Trust Region Policy Optimization

Reinforcement Learning

Sai Sampath Kedari
sampath@umich.edu

Contents

1 From Natural Policy Gradient to TRPO

Natural policy gradient constructs a policy update by solving the local trust-region problem

$$\boxed{\begin{aligned}\Delta\theta_{\mathrm{NPG}} = \arg\max_{\Delta\theta} \quad & g_{\mathrm{old}}^{\top}\Delta\theta \\ \text{subject to} \quad & \frac{1}{2}\Delta\theta^{\top}F_{\mathrm{old}}\Delta\theta \le \delta.\end{aligned}}$$

Here, the true objective is replaced by its first-order approximation,

$$J(\theta_{\mathrm{old}} + \Delta\theta) \approx J(\theta_{\mathrm{old}}) + g_{\mathrm{old}}^{\top}\Delta\theta,$$

and the average KL divergence is replaced by its second-order approximation,

$$\bar{D}_{\mathrm{KL}}\left(\pi_{\mathrm{old}} \| \pi_{\theta_{\mathrm{old}}+\Delta\theta}\right) \approx \frac{1}{2}\Delta\theta^{\top}F_{\mathrm{old}}\Delta\theta.$$

Solving the local problem gives

$$\boxed{\Delta\theta_{\mathrm{NPG}} = \sqrt{\frac{2\delta}{g_{\mathrm{old}}^{\top}F_{\mathrm{old}}^{-1}g_{\mathrm{old}}}}F_{\mathrm{old}}^{-1}g_{\mathrm{old}}.}$$

Thus, the local models predict

$$g_{\mathrm{old}}^{\top}\Delta\theta_{\mathrm{NPG}} > 0$$

and

$$\frac{1}{2}\Delta\theta_{\mathrm{NPG}}^{\top}F_{\mathrm{old}}\Delta\theta_{\mathrm{NPG}} = \delta.$$

However, these statements concern the local approximations, not necessarily the nonlinear quantities evaluated at the finite candidate

$$\theta_{\mathrm{candidate}} = \theta_{\mathrm{old}} + \Delta\theta_{\mathrm{NPG}}.$$

Two failures are therefore possible.

Objective-model failure. The linear model predicts an improvement,

$$g_{\mathrm{old}}^{\top}\Delta\theta_{\mathrm{NPG}} > 0,$$

but this does not guarantee

$$J(\theta_{\mathrm{old}} + \Delta\theta_{\mathrm{NPG}}) > J(\theta_{\mathrm{old}}).$$

The neglected higher-order terms may make the finite update less effective than predicted or may even cause the true return to decrease.

Constraint-model failure. The NPG step satisfies the quadratic KL constraint,

$$\frac{1}{2}\Delta\theta_{\mathrm{NPG}}^{\top} F_{\mathrm{old}} \Delta\theta_{\mathrm{NPG}} = \delta,$$

but the actual average KL divergence is

$$\bar{D}_{\mathrm{KL}}\left(\pi_{\mathrm{old}} \| \pi_{\theta_{\mathrm{old}}+\Delta\theta_{\mathrm{NPG}}}\right) = \delta + O\left(\|\Delta\theta_{\mathrm{NPG}}\|_2^3\right).$$

Since the neglected remainder need not be negative, the finite candidate may violate the original constraint:

$$\bar{D}_{\mathrm{KL}}\left(\pi_{\mathrm{old}} \| \pi_{\theta_{\mathrm{old}}+\Delta\theta_{\mathrm{NPG}}}\right) > \delta.$$

Therefore,

$$\boxed{\text{NPG constructs a step inside the locally approximated trust region,}}$$

but it does not verify that the finite candidate satisfies the nonlinear objective and KL models.

TRPO addresses this limitation by using the NPG step as an initial proposal. It then evaluates the nonlinear empirical surrogate objective and the actual empirical average KL divergence on the old-policy rollout data. If the candidate fails either test, TRPO shortens the step through backtracking line search.

TRPO cannot directly test whether the true return J has improved without collecting new trajectories from every candidate policy. It instead checks a surrogate objective whose connection to true policy performance is established through the performance-difference identity.

