Natural Policy Gradient

Sham Kakade and Kianté Brantley

CS 6789: Foundations of Reinforcement Learning

Today:

Natural policy optimization

History:

A Natural Policy Gradient

Sham Kakade
Gatsby Computational Neuroscience Unit
17 Queen Square, London, UK WC1N 3AR
http://www.gatsby.ucl.ac.uk
sham@gatsby.ucl.ac.uk

NeurIPS 2002

Covariant Policy Search

J. Andrew Bagnell and Jeff Schneider
Robotics Institute
Carnegie-Mellon University
Pittsburgh, PA 15213
{*dbagnell,schneide*}*@ri.cmu.edu*

IJCAI 2003

Trust Region Policy Optimization

John Schulman JOSCHU@EECS.BERKELEY.EDU
Sergey Levine SLEVINE@EECS.BERKELEY.EDU
Philipp Moritz PCMORITZ@EECS.BERKELEY.EDU
Michael Jordan JORDAN@CS.BERKELEY.EDU
Pieter Abbeel PABBEEL@CS.BERKELEY.EDU
University of California, Berkeley, Department of Electrical Engineering and Computer Sciences

ICML 2015

Notations and Settings:

Finite horizon setting: $\mathscr{M} = \{S, A, H, r, P, \rho\}$

Notations and Settings:

Finite horizon setting: $\mathscr{M} = \{S, A, H, r, P, \rho\}$

Average state-action distribution:

$$d^{\pi}(s,a) = \frac{1}{H}\sum_{h=0}^{H-1} \mathbb{P}_h^{\pi}(s,a)$$

Notations and Settings:

Finite horizon setting: $\mathscr{M} = \{S, A, H, r, P, \rho\}$

Average state-action distribution:

$$d^{\pi}(s,a) = \frac{1}{H}\sum_{h=0}^{H-1} \mathbb{P}_h^{\pi}(s,a)$$

Policy class:

$$\Pi = \{\pi : S \mapsto A\} \subset S \mapsto A$$

$$\pi^{\star} = \arg\max_{\pi \in \Pi} V^{\pi}(\rho)$$

Notations and Settings:

Finite horizon setting: $\mathscr{M} = \{S, A, H, r, P, \rho\}$

Average state-action distribution:

$$d^{\pi}(s,a) = \frac{1}{H}\sum_{h=0}^{H-1} \mathbb{P}_h^{\pi}(s,a)$$

Policy class:

$$\Pi = \{\pi : S \mapsto A\} \subset S \mapsto A$$

$$\pi^{\star} = \arg\max_{\pi \in \Pi} V^{\pi}(\rho)$$

Trajectory distribution:

$$\Pr^{\pi}(\tau) = \rho(s_0)\pi(a_0 \,|\, s_0)P(s_1 \,|\, s_0, a_0)\pi(a_1 \,|\, s_1)\ldots P(s_{H-1} \,|\, s_{H-2}, a_{H-2})\pi(a_{H-1} \,|\, s_{H-1})$$

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

In other words:

$$\min_\theta \nabla \ell(\theta_0)^\top (\theta - \theta_0), \text{ subject to } \|\theta - \theta_0\|_2^2 \leq \delta,$$

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

In other words:

$$\min_\theta \nabla \ell(\theta_0)^\top (\theta - \theta_0), \text{ subject to } \|\theta - \theta_0\|_2^2 \leq \delta,$$

We in default are using Euclidean distance in the parameter θ space

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

In other words:

$$\min_\theta \nabla \ell(\theta_0)^\top (\theta - \theta_0), \text{ subject to } \|\theta - \theta_0\|_2^2 \leq \delta,$$

We in default are using Euclidean distance in the parameter θ space

Different re-parameterization (scaling & translation) can lead to a quite different GD path

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

by the chain rule,

$$\nabla_\phi \ell = A^\top \nabla_\theta \ell \, .$$

In other words:

$$\min_\theta \nabla \ell(\theta_0)^\top (\theta - \theta_0), \text{ subject to } \|\theta - \theta_0\|_2^2 \leq \delta,$$

$\theta = A\phi + b$ $\quad$ ϕ – coordinates

$\phi_{t+1} = \phi_t - \eta \nabla_\phi \ell$ $\quad$ gradient descent in the ϕ-coordinates

