Natural Policy Gradient

Reinforcement Learning

Sai Sampath Kedari
sampath@umich.edu

Contents

1 Local Models, Trust Regions, and Vanilla Policy Gradient

1.1 Trust-Region Viewpoint

Suppose the objective is

$$J(\theta).$$

Directly solving

$$\theta^\star = \arg\max_\theta J(\theta)$$

is generally difficult because J may be highly nonlinear and only accessible through interaction with the environment.

Trust-region optimization instead proceeds locally. At the current parameters θ_{old}, we construct a simpler surrogate model of the objective using local information available at θ_{old}.

Define the candidate displacement

$$\Delta\theta := \theta - \theta_{\text{old}}, \qquad \theta = \theta_{\text{old}} + \Delta\theta.$$

A local model

$$m_{\text{old}}(\Delta\theta) \approx J(\theta_{\text{old}} + \Delta\theta)$$

is expected to be accurate near $\Delta\theta = 0$, but there is no reason to trust the approximation arbitrarily far from the expansion point.

We therefore optimize the surrogate only inside a neighborhood around the current parameters:

$$\boxed{\begin{aligned} \Delta\theta^\star = \arg\max_{\Delta\theta} \quad & m_{\text{old}}(\Delta\theta) \\ \text{subject to} \quad & \text{size}(\Delta\theta) \le \delta. \end{aligned}}$$

The constraint defines the *trust region*: the region in which the local surrogate is trusted as an approximation to the true objective.

The particular choice of $\text{size}(\Delta\theta)$ determines the geometry of the trust region.

1.2 Euclidean Trust-Region Model

From the previous report, the policy objective is

$$J(\theta) = \mathbb{E}_{S_0\sim\rho_0}\left[v^{\pi_\theta}(S_0)\right],$$

with policy gradient

$$\nabla_\theta J(\theta) = \mathbb{E}_{\substack{S_t\sim d^{\pi_\theta,\rho_0} \\ A_t\sim\pi_\theta(\cdot|S_t)}}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t)\, A^{\pi_\theta}(S_t, A_t)\right].$$

At θ_{old}, define

$$g_{\text{old}} := \nabla_\theta J(\theta)|_{\theta=\theta_{\text{old}}}.$$

The first-order Taylor expansion is

$$J(\theta_{\text{old}} + \Delta\theta) = J(\theta_{\text{old}}) + g_{\text{old}}^\top \Delta\theta + o(\|\Delta\theta\|_2).$$

Dropping the higher-order term gives the first-order surrogate

$$\boxed{m_{\text{old}}(\Delta\theta) = J(\theta_{\text{old}}) + g_{\text{old}}^{\top}\Delta\theta.}$$

Since $J(\theta_{\text{old}})$ is constant with respect to $\Delta\theta$, maximizing the surrogate is equivalent to maximizing

$$g_{\text{old}}^{\top}\Delta\theta.$$

Using the Euclidean norm to define the trust region gives

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

For $g_{\text{old}} \neq 0$, the optimum lies on the boundary,

$$\frac{1}{2}\Delta\theta^{\star\top}\Delta\theta^{\star} = \delta.$$

1.3 Lagrangian Solution

Define

$$\mathcal{L}(\Delta\theta, \eta) = g_{\text{old}}^{\top}\Delta\theta - \eta\left(\frac{1}{2}\Delta\theta^{\top}\Delta\theta - \delta\right), \qquad \eta \ge 0.$$

Taking the total differential,

$$\begin{aligned} d\mathcal{L} &= d(g_{\text{old}}^{\top}\Delta\theta) - d\left[\eta\left(\frac{1}{2}\Delta\theta^{\top}\Delta\theta - \delta\right)\right] \\ &= g_{\text{old}}^{\top}d(\Delta\theta) - \left(\frac{1}{2}\Delta\theta^{\top}\Delta\theta - \delta\right)d\eta - \frac{\eta}{2}d(\Delta\theta^{\top}\Delta\theta) \\ &= g_{\text{old}}^{\top}d(\Delta\theta) - \left(\frac{1}{2}\Delta\theta^{\top}\Delta\theta - \delta\right)d\eta \\ &\quad - \frac{\eta}{2}\left[d(\Delta\theta)^{\top}\Delta\theta + \Delta\theta^{\top}d(\Delta\theta)\right] \\ &= \left(g_{\text{old}}^{\top} - \eta\Delta\theta^{\top}\right)d(\Delta\theta) + \left(\delta - \frac{1}{2}\Delta\theta^{\top}\Delta\theta\right)d\eta, \end{aligned}$$