2 From True Performance to the TRPO Optimization Problem

2.1 Exact Policy Improvement

Section 1 showed that TRPO requires an objective that can be evaluated for a candidate policy using trajectories collected by the old policy. Ideally, we would directly evaluate

$$J(\theta) - J(\theta_{\mathrm{old}}),$$

but evaluating $J(\theta)$ requires executing π_θ in the environment.

Using the normalized discounted state-visitation distribution, the performance-difference identity gives

$$\boxed{J(\theta) - J(\theta_{\mathrm{old}}) = \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\left[A^{\pi_{\mathrm{old}}}(S,A)\right]\right].}$$

Equivalently,

$$\boxed{J(\theta) = J(\theta_{\mathrm{old}}) + \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\left[A^{\pi_{\mathrm{old}}}(S,A)\right]\right].}$$

This identity is exact. Its proof is given in Appendix A.

The candidate policy appears in two places:

$$S \sim d^{\pi_\theta,\rho_0}, \qquad A \sim \pi_\theta(\cdot \mid S).$$

The first determines which states the candidate policy visits. The second determines which actions it selects at those states.

2.2 Freezing the State-Visitation Distribution

The available trajectories were generated by the old policy. Therefore, they contain states and actions distributed according to

$$S \sim d^{\pi_{\text{old}},\rho_0}, \qquad A \sim \pi_{\text{old}}(\cdot \mid S),$$

rather than

$$S \sim d^{\pi_\theta,\rho_0}, \qquad A \sim \pi_\theta(\cdot \mid S).$$

Before executing π_θ, its state-visitation distribution is unavailable. TRPO therefore makes the local approximation

$$\boxed{d^{\pi_\theta,\rho_0}(s) \approx d^{\pi_{\text{old}},\rho_0}(s).}$$

Applying this approximation to the exact performance-difference identity gives

$$J(\theta) \approx J(\theta_{\text{old}}) + \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_{\text{old}},\rho_0}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\left[A^{\pi_{\text{old}}}(S,A)\right]\right].$$

This motivates the population surrogate objective

$$\boxed{L_{\pi_{\text{old}}}(\theta) := J(\theta_{\text{old}}) + \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_{\text{old}},\rho_0}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\left[A^{\pi_{\text{old}}}(S,A)\right]\right].}$$

Thus,

$$J(\theta) \approx L_{\pi_{\text{old}}}(\theta).$$

The only approximation introduced so far is

$$d^{\pi_\theta,\rho_0} \longrightarrow d^{\pi_{\text{old}},\rho_0}.$$

2.3 Likelihood-Ratio Form of the Surrogate

The state expectation in $L_{\pi_{\text{old}}}(\theta)$ is now taken under the old policy, but the action expectation is still taken under the candidate policy.

For a fixed state s,

$$\mathbb{E}_{A\sim\pi_\theta(\cdot|s)}\left[A^{\pi_{\text{old}}}(s,A)\right] = \sum_a \pi_\theta(a \mid s)A^{\pi_{\text{old}}}(s,a).$$

Assuming that $\pi_{\text{old}}(a \mid s) > 0$ whenever $\pi_\theta(a \mid s) > 0$,

$$\begin{aligned}\sum_a \pi_\theta(a \mid s)A^{\pi_{\text{old}}}(s,a) &= \sum_a \pi_{\text{old}}(a \mid s)\frac{\pi_\theta(a \mid s)}{\pi_{\text{old}}(a \mid s)}A^{\pi_{\text{old}}}(s,a) \\ &= \mathbb{E}_{A\sim\pi_{\text{old}}(\cdot|s)}\left[\frac{\pi_\theta(A \mid s)}{\pi_{\text{old}}(A \mid s)}A^{\pi_{\text{old}}}(s,A)\right].\end{aligned}$$

Define the likelihood ratio

$$\boxed{r_\theta(S,A) := \frac{\pi_\theta(A \mid S)}{\pi_{\text{old}}(A \mid S)}.}$$

Substituting the likelihood-ratio identity into the surrogate gives

$$\boxed{L_{\pi_{\text{old}}}(\theta) = J(\theta_{\text{old}}) + \frac{1}{1-\gamma}\mathbb{E}_{\substack{S\sim d^{\pi_{\text{old}},\rho_0} \\ A\sim\pi_{\text{old}}(\cdot|S)}}\left[r_\theta(S,A)A^{\pi_{\text{old}}}(S,A)\right].}$$

The likelihood-ratio transformation is exact. It changes the action distribution from π_θ to π_{old} without introducing another population-level approximation.

