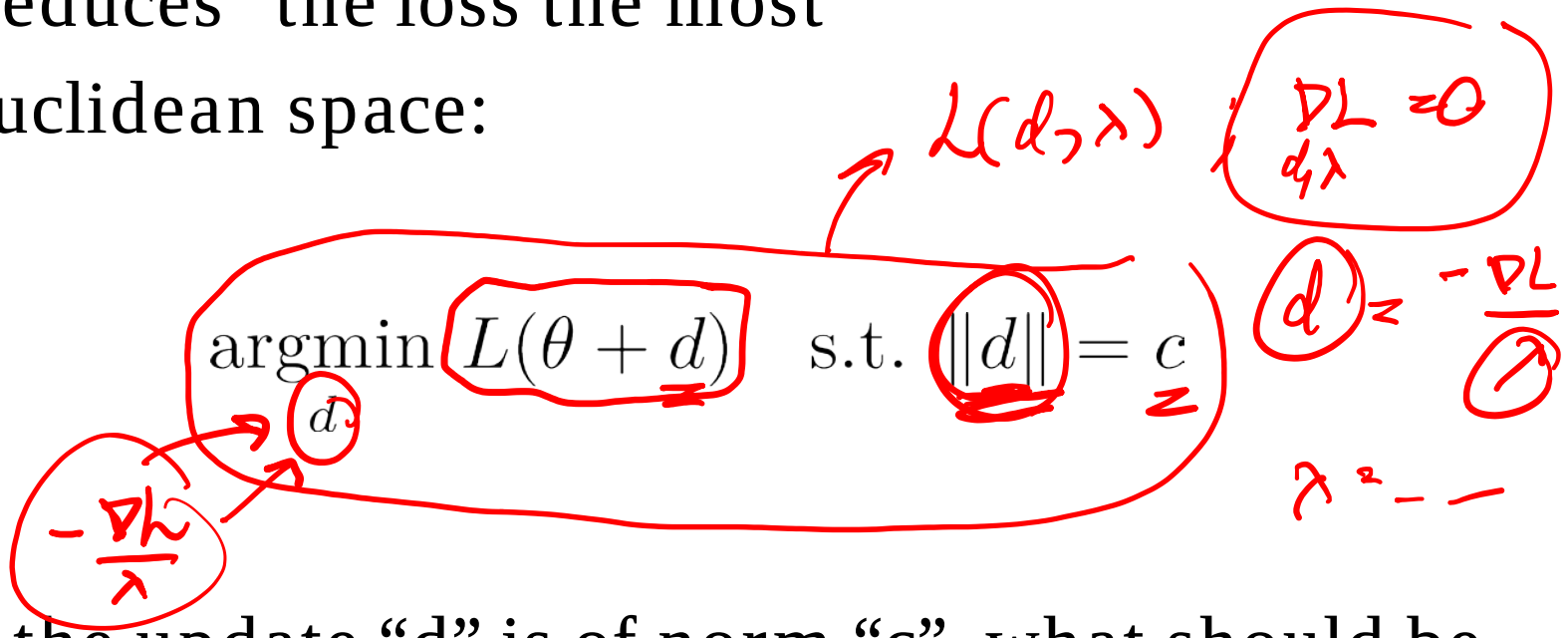
Natural gradients

Motivation

- Gradient descent is not always “steepest”
- Under a more realistic assumption, a better gradient could be derived
- **Leading to faster convergence**
- But higher computation

What is steepest descent?

- Given a fixed budget, what is the update that "reduces" the loss the most
- Euclidean space:



- If the update "d" is of norm "c", what should be its direction?

