Trust Region Policy Optimization & the Natural Policy Gradient

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2022

Today

- Announcements:
  If you are an undergraduate student at Harvard and are possibly interested in pursuing research, formally or informally, with the ML foundations group, please fill in the following form: https://forms.gle/yCiTfbXn31x2RQtHA

- Recap

- Today:
  1. ~~Convergence of Fitted Policy Iteration~~
  2. Trust Region Policy Optimization

     & the natural gradient

Recap

[Policy Eval Subroutine]: TD Learning for "tabular" case

[Iterative Policy Eval Subroutine/TD]

input: policy π, sample size N, init w_0

1. Sample trajectories $\tau_1, \ldots \tau_N \sim \rho_\pi$ which gives us a dataset D

   (each trajectory is of the form $\tau_i = \{s_0, a_0, r_0, \ldots s_{H-1}, a_{H-1}, r_{H-1},\}$)
2. For k = $0, \ldots, K$:
   1. Sample a transition $(s_h, r_h, s_{h+1},) \in D$ and update:

$$V_{k+1}(s_h) = V_k(s_h) - \eta_k \Big(V_k(s_h) - \big(r_h + V_k(s_{h+1})\big)\Big)$$

3. Return the function V_K as an estimate of V^π

Policy eval method.

Q-Learning for "tabular" case

[Iterative Policy Eval Subroutine/TD]

input: **offline dataset** $\tau_1, \ldots \tau_N \sim \rho_{\pi_{data}}$

1. For k = $0, \ldots, K$:
    1. Sample a transition $(s_h, a_h, r_h, s_{h+1}) \in D$ and update:

$$Q_{k+1}(s_h, a_h) = Q_k(s_h, a_h) - \eta_k \Big(Q_k(s_h, a_h) - \big(r_h + \max_{a'} Q_k(s_{h+1}, a') \big) \Big)$$

2. Return the function Q_K as an estimate of $Q^\star$

to learn/estimate Q^*

again the idea was to enforce
B. consistency.

Fitted Policy Iteration:
(aka Approximate Policy Iteration API)

1. Initialize staring policy π_0, samples size M
2. For k = 0, … :
    1. [Q-Evaluation Subroutine]
       Using M sampled trajectories, $\tau_1, \ldots \tau_N \sim \rho_{\pi_k}$,
       $\widetilde{Q}_k(s,a) \approx Q_h^{\pi_k}(s,a)$
    2. Policy Update
       $\pi_{k+1}(s) := \arg\max_a \widetilde{Q}^{\pi_k}(s,a)$
3. Return $\widetilde{Q}_K$ and π_K as an estimate of $Q^\star$ and $\pi^\star$

Recap on Supervised Learning: regression

We have a data distribution $\mathscr{D}$, $x_i \sim \mathscr{D}, y_i = f^\star(x_i) + \epsilon_i$, where noise $\mathbb{E}[\epsilon_i] = 0, |\epsilon_i| \leq c$

$$\hat{f} = \arg\min_{f \in \mathscr{F}} \sum_{i=1}^{N} \left(f(x_i) - y_i\right)^2$$

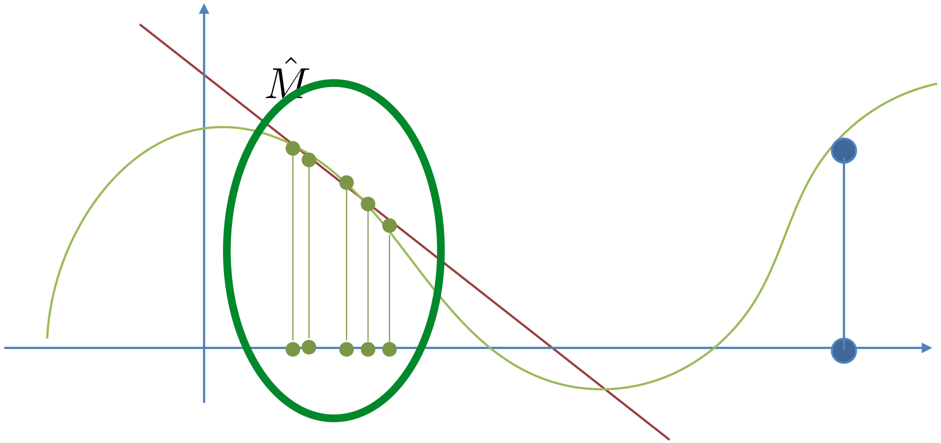
Supervised learning theory (e.g., VC theory) says that we can indeed **generalize**, i.e., we can predict well **under the same distribution:**

Assume $f^\star \in \mathscr{F}$ (this is called realizability), we can expect:

$$\mathbb{E}_{x\sim\mathscr{D}} \left(\hat{f}(x) - f^\star(x)\right)^2 \leq \delta$$

Supervise Learning can fail if there is train-test distribution mismatch

However, for some $\mathscr{D}' \neq \mathscr{D}$, $\mathbb{E}_{x \sim \mathscr{D}'} (f(x) - f^{\star}(x))^2$ might be arbitrarily large



Deeper neural nets and larger datasets are typically not enough to address "distribution shift"

Fitted Policy Improvement Guarantees

Fitted Policy Improvement Guarantees

- For all k, suppose that:

$$E_{\tau \sim \rho_{\pi_k}}\Big[\sum_{h=1}^{H}\big(\widetilde{Q}_k(s_h,a_h) - Q_h^{\pi_k}(s_h,a_h)\big)^2\Big] \le \delta, \text{ and } \max_{s,a} |\widetilde{Q}_k(s_h,a_h) - Q_h^{\pi_k}(s_h,a_h)| \le \delta_\infty$$

- δ: the average case supervised learning error (reasonable to expect this can be made small)

  δ_∞: the worse case error (often unreasonable to expect to be small)

Fitted Policy Improvement Guarantees

- For all k, suppose that:

$$E_{\tau\sim\rho_{\pi_k}}\Big[\sum_{h=1}^{H}\big(\widetilde{Q}_k(s_h,a_h)-Q_h^{\pi_k}(s_h,a_h)\big)^2\Big]\le\delta, \text{ and } \max_{s,a}|\widetilde{Q}_k(s_h,a_h)-Q_h^{\pi_k}(s_h,a_h)|\le\delta_\infty$$

- δ: the average case supervised learning error (reasonable to expect this can be made small)

  δ_∞: the worse case error (often unreasonable to expect to be small)

[Theorem:] We have that:

Fitted Policy Improvement Guarantees

- For all k, suppose that:

$$E_{\tau\sim\rho_{\pi_k}}\left[\sum_{h=1}^{H}\left(\widetilde{Q}_k(s_h,a_h)-Q_h^{\pi_k}(s_h,a_h)\right)^2\right]\le\delta, \text{ and } \quad \max_{s,a}|\widetilde{Q}_k(s_h,a_h)-Q_h^{\pi_k}(s_h,a_h)|\le\delta_\infty$$

- δ: the average case supervised learning error (reasonable to expect this can be made small)

  δ_∞: the worse case error (often unreasonable to expect to be small)

[Theorem:] We have that:

- One step performance degradation is bounded by the worst case error:

$$Q^{k+1}(s)\ge Q^k(s)-2H\delta_\infty$$ (and equality possible in some examples).