Since $J(\theta_{\text{old}})$ is constant with respect to θ and $1/(1-\gamma) > 0$, neither quantity changes the maximizer. Define the reduced surrogate objective

$$\boxed{\mathcal{L}^{\text{sur}}_{\pi_{\text{old}}}(\theta) := \mathbb{E}_{\substack{S\sim d^{\pi_{\text{old}},\rho_0} \\ A\sim\pi_{\text{old}}(\cdot|S)}}\left[r_\theta(S,A)A^{\pi_{\text{old}}}(S,A)\right].}$$

Therefore,

$$\boxed{\arg\max_\theta L_{\pi_{\text{old}}}(\theta) = \arg\max_\theta \mathcal{L}^{\text{sur}}_{\pi_{\text{old}}}(\theta).}$$

2.4 Local Agreement with the True Objective

Although the surrogate freezes the state-visitation distribution, it agrees with the true objective at the old policy.

At $\theta = \theta_{\text{old}}$,

$$\begin{aligned} L_{\pi_{\text{old}}}(\theta_{\text{old}}) &= J(\theta_{\text{old}}) + \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_{\text{old}},\rho_0}}\left[\mathbb{E}_{A\sim\pi_{\text{old}}(\cdot|S)}\left[A^{\pi_{\text{old}}}(S,A)\right]\right] \\ &= J(\theta_{\text{old}}), \end{aligned}$$

because

$$\mathbb{E}_{A\sim\pi_{\text{old}}(\cdot|S)}\left[A^{\pi_{\text{old}}}(S,A)\right] = 0.$$

Hence,

$$\boxed{L_{\pi_{\text{old}}}(\theta_{\text{old}}) = J(\theta_{\text{old}}).}$$

The gradients also agree. Since

$$r_\theta(S,A) = \frac{\pi_\theta(A \mid S)}{\pi_{\text{old}}(A \mid S)},$$

we have

$$\nabla_\theta r_\theta(S,A)|_{\theta=\theta_{\text{old}}} = \nabla_\theta \log \pi_\theta(A \mid S)|_{\theta=\theta_{\text{old}}}.$$

Therefore,

$$\begin{aligned} \nabla_\theta L_{\pi_{\text{old}}}(\theta)|_{\theta=\theta_{\text{old}}} &= \frac{1}{1-\gamma}\mathbb{E}_{\substack{S\sim d^{\pi_{\text{old}},\rho_0} \\ A\sim\pi_{\text{old}}(\cdot|S)}}\left[\nabla_\theta \log \pi_\theta(A \mid S)|_{\theta=\theta_{\text{old}}} A^{\pi_{\text{old}}}(S,A)\right] \\ &= \nabla_\theta J(\theta)|_{\theta=\theta_{\text{old}}} \\ &= g_{\text{old}}. \end{aligned}$$

Consequently,

$$\boxed{L_{\pi_{\text{old}}}(\theta_{\text{old}}) = J(\theta_{\text{old}}), \qquad \nabla_\theta L_{\pi_{\text{old}}}(\theta)|_{\theta=\theta_{\text{old}}} = \nabla_\theta J(\theta)|_{\theta=\theta_{\text{old}}}.}$$

Thus, $L_{\pi_{\text{old}}}$ has the same value and the same gradient as the true objective at θ_{old}. The surrogate is therefore a first-order accurate local model of J around the old policy.

The reduced surrogate satisfies

$$\mathcal{L}^{\text{sur}}_{\pi_{\text{old}}}(\theta_{\text{old}}) = 0$$

at the population level.

2.5 The Trust-Region Constraint

The surrogate was constructed by replacing

$$d^{\pi_\theta,\rho_0} \quad \text{with} \quad d^{\pi_{\text{old}},\rho_0}.$$

This approximation is reliable only when the candidate policy remains sufficiently close to the old policy. TRPO therefore constrains how far π_θ may move from $\pi_{\text{old}} = \pi_{\theta_{\text{old}}}$.

