

# Natural Policy Gradient

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

# Today:

A trust region optimization perspective of NPG (also recovers the TRPO algorithm)

# 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](mailto: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**  
**Sergey Levine**  
**Philipp Moritz**  
**Michael Jordan**  
**Pieter Abbeel**

University of California, Berkeley, Department of Electrical Engineering and Computer Sciences

JOSCHU@EECS.BERKELEY.EDU  
SLEVINE@EECS.BERKELEY.EDU  
PCMORITZ@EECS.BERKELEY.EDU  
JORDAN@CS.BERKELEY.EDU  
PABBEEL@CS.BERKELEY.EDU

ICML 2015

# Notations and Settings:

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

# Notations and Settings:

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

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:  $\mathcal{M} = \{S, A, H, r, P, \rho\}$

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:  $\mathcal{M} = \{S, A, H, r, P, \rho\}$

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:

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

# Revisit Gradient Descent:

$$\theta = \theta_0 - \eta \nabla_{\theta} \mathcal{L}(\theta_0)$$

## 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  $\theta$  space

## Revisit Gradient Descent:

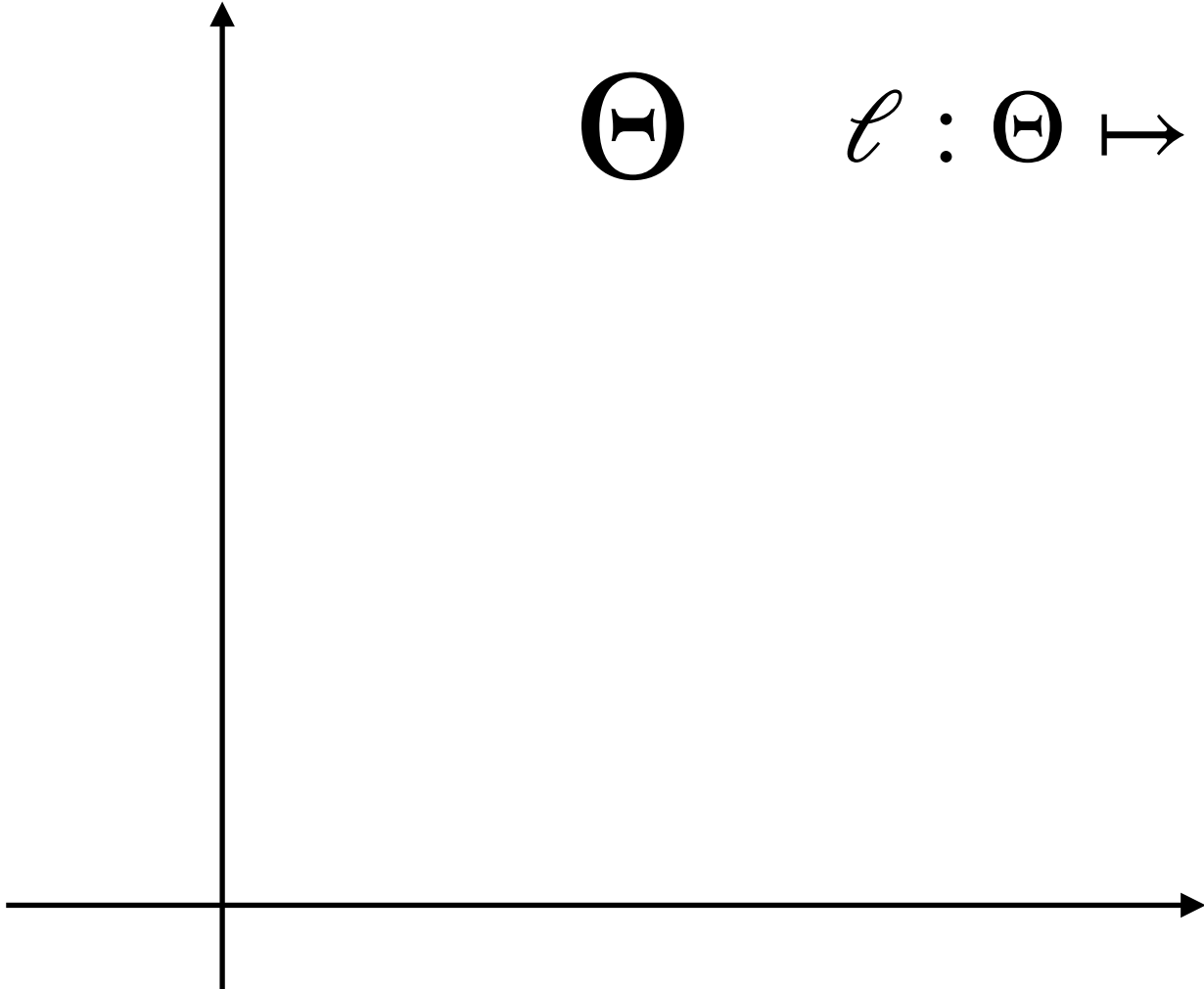
$$\theta = \theta_0 - \eta \nabla_{\theta} \ell(\theta_0)$$

In other words:

