ICML | 2022

A Parametric Class of Approximate Gradient Updates for Policy Optimization

Presenter:

Ramki Gummadi Google Research

Joint work with:

Saurabh Kumar Stanford University

Junfeng Wen layer6

Dale Schuurmans Google Research

Introduction

A novel gradient perspective on several objectives in RL enabling:

Conceptual insights

New relationships between classical algorithms:

- Policy Gradients
- Q-learning
- Other surrogate objectives with off-policy corrections

Practical algorithms

A parametric class of update rules that:

- Recovers classical baselines as special cases.
- Enables efficient search over a structured space of updates.
- Delivers gains on both final returns and speed of convergence.

Gradient Updates: Form-Axis Variants

Let Δ_r be the "prediction error"

For 1-Step Q-learning, Bellman error

$$\Delta_r \triangleq \mathcal{T}^\star Q^\pi(s,a) - Q_\theta(s,a)$$

For PG, Monte-Carlo prediction wrt policy logits

$$\Delta_r \triangleq \sum_{t=0}^{\infty} \gamma^t \hat{r}_t - Q_\theta(s,a)$$

Both definitions match for bandits

Theorem: Consider a state-action baseline equal to the policy logits. Then, we can contrast the unbiased gradient estimate as;

$$-\widehat{\nabla_\theta L^{QL}}(s,a) = \Delta_r \nabla_\theta Q_\theta(s,a)$$

$$-\widehat{\nabla_\theta L^{PG}}(s,a) = \Delta_r \left(\nabla_\theta Q_\theta(s,a) - \mathbb{E}_{u\sim\pi_\theta(.|s)} \nabla_\theta Q_\theta(s,u)\right) + \nabla_\theta \mathbb{E}_{u\sim\pi_\theta(.|s)} \hat{Q}_\theta(s,u)$$

Bias correction term for policy logit baseline.

No dependence on data in either term!

Off-policy Corrected Scale-Axis Variant

From each sample, (s, a, r, z) , the gradients depend on two scalar *learning signals*:

$\Delta_r \triangleq \hat{r}_{target} - Q_\theta(s, a)$ ⟹ Prediction error/ bootstrapped Bellman error

$\Delta_O \triangleq \log \frac{\pi_\theta(a|s)}{\pi_b(a|s)}$ ⟹ The usual importance weight ratio with $\pi_\theta(a|s) \sim e^{Q_\theta(s,a)}$

$$-\widehat{\nabla_\theta L^{QL}}(s, a) = \boxed{e^{\Delta_O}\Delta_r} \nabla_\theta Q_\theta(s, a)$$

$$-\widehat{\nabla_\theta L^{PG}}(s, a) = \boxed{e^{\Delta_O}\Delta_r} \left(\nabla_\theta Q_\theta(s, a) - \mathbb{E}_{u\sim\pi_\theta(.|s)} \nabla_\theta Q_\theta(s, u)\right) + \nabla_\theta \mathbb{E}_{u\sim\pi_\theta(.|s)} \hat{Q}_\theta(s, u)$$

⇓

Gradient scaling function

$$f(\Delta_O, \Delta_r) = e^{\Delta_O}\Delta_R$$

A Maximum Likelihood Scale-Axis Variant

$\mathrm{KL}(\hat{p} \| \pi_\theta)$ ⟸ Max-Lik loss for getting π_θ to imitate $\hat{p}$

$\mathrm{KL}(\pi_\theta \| \hat{p})$ ⟸ Entropy regularized expected reward for $\hat{p} \sim e^{\hat{r}}$

Observation: $\Delta_r \approxeq \log \frac{\hat{p}}{\pi_\theta}$ when $\hat{p} \propto e^{\lambda \hat{r} + (1-\lambda) Q_\theta(s,a)}$

⇓

Theorem: $-\nabla_\theta \mathrm{KL}(\hat{p} \| \pi_\theta) = e^{\Delta_O}(e^{\Delta_r} - 1) \nabla_\theta Q_\theta(s, a)$