Define the old-policy state-averaged KL divergence by

$$\boxed{\bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_\theta\right) := \mathbb{E}_{S\sim d^{\pi_{\text{old}},\rho_0}}\left[D_{\text{KL}}\left(\pi_{\text{old}}(\cdot \mid S) \| \pi_\theta(\cdot \mid S)\right)\right].}$$

The trust-region constraint is

$$\boxed{\bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_\theta\right) \le \delta.}$$

Keeping the candidate policy close to the old policy limits the change in its state-visitation distribution and keeps the surrogate approximation local.

Combining the surrogate objective with the KL constraint gives the nonlinear TRPO optimization problem:

$$\boxed{\begin{aligned} \theta^\star_{\text{TRPO}} = \arg\max_\theta \quad & \mathbb{E}_{\substack{S\sim d^{\pi_{\text{old}},\rho_0} \\ A\sim\pi_{\text{old}}(\cdot|S)}}\left[\frac{\pi_\theta(A \mid S)}{\pi_{\text{old}}(A \mid S)} A^{\pi_{\text{old}}}(S,A)\right] \\ \text{subject to} \quad & \bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_\theta\right) \le \delta. \end{aligned}}$$

Equivalently,

$$\boxed{\begin{aligned} \theta^\star_{\text{TRPO}} = \arg\max_\theta \quad & L_{\pi_{\text{old}}}(\theta) \\ \text{subject to} \quad & \bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_\theta\right) \le \delta. \end{aligned}}$$

At this point, both the surrogate objective and the KL constraint remain nonlinear functions of θ.

2.6 Ideal Policy Optimization, NPG, and TRPO

The ideal policy-optimization problem is simply

$$\boxed{\theta^\star = \arg\max_\theta J(\theta).}$$

There is no trust-region constraint in the original problem. The objective is to find the policy parameters that maximize the true expected return.

NPG constructs a tractable local update by replacing the true objective with its first-order approximation. Because this approximation is reliable only near θ_{old}, NPG restricts the update using the quadratic approximation of the average KL divergence:

$$\boxed{\begin{aligned} \Delta\theta^{\star}_{\text{NPG}} = \arg\max_{\Delta\theta} \quad & g_{\text{old}}^{\top}\Delta\theta \\ \text{subject to} \quad & \frac{1}{2}\Delta\theta^{\top}F_{\text{old}}\Delta\theta \le \delta. \end{aligned}}$$

TRPO instead optimizes the nonlinear surrogate objective while constraining the nonlinear average KL divergence:

$$\boxed{\begin{aligned} \theta^{\star}_{\text{TRPO}} = \arg\max_{\theta} \quad & L_{\pi_{\text{old}}}(\theta) \\ \text{subject to} \quad & \bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_{\theta}\right) \le \delta. \end{aligned}}$$

The three optimization problems can be compared as follows:

Problem	Objective	Constraint
Ideal policy optimization	$J(\theta)$	none
NPG local problem	$g_{\text{old}}^{\top}\Delta\theta$	$\frac{1}{2}\Delta\theta^{\top}F_{\text{old}}\Delta\theta \le \delta$
TRPO	$L_{\pi_{\text{old}}}(\theta)$	$\bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_{\theta}\right) \le \delta$

Therefore, the progression is

$$\boxed{\text{Ideal policy optimization} \longrightarrow \text{NPG local problem} \longrightarrow \text{TRPO.}}$$

NPG trusts the linear objective model and the quadratic KL model. TRPO uses the NPG solution as a candidate step, but evaluates the nonlinear surrogate objective and the nonlinear average KL divergence before accepting that candidate.

The next section shows that locally approximating the TRPO problem recovers the NPG problem. This explains why the natural policy-gradient step becomes the search direction used by TRPO.

3 From TRPO to the NPG Candidate Step

The nonlinear TRPO problem derived in Section 2 is

$$\boxed{\begin{aligned} \theta^{\star}_{\text{TRPO}} = \arg\max_{\theta} \quad & L_{\pi_{\text{old}}}(\theta) \\ \text{subject to} \quad & \bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_{\theta}\right) \le \delta. \end{aligned}}$$

Write the candidate parameters as

$$\theta = \theta_{\text{old}} + \Delta\theta.$$

As established in Section 2, the surrogate and true objective have the same gradient at θ_{old}. Therefore,

$$L_{\pi_{\text{old}}}(\theta_{\text{old}} + \Delta\theta) \approx L_{\pi_{\text{old}}}(\theta_{\text{old}}) + g_{\text{old}}^{\top}\Delta\theta.$$