Fitted Policy Improvement Guarantees

- For all k, suppose that:

$$E_{\tau\sim\rho_{\pi_k}}\Big[\sum_{h=1}^{H}\big(\widetilde{Q}_k(s_h,a_h)-Q_h^{\pi_k}(s_h,a_h)\big)^2\Big]\le\delta, \text{ and } \max_{s,a}|\widetilde{Q}_k(s_h,a_h)-Q_h^{\pi_k}(s_h,a_h)|\le\delta_\infty$$

- δ: the average case supervised learning error (reasonable to expect this can be made small)
  δ_∞: the worse case error (often unreasonable to expect to be small)

[Theorem:] We have that:

- One step performance degradation is bounded by the worst case error:
  $Q^{k+1}(s)\ge Q^k(s)-2H\delta_\infty$ (and equality possible in some examples).
- For large enough K, final performance also governed by the worst case error:
  $Q^{\pi_K}(s)\ge Q^\star(s)-2H^2\delta_\infty$

Fitted Policy Improvement Guarantees

- For all k, suppose that:

$$E_{\tau\sim\rho_{\pi_k}}\Big[\sum_{h=1}^{H}\big(\widetilde{Q}_k(s_h,a_h)-Q_h^{\pi_k}(s_h,a_h)\big)^2\Big]\le\delta, \text{ and } \max_{s,a}|\widetilde{Q}_k(s_h,a_h)-Q_h^{\pi_k}(s_h,a_h)|\le\delta_\infty$$

- δ: the average case supervised learning error (reasonable to expect this can be made small)

  δ_∞: the worse case error (often unreasonable to expect to be small)

[Theorem:] We have that:

- One step performance degradation is bounded by the worst case error:

  $$Q^{k+1}(s)\ge Q^k(s)-2H\delta_\infty$$ (and equality possible in some examples).

- For large enough K, final performance also governed by the worst case error:

  $$Q^{\pi_K}(s)\ge Q^\star(s)-2H^2\delta_\infty$$

- (Intuition) If it somehow turns out that, for all iterations k, the density under the next policy, uniformly does not differ from that of previous policy, i.e. that

  $$\max_{s,a,h}\left(\frac{\Pr(s_h=s,a_h=a\,|\,\pi_{k+1})}{\Pr(s_h=s,a_h=a\,|\,\pi_k)}\right)\le C_\infty$$

  then we can bound our sub-optimality by the average case error:

  $$Q^{\pi_K}(s)\ge Q^\star(s)-2H^2\cdot C_\infty\cdot\delta$$

Today:

~~Optimality in Markov Decision Processes~~

Outline:

1. Quick intro on KL-divergence
   & the visitation measure
2. A Trust-Region Formulation for Policy Optimization
3. Algorithm: Natural Policy Gradient

KL-divergence: measures the distance between two distributions

Given two distributions P & Q, where $P \in \Delta(X), Q \in \Delta(X)$,
KL Divergence is defined as:

$$KL(P|Q) = \mathbb{E}_{x \sim P}\left[\ln \frac{P(x)}{Q(x)}\right]$$

KL-divergence: measures the distance between two distributions

Given two distributions P & Q, where $P \in \Delta(X), Q \in \Delta(X)$,
KL Divergence is defined as:

$$KL(P|Q) = \mathbb{E}_{x \sim P}\left[\ln \frac{P(x)}{Q(x)}\right]$$

always

$KL(P|Q) \geq 0$

Examples:

If $Q = P$, then $KL(P|Q) = KL(Q|P) = 0$

KL-divergence: measures the distance between two distributions

Given two distributions P & Q, where $P \in \Delta(X), Q \in \Delta(X)$,
KL Divergence is defined as:

$$KL(P|Q) = \mathbb{E}_{x \sim P}\left[\ln \frac{P(x)}{Q(x)}\right]$$

Examples:

If $Q = P$, then $KL(P|Q) = KL(Q|P) = 0$

If $P = \mathcal{N}(\mu_1, \sigma^2 I), Q = \mathcal{N}(\mu_2, \sigma^2 I)$, then $KL(P|Q) = \|\mu_1 - \mu_2\|_2^2 / \sigma^2$

KL-divergence: measures the distance between two distributions

Given two distributions P & Q, where $P \in \Delta(X), Q \in \Delta(X)$,
KL Divergence is defined as:

$$KL(P \mid Q) = \mathbb{E}_{x \sim P}\left[\ln \frac{P(x)}{Q(x)}\right]$$

Examples:

If $Q = P$, then $KL(P \mid Q) = KL(Q \mid P) = 0$

If $P = \mathcal{N}(\mu_1, \sigma^2 I), Q = \mathcal{N}(\mu_2, \sigma^2 I)$, then $KL(P \mid Q) = \|\mu_1 - \mu_2\|_2^2 / \sigma^2$

Fact:

$KL(P \mid Q) \geq 0$, and being 0 if and only if $P = Q$

Outline:

1. Quick intro on KL-divergence
2. A Trust-Region Formulation for Policy Optimization (TRPO)
3. Algorithm: Natural Policy Gradient

A trust region formulation for policy update:

At iteration t, with π_{θ_t} at hand, we compute θ_{t+1} as follows:

A trust region formulation for policy update:

At iteration t, with π_{θ_t} at hand, we compute θ_{t+1} as follows:

$$\max_{\theta} \mathbb{E}_{s_0,\ldots s_{H-1}\sim\rho_{\theta_t}} \left[\sum_{h=0}^{H-1} \mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a) \right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$$

A trust region formulation for policy update:

$A^{\pi}(s,a)$

$= Q^{\pi}(s,a) - V^{\pi}(s)$

At iteration t, with π_{θ_t} at hand, we compute θ_{t+1} as follows:

$$\max_{\theta} \mathbb{E}_{s_0,\ldots s_{H-1}\sim\rho_{\theta_t}} \left[\sum_{h=0}^{H-1} \mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a) \right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$$

We want to maximize local advantage against π_{θ_t}, but we want the new policy to be close to π_{θ_t} (in the KL sense)

Some Helpful Notation: Visitation Measures

$s_0 \sim \mu$, following π

- Visitation probability at time h: $\mathbb{P}_h(s_h, a_h \,|\, \mu, \pi)$
  (recall that we absorb h, into the state, i.e. $s \leftarrow (s, h)$)
- Average Visitation Measure:

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

$$d_\mu^\pi(s) = \frac{1}{H} \sum_{h=0}^{H-1} P_h(s \,|\, \mu, \pi)$$

- With this def, we have:

$$J(\theta) := E_{s_0 \sim \mu_0}\left[V^{\pi_\theta}(s_0)\right]$$

$$= E\left[\sum_{h=0}^{H-1} r(s_h, a_h) \,\middle|\, \mu_0, \pi_\theta\right] = H \underset{s,a \sim d_\mu^\pi}{E} \left[r(s,a)\right]$$

$$= H \cdot \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi_\theta(s)}\left[r(s,a)\right]$$

$$\boxed{\nabla J(\theta) = H \underset{s,a \sim d_\mu^\pi}{E} \left[\nabla \log \pi_\theta(a|s)\, A^\theta(s,a)\right]}$$

Visitation Measures: the discounted case

- Visitation probability at time h: $\mathbb{P}_h(s_h, a_h \mid \mu, \pi)$
  (recall that we absorb h, into the state, i.e. $s \leftarrow (s, h)$)
- Average Visitation Measure:

$$d_\mu^\pi(s, a) = (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h(s, a \mid \mu, \pi)$$

- With this def, we have:

$$J(\theta) := E_{s_0 \sim \mu_0} \left[V^{\pi_\theta}(s_0) \right]$$

$$= E\left[\sum_{h=0}^{\infty} \gamma^h r_h \middle| \mu_0, \pi_\theta \right]$$

$$= \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi(s)} \left[r(s, a) \right]$$

Equivalently,

At iteration t, with π_{θ_t} at hand, we compute θ_{t+1} as follows:

Equivalently,

At iteration t, with π_{θ_t} at hand, we compute θ_{t+1} as follows:

$$\max_{\theta} H \cdot \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(s)} A^{\pi_{\theta_t}}(s, a) \right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$$

We want to maximize local advantage against π_{θ_t}, but we want the new policy to be close to π_{θ_t} (in the KL sense)

Equivalently,

At iteration t, with π_{θ_t} at hand, we compute θ_{t+1} as follows:

$$\max_{\theta} H \cdot \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a) \right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \le \delta$$

We want to maximize local advantage against π_{θ_t}, but we want the new policy to be close to π_{θ_t} (in the KL sense)

How we can actually do the optimization here?
After all, we don't even know the analytical form of trajectory likelihood…

Outline:

1. Quick intro on KL-divergence
2. A Trust-Region Formulation for Policy Optimization
3. Algorithm: Natural Policy Gradient

A trust region formulation for policy update:

At iteration t, with π_{θ_t} at hand, we compute θ_{t+1} as follows:

$$\max_{\theta} H \cdot \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(s)} A^{\pi_{\theta_t}}(s,a) \right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta} \right) \leq \delta$$

High-level strategy:

1. First-order Taylor expansion on the objective at θ_t

2.second-order Taylor expansion of the constraint at θ_t

Simplify Objective Function

H: $$\max_{\theta} \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$

Simplify Objective Function

$$\max_{\theta} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_{\theta}(s)} A^{\pi_{\theta_t}}(s, a) \right]$$

Since the objective is also non-linear,
let's do first order-talyor expansion on it:

Simplify Objective Function

$\nabla_\theta \mathbb{E}_{a\sim\pi_\theta}[f(a)]$

$= \mathbb{E}_{a\sim\pi_\theta}[\nabla \log \pi^\theta(a) f(a)]$

$$H \cdot \max_\theta \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$

Since the objective is also non-linear,
let's do first order-talyor expansion on it:

$\underbrace{\nabla_\theta \log \pi_\theta(a|s)\Big|_{\theta=\theta_t}}$

$$H\,\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right] \approx H\,\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)} A^{\pi_{\theta_t}}(s,a)\right] + \underbrace{\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)} \boxed{\nabla_\theta \ln \pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a)\right]}_{\nabla_\theta J(\pi_{\theta_t})} \cdot (\theta - \theta_t)$$

$$f_{\theta_t}(\theta) = f_{\theta_t}(\theta_t) + \left[\nabla f_{\theta_t}(\theta)\right]_{\theta=\theta_t} (\theta - \theta_t) + o(\quad)$$

because

$E_{a\sim\pi(s)}[A^{\pi}(s,a)]=0$

$J(\theta) = H \cdot E_{s\sim d^{\theta}} E_{a\sim\pi^{\theta}} [r(s,a)]$

Simplify Objective Function

$$\max_{\theta} \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_{\theta}(s)} A^{\pi_{\theta_t}}(s,a) \right]$$

Since the objective is also non-linear,
let's do first order-talyor expansion on it:

$$H\ \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_{\theta}(s)} A^{\pi_{\theta_t}}(s,a) \right] \approx H\ \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)} A^{\pi_{\theta_t}}(s,a) \right] + \underbrace{\mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)} \nabla_{\theta} \ln \pi_{\theta_t}(a\,|\,s) A^{\pi_{\theta_t}}(s,a) \right]}_{\nabla_{\theta} J(\pi_{\theta_t})} \cdot (\theta - \theta_t)$$

$$= \nabla_{\theta} J(\pi_{\theta_t})^{\top} (\theta - \theta_t)$$

$+$

Simplify Constraint via second-order Taylor Expansion:

Simplify Constraint via second-order Taylor Expansion:

$$KL(\rho_{\theta_t} | \rho_{\theta}) := \ell(\theta)$$

Simplify Constraint via second-order Taylor Expansion:

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta)$$

$$\ell(\theta) \approx \ell(\theta_t) + \nabla \ell(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

Simplify Constraint via second-order Taylor Expansion:

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta)$$

$$\ell(\theta) \approx \ell(\theta_t) + \nabla \ell(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

$$\ell(\theta_t) = KL(\rho_{\theta_t} | \rho_{\theta_t}) = 0$$

Simplify Constraint via second-order Taylor Expansion:

$$KL(\rho_{\theta_t} | \rho_\theta) := \ell(\theta)$$

$$\ell(\theta) \approx \ell(\theta_t) + \nabla \ell(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \ell(\theta_t)(\theta - \theta_t)$$

$$\ell(\theta_t) = KL(\rho_{\theta_t} | \rho_{\theta_t}) = 0$$

We will show that $\nabla_\theta \ell(\theta_t) = 0$, and $\nabla^2 \ell(\theta_t)$ has a nice form!