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

by the chain rule,

$$\nabla_\phi \ell = A^\top \nabla_\theta \ell \,.$$

In other words:

$$\min_\theta \nabla \ell(\theta_0)^\top (\theta - \theta_0), \text{ subject to } \|\theta - \theta_0\|_2^2 \le \delta,$$

$\theta = A\phi + b$ ϕ – coordinates

$\phi_{t+1} = \phi_t - \eta \nabla_\phi \ell$ gradient descent in the ϕ-coordinates

$$\theta_{t+1} = A\left(\phi_t - \eta \nabla_\phi \ell\right) + b \,.$$

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

by the chain rule,

$$\nabla_\phi \ell = A^\top \nabla_\theta \ell \, .$$

In other words:

$$\min_\theta \nabla \ell(\theta_0)^\top (\theta - \theta_0), \text{ subject to } \|\theta - \theta_0\|_2^2 \leq \delta,$$

$\theta = A\phi + b$ $\quad$ ϕ – coordinates

$\phi_{t+1} = \phi_t - \eta \nabla_\phi \ell$ $\quad$ gradient descent in the ϕ-coordinates

$$\theta_{t+1} = A\left(\phi_t - \eta \nabla_\phi \ell\right) + b \, .$$

$$\theta_{t+1} = (A\phi_t + b) - \eta A \nabla_\phi \ell \, .$$

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

by the chain rule,

$$\nabla_\phi \ell = A^\top \nabla_\theta \ell \, .$$

In other words:

$$\min_\theta \nabla \ell(\theta_0)^\top (\theta - \theta_0), \text{ subject to } \|\theta - \theta_0\|_2^2 \leq \delta,$$

$\theta = A\phi + b$ $\quad$ ϕ – coordinates

$\phi_{t+1} = \phi_t - \eta \nabla_\phi \ell$ $\quad$ gradient descent in the ϕ-coordinates

$$\theta_{t+1} = A\left(\phi_t - \eta \nabla_\phi \ell\right) + b \, .$$

$$\theta_{t+1} = (A\phi_t + b) - \eta A \nabla_\phi \ell \, .$$

$$\theta_{t+1} = \theta_t - \eta A \nabla_\phi \ell \, .$$

Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_\theta \ell(\theta_0)$$

by the chain rule,

$$\nabla_\phi \ell = A^\top \nabla_\theta \ell \, .$$

In other words:

$$\min_\theta \nabla \ell(\theta_0)^\top (\theta - \theta_0), \text{ subject to } \|\theta - \theta_0\|_2^2 \le \delta,$$

$\theta = A\phi + b$ $\quad$ ϕ – coordinates

$\phi_{t+1} = \phi_t - \eta \nabla_\phi \ell$ $\quad$ gradient descent in the ϕ-coordinates

$$\theta_{t+1} = A\left(\phi_t - \eta \nabla_\phi \ell\right) + b \, .$$

$$\theta_{t+1} = (A\phi_t + b) - \eta A \nabla_\phi \ell \, .$$

$$\theta_{t+1} = \theta_t - \eta A \nabla_\phi \ell \, .$$

$$\theta_{t+1} = \theta_t - \eta A (A^\top \nabla_\theta \ell) \, .$$

Policy Optimization:

$$\max_{\pi_\theta} V^{\pi_\theta}(\rho)$$

$$\text{s.t., } KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta$$

Policy Optimization:

$$\max_{\pi_\theta} V^{\pi_\theta}(\rho)$$

$$\text{s.t., } KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta$$

Sequential convex programming:

We linearize the objective function & quadratize the KL constraint

Policy Optimization:

$$\max_{\pi_\theta} V^{\pi_\theta}(\rho)$$

$$\text{s.t., } KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta$$

Sequential convex programming:
We linearize the objective function & quadratize the KL constraint

We know the first order Taylor expansion of $V^{\pi_\theta}(\rho)$

$$V^{\pi_{\theta_0}}(\rho) + \nabla V^{\pi_{\theta_0}}(\rho)^\top \left(\theta - \theta_0\right)$$

Policy Optimization:

$$\max_{\pi_\theta} V^{\pi_\theta}(\rho)$$

$$\text{s.t., } KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta$$

Sequential convex programming:
We linearize the objective function & quadratize the KL constraint

We know the first order Taylor expansion of $V^{\pi_\theta}(\rho)$

$$V^{\pi_{\theta_0}}(\rho) + \nabla V^{\pi_{\theta_0}}(\rho)^\top \left(\theta - \theta_0\right)$$

Q: How to do second-order Taylor expansion on the KL constraint?