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

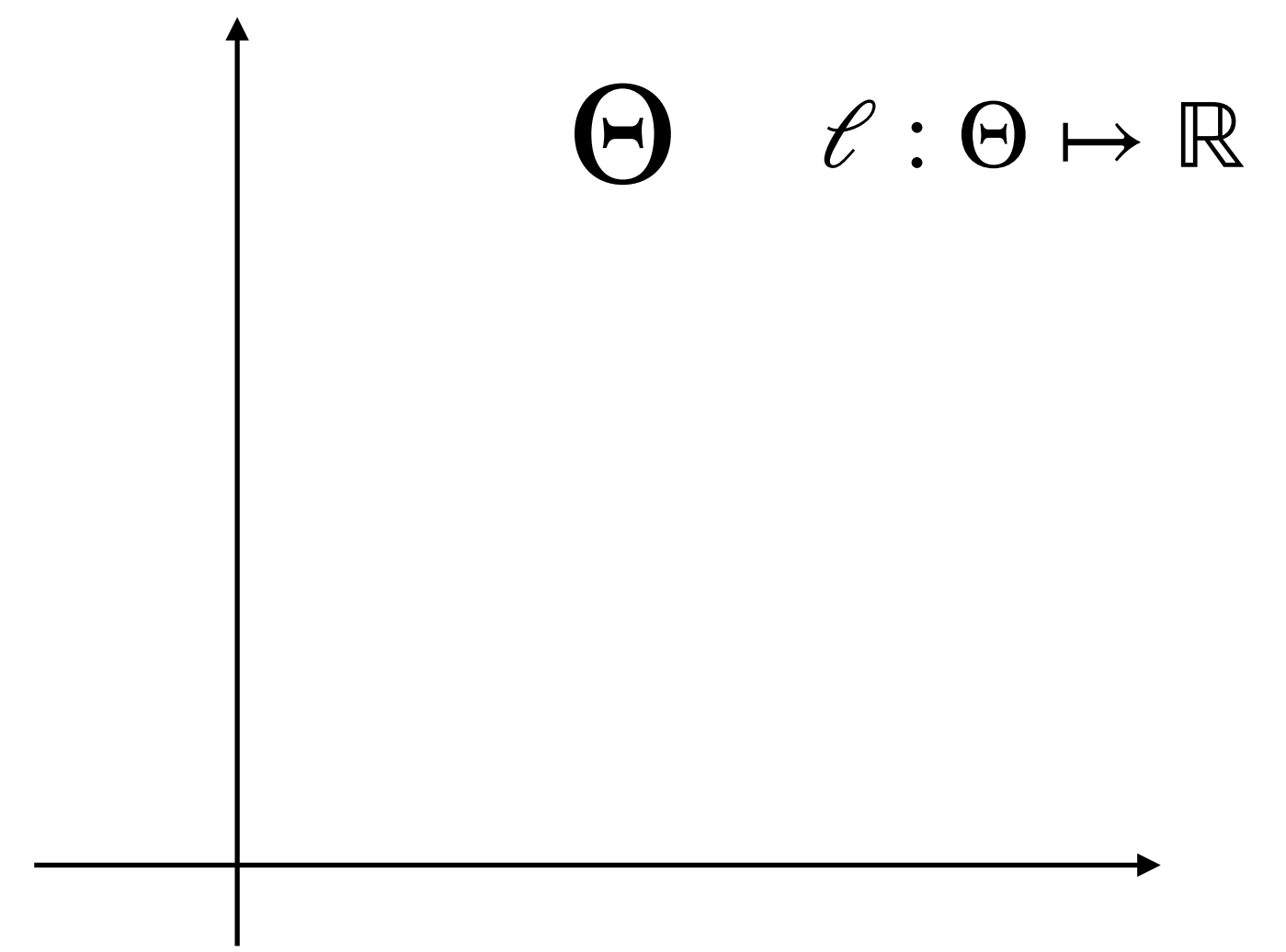
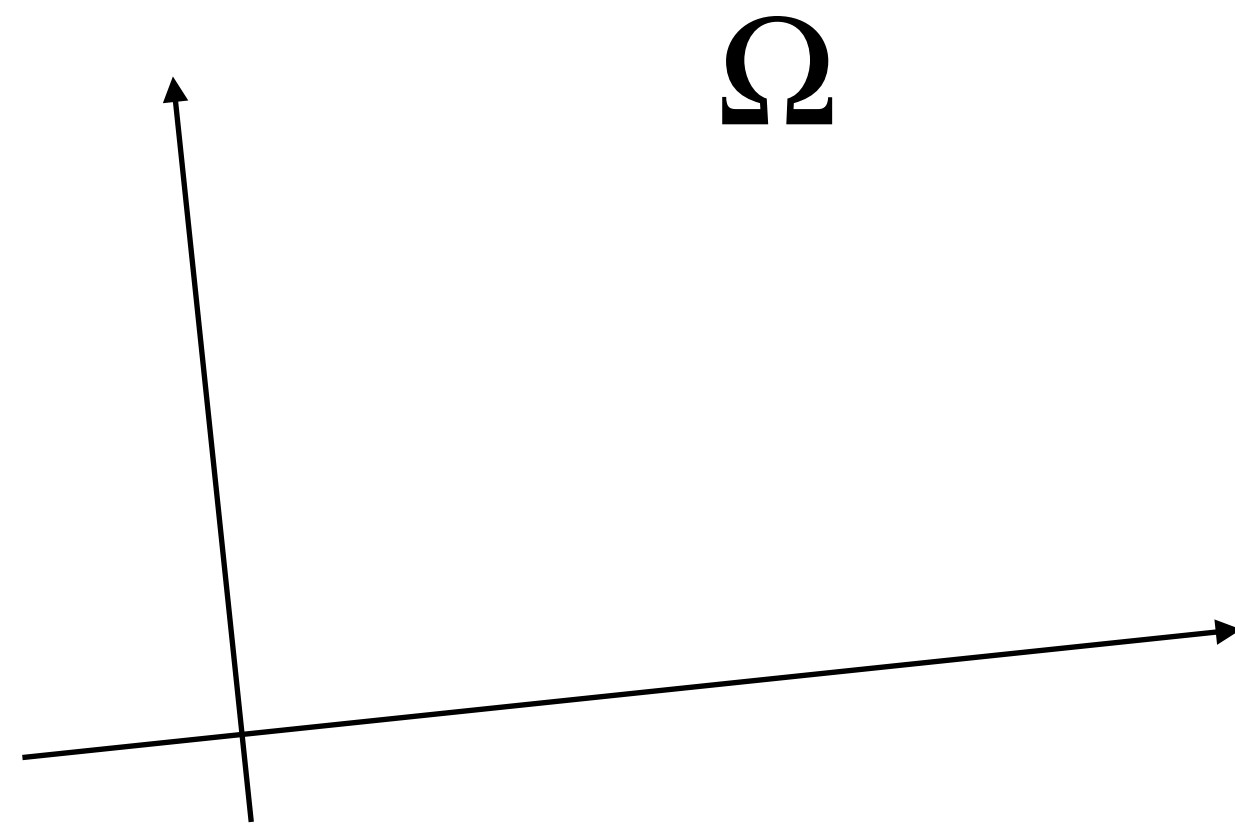
We in default are using Euclidean distance in the parameter  $\theta$  space

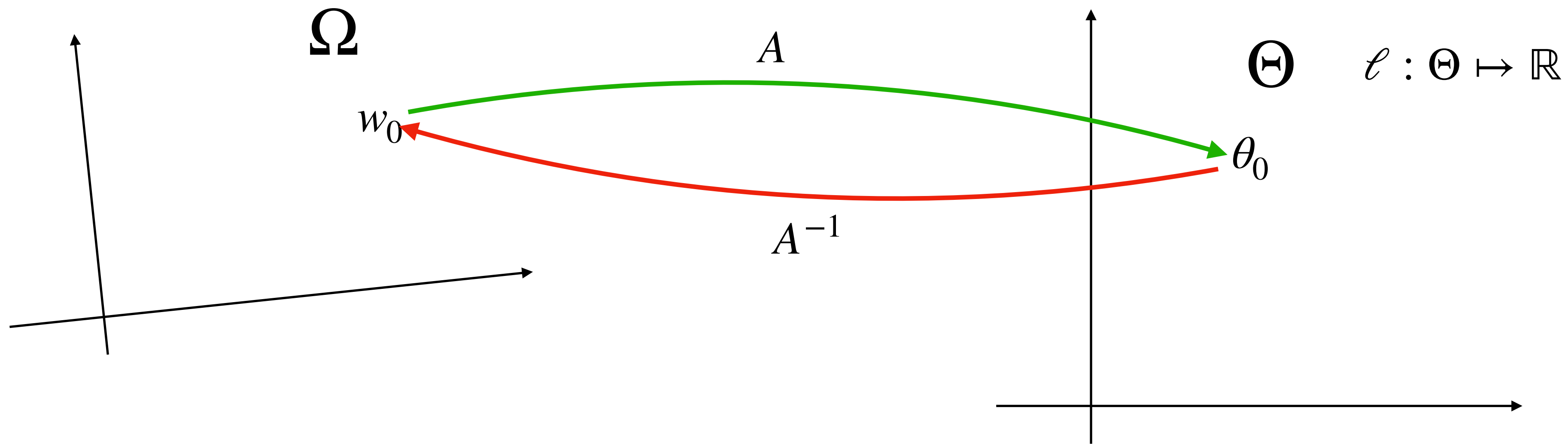
Different re-parameterization (scaling & translation) can lead to a quite different GD path