Gradient descent is steepest in Euclidean

- d is steepest if it reduces the loss fastest

$$d^* = \underset{d}{\operatorname{argmin}}\ L(\theta + d) \quad \text{s.t. } \|d\| = c$$

- Lagrangian

$$\mathcal{L}(d, \lambda) = L(\theta + d) + \lambda(d^T d - c^2)$$

- Solve for critical point:

$$\nabla_d \mathcal{L} = \nabla_d L(\theta + d) + \lambda d = 0$$

$$\nabla_\lambda \mathcal{L} = d^T d - c^2 = 0$$

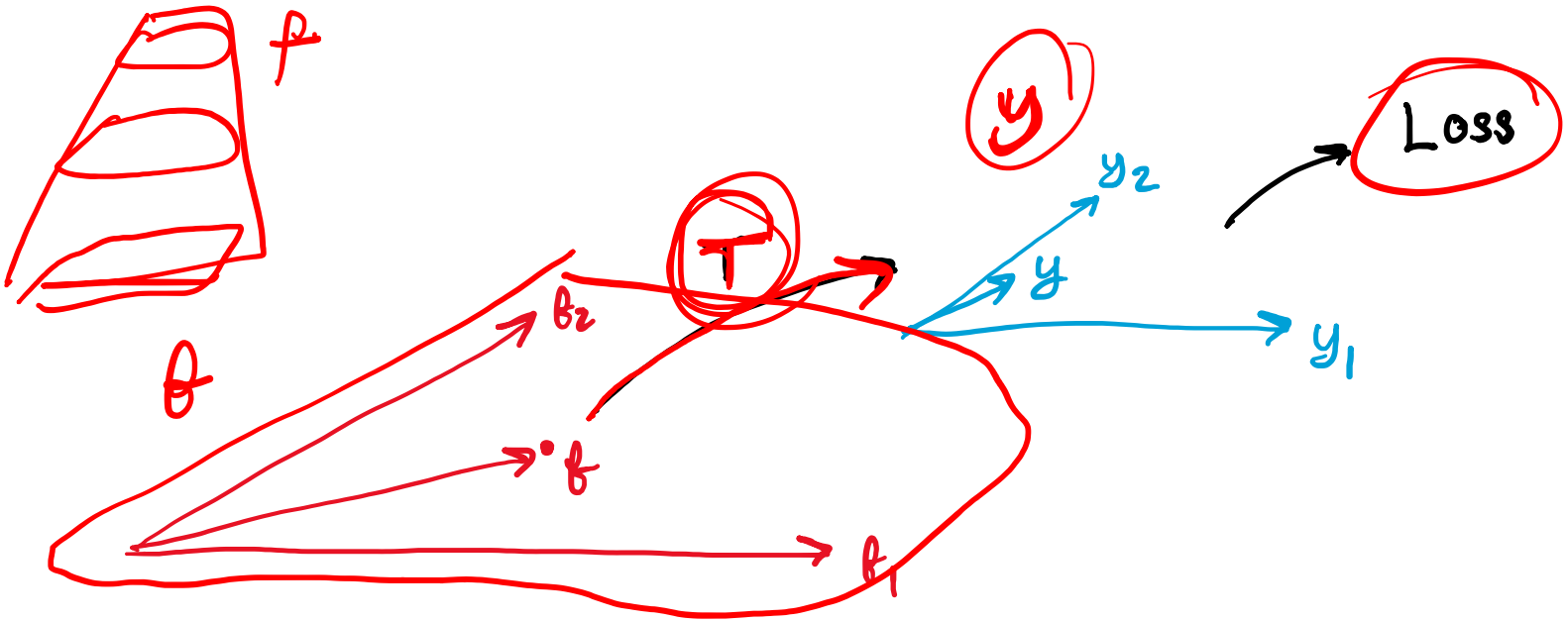
Gradient descent is steepest in Euclidean

Solve for critical point:

$$\nabla_d \mathcal{L} = \nabla_d L(\theta + d) + \lambda d = 0$$

$$\nabla_\lambda \mathcal{L} = d^T d - c^2 = 0$$

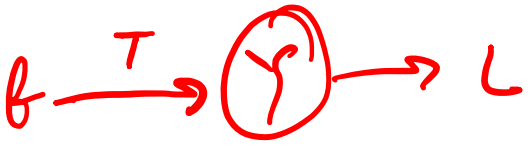
If we have another space above



- What is now the steepest descent wrt. the space above?