Let’s do second order Taylor Expansion on the KL-divergence

Let's do second order Taylor Expansion on the KL-divergence

$$\frac{1}{H} KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) = \frac{1}{H} \sum_{\tau} \Pr^{\theta_0}(\tau) \ln \frac{\Pr^{\theta_0}(\tau)}{\Pr^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \Pr^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h \mid s_h)}{\pi_{\theta}(a_h \mid s_h)}$$

Let's do second order Taylor Expansion on the KL-divergence

$$\frac{1}{H} KL\left(\mathsf{Pr}^{\pi_{\theta_0}} || \mathsf{Pr}^{\pi_\theta}\right) = \frac{1}{H} \sum_{\tau} \mathsf{Pr}^{\theta_0}(\tau) \ln \frac{\mathsf{Pr}^{\theta_0}(\tau)}{\mathsf{Pr}^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \mathsf{Pr}^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h \mid s_h)}{\pi_{\theta}(a_h \mid s_h)}$$

$$= \mathbb{E}_{s_h, a_h \sim d^{\pi_{\theta_0}}} \left[\ln \frac{\pi_{\theta_0}(a_h \mid s_h)}{\pi_{\theta}(a_h \mid s_h)} \right] := \ell(\theta)$$

Let's do second order Taylor Expansion on the KL-divergence

$$\frac{1}{H} KL\left(\mathsf{Pr}^{\pi_{\theta_0}} || \mathsf{Pr}^{\pi_\theta}\right) = \frac{1}{H} \sum_{\tau} \mathsf{Pr}^{\theta_0}(\tau) \ln \frac{\mathsf{Pr}^{\theta_0}(\tau)}{\mathsf{Pr}^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \mathsf{Pr}^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h \mid s_h)}{\pi_{\theta}(a_h \mid s_h)}$$

$$= \mathbb{E}_{s_h, a_h \sim d^{\pi_{\theta_0}}} \left[\ln \frac{\pi_{\theta_0}(a_h \mid s_h)}{\pi_{\theta}(a_h \mid s_h)} \right] := \ell(\theta) \qquad \ell(\theta_0) = 0$$

Let's do second order Taylor Expansion on the KL-divergence

$$\frac{1}{H} KL\left(\mathsf{Pr}^{\pi_{\theta_0}} || \mathsf{Pr}^{\pi_\theta}\right) = \frac{1}{H} \sum_{\tau} \mathsf{Pr}^{\theta_0}(\tau) \ln \frac{\mathsf{Pr}^{\theta_0}(\tau)}{\mathsf{Pr}^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \mathsf{Pr}^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h \,|\, s_h)}{\pi_{\theta}(a_h \,|\, s_h)}$$

$$= \mathbb{E}_{s_h, a_h \sim d^{\pi_{\theta_0}}} \left[\ln \frac{\pi_{\theta_0}(a_h \,|\, s_h)}{\pi_{\theta}(a_h \,|\, s_h)} \right] := \ell(\theta) \qquad \ell(\theta_0) = 0$$

$$\nabla_\theta \ell(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_{a} \pi_{\theta_0}(a \,|\, s) \left(- \nabla_\theta \ln \pi_\theta(a_h \,|\, s_h)\,|_{\theta=\theta_0} \right)$$

Let's do second order Taylor Expansion on the KL-divergence

$$\frac{1}{H} KL\left(\mathsf{Pr}^{\pi_{\theta_0}} || \mathsf{Pr}^{\pi_\theta}\right) = \frac{1}{H} \sum_{\tau} \mathsf{Pr}^{\theta_0}(\tau) \ln \frac{\mathsf{Pr}^{\theta_0}(\tau)}{\mathsf{Pr}^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \mathsf{Pr}^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h \,|\, s_h)}{\pi_\theta(a_h \,|\, s_h)}$$

$$= \mathbb{E}_{s_h, a_h \sim d^{\pi_{\theta_0}}} \left[\ln \frac{\pi_{\theta_0}(a_h \,|\, s_h)}{\pi_\theta(a_h \,|\, s_h)} \right] := \ell(\theta) \qquad \ell(\theta_0) = 0$$

$$\nabla_\theta \ell(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a \,|\, s) \left(- \nabla_\theta \ln \pi_\theta(a_h \,|\, s_h)\,|_{\theta=\theta_0} \right)$$

$$= - \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a \,|\, s) \frac{\nabla_\theta \pi_{\theta_0}(a \,|\, s)}{\pi_{\theta_0}(a \,|\, s)}$$