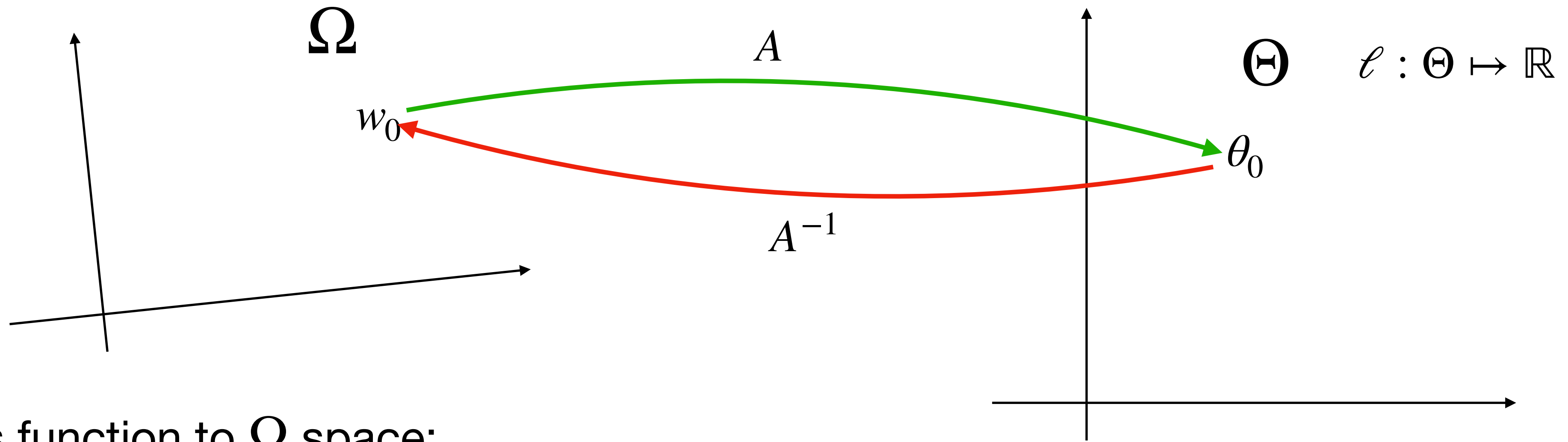


$\Theta$

$\ell : \Theta \mapsto \mathbb{R}$

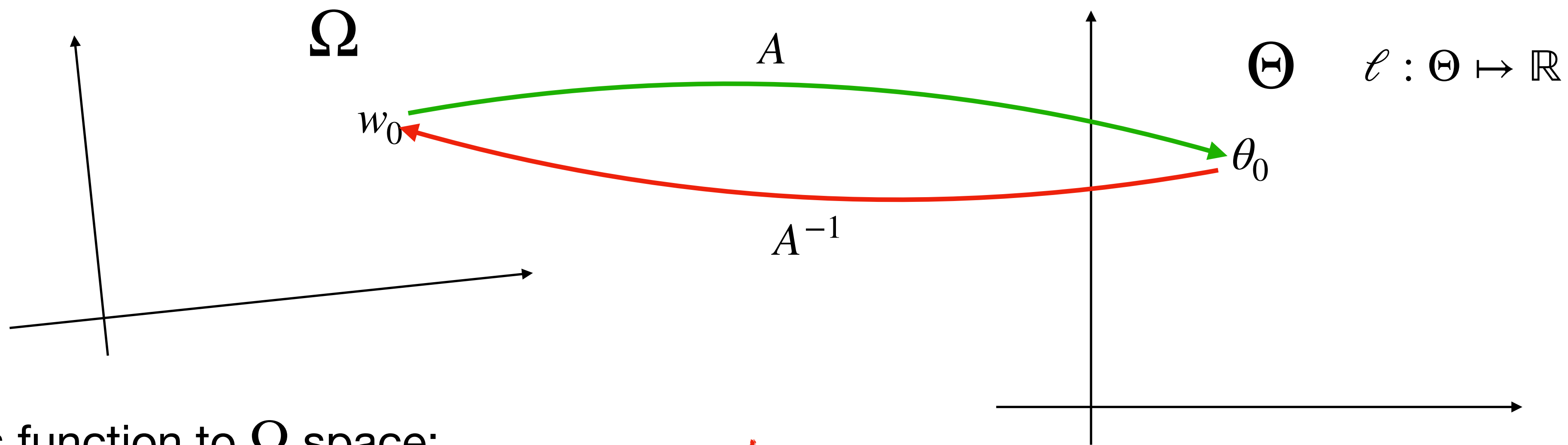






Map the loss function to  $\Omega$  space:

$$g(w) := \ell(Aw)$$



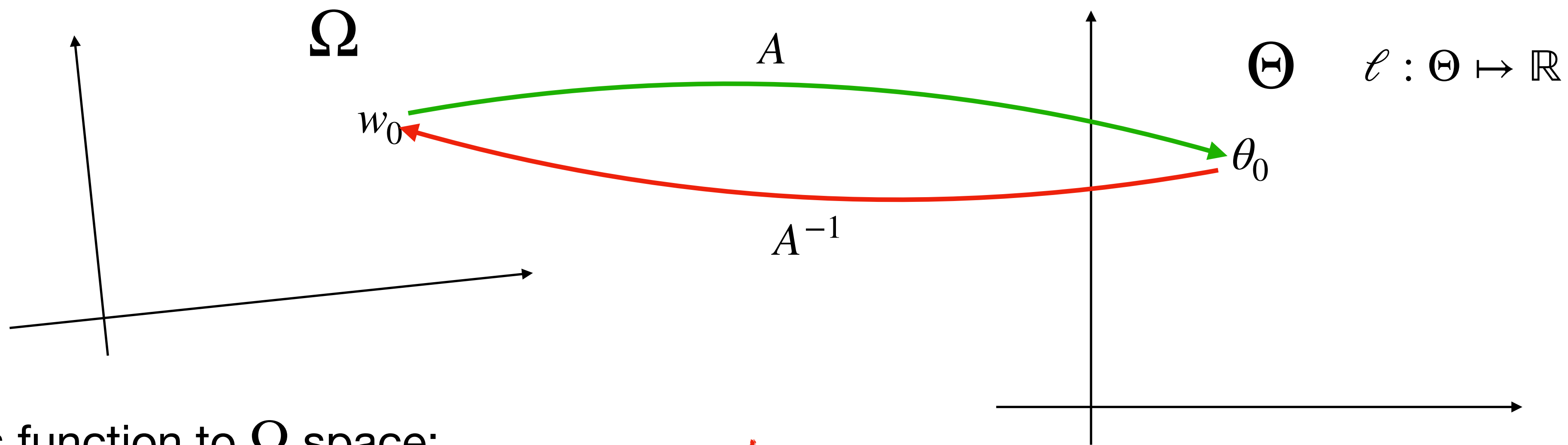
Map the loss function to  $\Omega$  space:

$$g(w) := \ell(Aw)$$

$$w = w_0 - \eta \nabla g(w_0)$$

$$\ell(\theta)$$

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



Map the loss function to  $\Omega$  space:

$$g(w) := \ell(Aw)$$

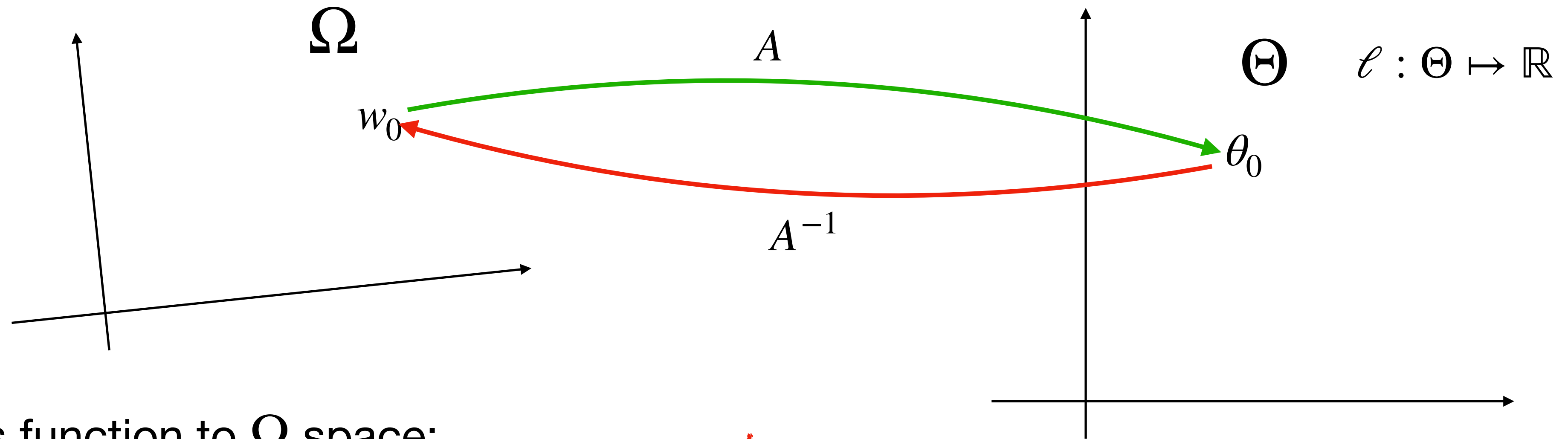
$$w = w_0 - \eta \nabla g(w_0)$$

$$= A^{-1}\theta_0 - \eta \nabla g(w_0)$$

$$\ell(\theta)$$

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





Map the loss function to  $\Omega$  space:

$$g(w) := \ell(Aw)$$

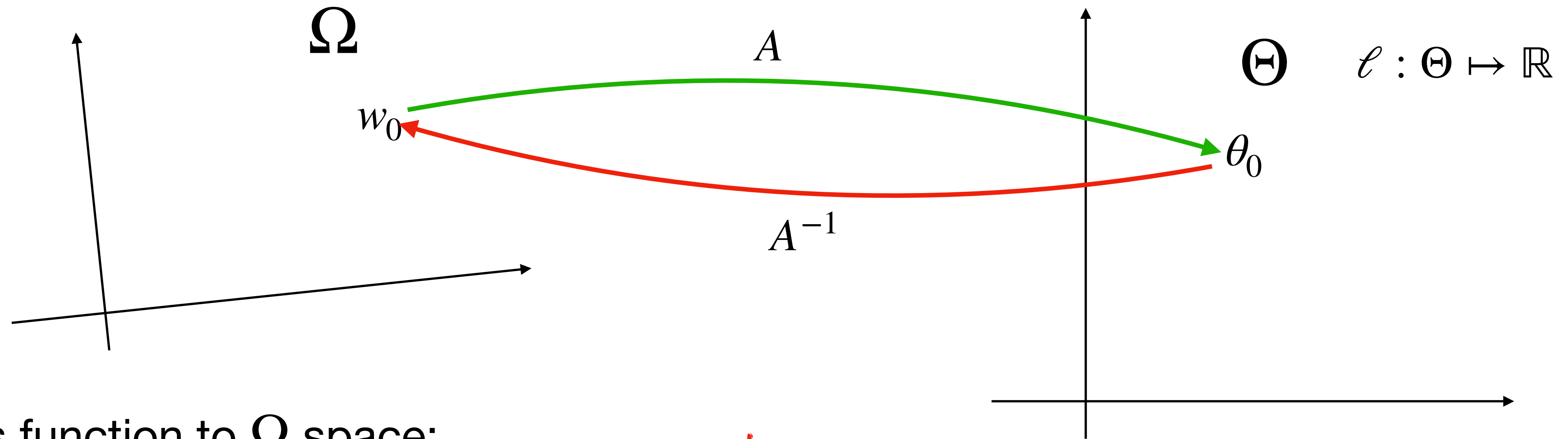
$$w = w_0 - \eta \nabla g(w_0)$$

$$= A^{-1}\theta_0 - \eta \nabla g(w_0)$$

$$(\nabla_w g(w) = \nabla_w \ell(Aw) = A \nabla_{\theta} \ell(\theta) |_{\theta=Aw})$$

$$\ell(\theta)$$

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



Map the loss function to  $\Omega$  space:

$$g(w) := \ell(Aw)$$

$$w = w_0 - \eta \nabla g(w_0)$$

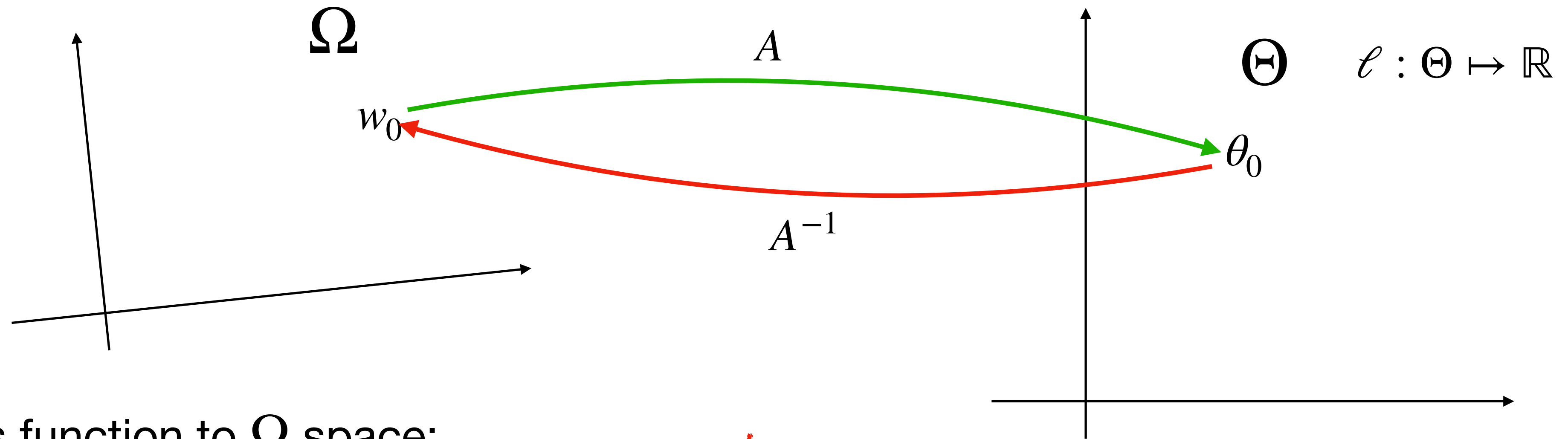
$$= A^{-1}\theta_0 - \eta \nabla g(w_0)$$

$$(\nabla_w g(w) = \nabla_w \ell(Aw) = A \nabla_{\theta} \ell(\theta) |_{\theta=Aw})$$

$$= A^{-1}\theta_0 - \eta A \nabla_{\theta} \ell(\theta_0)$$

$$\ell(\theta)$$

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



Map the loss function to  $\Omega$  space:

$$g(w) := \ell(Aw)$$

$$w = w_0 - \eta \nabla g(w_0)$$

$$= A^{-1}\theta_0 - \eta \nabla g(w_0)$$

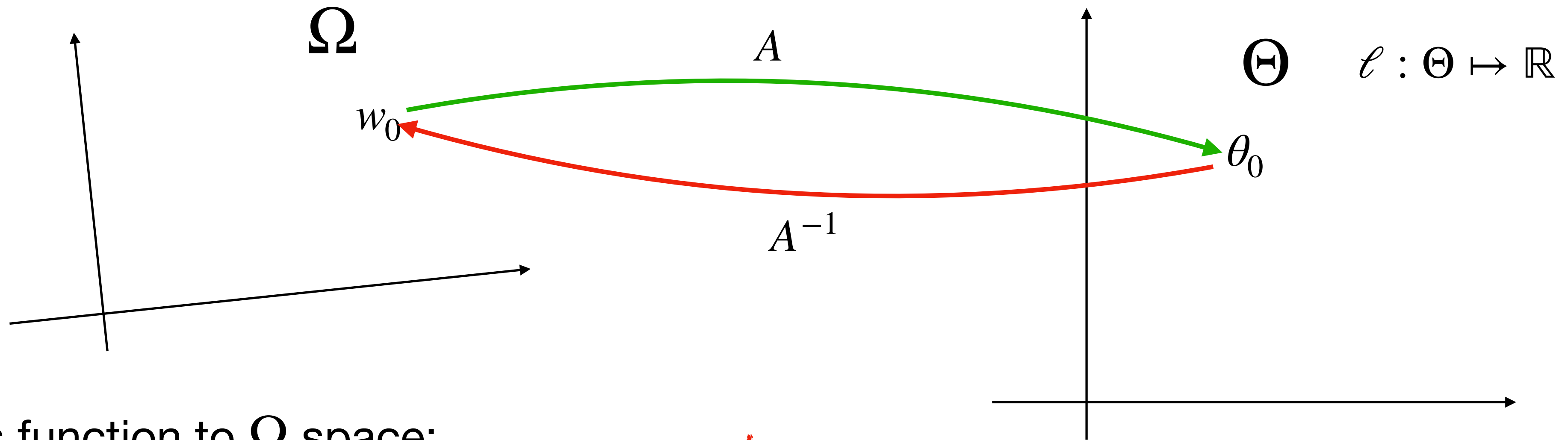
$$(\nabla_w g(w) = \nabla_w \ell(Aw) = A \nabla_\theta \ell(\theta) |_{\theta=Aw})$$

$$= A^{-1}\theta_0 - \eta A \nabla_\theta \ell(\theta_0)$$

$$\theta' = \theta_0 - \eta A^2 \nabla_\theta \ell(\theta_0)$$

$$\ell(\theta)$$

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



Map the loss function to  $\Omega$  space:

$$g(w) := \ell(Aw)$$

$$w = w_0 - \eta \nabla g(w_0)$$

$$= A^{-1}\theta_0 - \eta \nabla g(w_0)$$

$$(\nabla_w g(w) = \nabla_w \ell(Aw) = A \nabla_{\theta} \ell(\theta) |_{\theta=Aw})$$

$$= A^{-1}\theta_0 - \eta A \nabla_{\theta} \ell(\theta_0)$$

$$\theta' = \theta_0 - \eta A^2 \nabla_{\theta} \ell(\theta_0)$$

$$\ell(\theta)$$

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

Linear transformation  $A$  makes the GD path different!  
i.e., not invariant wrt linear transformation (scaling, rotations, etc)