where

$$d(\Delta\theta)^{\top}\Delta\theta = \Delta\theta^{\top}d(\Delta\theta),$$

since both sides are scalars.

Comparing with

$$d\mathcal{L} = \nabla_{\Delta\theta}\mathcal{L}^{\top}d(\Delta\theta) + \frac{\partial\mathcal{L}}{\partial\eta}\,d\eta,$$

gives

$$\nabla_{\Delta\theta}\mathcal{L} = g_{\text{old}} - \eta\Delta\theta, \qquad \frac{\partial\mathcal{L}}{\partial\eta} = \delta - \frac{1}{2}\Delta\theta^{\top}\Delta\theta.$$

Setting both equal to zero,

$$\Delta\theta^\star = \frac{1}{\eta} g_{\text{old}}, \qquad \frac{1}{2}\Delta\theta^{\star\top}\Delta\theta^\star = \delta.$$

Therefore,

$$\frac{1}{2\eta^2} g_{\text{old}}^\top g_{\text{old}} = \delta,$$

and hence

$$\eta^\star = \sqrt{\frac{g_{\text{old}}^\top g_{\text{old}}}{2\delta}}.$$

Substituting back,

$$\boxed{\Delta\theta^\star = \sqrt{\frac{2\delta}{g_{\text{old}}^\top g_{\text{old}}}} g_{\text{old}} = \sqrt{2\delta}\, \frac{g_{\text{old}}}{\|g_{\text{old}}\|_2}.}$$

Thus,

$$\frac{g_{\text{old}}}{\|g_{\text{old}}\|_2}$$

is the normalized Euclidean steepest-ascent direction, while

$$g_{\text{old}}$$

is the corresponding unnormalized direction.

1.4 Vanilla Policy-Gradient Update

Rather than fixing the Euclidean trust-region radius explicitly, ordinary gradient ascent controls the step magnitude with a learning rate $\alpha > 0$:

$$\Delta\theta = \alpha g_{\text{old}}.$$

Therefore,

$$\boxed{\theta_{\text{new}} = \theta_{\text{old}} + \alpha g_{\text{old}}.}$$

Using the sample policy-gradient estimate

$$\widehat{g}_{\text{old}} = \widehat{\mathbb{E}}_t \left[\nabla_\theta \log \pi_\theta(A_t \mid S_t)|_{\theta=\theta_{\text{old}}} \widehat{A}_t^{\text{GAE}(\gamma,\lambda)} \right],$$

the practical vanilla policy-gradient update is

$$\boxed{\theta_{\text{new}} = \theta_{\text{old}} + \alpha \widehat{g}_{\text{old}}.}$$

1.5 Why Euclidean Geometry Is Insufficient

The trust-region idea itself is not the problem. The questionable choice is the Euclidean measure

$$\frac{1}{2}\Delta\theta^\top \Delta\theta.$$

It controls how far the *parameters* move, but the object being optimized is the policy distribution

$$\pi_\theta(\cdot \mid s).$$

Equal Euclidean parameter displacements can induce very different changes in the corresponding action distributions. Therefore, Euclidean parameter distance does not provide a reliable measure of how much the policy itself has changed.

Natural policy gradient keeps the same local-model and trust-region viewpoint, but replaces Euclidean parameter-space geometry with a geometry defined directly by changes in the policy distribution.