Let’s compute the Hessian of the KL-divergence

Let’s compute the Hessian of the KL-divergence

$$\mathbb{E}_{s,a\sim d^{\pi_{\theta_0}}}\left[\ln\frac{\pi_{\theta_0}(a_h\,|\,s_h)}{\pi_{\theta}(a_h\,|\,s_h)}\right] := \ell(\theta)$$

Let’s compute the Hessian of the KL-divergence

$$\mathbb{E}_{s,a\sim d^{\pi_{\theta_0}}}\left[\ln\frac{\pi_{\theta_0}(a_h\,|\,s_h)}{\pi_\theta(a_h\,|\,s_h)}\right] := \ell(\theta)$$

$$\nabla^2_\theta \ell(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{s\sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a\,|\,s)\Big(-\nabla^2_\theta \ln \pi_\theta(a\,|\,s)\,|_{\theta=\theta_0}\Big)$$

Let's compute the Hessian of the KL-divergence

$$\mathbb{E}_{s,a\sim d^{\pi_{\theta_0}}}\left[\ln\frac{\pi_{\theta_0}(a_h\,|\,s_h)}{\pi_\theta(a_h\,|\,s_h)}\right] := \ell(\theta)$$

$$\nabla^2_\theta \ell(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{s\sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a\,|\,s)\Big(-\nabla^2_\theta \ln \pi_\theta(a\,|\,s)\,|_{\theta=\theta_0}\Big)$$

$$= -\,\mathbb{E}_{s\sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a\,|\,s)\left(\frac{\nabla^2_\theta \pi_{\theta_0}(a\,|\,s)}{\pi_{\theta_0}(a\,|\,s)} - \frac{\nabla_\theta \pi_{\theta_0}(a\,|\,s)\,\nabla_\theta \pi_{\theta_0}(a\,|\,s)^\top}{\pi^2_{\theta_0}(a\,|\,s)}\right)$$

Let's compute the Hessian of the KL-divergence

$$\mathbb{E}_{s,a\sim d^{\pi_{\theta_0}}}\left[\ln\frac{\pi_{\theta_0}(a_h\,|\,s_h)}{\pi_\theta(a_h\,|\,s_h)}\right] := \ell(\theta)$$

$$\nabla^2_\theta \ell(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{s\sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a\,|\,s)\Big(-\nabla^2_\theta \ln \pi_\theta(a\,|\,s)\,|_{\theta=\theta_0}\Big)$$

$$= -\,\mathbb{E}_{s\sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a\,|\,s)\left(\frac{\nabla^2_\theta \pi_{\theta_0}(a\,|\,s)}{\pi_{\theta_0}(a\,|\,s)} - \frac{\nabla_\theta \pi_{\theta_0}(a\,|\,s)\,\nabla_\theta \pi_{\theta_0}(a\,|\,s)^\top}{\pi^2_{\theta_0}(a\,|\,s)}\right)$$

$$= \mathbb{E}_{s,a\sim d^{\pi_{\theta_0}}}\left[\nabla_\theta \ln \pi_{\theta_0}(a\,|\,s)\Big(\nabla_\theta \ln \pi_{\theta_0}(a\,|\,s)\Big)^\top\right]$$

Let's compute the Hessian of the KL-divergence

$$\mathbb{E}_{s,a\sim d^{\pi_{\theta_0}}}\left[\ln\frac{\pi_{\theta_0}(a_h\,|\,s_h)}{\pi_\theta(a_h\,|\,s_h)}\right] := \ell(\theta)$$

$$\nabla^2_\theta \ell(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{s\sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a\,|\,s)\Big(-\nabla^2_\theta \ln \pi_\theta(a\,|\,s)\,|_{\theta=\theta_0}\Big)$$

$$= -\,\mathbb{E}_{s\sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a\,|\,s)\left(\frac{\nabla^2_\theta \pi_{\theta_0}(a\,|\,s)}{\pi_{\theta_0}(a\,|\,s)} - \frac{\nabla_\theta \pi_{\theta_0}(a\,|\,s)\,\nabla_\theta \pi_{\theta_0}(a\,|\,s)^\top}{\pi^2_{\theta_0}(a\,|\,s)}\right)$$

$$= \mathbb{E}_{s,a\sim d^{\pi_{\theta_0}}}\left[\nabla_\theta \ln \pi_{\theta_0}(a\,|\,s)\Big(\nabla_\theta \ln \pi_{\theta_0}(a\,|\,s)\Big)^\top\right]$$

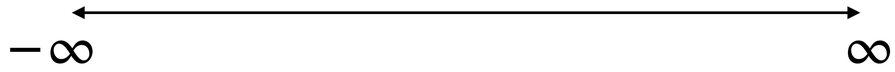
Fisher Information Matrix!