The gradient of the KL-divergence is zero at θ_t

Change from trajectory distribution to state-action distribution:

The gradient of the KL-divergence is zero at θ_t

Change from trajectory distribution to state-action distribution:

$$KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) = \mathbb{E}_{\tau\sim\rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)} = \mathbb{E}_{\tau\sim\rho_{\pi_{\theta_t}}} \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)}$$

$$\rho_{\pi}(\tau) = \mu_0(s_0)\, \pi_\theta(a_0|s_0)\, P(s_1|s_0, a_0) \cdots$$

The gradient of the KL-divergence is zero at θ_t

Change from trajectory distribution to state-action distribution:

$$KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)} = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)}$$

$$= H\mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right] := \ell(\theta)$$

The gradient of the KL-divergence is zero at θ_t

Change from trajectory distribution to state-action distribution:

$$KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)} = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)}$$

$$= H \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right] := \ell(\theta)$$

$$H \cdot \mathbb{E}_{s \sim d_\mu^{\theta_t}} \sum_a \pi_{\theta_t}(a|s) \log \frac{\pi_{\theta_t}(a|s)}{\pi_\theta(a|s)}$$

$$\nabla_\theta \ell(\theta) |_{\theta=\theta_t} = H \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(- \nabla_\theta \ln \pi_\theta(a_h | s_h) |_{\theta=\theta_t} \right)$$

The gradient of the KL-divergence is zero at θ_t

Change from trajectory distribution to state-action distribution:

$$KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)} = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)}$$

$$= H \mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right] := \ell(\theta)$$

$$\nabla_\theta \ell(\theta) |_{\theta=\theta_t} = H \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(- \nabla_\theta \ln \pi_\theta(a_h | s_h) |_{\theta=\theta_t} \right)$$

$$= - H \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \frac{\nabla_\theta \pi_{\theta_t}(a | s)}{\pi_{\theta_t}(a | s)}$$

The gradient of the KL-divergence is zero at θ_t

Change from trajectory distribution to state-action distribution:

$$KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \ln \frac{\rho_{\pi_{\theta_t}}(\tau)}{\rho_{\pi_\theta}(\tau)} = \mathbb{E}_{\tau \sim \rho_{\pi_{\theta_t}}} \sum_{h=0}^{H-1} \ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)}$$

$$= H\mathbb{E}_{s_h, a_h \sim d_\mu^{\pi_{\theta_t}}} \left[\ln \frac{\pi_{\theta_t}(a_h | s_h)}{\pi_\theta(a_h | s_h)} \right] := \ell(\theta)$$

$$= \nabla \sum_a \pi(a|s)$$

$$= \nabla 1$$

$$= 0$$

$$\nabla_\theta \ell(\theta) |_{\theta=\theta_t} = H\mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \left(-\nabla_\theta \ln \pi_\theta(a_h | s_h) |_{\theta=\theta_t} \right)$$

$$= -H\mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \frac{\nabla_\theta \pi_{\theta_t}(a | s)}{\pi_{\theta_t}(a | s)} = 0$$

Let's compute the Hessian of the KL-divergence at θ_t

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

Let's compute the Hessian of the KL-divergence at θ_t

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

$$\frac{1}{H} \nabla_\theta^2 \ell(\theta) |_{\theta=\theta_t} = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \sum_a \pi_{\theta_t}(a | s) \Big(- \nabla_\theta^2 \ln \pi_\theta(a | s) |_{\theta=\theta_t} \Big)$$

Let's compute the Hessian of the KL-divergence at θ_t

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

$$\nabla \log \pi_\theta = \frac{\nabla \pi_\theta}{\pi_\theta}$$

$$\frac{1}{H}\nabla_\theta^2 \ell(\theta)|_{\theta=\theta_t} = \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\sum_a \pi_{\theta_t}(a|s)\left(-\nabla_\theta^2 \ln \pi_\theta(a|s)|_{\theta=\theta_t}\right)$$

$$\nabla^2 \log \pi_\theta = \frac{\nabla^2 \pi_\theta}{\pi_\theta} + \nabla \pi_\theta \nabla\left(\frac{1}{\pi_\theta}\right)$$

$$= -\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\sum_a \underbrace{\pi_{\theta_t}(\cancel{a|s})\left(\frac{\nabla_\theta^2 \pi_{\theta_t}(a|s)}{\pi_{\theta_t}(\cancel{a|s})}\right.}_{0} - \left.\frac{\nabla_\theta \pi_{\theta_t}(a|s)\nabla_\theta \pi_{\theta_t}(a|s)^\top}{\pi_{\theta_t}^2(a|s)}\right)$$

$$\sum_a \nabla^2 \pi_\theta(a|s) = \nabla^2 \sum_a \pi_\theta(a|s) = \nabla^2 1 = 0$$

Let’s compute the Hessian of the KL-divergence at θ_t

$$H \cdot \mathbb{E}_{s,a\sim d_\mu^{\pi_{\theta_t}}}\left[\ln \frac{\pi_{\theta_t}(a_h \mid s_h)}{\pi_\theta(a_h \mid s_h)}\right] := \ell(\theta)$$

$$\frac{1}{H}\nabla_\theta^2 \ell(\theta)\,|_{\theta=\theta_t} = \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\sum_a \pi_{\theta_t}(a \mid s)\Big(-\nabla_\theta^2 \ln \pi_\theta(a \mid s)\,|_{\theta=\theta_t}\Big)$$

$$= -\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\sum_a \pi_{\theta_t}(a \mid s)\left(\frac{\nabla_\theta^2 \pi_{\theta_t}(a \mid s)}{\pi_{\theta_t}(a \mid s)} - \frac{\nabla_\theta \pi_{\theta_t}(a \mid s)\,\nabla_\theta \pi_{\theta_t}(a \mid s)^\top}{\pi_{\theta_t}^2(a \mid s)}\right)$$

$$= \mathbb{E}_{s,a\sim d_\mu^{\pi_{\theta_t}}}\left[\nabla_\theta \ln \pi_{\theta_t}(a \mid s)\Big(\nabla_\theta \ln \pi_{\theta_t}(a \mid s)\Big)^\top\right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