2 Policy-Space Geometry and the KL Constraint

2.1 KL-Constrained Policy Update

At the current parameters θ_{old}, consider

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

Define the average KL divergence

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

For a fixed state S,

$$D_{\text{KL}}\left(\pi_{\text{old}}(\cdot \mid S) \| \pi_\theta(\cdot \mid S)\right) = \mathbb{E}_{A\sim\pi_{\text{old}}(\cdot|S)}\left[\log \frac{\pi_{\text{old}}(A \mid S)}{\pi_\theta(A \mid S)}\right].$$

The policy-space trust-region problem is

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

2.2 First-Order Approximation of the Objective

Using the first-order model already derived,

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

Since $J(\theta_{\text{old}})$ is constant,

$$\boxed{\begin{aligned} \Delta\theta^\star \approx \arg\max_{\Delta\theta} \quad & g_{\text{old}}^\top \Delta\theta \\ \text{subject to} \quad & \bar{D}_{\text{KL}}\left(\pi_{\text{old}} \| \pi_{\theta_{\text{old}}+\Delta\theta}\right) \le \delta. \end{aligned}}$$

2.3 Second-Order Approximation of the KL Constraint

For compactness, define

$$\bar{D}_{\mathrm{KL}}(\theta) := \bar{D}_{\mathrm{KL}}\left(\pi_{\mathrm{old}} \| \pi_\theta\right).$$

Expanding around θ_{old},

$$\begin{aligned}\bar{D}_{\mathrm{KL}}(\theta_{\mathrm{old}} + \Delta\theta) \approx\ & \underbrace{\bar{D}_{\mathrm{KL}}(\theta_{\mathrm{old}})}_{\text{0th-order term}} \\ & + \underbrace{\nabla_\theta \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}}^{\top} \Delta\theta}_{\text{1st-order term}} \\ & + \underbrace{\frac{1}{2}\Delta\theta^\top \nabla_\theta^2 \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}} \Delta\theta}_{\text{2nd-order term}}.\end{aligned}$$

0th-order term. At the expansion point, $\theta = \theta_{\mathrm{old}}$, so the candidate policy π_θ becomes the old policy π_{old}. Therefore,

$$\begin{aligned}\bar{D}_{\mathrm{KL}}(\theta_{\mathrm{old}}) &= \bar{D}_{\mathrm{KL}}\left(\pi_{\mathrm{old}} \| \pi_{\mathrm{old}}\right) \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}}}, \rho_0}\left[D_{\mathrm{KL}}\left(\pi_{\mathrm{old}}(\cdot \mid S) \| \pi_{\mathrm{old}}(\cdot \mid S)\right)\right] \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}}}, \rho_0}\left[0\right] \\ &= 0.\end{aligned}$$

Hence, the zeroth-order term in the Taylor expansion is

$$\boxed{\bar{D}_{\mathrm{KL}}(\theta_{\mathrm{old}}) = 0.}$$

1st-order term. The gradient of the average KL divergence at the expansion point is

$$\begin{aligned}\nabla_\theta \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}} &= \nabla_\theta \bar{D}_{\mathrm{KL}}\left(\pi_{\mathrm{old}} \| \pi_\theta\right)\big|_{\theta=\theta_{\mathrm{old}}} \\ &= \nabla_\theta \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}}}, \rho_0}\left[D_{\mathrm{KL}}\left(\pi_{\mathrm{old}}(\cdot \mid S) \| \pi_\theta(\cdot \mid S)\right)\right]\big|_{\theta=\theta_{\mathrm{old}}} \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}}}, \rho_0}\left[\nabla_\theta D_{\mathrm{KL}}\left(\pi_{\mathrm{old}}(\cdot \mid S) \| \pi_\theta(\cdot \mid S)\right)|_{\theta=\theta_{\mathrm{old}}}\right] \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}}}, \rho_0}\left[\nabla_\theta \mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}\left[\log \frac{\pi_{\mathrm{old}}(A \mid S)}{\pi_\theta(A \mid S)}\right]\Big|_{\theta=\theta_{\mathrm{old}}}\right] \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}}}, \rho_0}\left[\mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}\left[\nabla_\theta \left(\log \pi_{\mathrm{old}}(A \mid S) - \log \pi_\theta(A \mid S)\right)|_{\theta=\theta_{\mathrm{old}}}\right]\right] \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}}}, \rho_0}\left[\mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}\left[\underbrace{\nabla_\theta \log \pi_{\mathrm{old}}(A \mid S)}_{0} - \nabla_\theta \log \pi_\theta(A \mid S)|_{\theta=\theta_{\mathrm{old}}}\right]\right] \\ &= -\mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}}}, \rho_0}\left[\mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}\left[\nabla_\theta \log \pi_\theta(A \mid S)|_{\theta=\theta_{\mathrm{old}}}\right]\right].\end{aligned}$$