**What would happen if we use a different distance metric...**

$$\begin{aligned} & \min_w \nabla_w g(w_0)^\top (w - w_0) \\ & \text{s.t., } (w - w_0)^\top (AA)(w - w_0) \leq \delta \end{aligned}$$

## What would happen if we use a different distance metric...

$$\begin{aligned} & \min_w \nabla_w g(w_0)^\top (w - w_0) \\ & \text{s.t., } (w - w_0)^\top (AA)(w - w_0) \leq \delta \end{aligned}$$

This gives us:

$$w = w_0 - \eta A^{-2} \nabla g(w_0)$$

## What would happen if we use a different distance metric...

$$\begin{aligned} & \min_w \nabla_w g(w_0)^\top (w - w_0) \\ & \text{s.t., } (w - w_0)^\top (AA)(w - w_0) \leq \delta \end{aligned}$$

This gives us:

$$w = w_0 - \eta A^{-2} \nabla g(w_0) \quad (\nabla_w g(w_0) = \nabla_w \ell(Aw_0) = A \nabla_\theta \ell(\theta) |_{\theta=Aw_0})$$

## What would happen if we use a different distance metric...

$$\begin{aligned} & \min_w \nabla_w g(w_0)^\top (w - w_0) \\ & \text{s.t., } (w - w_0)^\top (AA)(w - w_0) \leq \delta \end{aligned}$$

This gives us:

$$w = w_0 - \eta A^{-2} \nabla g(w_0) \quad (\nabla_w g(w_0) = \nabla_w \ell(Aw_0) = A \nabla_\theta \ell(\theta) |_{\theta=Aw_0})$$

$$w = A^{-1} \theta_0 - \eta A^{-1} \nabla_\theta \ell(\theta_0)$$



## What would happen if we use a different distance metric...

$$\begin{aligned} & \min_w \nabla_w g(w_0)^\top (w - w_0) \\ & \text{s.t., } (w - w_0)^\top (AA)(w - w_0) \leq \delta \end{aligned}$$

This gives us:

$$w = w_0 - \eta A^{-2} \nabla g(w_0) \quad (\nabla_w g(w_0) = \nabla_w \ell(Aw_0) = A \nabla_\theta \ell(\theta) |_{\theta=Aw_0})$$

$$w = A^{-1} \theta_0 - \eta A^{-1} \nabla_\theta \ell(\theta_0)$$

$$\Rightarrow \theta = \theta_0 - \eta \nabla \ell(\theta_0)$$

# Back to Policy Optimization:

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

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

# Back to Policy Optimization:

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

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

Sequential convex programming:

We linearize the objective function & quadratize the KL constraint

# Back to Policy Optimization:

$$\begin{aligned} & \max_{\pi_{\theta}} V^{\pi_{\theta}}(\rho) \\ & \text{s.t., } KL(\text{Pr}^{\pi_{\theta_0}} || \text{Pr}^{\pi_{\theta}}) \leq \delta \end{aligned}$$

Sequential convex programming:

We linearize the objective function & quadratize the KL constraint

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

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

# Back to Policy Optimization:

$$\begin{aligned} & \max_{\pi_{\theta}} V^{\pi_{\theta}}(\rho) \\ & \text{s.t., } KL(\text{Pr}^{\pi_{\theta_0}} || \text{Pr}^{\pi_{\theta}}) \leq \delta \end{aligned}$$

Sequential convex programming:

We linearize the objective function & quadratize the KL constraint

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

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

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

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

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

$$\begin{aligned} \frac{1}{H} KL (\Pr^{\pi_{\theta_0}} || \Pr^{\pi_{\theta}}) &= \frac{1}{H} \sum_{\tau} \Pr^{\theta_0}(\tau) \ln \frac{\Pr^{\theta_0}(\tau)}{\Pr^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \Pr^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h | s_h)}{\pi_{\theta}(a_h | s_h)} \\ &= \mathbb{E}_{s_h, a_h \sim d^{\pi_{\theta_0}}} \left[ \ln \frac{\pi_{\theta_0}(a_h | s_h)}{\pi_{\theta}(a_h | s_h)} \right] := \ell(\theta) \end{aligned}$$



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

$$\begin{aligned} \frac{1}{H} KL (\text{Pr}^{\pi_{\theta_0}} || \text{Pr}^{\pi_{\theta}}) &= \frac{1}{H} \sum_{\tau} \text{Pr}^{\theta_0}(\tau) \ln \frac{\text{Pr}^{\theta_0}(\tau)}{\text{Pr}^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \text{Pr}^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h | s_h)}{\pi_{\theta}(a_h | s_h)} \\ &= \mathbb{E}_{s_h, a_h \sim d^{\pi_{\theta_0}}} \left[ \ln \frac{\pi_{\theta_0}(a_h | s_h)}{\pi_{\theta}(a_h | s_h)} \right] := \ell(\theta) \quad \ell(\theta_0) = 0 \end{aligned}$$

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

$$\begin{aligned}\frac{1}{H}KL(\text{Pr}^{\pi_{\theta_0}} || \text{Pr}^{\pi_{\theta}}) &= \frac{1}{H} \sum_{\tau} \text{Pr}^{\theta_0}(\tau) \ln \frac{\text{Pr}^{\theta_0}(\tau)}{\text{Pr}^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \text{Pr}^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h | s_h)}{\pi_{\theta}(a_h | s_h)} \\ &= \mathbb{E}_{s_h, a_h \sim d^{\pi_{\theta_0}}} \left[ \ln \frac{\pi_{\theta_0}(a_h | s_h)}{\pi_{\theta}(a_h | s_h)} \right] := \ell(\theta) \quad \ell(\theta_0) = 0\end{aligned}$$

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

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

$$\begin{aligned}\frac{1}{H}KL(\text{Pr}^{\pi_{\theta_0}} || \text{Pr}^{\pi_{\theta}}) &= \frac{1}{H} \sum_{\tau} \text{Pr}^{\theta_0}(\tau) \ln \frac{\text{Pr}^{\theta_0}(\tau)}{\text{Pr}^{\theta}(\tau)} = \frac{1}{H} \sum_{\tau} \text{Pr}^{\theta_0}(\tau) \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_0}(a_h | s_h)}{\pi_{\theta}(a_h | s_h)} \\ &= \mathbb{E}_{s_h, a_h \sim d^{\pi_{\theta_0}}} \left[ \ln \frac{\pi_{\theta_0}(a_h | s_h)}{\pi_{\theta}(a_h | s_h)} \right] := \ell(\theta) \quad \ell(\theta_0) = 0\end{aligned}$$

$$\begin{aligned}\nabla_{\theta} \ell(\theta) |_{\theta=\theta_0} &= \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a | s) \left( -\nabla_{\theta} \ln \pi_{\theta}(a | s) |_{\theta=\theta_0} \right) \\ &= -\mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a | s) \frac{\nabla_{\theta} \pi_{\theta_0}(a | s)}{\pi_{\theta_0}(a | s)}\end{aligned}$$

**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

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

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

## Let's compute the Hessian of the KL-divergence

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

$$\begin{aligned} \nabla_{\theta}^2 \ell(\theta) |_{\theta=\theta_0} &= \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a | s) \left( - \nabla_{\theta}^2 \ln \pi_{\theta}(a | s) |_{\theta=\theta_0} \right) \\ &= - \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a | s) \left( \frac{\nabla_{\theta}^2 \pi_{\theta_0}(a | s)}{\pi_{\theta_0}(a | s)} - \frac{\nabla_{\theta} \pi_{\theta_0}(a | s) \nabla_{\theta} \pi_{\theta_0}(a | s)^{\top}}{\pi_{\theta_0}^2(a | s)} \right) \end{aligned}$$