Second-order Taylor Expansion of KL at θ_0

$$\frac{1}{H} KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta \Rightarrow \frac{1}{2}(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$

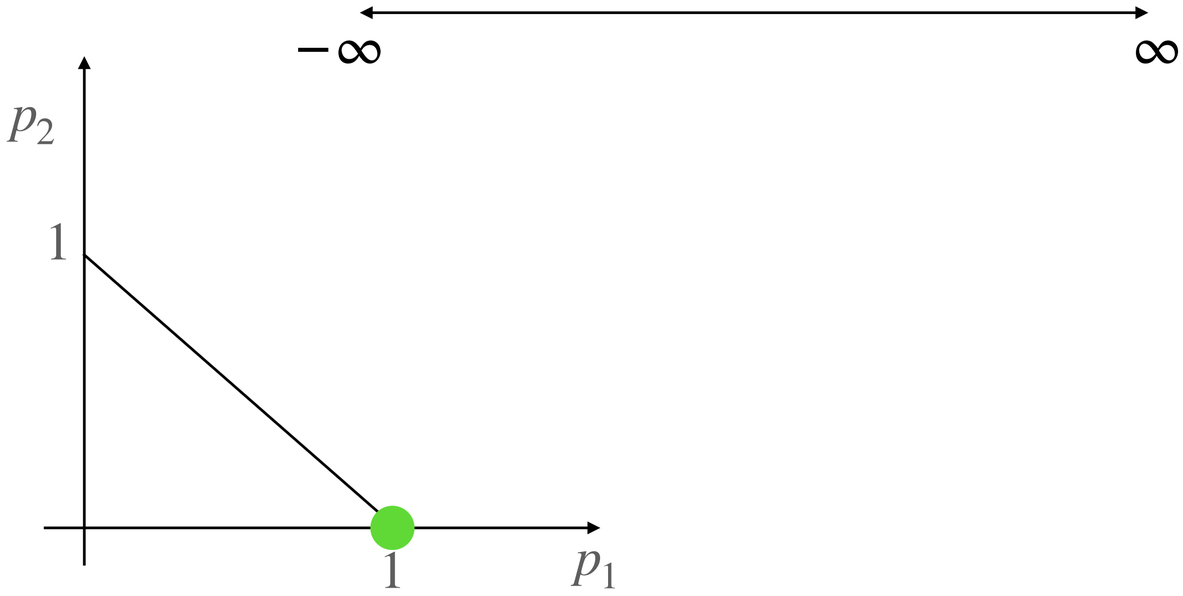
Second-order Taylor Expansion of KL at θ_0

$$\frac{1}{H} KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta \Rightarrow \frac{1}{2}(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$