By the zero-mean score-function identity,

$$\mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}\left[\nabla_\theta \log \pi_\theta(A \mid S)|_{\theta=\theta_{\mathrm{old}}}\right] = 0.$$

Therefore,

$$\begin{aligned}\nabla_\theta \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}} &= -\mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0}}[0] \\ &= 0.\end{aligned}$$

Hence,

$$\boxed{\nabla_\theta \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}} = 0.}$$

Consequently, the complete first-order Taylor term vanishes:

$$\boxed{\nabla_\theta \bar{D}_{\mathrm{KL}}(\theta)^\top\Big|_{\theta=\theta_{\mathrm{old}}} \Delta\theta = 0.}$$

The zero-mean score identity also has a maximum-likelihood interpretation. For each fixed state S,

$$\mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}[\log \pi_\theta(A \mid S)]$$

is the population expected log-likelihood under actions generated by π_{old}. It is maximized when the candidate distribution matches the data-generating distribution, which occurs at $\theta = \theta_{\mathrm{old}}$. Therefore, its gradient at θ_{old} is zero.

2nd-order term. The Hessian of the average KL divergence at the expansion point is

$$\begin{aligned}\nabla^2_\theta \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}} &= \nabla^2_\theta \bar{D}_{\mathrm{KL}}(\pi_{\mathrm{old}}\|\pi_\theta)\big|_{\theta=\theta_{\mathrm{old}}} \\ &= \nabla^2_\theta \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0}}[D_{\mathrm{KL}}(\pi_{\mathrm{old}}(\cdot \mid S)\|\pi_\theta(\cdot \mid S))]\big|_{\theta=\theta_{\mathrm{old}}} \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0}}\left[\nabla^2_\theta D_{\mathrm{KL}}(\pi_{\mathrm{old}}(\cdot \mid S)\|\pi_\theta(\cdot \mid S))\big|_{\theta=\theta_{\mathrm{old}}}\right] \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0}}\left[\nabla^2_\theta \mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}\left[\log\frac{\pi_{\mathrm{old}}(A \mid S)}{\pi_\theta(A \mid S)}\right]\Big|_{\theta=\theta_{\mathrm{old}}}\right] \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0}}\left[\mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}\left[\nabla^2_\theta(\log\pi_{\mathrm{old}}(A \mid S) - \log\pi_\theta(A \mid S))\big|_{\theta=\theta_{\mathrm{old}}}\right]\right] \\ &= \mathbb{E}_{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0}}\left[\mathbb{E}_{A\sim\pi_{\mathrm{old}}(\cdot|S)}\left[\underbrace{\nabla^2_\theta \log\pi_{\mathrm{old}}(A \mid S)}_{0} - \nabla^2_\theta \log\pi_\theta(A \mid S)\big|_{\theta=\theta_{\mathrm{old}}}\right]\right] \\ &= -\mathbb{E}_{\substack{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0} \\ A\sim\pi_{\mathrm{old}}(\cdot|S)}}\left[\nabla^2_\theta \log\pi_\theta(A \mid S)\big|_{\theta=\theta_{\mathrm{old}}}\right].\end{aligned}$$

By the Fisher information identity,

$$\begin{aligned}&-\mathbb{E}_{\substack{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0} \\ A\sim\pi_{\mathrm{old}}(\cdot|S)}}\left[\nabla^2_\theta \log\pi_\theta(A \mid S)\big|_{\theta=\theta_{\mathrm{old}}}\right] \\ &\quad= \mathbb{E}_{\substack{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0} \\ A\sim\pi_{\mathrm{old}}(\cdot|S)}}\left[\nabla_\theta \log\pi_\theta(A \mid S)|_{\theta=\theta_{\mathrm{old}}}\ \nabla_\theta \log\pi_\theta(A \mid S)|^\top_{\theta=\theta_{\mathrm{old}}}\right].\end{aligned}$$