⇓

Gradient scaling function

$$f(\Delta_O, \Delta_r) = e^{\Delta_O}(e^{\Delta_r} - 1)$$

Matches Q-learning gradients upto first order when $\Delta_R \approxeq 0$

Combination Updates along the Two Axes

Log prob ratio of IW

Pred error

Form Axis

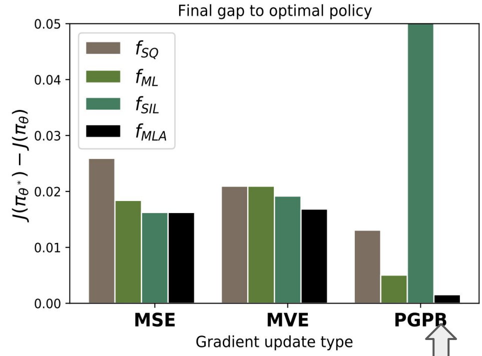
Scale Axis

$$\mathcal{U}_Q(f) \equiv f(\Delta_O, \Delta_R)\Big(\nabla_\theta q_\theta(s,a)\Big)$$

$$\mathcal{U}_V(f) \equiv f(\Delta_O, \Delta_R)\Big(\nabla_\theta q_\theta(s,a) - \mathbb{E}_{u|s\sim\pi}\left[\nabla_\theta q_\theta(s,u)\right]\Big)$$

$$\mathcal{U}_P(f) \equiv f(\Delta_O, \Delta_R)\Big(\nabla_\theta q_\theta(s,a) - \mathbb{E}_{u|s\sim\pi}\left[\nabla_\theta q_\theta(s,u)\right]\Big) + \nabla_\theta \mathbb{E}_{u|s\sim\pi_\theta}\left[\hat{q}_\theta(s,u)\right]$$

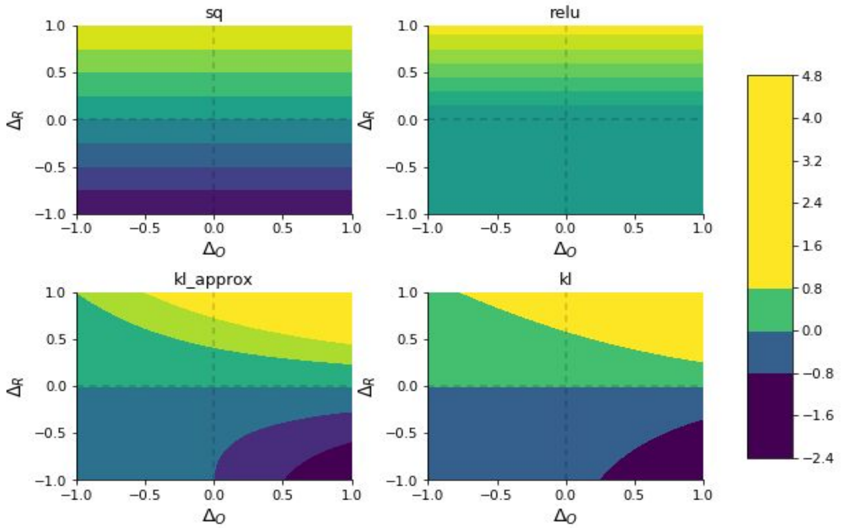
f	$\mathcal{U}_Q(f)$	$\mathcal{U}_V(f)$	$\mathcal{U}_P(f)$
$f_{SQ}(x,y) \triangleq e^x y$	$\hat{G}_Q$	$\hat{G}_V$	$\hat{G}_{PGPB}$
$f_{ML}(x,y) \triangleq e^x(e^y - 1)$	$\hat{G}_{ML}$	$\hat{G}_{ML,V}$	$\hat{G}_{ML,PGPB}$
$f_{SIL}(x,y) \triangleq e^x \max(y,0)$	$\hat{G}_{SIL}$	$\hat{G}_{SIL,V}$	$\hat{G}_{SIL,PGPB}$
$f_{MLA}(x,y) \triangleq \begin{cases} -\frac{1}{2}(1+x)^2, & \text{if } 1+x+y \le 0 < 1+x \\ y\max\left(1+x+\frac{y}{2}, 0\right), & \text{otherwise} \end{cases}$	$\hat{G}_{MLA}$	$\hat{G}_{MLA,V}$	$\hat{G}_{MLA,PGPB}$