## Let's compute the Hessian of the KL-divergence

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

$$\begin{aligned} \nabla_{\theta}^2 \ell(\theta) |_{\theta=\theta_0} &= \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a | s) \left( - \nabla_{\theta}^2 \ln \pi_{\theta}(a | s) |_{\theta=\theta_0} \right) \\ &= - \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a | s) \left( \frac{\nabla_{\theta}^2 \pi_{\theta_0}(a | s)}{\pi_{\theta_0}(a | s)} - \frac{\nabla_{\theta} \pi_{\theta_0}(a | s) \nabla_{\theta} \pi_{\theta_0}(a | s)^{\top}}{\pi_{\theta_0}^2(a | s)} \right) \\ &= \mathbb{E}_{s,a \sim d^{\pi_{\theta_0}}} \left[ \nabla_{\theta} \ln \pi_{\theta_0}(a | s) \left( \nabla_{\theta} \ln \pi_{\theta_0}(a | s) \right)^{\top} \right] \end{aligned}$$



# Let's compute the Hessian of the KL-divergence

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

$$\begin{aligned} \nabla_{\theta}^2 \ell(\theta) |_{\theta=\theta_0} &= \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a | s) \left( - \nabla_{\theta}^2 \ln \pi_{\theta}(a | s) |_{\theta=\theta_0} \right) \\ &= - \mathbb{E}_{s \sim d^{\pi_{\theta_0}}} \sum_a \pi_{\theta_0}(a | s) \left( \frac{\nabla_{\theta}^2 \pi_{\theta_0}(a | s)}{\pi_{\theta_0}(a | s)} - \frac{\nabla_{\theta} \pi_{\theta_0}(a | s) \nabla_{\theta} \pi_{\theta_0}(a | s)^{\top}}{\pi_{\theta_0}^2(a | s)} \right) \\ &= \mathbb{E}_{s,a \sim d^{\pi_{\theta_0}}} \left[ \nabla_{\theta} \ln \pi_{\theta_0}(a | s) \left( \nabla_{\theta} \ln \pi_{\theta_0}(a | s) \right)^{\top} \right] \end{aligned}$$

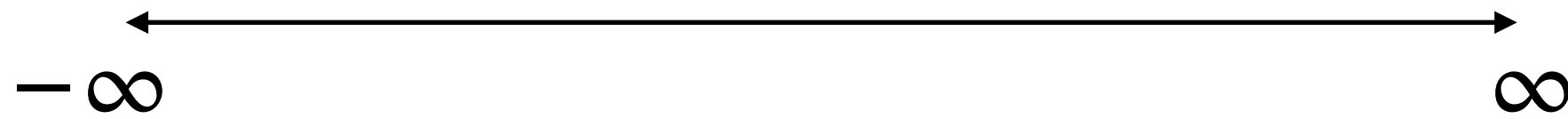
**Fisher Information Matrix!**

## Second-order Taylor Expansion of KL at $\theta_0$

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

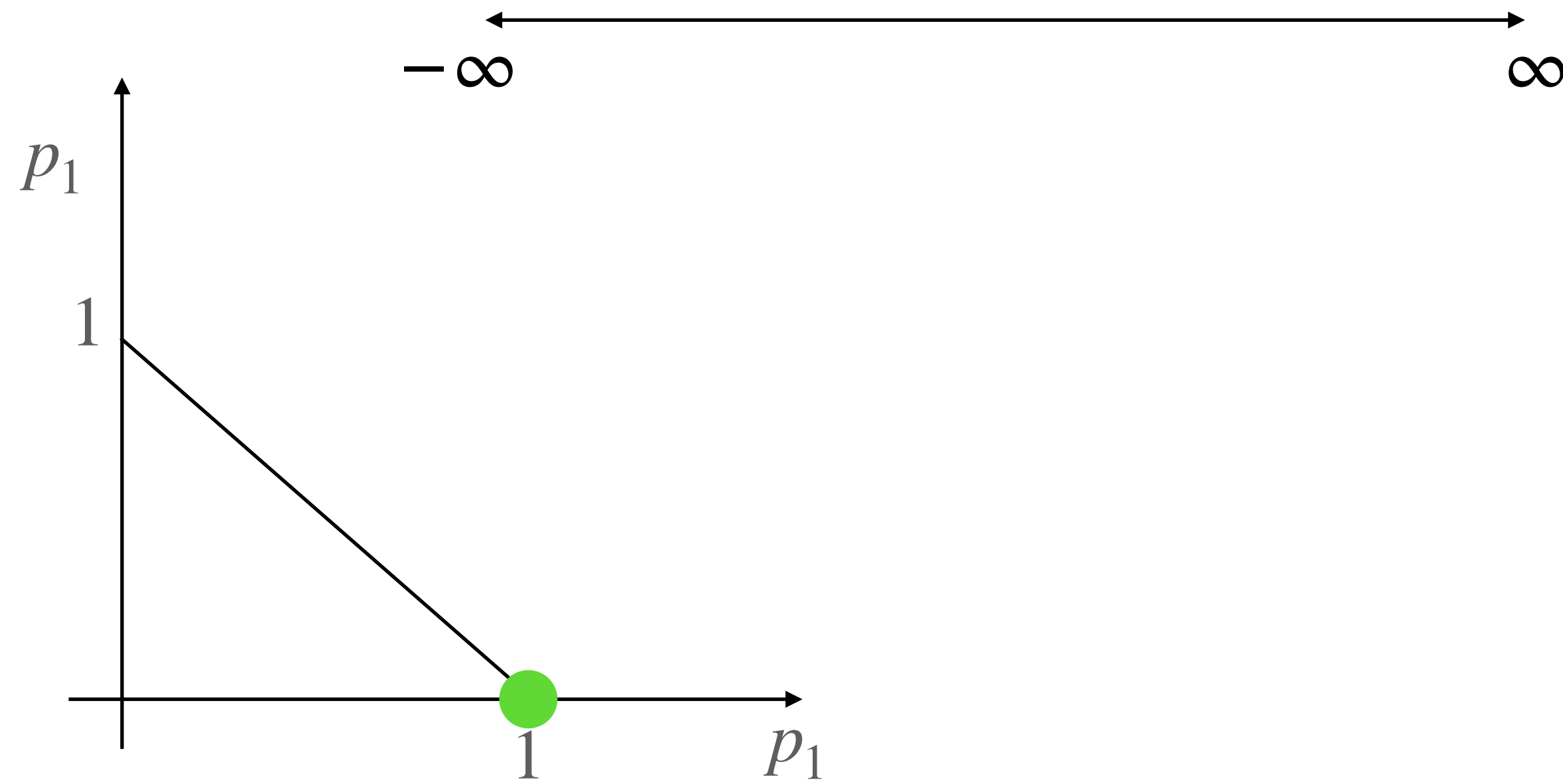
## Second-order Taylor Expansion of KL at $\theta_0$

