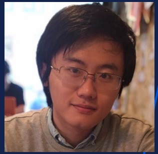
DeepMind

ICML 2020
Paper 941

# Taylor Expansion Policy Optimization



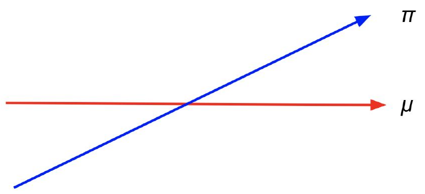
**Yunhao Tang (Columbia University)**, Michal Valko (DeepMind), Rémi Munos (DeepMind)

# Take-away messages

- Generalized formulation of TRPO
  - High-order objective $\rightarrow$ **new algorithm !!!**
  - First-order objective $\rightarrow$ TRPO
- Connections between TRPO vs. off-policy evaluation
  - TRPO $\longleftrightarrow$ special variant of Retrace $Q(\lambda)$
- Performance gains on large-scale algorithms
  - Distributed IMPALA & R2D2

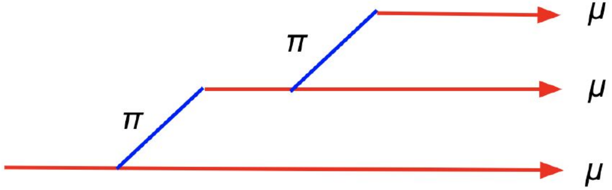
# Intuitions of high-order expansions

- Estimating value-function with off-policy data requires full IS




- First-order: one-step deviation (TRPO, PPO, MPO...)


- Second-order: two-step deviation



# Background: Taylor expansions

- Consider a real function $f(x), x \in \mathbb{R}$
- Fixing a reference point $x_0$
- Any point could be evaluated with the expansion

$$f(x) = \sum_{i=0}^{k} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i + o((x - x_0)^{k+1})$$

- Can we do Taylor expansion of Q-function and value-function?

# Notations

- State space and action space $x_t \in \mathcal{X}, a_t \in \mathcal{A}$
- Policy
  - Target policy $\pi$
  - Behavior policy $\mu$
- Matrix & vector quantities
  - Reward and Q-function $R, Q^\pi \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$
  - Matrix equality $Q^\pi = (I - \gamma P^\pi)^{-1} R$

# Taylor expansions of Q-function

- Useful matrix equality

$$(I - A)^{-1} = (I - B)^{-1} + (I - B)^{-1}(A - B)(I - A)^{-1}$$

- Expanding the Q-function equality w.r.t. $\mu$

$$\begin{aligned} Q^{\pi} &= (I - \gamma P^{\pi})^{-1} R \\ &= Q^{\mu} + (I - \gamma P^{\mu})^{-1}(P^{\pi} - P^{\mu})Q^{\mu} \end{aligned}$$

- Can recursively apply the above expansion

# Taylor expansion of Q-function

- **Theorem 1.** Generic Taylor expansion

$$Q^\pi - Q^\mu = \sum_{k=1}^{K}\Big(\gamma(I-\gamma P^\mu)^{-1}(P^\pi - P^\mu)\Big)^k Q^\mu$$

$$+ \Big(\gamma(I-\gamma P^\mu)^{-1}(P^\pi - P^\mu)\Big)^{K+1} Q^\pi$$

K-th order expansion

$(P^\pi - P^\mu)^K$

Residual term

# Taylor expansion of RL objective

- We care about policy optimization

$$\max_{\pi} V^{\pi}(x_0) = \sum_{a \in \mathcal{A}} \pi(a|x_0) Q^{\pi}(x_0, a)$$

- Can apply similar expansions to value function
  - Make use of results from the Q-function
  - K-th order expansion

$$V^{\pi}(x_0) = (\sum_{k=0}^{K} L_k(\pi, \mu)) + o(|\pi - \mu|^{K+1})$$

# Example: Zero-order expansion

- Zero-order

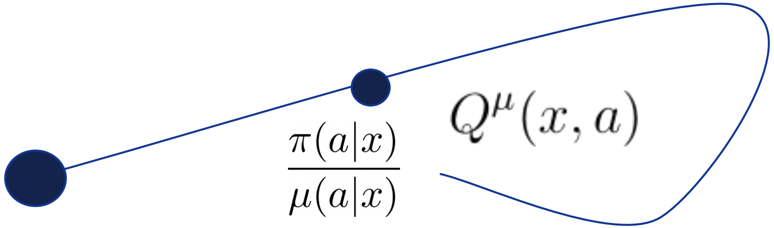
$$L_0(\pi, \mu) = V^{\mu}(x_0)$$

# Example: First-order expansion

- First-order

$$L_1(\pi, \mu) = \mathbb{E}_{(x,a)\sim\mu|x_0}\left[\left(\frac{\pi(a|x)}{\mu(a|x)} - 1\right) Q^{\mu}(x,a)\right]$$

- Can be estimated by samples $(x, a) \sim \mu|x_0$
  - Surrogate objective for TRPO, PPO, MPO...