Let's compute the Hessian of the KL-divergence at θ_t

$$H \cdot \mathbb{E}_{s,a\sim d_\mu^{\pi_{\theta_t}}}\left[\ln \frac{\pi_{\theta_t}(a_h \mid s_h)}{\pi_\theta(a_h \mid s_h)}\right] := \ell(\theta)$$

$$\frac{1}{H}\nabla_\theta^2 \ell(\theta)\mid_{\theta=\theta_t} = \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\sum_a \pi_{\theta_t}(a \mid s)\left(-\nabla_\theta^2 \ln \pi_\theta(a \mid s)\mid_{\theta=\theta_t}\right)$$

$$= -\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\sum_a \pi_{\theta_t}(a \mid s)\left(\frac{\nabla_\theta^2 \pi_{\theta_t}(a \mid s)}{\pi_{\theta_t}(a \mid s)} - \frac{\nabla_\theta \pi_{\theta_t}(a \mid s)\,\nabla_\theta \pi_{\theta_t}(a \mid s)^\top}{\pi_{\theta_t}^2(a \mid s)}\right)$$

$$= \mathbb{E}_{s,a\sim d_\mu^{\pi_{\theta_t}}}\left[\nabla_\theta \ln \pi_{\theta_t}(a \mid s)\left(\nabla_\theta \ln \pi_{\theta_t}(a \mid s)\right)^\top\right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

It's called fisher Information Matrix!

M

$\forall x$ PSD matrix

$x^\top M x \geq 0$

Family of dist.

$\{\pi(\cdot \mid s) \mid s \in S\}$

Summary so far:

We did second-order Taylor expansion on the KL constraint, and we get:

$$\frac{1}{H} KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \approx \frac{1}{2}(\theta - \theta_t)^\top F_{\theta_t}(\theta - \theta_t)$$

$$F_{\theta_t} := \mathbb{E}_{s,a \sim d_\mu^{\pi_{\theta_t}}}\left[\nabla_\theta \ln \pi_{\theta_t}(a|s)\left(\nabla_\theta \ln \pi_{\theta_t}(a|s)\right)^\top\right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

Summary so far:

We did second-order Taylor expansion on the KL constraint, and we get:

$$\frac{1}{H} KL \left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta} \right) \approx \frac{1}{2} (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t)$$

$$F_{\theta_t} := \mathbb{E}_{s,a \sim d_\mu^{\pi_{\theta_t}}} \left[\nabla_\theta \ln \pi_{\theta_t}(a|s) \left(\nabla_\theta \ln \pi_{\theta_t}(a|s) \right)^\top \right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

This leads to the following simplified constrained optimization:

Summary so far:

We did second-order Taylor expansion on the KL constraint, and we get:

$$\frac{1}{H} KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \approx \frac{1}{2}(\theta - \theta_t)^\top F_{\theta_t}(\theta - \theta_t)$$

$$F_{\theta_t} := \mathbb{E}_{s,a\sim d_\mu^{\pi_{\theta_t}}}\left[\nabla_\theta \ln \pi_{\theta_t}(a|s)\left(\nabla_\theta \ln \pi_{\theta_t}(a|s)\right)^\top\right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

This leads to the following simplified constrained optimization: "local"

$$\max_\theta \nabla_\theta J(\pi_{\theta_t})^\top(\theta - \theta_t)$$ ← TRPO objective.

s.t. $(\theta - \theta_t)^\top F_{\theta_t}(\theta - \theta_t) \le \delta$

Outlines

1. Quick intro on KL-divergence

2. A Trust-Region Formulation for Policy Optimization

3. Algorithm: Natural Policy Gradient

Put everything together, we get:

(dropping the H factors) At iteration t, we update to θ_{t+1} via:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \le \delta$$

Put everything together, we get:

(dropping the H factors) At iteration t, we update to θ_{t+1} via:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

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

Linear objective and quadratic convex constraint, we can solve it optimally!

Put everything together, we get:

(dropping the H factors) At iteration t, we update to θ_{t+1} via:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \le \delta$$

$\longrightarrow$ solution is θ_{t+1}

Linear objective and quadratic convex constraint, we can solve it optimally!

Indeed this gives us:

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})$$

Put everything together, we get:

(dropping the H factors) At iteration t, we update to θ_{t+1} via:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

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

F has units of $\frac{1}{\theta^2}$

Linear objective and quadratic convex constraint, we can solve it optimally!

Indeed this gives us:

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})$$

Where $\eta = \sqrt{\dfrac{\delta}{\nabla_\theta J(\pi_{\theta_t})^\top F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})}}$

Algorithm: Natural Policy Gradient

Initialize θ_0

For t = 0, …

Algorithm: Natural Policy Gradient

Initialize θ_0

For t = 0, …

Estimate PG $\nabla_\theta J(\pi_{\theta_t})$

Algorithm: Natural Policy Gradient

Initialize θ_0

For t = 0, …

Estimate PG $\nabla_\theta J(\pi_{\theta_t})$

Estimate Fisher info-matrix $F_{\theta_t} := \mathbb{E}_{s,a \sim d_\mu^{\pi_{\theta_t}}} \nabla_\theta \ln \pi_{\theta_t}(a \mid s) (\nabla_\theta \ln \pi_{\theta_t}(a \mid s))^\top$

Algorithm: Natural Policy Gradient

Initialize θ_0

For t = 0, …

Estimate PG $\nabla_\theta J(\pi_{\theta_t})$

Estimate Fisher info-matrix $F_{\theta_t} := \mathbb{E}_{s,a \sim d_\mu^{\pi_{\theta_t}}} \nabla_\theta \ln \pi_{\theta_t}(a \mid s)(\nabla_\theta \ln \pi_{\theta_t}(a \mid s))^\top$

Natural Gradient Ascent: $\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})$