Therefore, define the Fisher information matrix at the old policy as

$$\boxed{\begin{aligned}F_{\mathrm{old}} &:= \nabla^2_\theta \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}} \\ &= \mathbb{E}_{\substack{S\sim d^{\pi_{\theta_{\mathrm{old}}},\rho_0} \\ A\sim\pi_{\mathrm{old}}(\cdot|S)}}\left[\nabla_\theta \log\pi_\theta(A \mid S)|_{\theta=\theta_{\mathrm{old}}}\ \nabla_\theta \log\pi_\theta(A \mid S)|^\top_{\theta=\theta_{\mathrm{old}}}\right].\end{aligned}}$$

Substituting the 0th, 1st, and 2nd order terms into the Taylor expansion gives the following.

$$\begin{aligned}\bar{D}_{\mathrm{KL}}\left(\pi_{\mathrm{old}} \| \pi_{\theta_{\mathrm{old}}+\Delta\theta}\right) &\approx \underbrace{\bar{D}_{\mathrm{KL}}(\theta_{\mathrm{old}})}_{0} + \underbrace{\nabla_\theta \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}}^{\top}}_{0} \Delta\theta + \frac{1}{2}\Delta\theta^\top \underbrace{\nabla_\theta^2 \bar{D}_{\mathrm{KL}}(\theta)\big|_{\theta=\theta_{\mathrm{old}}}}_{F_{\mathrm{old}}} \Delta\theta \\ &= \frac{1}{2}\Delta\theta^\top F_{\mathrm{old}} \Delta\theta.\end{aligned}$$

Hence, the second-order local approximation of the average KL divergence is

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

2.4 Local Natural Policy-Gradient Problem

The problem we would like to solve is

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

$\Downarrow$ construct a local trust region in policy-distribution space

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

$\Downarrow$ first-order model of J and second-order model of $\bar{D}_{\mathrm{KL}}$

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

$\Downarrow$ $J(\theta_{\mathrm{old}})$ is constant with respect to $\Delta\theta$

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

This local problem has the same structure as the Euclidean trust-region problem of Section 1. The matrix defining the local trust-region geometry changes from the identity matrix to the Fisher information matrix:

$$\boxed{I \quad \longrightarrow \quad F_{\mathrm{old}}.}$$

3 Solving the Natural Policy-Gradient Problem

3.1 Lagrangian Solution

For $g_{\text{old}} \neq 0$, the constraint is active:

$$\frac{1}{2}\Delta\theta^{\star\top} F_{\text{old}} \Delta\theta^{\star} = \delta.$$

Define

$$\mathcal{L}(\Delta\theta, \eta) = g_{\text{old}}^{\top}\Delta\theta - \eta\left(\frac{1}{2}\Delta\theta^{\top} F_{\text{old}}\Delta\theta - \delta\right), \qquad \eta \geq 0.$$

Taking the total differential,

$$\begin{aligned}
d\mathcal{L} &= d(g_{\text{old}}^{\top}\Delta\theta) - d\left[\eta\left(\frac{1}{2}\Delta\theta^{\top} F_{\text{old}}\Delta\theta - \delta\right)\right] \\
&= g_{\text{old}}^{\top} d(\Delta\theta) - \left(\frac{1}{2}\Delta\theta^{\top} F_{\text{old}}\Delta\theta - \delta\right) d\eta - \frac{\eta}{2} d\left(\Delta\theta^{\top} F_{\text{old}}\Delta\theta\right) \\
&= g_{\text{old}}^{\top} d(\Delta\theta) - \left(\frac{1}{2}\Delta\theta^{\top} F_{\text{old}}\Delta\theta - \delta\right) d\eta \\
&\quad - \frac{\eta}{2}\left[d(\Delta\theta)^{\top} F_{\text{old}}\Delta\theta + \Delta\theta^{\top} F_{\text{old}} d(\Delta\theta)\right] \\
&= \left(g_{\text{old}}^{\top} - \eta\Delta\theta^{\top} F_{\text{old}}\right) d(\Delta\theta) + \left(\delta - \frac{1}{2}\Delta\theta^{\top} F_{\text{old}}\Delta\theta\right) d\eta,
\end{aligned}$$

where $F_{\text{old}}^{\top} = F_{\text{old}}$.