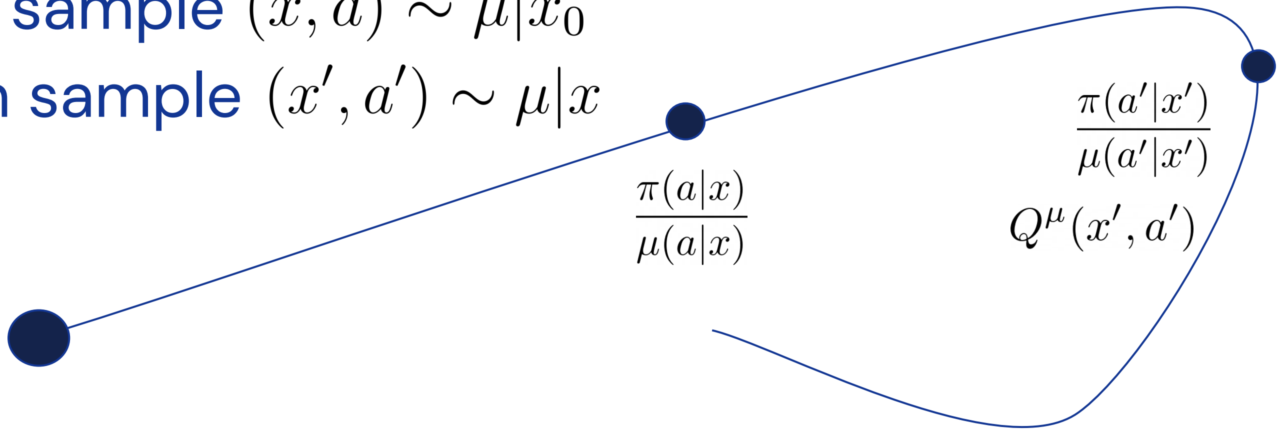

Schulman et al 2015, 2017; Abdolmaleki et al, 2018

# Example: Second-order expansion

- Second-order

$$L_2(\pi,\mu) = \mathbb{E}_{\substack{(x,a)\sim\mu|x_0 \\ (x',a')\sim\mu|x}}\left[\left(\frac{\pi(a|x)}{\mu(a|x)}-1\right)\left(\frac{\pi(a'|x')}{\mu(a'|x')}-1\right)Q^\mu(x',a')\right]$$

- Nested expectation
  - First sample $(x,a)\sim\mu|x_0$
  - Then sample $(x',a')\sim\mu|x$



# Example: K-th order expansion

- General K-th order

$$L_K(\pi,\mu) = \mathbb{E}_{(x^{(i)},a^{(i)})_{1\le i\le K}}\left[\Pi_{i=1}^{K}\left(\frac{\pi(a^{(i)}|x^{(i)})}{\mu(a^{(i)}|x^{(i)})} - 1\right)Q^{\mu}(x^{(K)},a^{(K)})\right]$$

- Nested expectation
  - Sample all pairs sequentially
  - Can be estimated from a single trajectory

# Generalized TRPO

- Generalized objective

$$\max_{\pi} \sum_{k=1}^{K} L_k(\pi, \mu), \ \ |\pi - \mu| < \epsilon$$

- With general K
    - Optimize via backprop and first-order SGD
    - **Theorem 2**. Monotonic improvement
- With large K, optimize the exact objective

$$\lim_{K\to\infty} \sum_{k=1}^{K} L_k(\pi, \mu) = V^{\pi}(x_0) - V^{\mu}(x_0)$$

# Trade-off of K

$$\max_{\pi} \sum_{k=1}^{K} L_k(\pi, \mu), \ |\pi - \mu| < \epsilon$$

| | |
|---|---|
| Large bias | Small bias |
| Small variance | Large variance ? |
| Small K | Large K |