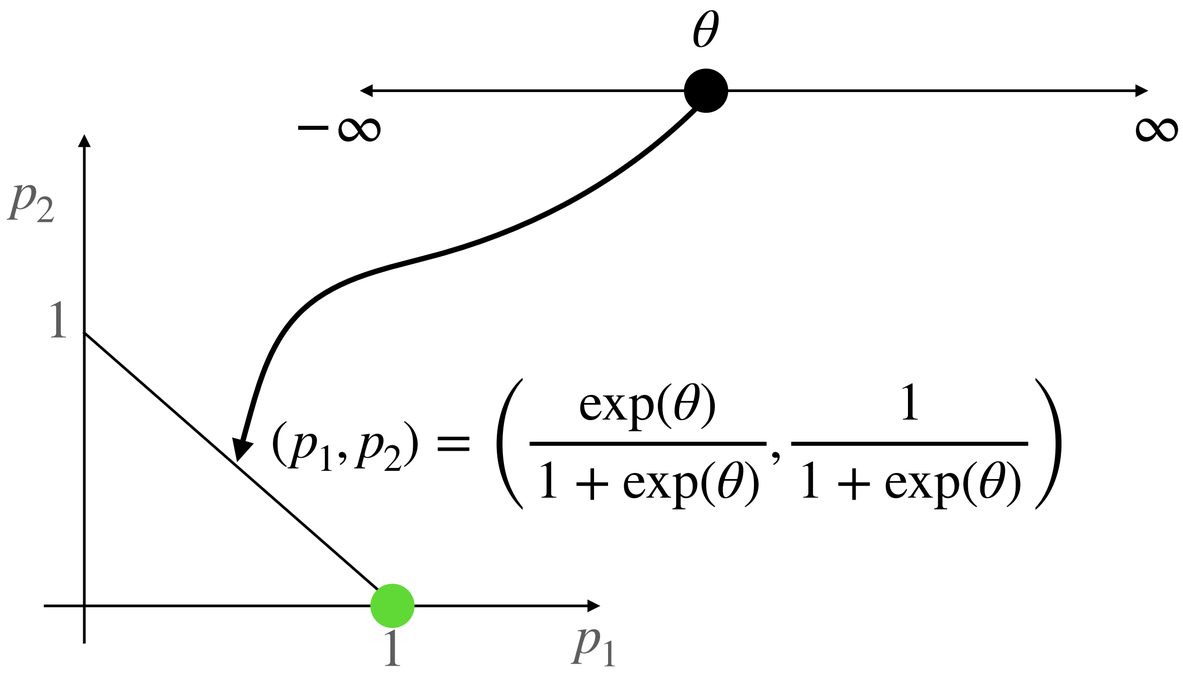
Second-order Taylor Expansion of KL at θ_0

$$\frac{1}{H} KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta \Rightarrow \frac{1}{2}(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$



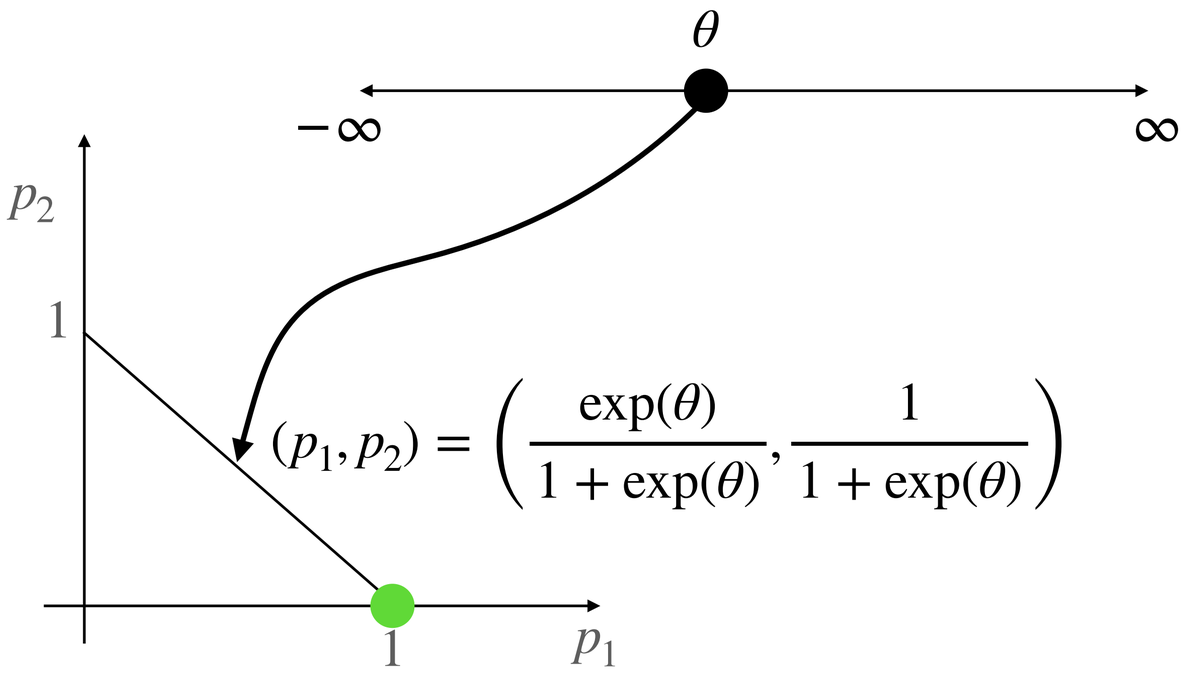
Second-order Taylor Expansion of KL at θ_0

$$\frac{1}{H} KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta \Rightarrow \frac{1}{2}(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$