Comparing with

$$d\mathcal{L} = \nabla_{\Delta\theta}\mathcal{L}^{\top} d(\Delta\theta) + \frac{\partial\mathcal{L}}{\partial\eta}\, d\eta,$$

gives

$$\nabla_{\Delta\theta}\mathcal{L} = g_{\text{old}} - \eta F_{\text{old}}\Delta\theta,$$

and

$$\frac{\partial\mathcal{L}}{\partial\eta} = \delta - \frac{1}{2}\Delta\theta^{\top} F_{\text{old}}\Delta\theta.$$

Setting both equal to zero,

$$g_{\text{old}} - \eta F_{\text{old}}\Delta\theta^{\star} = 0,$$

so

$$\boxed{\Delta\theta^{\star} = \frac{1}{\eta} F_{\text{old}}^{-1} g_{\text{old}}.}$$

Hence,

$$\boxed{\widetilde{g}_{\text{old}} := F_{\text{old}}^{-1} g_{\text{old}}}$$

is the unnormalized natural policy-gradient direction.

Substituting into the active constraint,

$$\frac{1}{2\eta^2} g_{\text{old}}^{\top} F_{\text{old}}^{-1} g_{\text{old}} = \delta,$$

which gives

$$\eta^\star = \sqrt{\frac{g_{\text{old}}^\top F_{\text{old}}^{-1} g_{\text{old}}}{2\delta}}.$$

Therefore,

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

This is the normalized natural policy-gradient step.

3.2 Natural Policy-Gradient Updates

With an arbitrary learning rate $\alpha > 0$,

$$\boxed{\theta_{\text{new}} = \theta_{\text{old}} + \alpha F_{\text{old}}^{-1} g_{\text{old}}.}$$

Using sample estimates,

$$\boxed{\theta_{\text{new}} = \theta_{\text{old}} + \alpha \widehat{F}_{\text{old}}^{-1} \widehat{g}_{\text{old}}.}$$

If the KL trust-region radius directly determines the step magnitude,

$$\boxed{\theta_{\text{new}} = \theta_{\text{old}} + \sqrt{\frac{2\delta}{\widehat{g}_{\text{old}}^\top \widehat{F}_{\text{old}}^{-1} \widehat{g}_{\text{old}}}} \widehat{F}_{\text{old}}^{-1} \widehat{g}_{\text{old}}.}$$

4 Practical Computation of the NPG Update

Consider a fixed rollout batch generated by π_{old}:

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

The empirical policy-gradient estimate is

$$\boxed{\widehat{g}_{\text{old}} = \frac{1}{B} \sum_{i=1}^{B} \nabla_\theta \log \pi_\theta(A_i \mid S_i)|_{\theta=\theta_{\text{old}}} \widehat{A}_i^{\pi_{\text{old}}}.}$$

The empirical average KL divergence is

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

Its Hessian at the old policy defines the empirical Fisher matrix:

$$\boxed{\widehat{F}_{\text{old}} = \nabla_\theta^2 \widehat{\bar{D}_{\text{KL}}}(\theta)\Big|_{\theta=\theta_{\text{old}}}.}$$

4.1 Damped Natural-Gradient Direction

The empirical Fisher matrix may be singular or poorly conditioned. Define

$$\boxed{\widetilde{F}_{\text{old}} = \widehat{F}_{\text{old}} + \lambda_{\text{damp}} I, \qquad \lambda_{\text{damp}} > 0.}$$

Rather than explicitly computing $\widetilde{F}_{\text{old}}^{-1}$, solve

$$\boxed{\widetilde{F}_{\text{old}} x = \widehat{g}_{\text{old}}.}$$

The solution

$$x = \widetilde{F}_{\text{old}}^{-1} \widehat{g}_{\text{old}}$$

is the damped natural-gradient direction.