# Variance reduction for K-th order

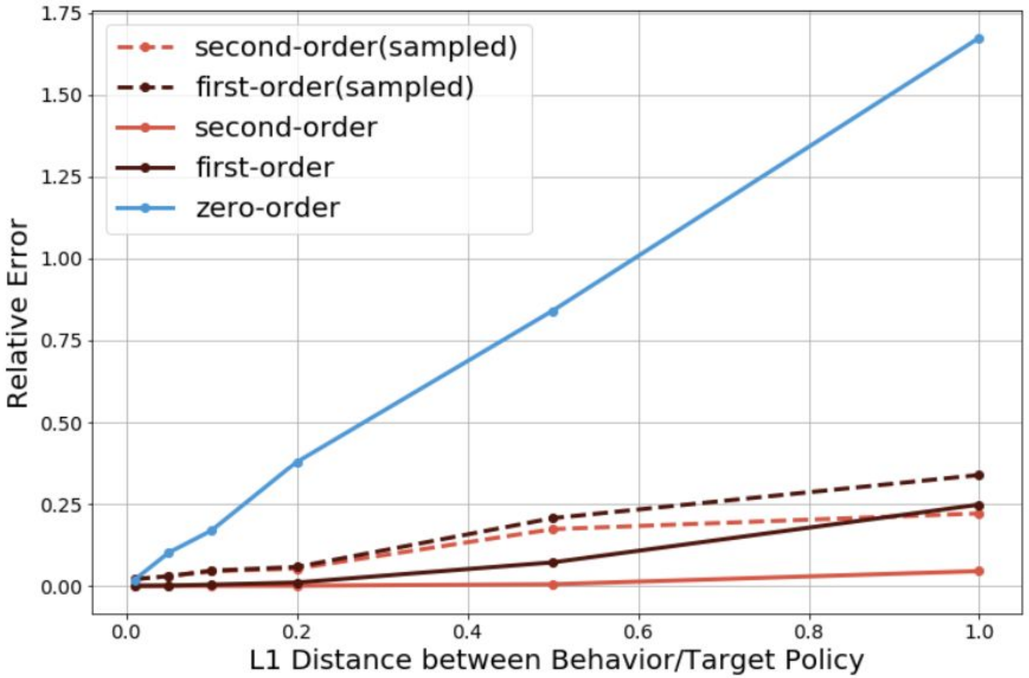
- Replace Q-function estimate by advantage estimate
  - **Theorem 3**. For general K

$$\mathbb{E}_{(x^{(i)},a^{(i)})_{1\le i\le K}}\left[\prod_{i=1}^{K}\left(\frac{\pi(a^{(i)}|x^{(i)})}{\mu(a^{(i)}|x^{(i)})}-1\right)A^{\mu}(x^{(K)},a^{(K)})\right]$$

$$Q^{\mu}(x^{(K)},a^{(K)})$$

# Effect of high-order expansions

- Tabular MDP
  - Can calculate exact error
- Measure the error
  - Zero-order
  - First-order
  - Second-order
- Exact vs. Sample



# TRPO as off-policy evaluation

- Taylor expansions naturally relate to off-policy evaluation

$$\sum_{k=1}^{K} L_k(\pi, \mu) + V^{\mu}(x_0) \approx V^{\pi}(x_0)$$

- All quantities on LHS are from behavior policy
- LHS becomes more accurate with large K

# Background on off-policy evaluation

- Return-based off-policy evaluation
  - Retrace operator $\mathcal{R}_c^{\pi,\mu}$
  - Evaluate by iterating the operator

$$\lim_{K\to\infty} (\mathcal{R}_c^{\pi,\mu})^K Q = Q^\pi$$

- Trace coefficient $c(x,a)$
  - Special case $c(x,a) = \lambda$
  - Converge only when $|\pi - \mu| < \epsilon$

Harutyunyan et al, 2016; Munos et al, 2016

# Connections to off-policy evaluation

- K-th order Taylor expansion is off-policy evaluation
  - **Theorem 4**. Equivalence

K-th order expansion Of Q-func

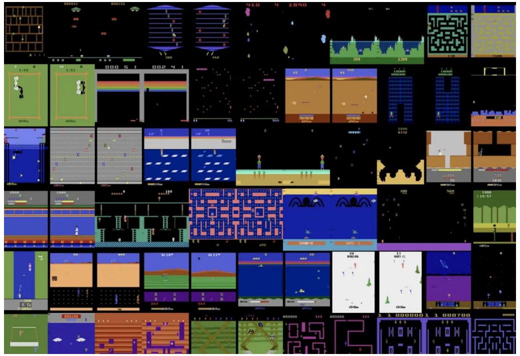
$$Q^{\mu} + \sum_{k=1}^{K} U_k = (\mathcal{R}_1^{\pi,\mu})^K Q^{\mu}$$