$\underset{d}{\operatorname{argmin}} \; L(\beta + d) \quad \text{s.t.} \quad \|d\| \le c$

Norm of the new space



$\|\beta\| = \langle \beta, \beta \rangle = \beta^T \beta$

$T : \beta \to y \qquad y = T\beta$ Linear map

$\|y\| = \langle y, y \rangle = y^T y$

$= (T\beta)^T (T\beta)$

$= \beta^T T^T T \beta \neq \beta^T \beta$

$\|\beta\|_y \neq \|\beta\|_\beta$

Norm of the new space

Steepest is “subjective” because “norm” is subjective

$$d^* = \underset{d}{\operatorname{argmin}}\ L(\theta + d) \quad \text{s.t. } \|d\| = c$$

different meaning

$d \neq -\nabla L$

Is there a more "natural" space than parameter space?

- Policy is a probability function
- It is more natural to think in **"space of probability functions"**
- What is the steepest descent in the probability function space?

$$(\nabla_\theta^2 KL)^{-1}$$

Steepest in prob. fn. space

- Define the "distance" in the function space
- KL Divergence comes into mind:

$$\mathrm{KL}(P\|Q) = \sum_x P(x)[\log P(x) - \log Q(x)]$$

- Steepest descent can get from solving:

$$d^* = \underset{d}{\operatorname{argmax}}\ J(\theta + d) \quad \text{s.t. } \mathrm{KL}(\theta\|\theta + d) = c$$