Using the second-order approximation of the average KL divergence derived in the NPG report,

$$\bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_{\theta_{\text{old}}+\Delta\theta}\right) \approx \frac{1}{2}\Delta\theta^{\top} F_{\text{old}} \Delta\theta.$$

After removing the constant $L_{\pi_{\text{old}}}(\theta_{\text{old}})$, the local TRPO problem becomes

$$\boxed{\begin{aligned} \Delta\theta^{\star} = \arg\max_{\Delta\theta} \quad & g_{\text{old}}^{\top}\Delta\theta \\ \text{subject to} \quad & \frac{1}{2}\Delta\theta^{\top} F_{\text{old}} \Delta\theta \le \delta. \end{aligned}}$$

This is exactly the NPG trust-region problem. Its solution is

$$\boxed{\Delta\theta_{\text{NPG}} = \sqrt{\frac{2\delta}{g_{\text{old}}^{\top} F_{\text{old}}^{-1} g_{\text{old}}}} F_{\text{old}}^{-1} g_{\text{old}}.}$$

TRPO therefore uses

$$\boxed{\theta_{\text{proposal}} = \theta_{\text{old}} + \Delta\theta_{\text{NPG}}}$$

as its initial candidate. Unlike NPG, TRPO does not automatically accept this finite step. It next evaluates the nonlinear surrogate objective and the nonlinear average KL divergence through backtracking line search.

4 Backtracking Line Search

Section 3 produced the NPG proposal

$$\theta_{\text{proposal}} = \theta_{\text{old}} + \Delta\theta_{\text{NPG}}.$$

Because this proposal was obtained using local approximations, TRPO evaluates the corresponding nonlinear surrogate objective and nonlinear average KL divergence before accepting it.

Choose a backtracking coefficient

$$0 < \beta < 1$$

and consider

$$\boxed{\theta_j = \theta_{\text{old}} + \beta^j \Delta\theta_{\text{NPG}}, \qquad j = 0, 1, \ldots, J-1.}$$

Starting with $j = 0$, accept the first candidate satisfying both

$$\boxed{L_{\pi_{\text{old}}}(\theta_j) > L_{\pi_{\text{old}}}(\theta_{\text{old}})}$$

and

$$\boxed{\bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_{\theta_j}\right) \le \delta.}$$

If θ_j fails either condition, increase j and test the shorter step

$$\beta^{j+1}\Delta\theta_{\mathrm{NPG}}.$$

If $j^\star$ is the first accepted index, then

$$\boxed{\theta_{\mathrm{new}} = \theta_{\mathrm{old}} + \beta^{j^\star}\Delta\theta_{\mathrm{NPG}}.}$$

If no candidate satisfies both conditions, reject the update and set

$$\boxed{\theta_{\mathrm{new}} = \theta_{\mathrm{old}}.}$$

The NPG direction remains fixed throughout the line search. TRPO changes only the step length until the surrogate improves and the KL constraint is satisfied.

5 Practical Computation of the TRPO Update

5.1 Rollout Batch

At iteration k, freeze the current policy as

$$\pi_{\mathrm{old}} = \pi_{\theta_k}.$$

Run N parallel environments for T time steps using π_{old}. The rollout buffer therefore contains

$$\boxed{B = TN}$$

transitions.

For environment $n \in \{1, \dots, N\}$ and time $t \in \{0, \dots, T-1\}$, store

$$\left(S_{t,n}, A_{t,n}, R_{t+1,n}, S_{t+1,n}, d^{\mathrm{term}}_{t+1,n}, d^{\mathrm{trunc}}_{t+1,n}, \ell^{\mathrm{old}}_{t,n}, V^{\pi_{\mathrm{old}}}(S_{t,n})\right),$$

where

$$\ell^{\mathrm{old}}_{t,n} := \log \pi_{\mathrm{old}}(A_{t,n} \mid S_{t,n}).$$

The old action-distribution parameters at each state must also be available for computing the KL divergence. They may either be stored in the rollout buffer or recomputed using a frozen copy of π_{old}.