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



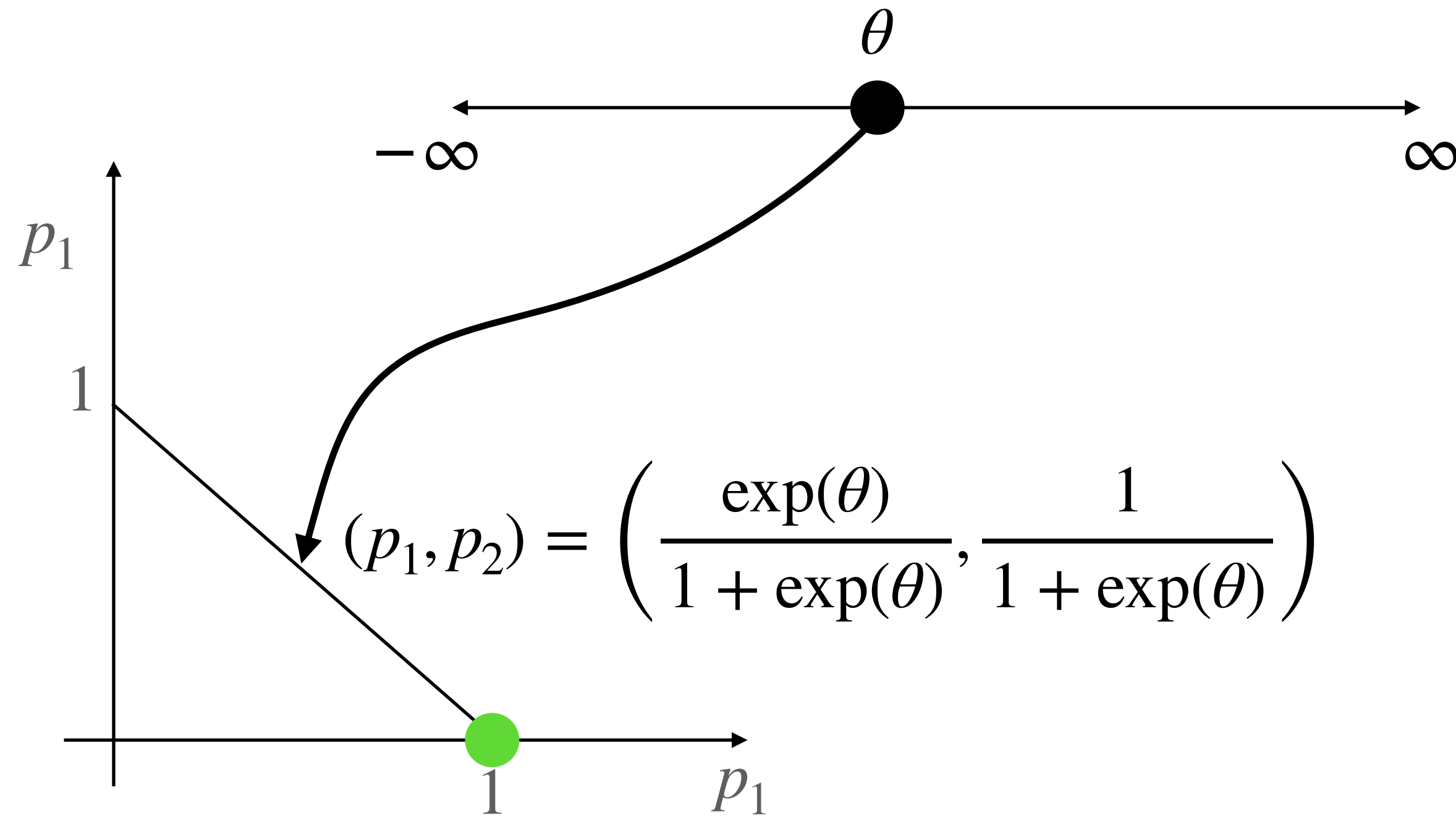
## Second-order Taylor Expansion of KL at $\theta_0$

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



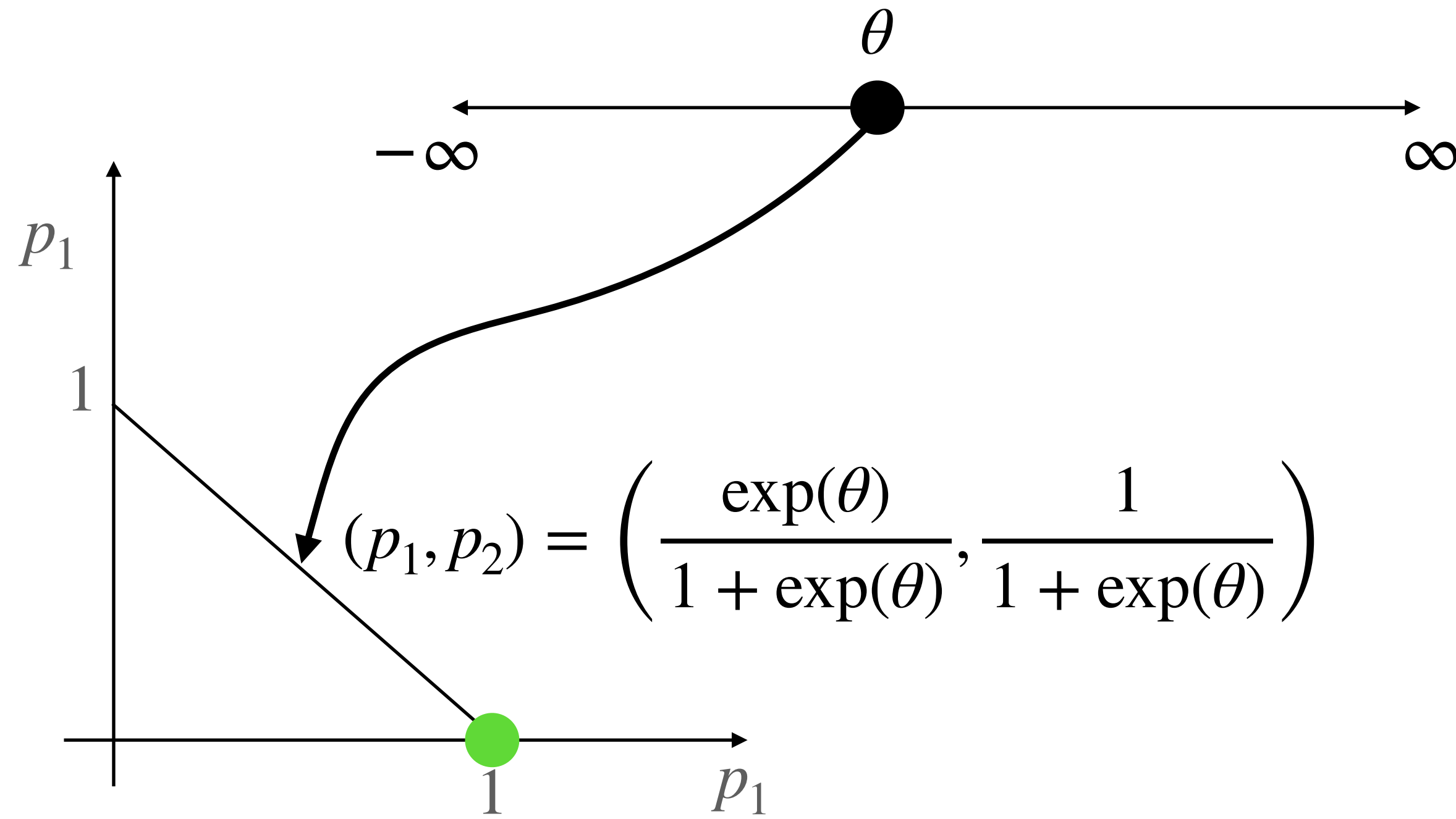
## Second-order Taylor Expansion of KL at $\theta_0$