4.2 Fisher-Vector Products and Conjugate Gradient

Constructing $\widehat{F}_{\text{old}}$ explicitly is impractical for a neural-network policy. Conjugate gradient avoids this by requiring only products of the form $\widehat{F}_{\text{old}} v$.

The Fisher-vector product is

$$\boxed{\widehat{F}_{\text{old}} v = \nabla_\theta \left[\left(\nabla_\theta \widehat{\bar{D}_{\text{KL}}}(\theta) \right)^\top v \right] \Big|_{\theta = \theta_{\text{old}}}.}$$

Therefore, the damped matrix-vector product used by conjugate gradient is

$$\boxed{\widetilde{F}_{\text{old}} v = \widehat{F}_{\text{old}} v + \lambda_{\text{damp}} v.}$$

After K conjugate-gradient iterations,

$$\boxed{x_K \approx \widetilde{F}_{\text{old}}^{-1} \widehat{g}_{\text{old}}.}$$

No Fisher matrix is constructed or inverted explicitly.

4.3 Choosing the Step Length

For fixed-learning-rate NPG, the update is

$$\boxed{\theta_{\text{new}} = \theta_{\text{old}} + \alpha x_K.}$$

For KL-normalized NPG, write

$$\Delta\theta_{\text{NPG}} = \alpha_\delta x_K.$$

The original quadratic KL model is

$$\frac{1}{2} \Delta\theta_{\text{NPG}}^\top \widehat{F}_{\text{old}} \Delta\theta_{\text{NPG}} \le \delta.$$

Placing the proposed step on its boundary gives

$$\frac{1}{2}\alpha_\delta^2 x_K^\top \widehat{F}_{\text{old}} x_K = \delta.$$

Therefore,

$$\boxed{\alpha_\delta = \sqrt{\frac{2\delta}{x_K^\top \widehat{F}_{\text{old}} x_K}}.}$$

The normalized NPG update is consequently

$$\boxed{\theta_{\text{new}} = \theta_{\text{old}} + \sqrt{\frac{2\delta}{x_K^\top \widehat{F}_{\text{old}} x_K}}\, x_K.}$$

Here, damping regularizes the direction computation, while the final scaling uses $\widehat{F}_{\text{old}}$ because $\widehat{F}_{\text{old}}$ defines the original quadratic KL model.

4.4 Practical NPG Algorithm

The following algorithm computes the KL-normalized NPG update on a fixed old-policy rollout batch.

Algorithm 1 KL-Normalized Natural Policy-Gradient Update

Require: Fixed rollout batch $\mathcal{B}_{\text{old}}$
Require: Old parameters θ_{old}
Require: KL radius δ
Require: Damping coefficient λ_{damp}
Require: Number of conjugate-gradient iterations K

Compute the empirical policy gradient $\widehat{g}_{\text{old}}$
Define $\text{FVP}(v) \leftarrow \widehat{F}_{\text{old}} v$
Define $\text{DampedFVP}(v) \leftarrow \text{FVP}(v) + \lambda_{\text{damp}} v$
$x_K \leftarrow \textsc{ConjugateGradient}(\text{DampedFVP}, \widehat{g}_{\text{old}}, K)$
$q_K \leftarrow x_K^\top \text{FVP}(x_K)$
$\alpha_\delta \leftarrow \sqrt{\dfrac{2\delta}{q_K}}$
$\theta_{\text{new}} \leftarrow \theta_{\text{old}} + \alpha_\delta x_K$
return θ_{new}

The conjugate-gradient solve produces

$$x_K \approx \left(\widehat{F}_{\text{old}} + \lambda_{\text{damp}} I\right)^{-1} \widehat{g}_{\text{old}},$$

while

$$q_K = x_K^\top \widehat{F}_{\text{old}} x_K$$

is computed using one additional undamped Fisher-vector product.

For fixed-learning-rate NPG, the trust-region scaling steps are omitted and the final update is simply

$$\boxed{\theta_{\text{new}} = \theta_{\text{old}} + \alpha x_K.}$$