Using the stored rewards, values, next states, and episode boundaries, compute the advantage estimates

$$\widehat{A}^{\pi_{\mathrm{old}}}_{t,n}.$$

Advantage estimation is performed separately along the time dimension of each environment. After computing the advantages, flatten the $T \times N$ rollout batch into

$$\boxed{\mathcal{B} = \left\{S_i,\ A_i,\ \widehat{A}^{\pi_{\mathrm{old}}}_i,\ \ell^{\mathrm{old}}_i\right\}_{i=1}^{B}.}$$

The batch $\mathcal{B}$ and all advantage estimates remain fixed throughout the policy update and backtracking line search.

5.2 Evaluating a Candidate Policy

For any candidate parameter vector θ, evaluate

$$\ell_i(\theta) := \log \pi_\theta(A_i \mid S_i)$$

on the stored state-action pairs.

The likelihood ratio is

$$\boxed{r_i(\theta) = \exp\left(\ell_i(\theta) - \ell_i^{\text{old}}\right) = \frac{\pi_\theta(A_i \mid S_i)}{\pi_{\text{old}}(A_i \mid S_i)}.}$$

The empirical surrogate objective is

$$\boxed{\widehat{\mathcal{L}}^{\text{sur}}_{\pi_{\text{old}}}(\theta) = \frac{1}{B}\sum_{i=1}^{B} r_i(\theta)\widehat{A}_i^{\pi_{\text{old}}}.}$$

For the same candidate, compute the empirical average KL divergence

$$\boxed{\widehat{\bar{D}_{\text{KL}}}\left(\pi_{\text{old}} \| \pi_\theta\right) = \frac{1}{B}\sum_{i=1}^{B} D_{\text{KL}}\left(\pi_{\text{old}}(\cdot \mid S_i) \| \pi_\theta(\cdot \mid S_i)\right).}$$

The old-policy surrogate value is

$$\widehat{\mathcal{L}}^{\text{sur}}_{\pi_{\text{old}}}(\theta_{\text{old}}) = \frac{1}{B}\sum_{i=1}^{B} \widehat{A}_i^{\pi_{\text{old}}}.$$

This sampled quantity need not be exactly zero. Therefore, the line search compares the candidate surrogate value with the old surrogate value:

$$\widehat{\mathcal{L}}^{\text{sur}}_{\pi_{\text{old}}}(\theta) > \widehat{\mathcal{L}}^{\text{sur}}_{\pi_{\text{old}}}(\theta_{\text{old}}).$$

Every backtracking candidate is evaluated using the same fixed batch $\mathcal{B}$. No new trajectories are collected during the line search.

5.3 TRPO Update on a Fixed Rollout Batch

Let $\mathcal{B}_{\text{old}}$ be a fixed rollout batch generated by π_{old}. The batch contains the sampled state-action pairs, old-policy quantities, and advantage estimates required to evaluate candidate policies.

For compactness, define

$$\widehat{L}(\theta) := \widehat{\mathcal{L}}^{\text{sur}}_{\pi_{\text{old}}}(\theta), \qquad \widehat{D}(\theta) := \widehat{\bar{D}_{\text{KL}}}\left(\pi_{\text{old}} \| \pi_\theta\right).$$

The batch and all old-policy quantities remain fixed throughout the following update.

Algorithm 1 TRPO Backtracking Update

Require: Fixed batch $\mathcal{B}_{\text{old}}$
Require: Old parameters θ_{old} and NPG proposal $\Delta\theta_{\text{NPG}}$
Require: KL radius δ, backtracking coefficient β, maximum steps J

$L_{\text{old}} \leftarrow \widehat{L}(\theta_{\text{old}})$
$\theta_{\text{new}} \leftarrow \theta_{\text{old}}$
for $j = 0, \ldots, J-1$ **do**
    $\theta_j \leftarrow \theta_{\text{old}} + \beta^j \Delta\theta_{\text{NPG}}$
    $L_j \leftarrow \widehat{L}(\theta_j)$
    $D_j \leftarrow \widehat{D}(\theta_j)$
    if $L_j > L_{\text{old}}$ **and** $D_j \leq \delta$ **then**
        $\theta_{\text{new}} \leftarrow \theta_j$
        break
    end if
end for
return θ_{new}

Appendix A Proof of the Performance-Difference Identity

Equivalent forms of the objective. For any policy π, the objective can be written in two equivalent forms:

$$\boxed{J(\pi) = \mathbb{E}_{S_0 \sim \rho_0}\left[V^{\pi}(S_0)\right] = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right].}$$