Algorithm: Natural Policy Gradient

Initialize θ_0

For t = 0, …

Estimate PG $\widehat{\nabla_\theta J}(\pi_{\theta_t})$

Estimate Fisher info-matrix $F_{\theta_t} := \hat{\mathbb{E}}_{s,a\sim d_\mu^{\pi_{\theta_t}}} \nabla_\theta \ln \pi_{\theta_t}(a \mid s) \left(\nabla_\theta \ln \pi_{\theta_t}(a \mid s) \right)^\top$

Natural Gradient Ascent: $\theta_{t+1} = \theta_t + \eta \hat{F}_{\theta_t}^{-1} \hat{\nabla}_\theta J(\pi_{\theta_t})$

Where $\eta = \sqrt{\dfrac{\delta}{\nabla_\theta J(\pi_{\theta_t})^\top F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})}}$

Algorithm: Natural Policy Gradient

Initialize θ_0

For t = 0, …

Estimate PG $\nabla_\theta J(\pi_{\theta_t})$

Estimate Fisher info-matrix $F_{\theta_t} := \mathbb{E}_{s,a\sim d_\mu^{\pi_{\theta_t}}} \nabla_\theta \ln \pi_{\theta_t}(a \mid s)(\nabla_\theta \ln \pi_{\theta_t}(a \mid s))^\top$

Natural Gradient Ascent: $\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})$

Where $\eta = \sqrt{\dfrac{\delta}{\nabla_\theta J(\pi_{\theta_t})^\top F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})}}$

(We will implement it in HW4 on Cartpole)

Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1 + \exp(\theta)}, \frac{1}{1 + \exp(\theta)} \right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Example of Natural Gradient on 1-d problem:

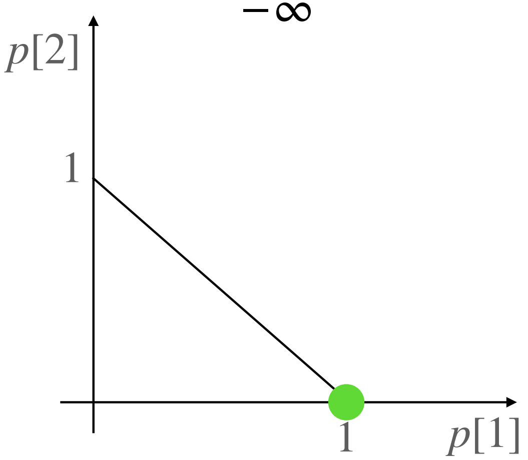
$$p_\theta = \left(\frac{\exp(\theta)}{1 + \exp(\theta)}, \frac{1}{1 + \exp(\theta)} \right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

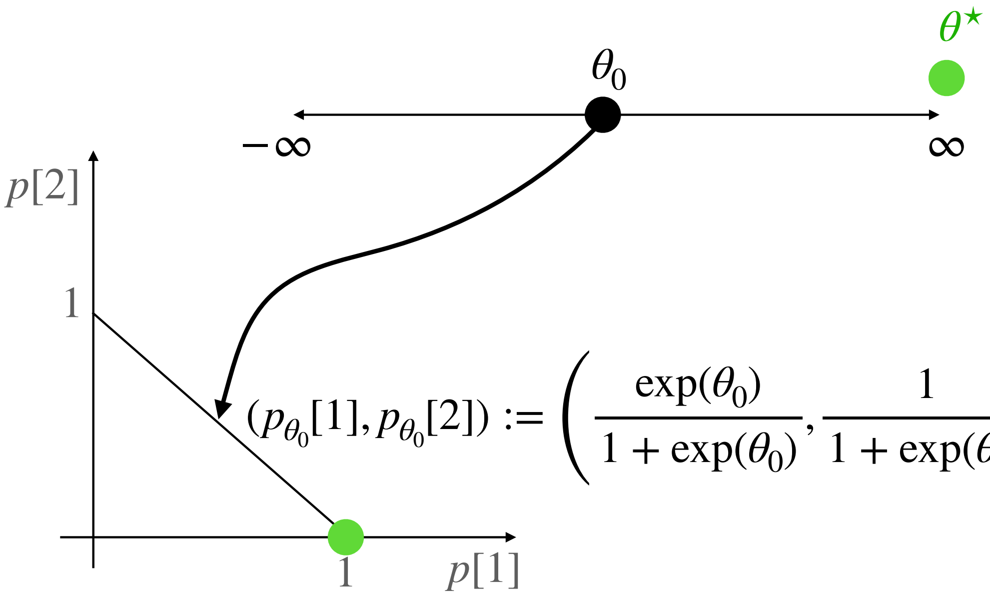
$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$



Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1 + \exp(\theta)}, \frac{1}{1 + \exp(\theta)} \right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$



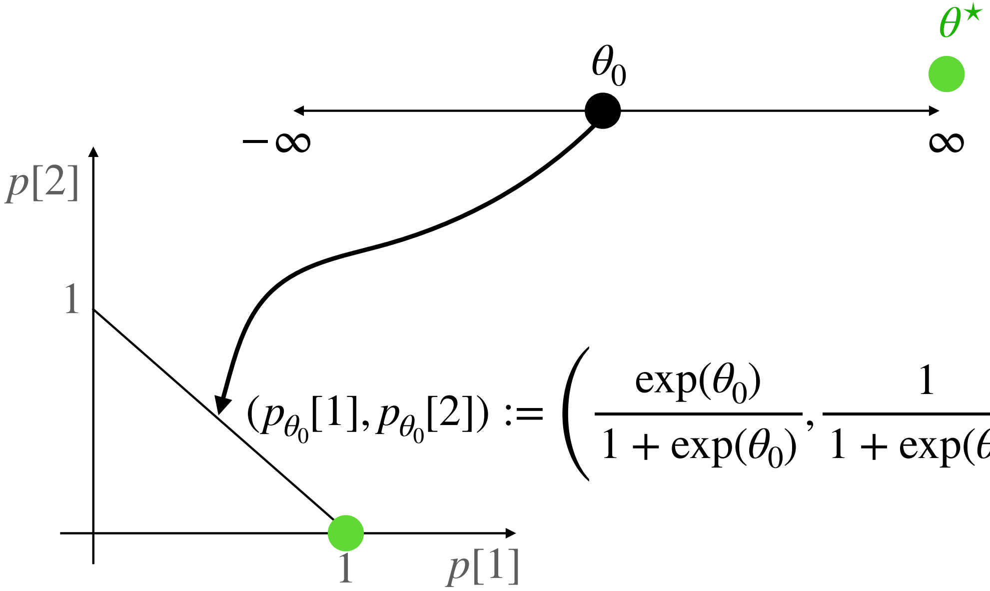
$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1 + \exp(\theta_0)}, \frac{1}{1 + \exp(\theta_0)} \right)$$

Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)} \right)$$

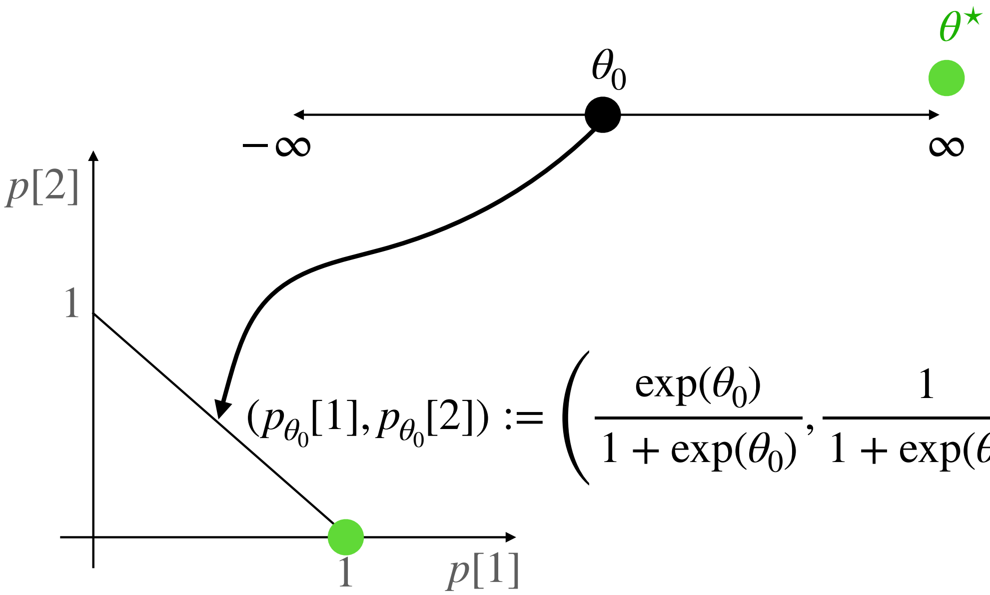
Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

Hence: $f_{\theta_0} \to 0^+$, as $\theta_0 \to \infty$

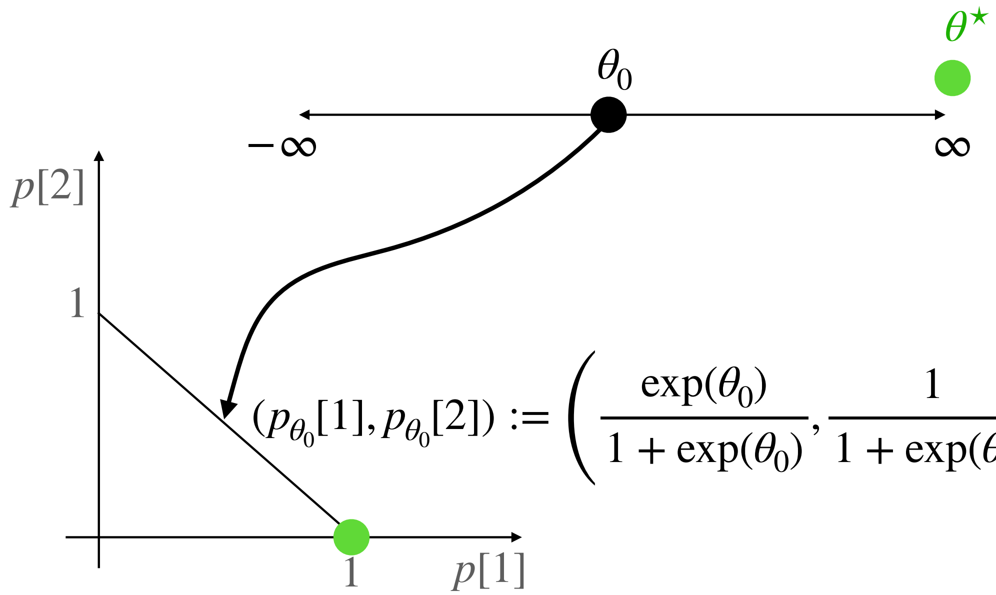


$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)} \right)$$

Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)}\right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)}\right)$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

Hence: $f_{\theta_0} \to 0^+$, as $\theta_0 \to \infty$

NPG: $\theta_1 = \theta_0 + \eta \dfrac{g'(\theta_0)}{f_{\theta_0}}$

Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

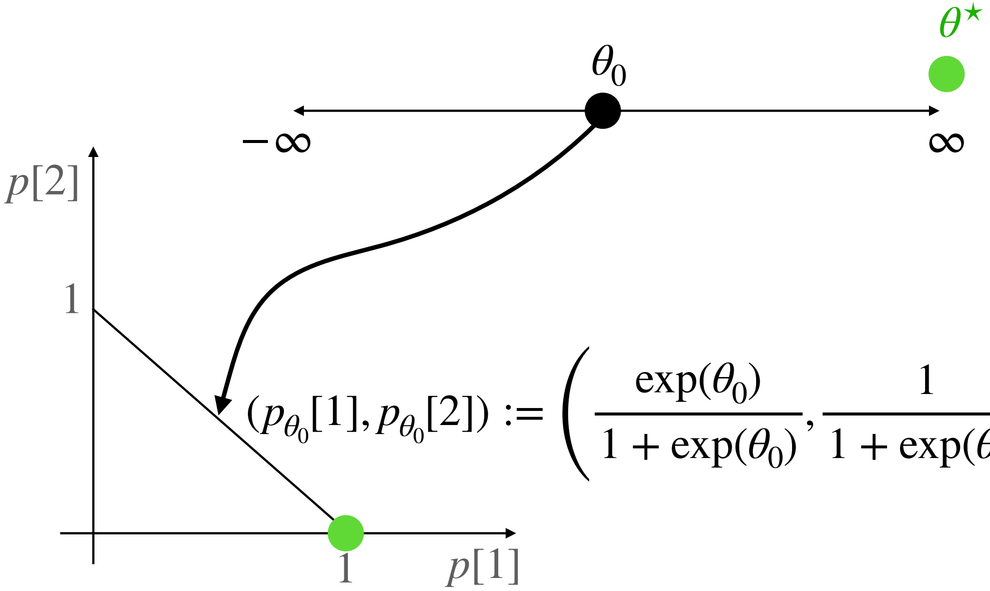
$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