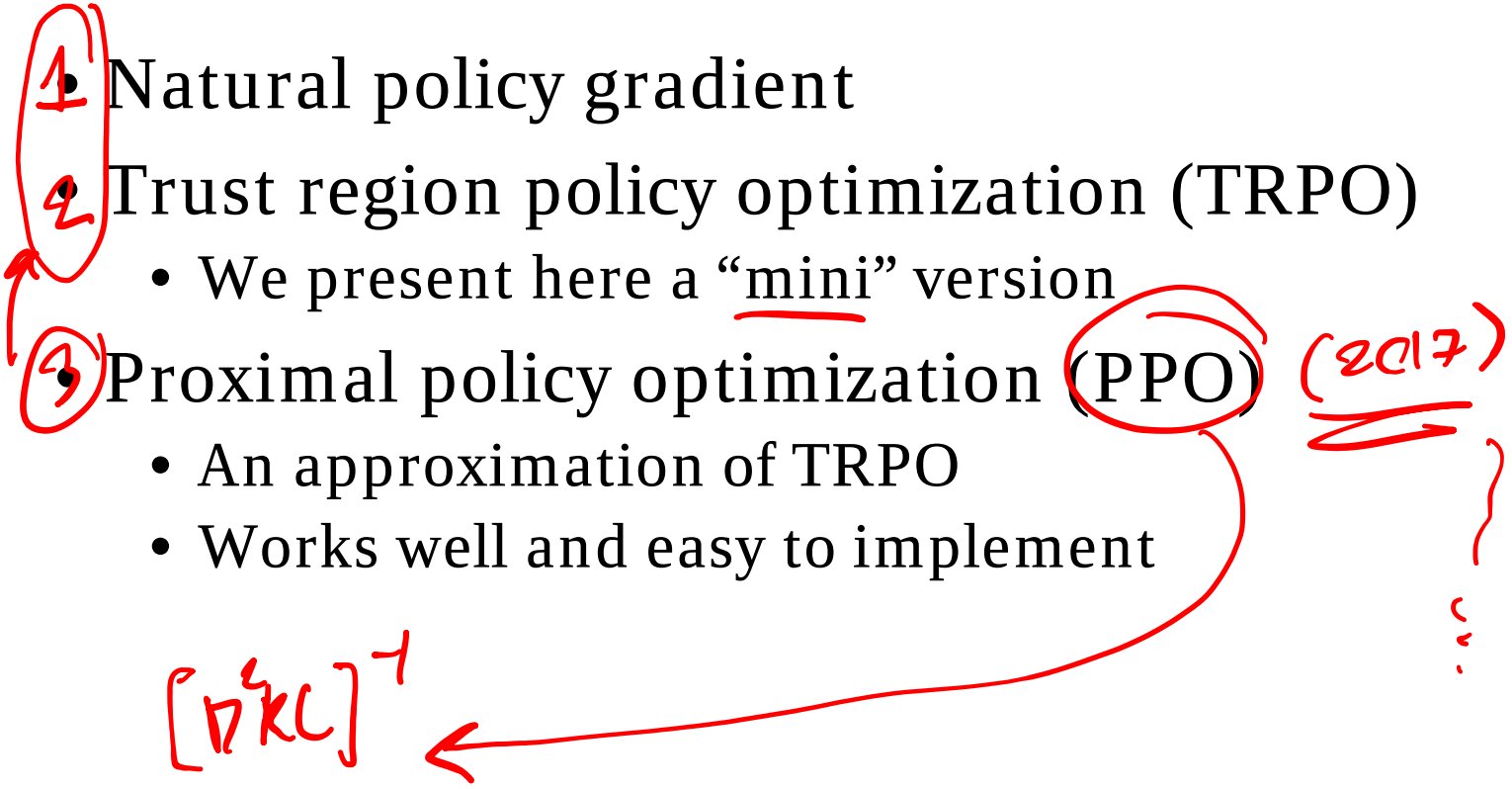
Connection

- Limited trust policy gradient
- Policy improvement guarantee
- Steepest descent on probability function space (Natural gradient)

They are doing the same thing

Related works

1. Natural policy gradient
2. Trust region policy optimization (TRPO)
   - We present here a "mini" version
3. Proximal policy optimization (PPO)
   - An approximation of TRPO
   - Works well and easy to implement



More on policy gradient

Action dependent baseline

- **PG** has high variance because it uses “indirect gradient”

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

- **DPG** has lower variance because it can “backprop”

$$\nabla J(\theta) = \sum_{s} d^{\pi}(s)\nabla_{a}Q_{\phi}(s,a)|_{a=\pi(s)}\nabla_{\theta}\pi_{\theta}(s)$$

Action dependent baseline

- **Can we combine the two?**
- Q-Prop

$$\nabla J(\theta) = \mathbb{E}_{s,a}\left[(Q^{\pi}(s,a) - \overline{Q}_{\theta}(s,a))\nabla_{\theta}\log\pi_{\theta}(a|s)\right] + \mathbb{E}_{s}\left[\nabla_{a}Q_{\phi}(s,a)|_{a=u_{\theta}(s)}\nabla_{\theta}u_{\theta}(s)\right]$$

$$u_{\theta}(s) = \sum_{a}\pi_{\theta}(a|s)a$$

- Taylor expansion (first order)

$$\overline{Q}_{\phi}(s,a) = Q_{\phi}(s,u_{\theta}(s)) + \nabla_{a}Q_{\phi}(s,a)|_{a=u_{\theta}(s)}(a - u_{\theta}(s))$$

Gu, Shixiang, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. 2016. “Q-Prop: Sample-Efficient Policy Gradient with An Off-Policy Critic.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1611.02247.

Policy gradient from minimizing KL

- If we look at Q as "unnormalized" policy
  - A little bit sharper of Q is exp(Q)
  - This is our target policy
- We could use a KL:

$$\pi = \underset{\pi \in \Pi}{\operatorname{argmin}} \, \mathrm{D}_{KL} \left(\pi(\cdot|s) \middle\| \frac{\exp(Q(\cdot, s))}{Z} \right)$$

- Minimizing KL is an optimization task

Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1801.01290.

Policy gradient from minimizing KL

- KL policy gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_s \left[\nabla_\theta \mathrm{D}_{KL} \left(\pi(\cdot|s) \middle\| \frac{\exp(Q(\cdot, s))}{Z} \right) \right]$$

- Z is a constant, ignored
- Policy improve to Q
- Policy eval: Q gets even sharper
- Repeat

Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. “Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1801.01290.