Second-order Taylor Expansion of KL at θ_0

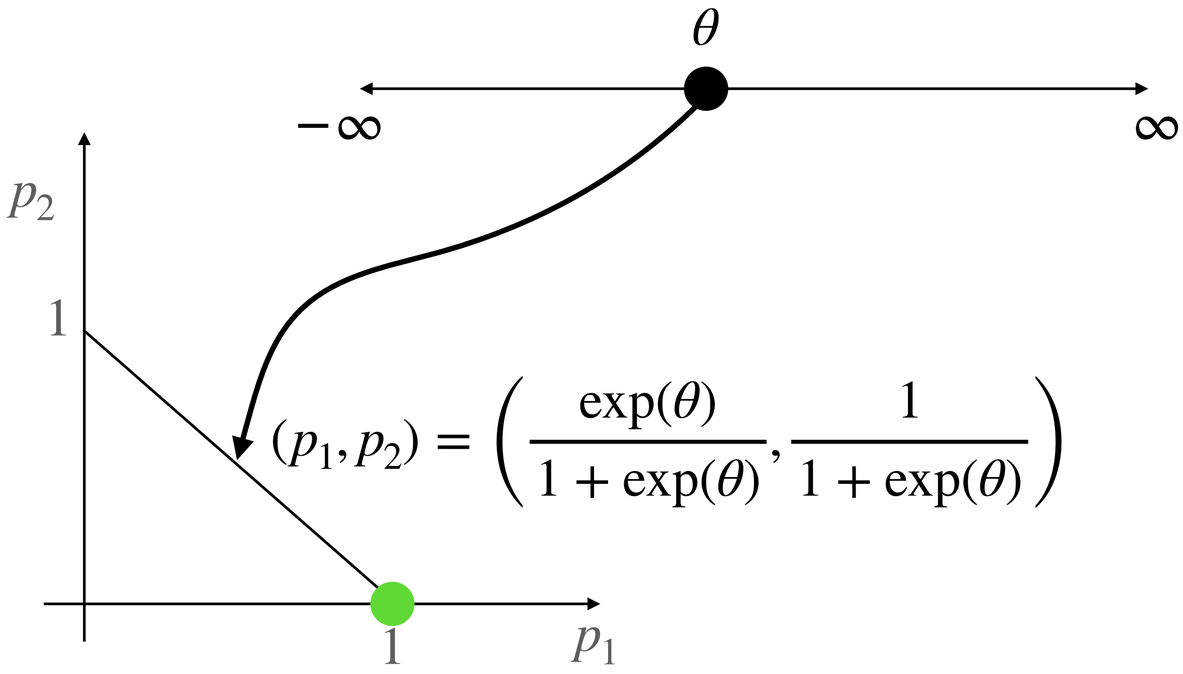
$$\frac{1}{H} KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta \Rightarrow \frac{1}{2}(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$



$F_\theta \to 0^+$, as $\theta \to \infty$

Second-order Taylor Expansion of KL at θ_0

$$\frac{1}{H} KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta \Rightarrow \frac{1}{2}(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$