Iterating operator K times

- Convergence
  - LHS: Taylor expansion convergence
  - RHS: operator contraction

# Experiments: Second-order new algorithm

- Benchmark: Atari-57 games
- Metric: mean normalized scores
  - See more in paper
- Baseline distributed algorithm
  - Centralized learner $\pi$
  - Distributed actors $\mu$
- Actors sync from learner periodically
  - Actors slightly lag behind learner
  - No explicit trust region (to ensure throughput)
  - Examples: IMPALA, R2D2



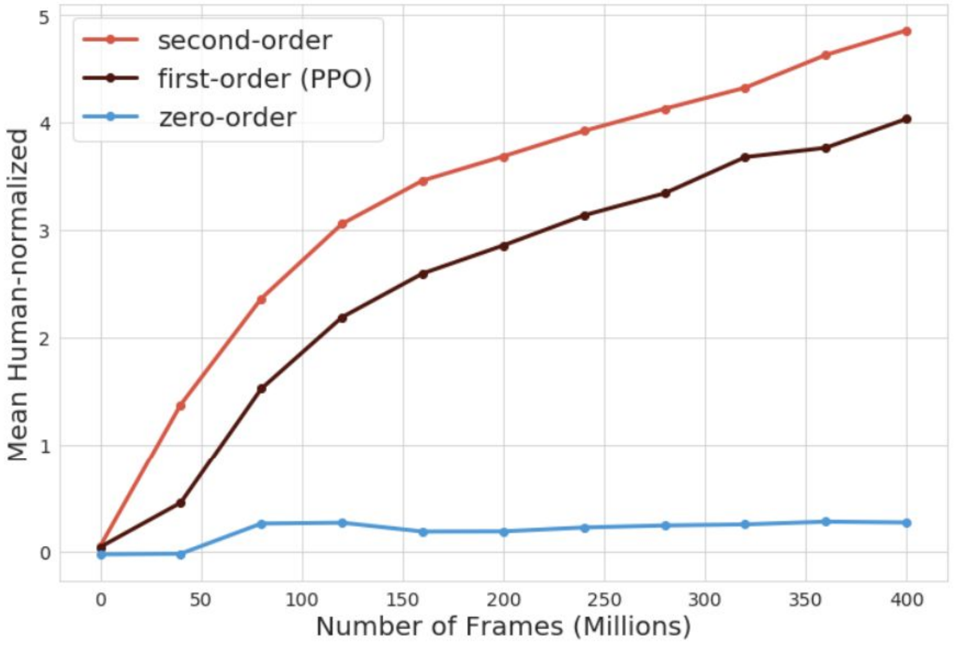
Espeholt et al, 2018; Kapturowski et al, 2018

# Asynchronous actor-critic

- Learner + actors both placed on same TPU
  - Near on-policy?

    $$\pi \approx \mu$$

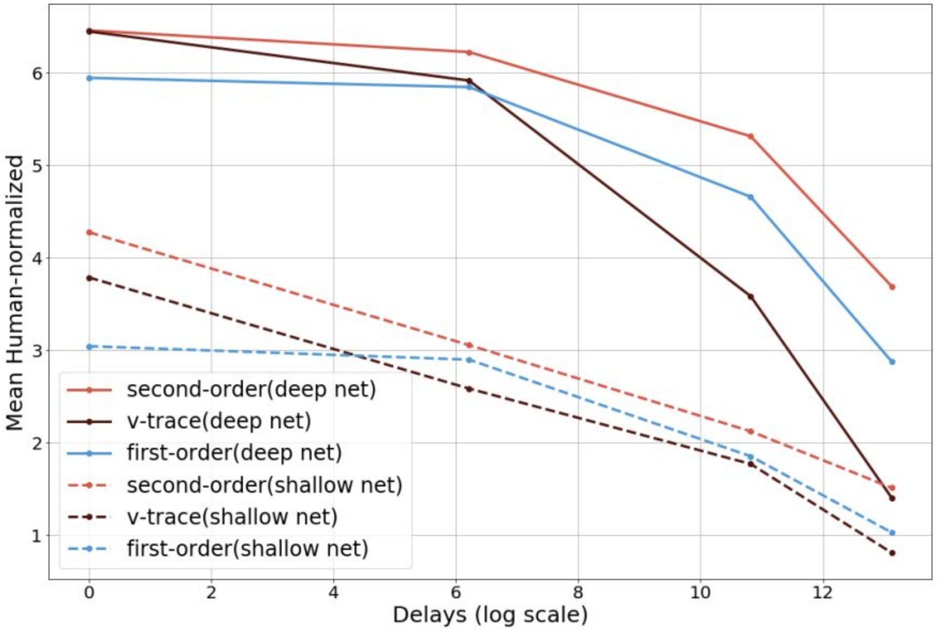
- Actor-critic updates
  - Zero-order
  - First-order (PPO)
  - Second-order