Novel updates

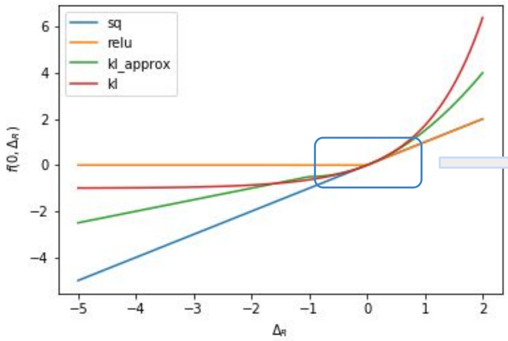
Scale function approximations

Description	$f(\Delta_O, \Delta_R)$
Q-learning	Δ_R
KL divergence exact	$e^{\Delta_O}(e^{\Delta_R} - 1)$
KL divergence approximation	$\Delta_R(1 + \Delta_O + \frac{\Delta_R}{2})_+$
Self imitation learning lower bound	$(\Delta_R)_+$
Hyper-parameterized	???

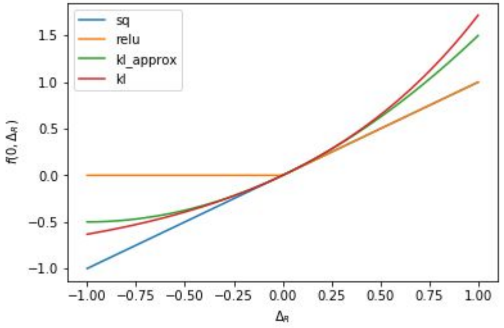


Constraints for $f : \mathbb{R}^2 \mapsto \mathbb{R}$

- $f(., 0) = 0$
- $|f(\Delta_O, .)|$ increasing in Δ_O
- $f(-\infty, .) = 0$
- $f(., \Delta_R)$ increasing in Δ_R



$\Delta_O = 0$



A Parametric Class of Approximate Gradient Updates

Unifies several objectives as differing forms of the novel gradient scale function

- Maximum Likelihood update
- Self Imitation Learning
- Robust losses (e.g. Huber Loss)
- PPO clipping surrogate objective

A simple and general parametric scale function:

$$f_{MLA(\alpha_o,\alpha_r)}(x,y) = y \max\left(1 + \alpha_o x + \alpha_r y, \frac{(1+\alpha_o x)_+}{2}\right)$$

Can recovers classical baselines for known parameters with diverse behaviors

- $\alpha_o = 0$, $\alpha_r = 0$ ⟵ MSE objective for value estimation
- $\alpha_o = 1$, $\alpha_r = 0$ ⟵ (Approximate) Importance weighted PG
- $\alpha_o = 1$, $\alpha_r = 1$ ⟵ (Approximate) Max-Likelihood IW variant

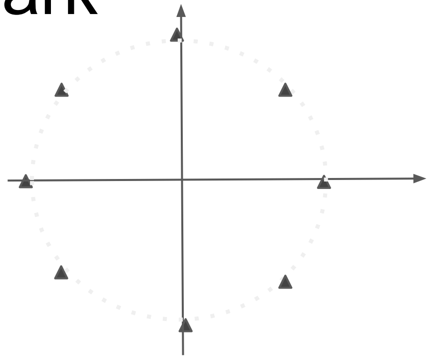
Empirical Analysis: A Diagnostic Benchmark

$$\phi_\theta(x) \triangleq (\theta_0 x_0, \theta_1 x_1) + \theta - \theta^*$$

$$Q_\theta(x, a) = \langle \phi_\theta(x), \psi(a) \rangle$$

Discrete action space embedded on unit circle

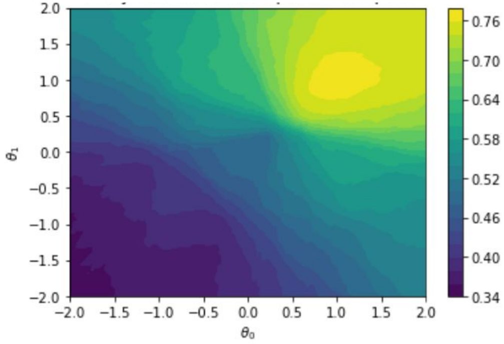
$$\psi(a) = (\cos(2\pi a/N), \sin(2\pi a/N))$$