Use the trajectory form for the candidate policy and the value-function form for the old policy:

$$J(\pi_\theta) - J(\pi_{\text{old}}) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right] - \mathbb{E}_{S_0 \sim \rho_0}\left[V^{\pi_{\text{old}}}(S_0)\right].$$

Because the candidate trajectory also begins with $S_0 \sim \rho_0$,

$$\mathbb{E}_{S_0 \sim \rho_0}\left[V^{\pi_{\text{old}}}(S_0)\right] = \mathbb{E}_{\tau \sim \pi_\theta}\left[V^{\pi_{\text{old}}}(S_0)\right].$$

Therefore,

$$\boxed{J(\pi_\theta) - J(\pi_{\text{old}}) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} - V^{\pi_{\text{old}}}(S_0)\right].}$$

Telescoping TD-error identity. Along the candidate trajectory, define the TD error computed using the old-policy value function:

$$\boxed{\delta_t^{\pi_{\text{old}}} := R_{t+1} + \gamma V^{\pi_{\text{old}}}(S_{t+1}) - V^{\pi_{\text{old}}}(S_t).}$$

For a finite T,

$$\begin{aligned}
\sum_{t=0}^{T} \gamma^t \delta_t^{\pi_{\text{old}}} &= \sum_{t=0}^{T} \gamma^t \left[R_{t+1} + \gamma V^{\pi_{\text{old}}}(S_{t+1}) - V^{\pi_{\text{old}}}(S_t)\right] \\
&= \left[R_1 + \gamma V^{\pi_{\text{old}}}(S_1) - V^{\pi_{\text{old}}}(S_0)\right] \\
&\quad + \left[\gamma R_2 + \gamma^2 V^{\pi_{\text{old}}}(S_2) - \gamma V^{\pi_{\text{old}}}(S_1)\right] \\
&\quad + \left[\gamma^2 R_3 + \gamma^3 V^{\pi_{\text{old}}}(S_3) - \gamma^2 V^{\pi_{\text{old}}}(S_2)\right] \\
&\quad + \cdots \\
&\quad + \left[\gamma^T R_{T+1} + \gamma^{T+1} V^{\pi_{\text{old}}}(S_{T+1}) - \gamma^T V^{\pi_{\text{old}}}(S_T)\right].
\end{aligned}$$

The intermediate value terms cancel, leaving

$$\sum_{t=0}^{T} \gamma^t \delta_t^{\pi_{\text{old}}} = \sum_{t=0}^{T} \gamma^t R_{t+1} - V^{\pi_{\text{old}}}(S_0) + \gamma^{T+1} V^{\pi_{\text{old}}}(S_{T+1}).$$

For bounded rewards and $0 \le \gamma < 1$,

$$\lim_{T \to \infty} \gamma^{T+1} V^{\pi_{\text{old}}}(S_{T+1}) = 0.$$

Hence,

$$\boxed{\sum_{t=0}^{\infty} \gamma^t R_{t+1} - V^{\pi_{\text{old}}}(S_0) = \sum_{t=0}^{\infty} \gamma^t \delta_t^{\pi_{\text{old}}}.}$$

This equality holds for each trajectory before taking any expectation. Substituting it into the objective difference gives

$$\boxed{J(\pi_\theta) - J(\pi_{\text{old}}) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \delta_t^{\pi_{\text{old}}} \right].}$$

Conditional expectation of the TD error. Define the history before selecting A_t as

$$\mathcal{H}_t := (S_0, A_0, R_1, S_1, \ldots, A_{t-1}, R_t, S_t)\,.$$

For each t,

$$\begin{aligned} \mathbb{E}_{\tau \sim \pi_\theta} \left[\delta_t^{\pi_{\text{old}}} \right] &= \mathbb{E}_{\mathcal{H}_t, A_t, R_{t+1}, S_{t+1}} \left[\delta_t^{\pi_{\text{old}}} \right] \\ &= \mathbb{E}_{\mathcal{H}_t, A_t} \left[\mathbb{E} \left[\delta_t^{\pi_{\text{old}}} \mid \mathcal{H}_t, A_t \right] \right]. \end{aligned}$$

The second equality is the law of iterated expectation.