# Distributed actor-critic: IMPALA agent

- Learner on GPU
- Actors on CPUs
- Create artificial updates

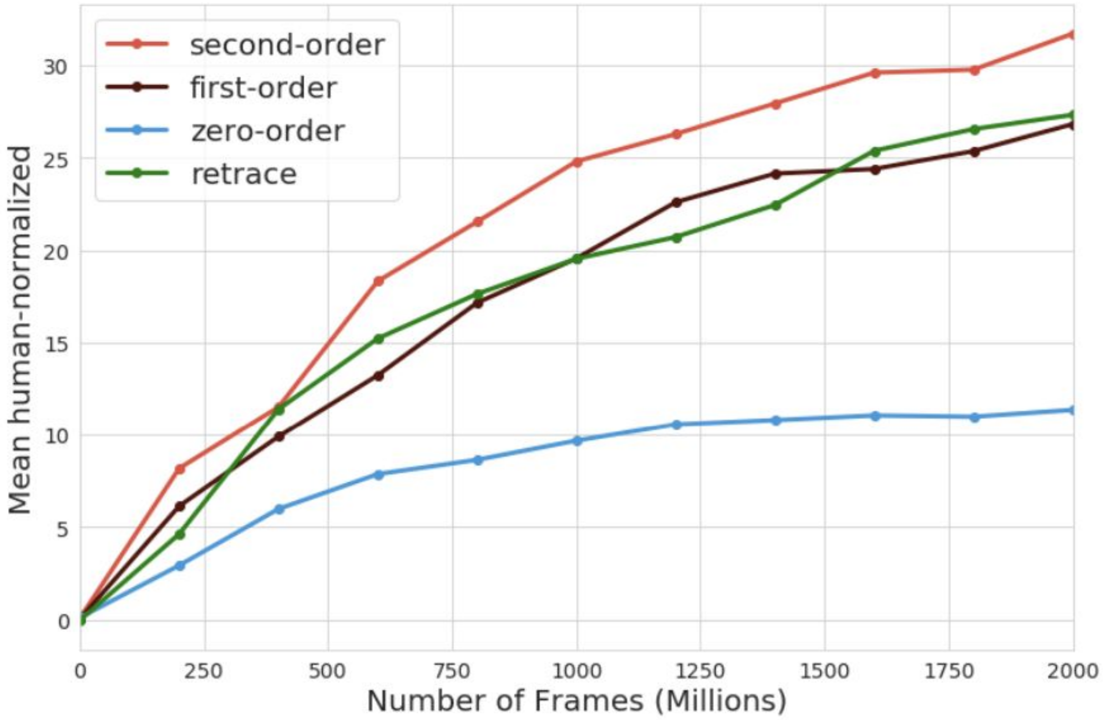
- Actor-critic updates
  - First-order
  - V-trace
  - Second-order



# Distributed Q-learning: R2D2 agent

- Learner on GPU
- Actors on CPUs

- Q-learning
  - Zero-order
  - First-order
  - Retrace
  - Second-order



# Take-home messages

- Taylor expansions generalize TRPO
  - Generalized policy optimization objective
  - Introduce non-linearity beyond first-order
- Taylor expansions $\rightarrow$ off-policy evaluation
  - Taylor expansions $\longleftrightarrow$ a special variant of Retrace
- Empirical gains on distributed algorithms

# Thank you! Please come to our poster


- Special thanks to **Mark Rowland** for insightful comments
- Many thanks to DeepMind teams for technical support
  - Special thanks to **Florent Altche**
  - Special thanks to other DeepMind teams for developments of great distributed agents