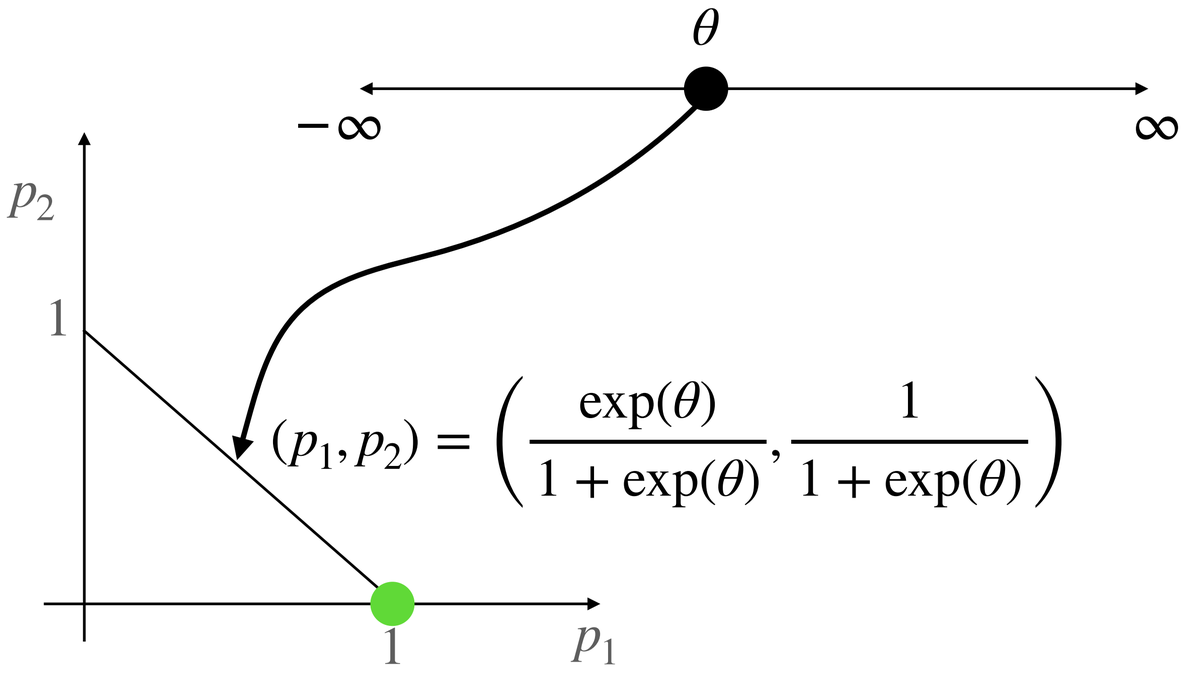


$$F_\theta \to 0^+, \text{as } \theta \to \infty$$

$$F_{\theta_0}(\theta - \theta_0)^2 \leq \delta \Rightarrow (\theta - \theta_0)^2 \leq \frac{\delta}{F_{\theta_0}} \to \infty, \text{as } \theta_0 \to \infty$$

Second-order Taylor Expansion of KL at θ_0

$$\frac{1}{H} KL\left(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_\theta}\right) \leq \delta \Rightarrow \frac{1}{2}(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$



$$F_\theta \to 0^+, \text{as } \theta \to \infty$$

$$F_{\theta_0}(\theta - \theta_0)^2 \leq \delta \Rightarrow (\theta - \theta_0)^2 \leq \frac{\delta}{F_{\theta_0}} \to \infty, \text{as } \theta_0 \to \infty$$

Plain GD in θ will move to $\theta = \infty$ at a constant speed, while Natural GD can traverse faster and faster when θ gets bigger (Infinitely fast when $\theta \to \infty$)

Now we can solve the following quadratic programming:

$$\max_{\theta} \nabla V^{\pi_{\theta_0}}(\rho)^{\top}\left(\theta - \theta_0\right)$$

$$\text{s.t. } \left(\theta - \theta_0\right)^{\top} F_{\theta_0}(\theta - \theta_0) \leq \delta$$

Now we can solve the following quadratic programming:

$$\max_{\theta} \nabla V^{\pi_{\theta_0}}(\rho)^{\top}\left(\theta - \theta_0\right)$$

$$\text{s.t. } \left(\theta - \theta_0\right)^{\top} F_{\theta_0}(\theta - \theta_0) \leq \delta$$

We have a closed form solution:

$$\theta = \theta_0 + \sqrt{\frac{\delta}{(\nabla V^{\pi_{\theta_0}})^{\top} F_{\theta_0}^{-1} \nabla V^{\pi_{\theta_0}}}} \cdot F_{\theta_0}^{-1} \nabla V^{\pi_{\theta_0}}$$

Now we can solve the following quadratic programming:

$$\max_{\theta} \nabla V^{\pi_{\theta_0}}(\rho)^{\top}\left(\theta - \theta_0\right)$$

$$\text{s.t. } \left(\theta - \theta_0\right)^{\top} F_{\theta_0}(\theta - \theta_0) \leq \delta$$

We have a closed form solution:

$$\theta = \theta_0 + \sqrt{\frac{\delta}{(\nabla V^{\pi_{\theta_0}})^{\top} F_{\theta_0}^{-1} \nabla V^{\pi_{\theta_0}}}} \cdot F_{\theta_0}^{-1} \nabla V^{\pi_{\theta_0}}$$

Self-normalized step-size
(Learning rate is adaptive)

Summary

Natural Policy Gradient invariant to linear transformation
(Trust region constraint in terms KL on trajectory distributions)

Second order Taylor expansion of $\ell(\theta) := KL\left(\text{Pr}^{\pi_{\theta_0}} || \text{Pr}^{\pi_\theta}\right)$ at θ_0 is $(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0)$