Because the environment is Markov,

$$(R_{t+1}, S_{t+1}) \mid \mathcal{H}_t, A_t \overset{d}{=} (R_{t+1}, S_{t+1}) \mid S_t, A_t.$$

Therefore,

$$\begin{aligned} \mathbb{E} \left[\delta_t^{\pi_{\text{old}}} \mid \mathcal{H}_t, A_t \right] &= \mathbb{E} \left[\delta_t^{\pi_{\text{old}}} \mid S_t, A_t \right] \\ &= \mathbb{E} \left[R_{t+1} + \gamma V^{\pi_{\text{old}}}(S_{t+1}) \mid S_t, A_t \right] - V^{\pi_{\text{old}}}(S_t) \\ &= Q^{\pi_{\text{old}}}(S_t, A_t) - V^{\pi_{\text{old}}}(S_t) \\ &= A^{\pi_{\text{old}}}(S_t, A_t). \end{aligned}$$

Substituting this inner conditional expectation into the outer expectation,

$$\begin{aligned} \mathbb{E}_{\tau \sim \pi_\theta} \left[\delta_t^{\pi_{\text{old}}} \right] &= \mathbb{E}_{\mathcal{H}_t, A_t} \left[A^{\pi_{\text{old}}}(S_t, A_t) \right] \\ &= \mathbb{E}_{\tau \sim \pi_\theta} \left[A^{\pi_{\text{old}}}(S_t, A_t) \right]. \end{aligned}$$

Thus,

$$\boxed{A^{\pi_{\text{old}}}(S_t, A_t) = \mathbb{E} \left[\delta_t^{\pi_{\text{old}}} \mid S_t, A_t \right].}$$

The TD error is a random one-step quantity, while the advantage is its conditional expectation given the current state and action.

Using this result,

$$\begin{aligned} J(\pi_\theta) - J(\pi_{\text{old}}) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\delta_t^{\pi_{\text{old}}}\right] \\ &= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{\tau\sim\pi_\theta}\left[\delta_t^{\pi_{\text{old}}}\right] \\ &= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{\tau\sim\pi_\theta}\left[A^{\pi_{\text{old}}}(S_t,A_t)\right] \\ &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_{\text{old}}}(S_t,A_t)\right]. \end{aligned}$$

Therefore, the trajectory form of the performance-difference identity is

$$\boxed{J(\pi_\theta) - J(\pi_{\text{old}}) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_{\text{old}}}(S_t,A_t)\right].}$$

Here, π_θ generates the trajectory, while $A^{\pi_{\text{old}}}$ evaluates the resulting state-action pairs relative to the old policy.

Discounted state-visitation form. Recall the normalized discounted state-visitation distribution

$$d^{\pi_\theta,\rho_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr_{\rho_0,\pi_\theta}(S_t = s).$$

Expanding the trajectory expectation,

$$\begin{aligned} &\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_{\text{old}}}(S_t,A_t)\right] \\ &= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{\tau\sim\pi_\theta}\left[A^{\pi_{\text{old}}}(S_t,A_t)\right] \\ &= \sum_{t=0}^{\infty}\gamma^t\sum_s \Pr_{\rho_0,\pi_\theta}(S_t=s)\sum_a \pi_\theta(a\mid s)A^{\pi_{\text{old}}}(s,a) \\ &= \sum_s\left[\sum_{t=0}^{\infty}\gamma^t \Pr_{\rho_0,\pi_\theta}(S_t=s)\right]\sum_a \pi_\theta(a\mid s)A^{\pi_{\text{old}}}(s,a) \\ &= \frac{1}{1-\gamma}\sum_s d^{\pi_\theta,\rho_0}(s)\sum_a \pi_\theta(a\mid s)A^{\pi_{\text{old}}}(s,a) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\left[A^{\pi_{\text{old}}}(S,A)\right]\right]. \end{aligned}$$

Combining this with the trajectory form gives

$$\boxed{J(\pi_\theta) - J(\pi_{\text{old}}) = \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\left[A^{\pi_{\text{old}}}(S,A)\right]\right].}$$

Equivalently,

$$\boxed{J(\pi_\theta) = J(\pi_{\text{old}}) + \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\left[A^{\pi_{\text{old}}}(S,A)\right]\right].}$$

For continuous state or action spaces, the corresponding sums are replaced by integrals.