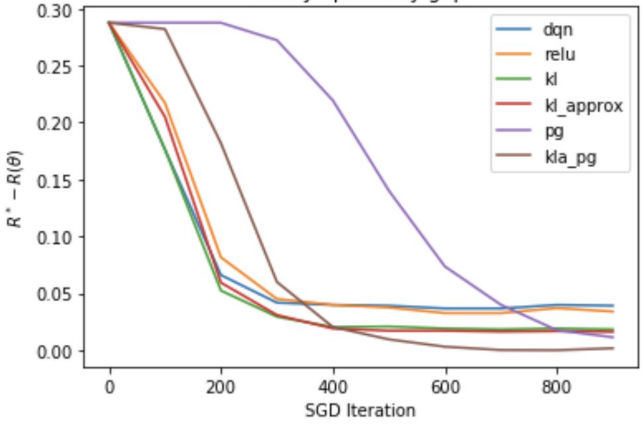


$$Q^*(x, a) = \sigma(Q_{\theta^*}(x, a))$$

- Model mismatch; Impossible to perfectly fit $Q_\theta(x, a)$ to $Q^*(x, a)$
- Optimal policy guaranteed to be $\theta = \theta^*$

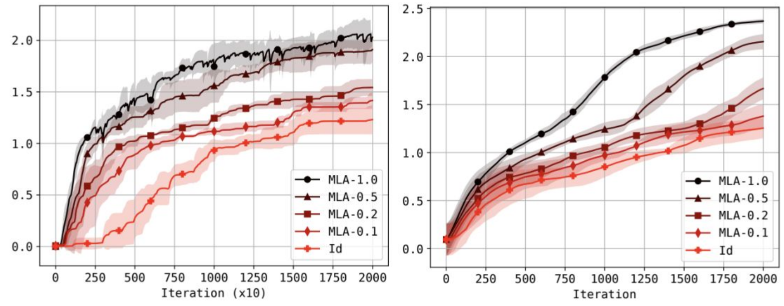
Greedy policy eval for problem setting $\theta^* = (1, 1)$, $N = 8$





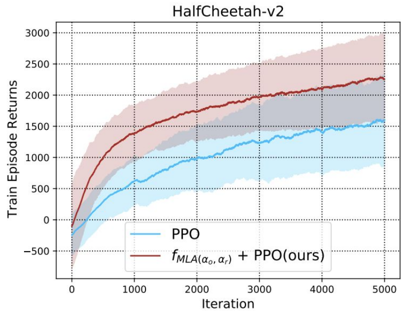
Empirical Results

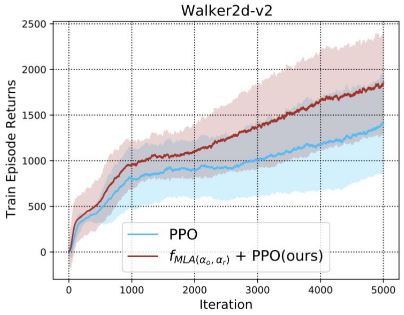
Four Room env (Tabular)

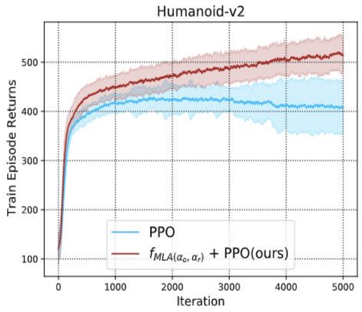


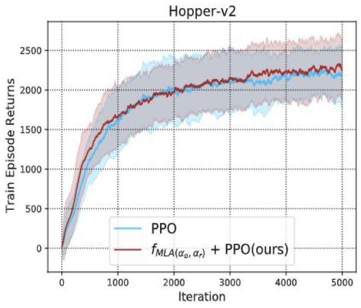
(a) Policy Gradient (b) Q-learning

Continuous Action Control: Mujoco









Summary: A novel perspective on Policy Optimization in RL

- Seemingly different learning objectives in RL have close connections:
  - Combining PG and Q-learning [*Donoghue et al.* 2017]
  - Equivalence between PG and Q-learning [*Schulman et al.* 2018]

- Our contributions:
  - New characterization of relations between gradients of classical objectives.
  - Several gradient updates organized into two novel axes of variation:
    - *Form Axis*: MSE, MVE, Policy Gradient objectives.
    - *Scale Axis*: Off-policy corrections & other surrogate objectives.
  - A simple and performant class of easy-to-tune update rules.