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



## Second-order Taylor Expansion of KL at $\theta_0$

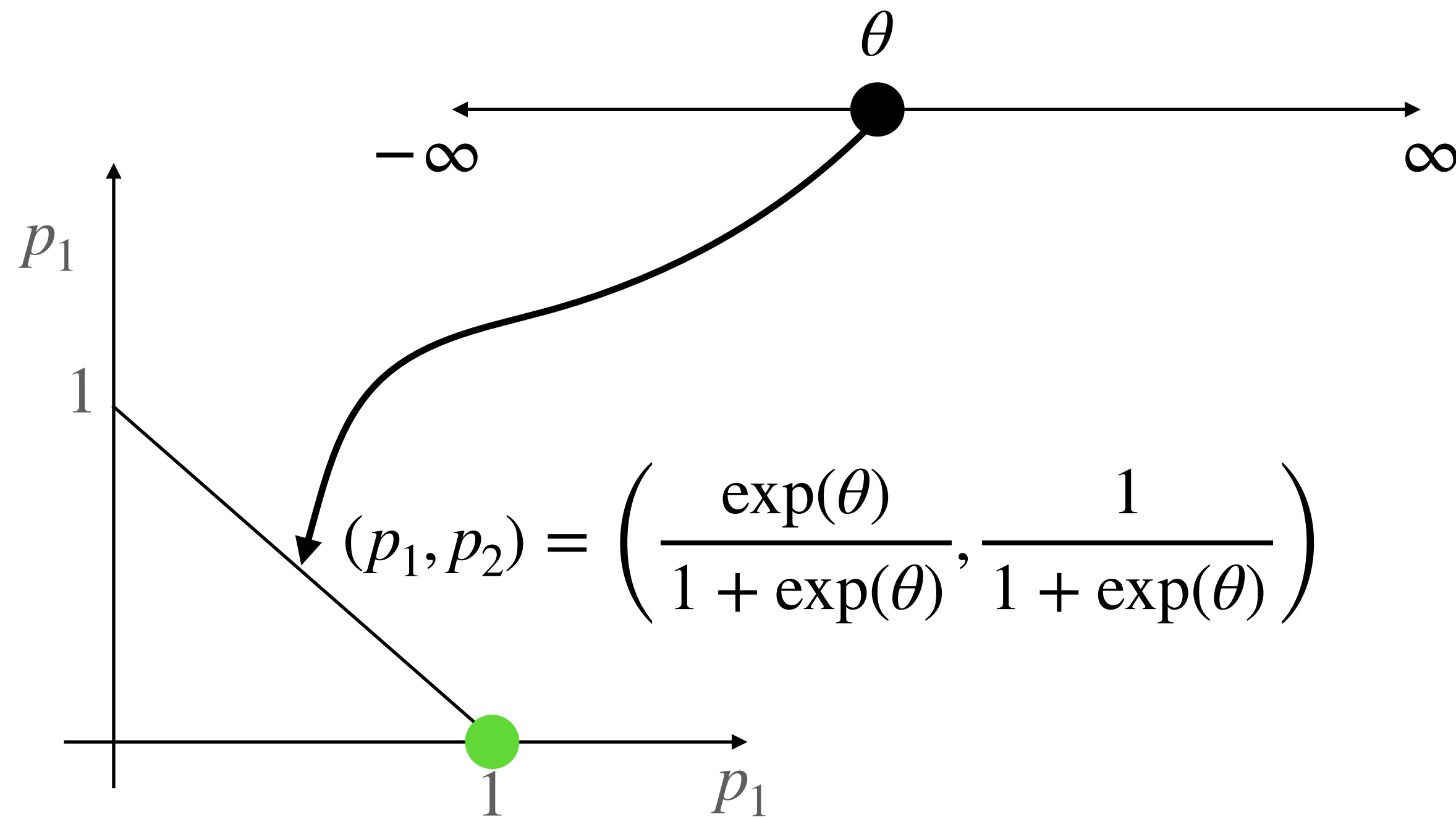
$$\frac{1}{H} KL(\text{Pr}^{\pi_{\theta_0}} || \text{Pr}^{\pi_{\theta}}) \leq \delta \Rightarrow \frac{1}{2}(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$



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

## Second-order Taylor Expansion of KL at $\theta_0$

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

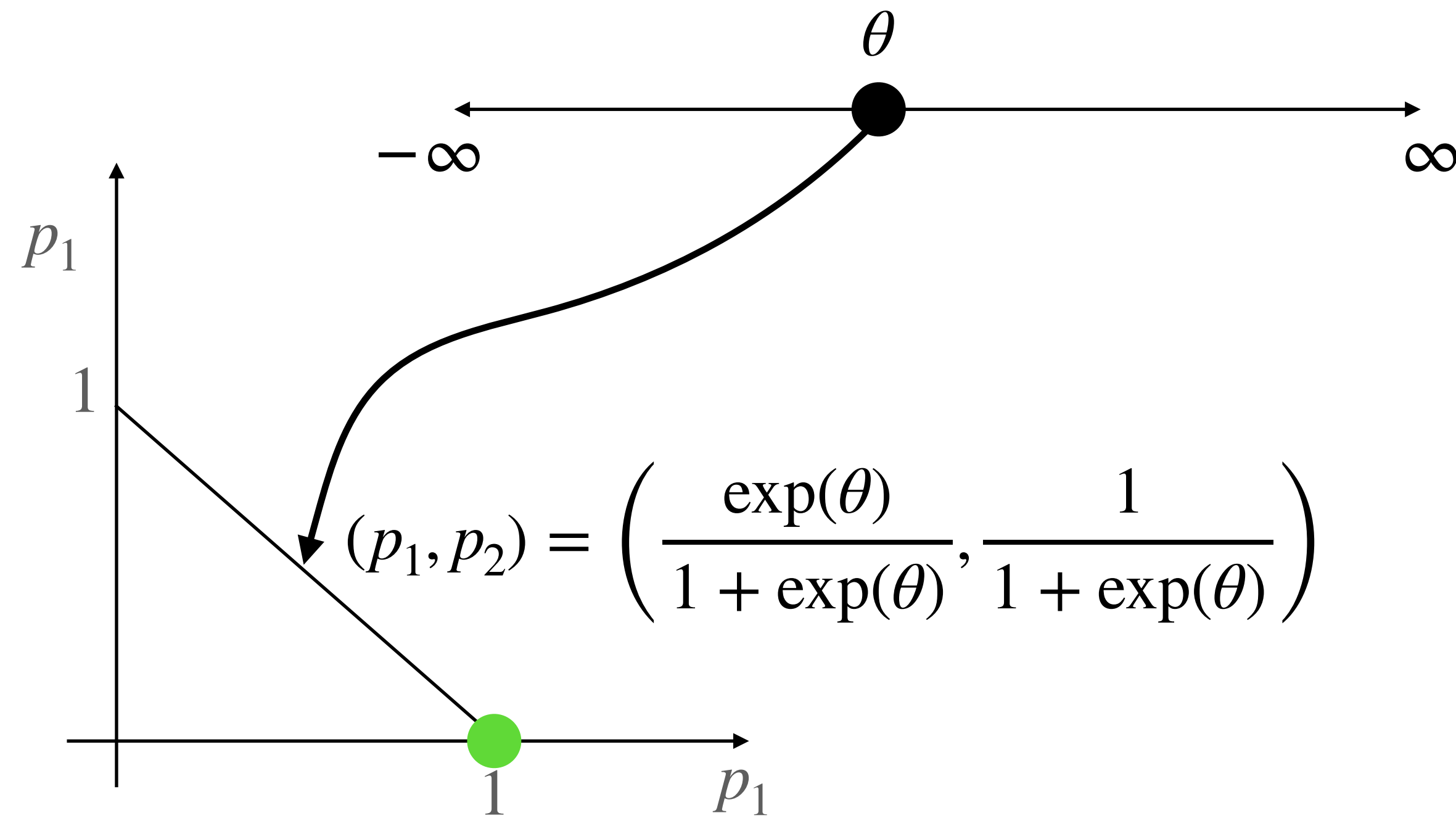


$$F_{\theta} \rightarrow 0^+, \text{ as } \theta \rightarrow \infty$$

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

## Second-order Taylor Expansion of KL at $\theta_0$

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



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

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

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



**Now we can solve the following quadratic programming:**

$$\begin{aligned} & \max_{\theta} \nabla V^{\pi_{\theta_0}}(\rho)^\top (\theta - \theta_0) \\ \text{s.t.} \quad & (\theta - \theta_0)^\top F_{\theta_0} (\theta - \theta_0) \leq \delta \end{aligned}$$

**Now we can solve the following quadratic programming:**

$$\begin{aligned} \max_{\theta} \quad & \nabla V^{\pi_{\theta_0}}(\rho)^\top (\theta - \theta_0) \\ \text{s.t.} \quad & (\theta - \theta_0)^\top F_{\theta_0} (\theta - \theta_0) \leq \delta \end{aligned}$$

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:**

$$\begin{aligned} \max_{\theta} \quad & \nabla V^{\pi_{\theta_0}}(\rho)^\top (\theta - \theta_0) \\ \text{s.t.} \quad & (\theta - \theta_0)^\top F_{\theta_0} (\theta - \theta_0) \leq \delta \end{aligned}$$

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  $\ell(\theta) := KL(\Pr^{\pi_{\theta_0}} || \Pr^{\pi_{\theta}})$  at  $\theta_0$  is  $(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0)$