Hence: $f_{\theta_0} \to 0^+$, as $\theta_0 \to \infty$

NPG: $\theta_1 = \theta_0 + \eta \dfrac{g'(\theta_0)}{f_{\theta_0}}$

GA: $\theta_1 = \theta_0 + \eta g'(\theta_0)$

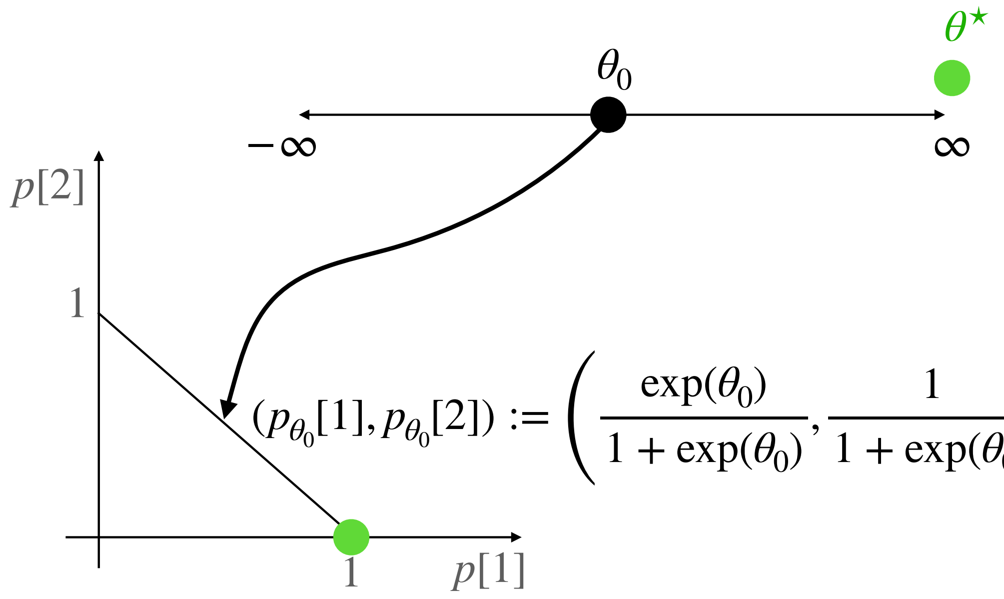


$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)} \right)$$

Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)} \right)$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

Hence: $f_{\theta_0} \to 0^+$, as $\theta_0 \to \infty$

NPG: $\theta_1 = \theta_0 + \eta \dfrac{g'(\theta_0)}{f_{\theta_0}}$

GA: $\theta_1 = \theta_0 + \eta g'(\theta_0)$

i.e., Plain GA in θ will move to $\theta = \infty$ at a constant speed,
while Natural GA can traverse faster and faster when θ gets bigger
(subject to the same learning rate)

Summary for NPG:

Trust Region Policy Optimization and NPG

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$$

Intuition: maximize local adv subject to being incremental (in KL);

Summary for NPG:

Trust Region Policy Optimization and NPG

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right] \longrightarrow$$ First-order Taylor expansion at θ_t

s.t., $KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta \longrightarrow$ second-order Taylor expansion at θ_t

Intuition: maximize local adv subject to being incremental (in KL);

Summary for NPG:

Trust Region Policy Optimization and NPG

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right] \longrightarrow$$ First-order Taylor expansion at θ_t

s.t., $KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta \longrightarrow$ second-order Taylor expansion at θ_t

Intuition: maximize local adv subject to being incremental (in KL);

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

s.t. $(\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$

Summary for NPG:

Trust Region Policy Optimization and NPG

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right] \longrightarrow$$ First-order Taylor expansion at θ_t

s.t., $KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta \longrightarrow$ second-order Taylor expansion at θ_t

Intuition: maximize local adv subject to being incremental (in KL);

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

s.t. $(\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$

(Exercise: work out the $\arg\max_\theta$)

Summary for NPG:

Trust Region Policy Optimization and NPG

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a) \right] \longrightarrow$$ First-order Taylor expansion at θ_t

s.t., $KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta \longrightarrow$ second-order Taylor expansion at θ_t

Intuition: maximize local adv subject to being incremental (in KL);

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

s.t. $(\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t}) \longleftarrow$$

NPG

(Exercise: work out the $\arg\max_\theta$)

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s, a) \right]$$

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \right] \underbrace{- \lambda \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[\mathrm{KL}\left(\pi_{\theta_t}(a|s) | \pi_\theta(a|s) \right) \right]}_{\text{regularization}}$$

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \right] \underbrace{- \lambda \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[\text{KL}\left(\pi_{\theta_t}(a\,|\,s) \,|\, \pi_\theta(a\,|\,s) \right) \right]}_{\text{regularization}}$$

Use importance weighting & expand KL divergence:

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right] \underbrace{-\lambda \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\mathrm{KL}\left(\pi_{\theta_t}(a\,|\,s)\,|\,\pi_\theta(a\,|\,s)\right)\right]}_{\text{regularization}}$$

Use importance weighting & expand KL divergence:

$$\ell(\theta) := \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \frac{\pi_\theta(a\,|\,s)}{\pi_{\theta_t}(a\,|\,s)} A^{\pi_{\theta_t}}(s,a)\right] - \lambda \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)}\left[-\ln \pi_\theta(a\,|\,s)\right]$$

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right] \underbrace{-\lambda \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\mathrm{KL}\left(\pi_{\theta_t}(a|s) | \pi_\theta(a|s)\right)\right]}_{\text{regularization}}$$

Use importance weighting & expand KL divergence:

$$\ell(\theta) := \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a)\right] - \lambda \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)}\left[-\ln \pi_\theta(a|s)\right]$$

PPO: Perform a few steps of mini-batch SGA on $\ell(\theta)$ to approximate $\arg\max_{\theta} \ell(\theta)$

Next a few lectures:

Imitation Learning
(Learning from Demonstrations)

Can we learn a good policy purely from expert demonstrations?

Summary:

1. Convergence of Fitted Policy Iteration
2. Trust Region Policy Optimization
    1. Quick intro on KL-divergence
    2. TRPO formulation

1-minute feedback form: https://bit.ly/3RHtlxy