NEURAL INFORMATION PROCESSING SYSTEMS

Ordering-based Conditions for Global Convergence of Policy Gradient Methods

Jincheng Mei, Bo Dai, Alekh Agarwal,

Mohammad Ghavamzadeh, Csaba Szepesvari, Dale Schuurmans

Google DeepMind, Google Research, University of Alberta





Key Message

Standard Softmax Policy Gradient (PG) and Natural Policy Gradient (NPG) can achieve global convergence with non-zero approximation errors

Key Message

Standard Softmax Policy Gradient (PG) and Natural Policy Gradient (NPG) can achieve **global convergence** with **non-zero approximation errors**

Problem and parameterization

Policy optimization: $\max_{\theta \in \mathbb{R}^{d}} \pi_\theta^\top r$ $\qquad r \in \mathbb{R}^{K}$

Problem and parameterization

Policy optimization: $\max_{\theta \in \mathbb{R}^{d}} \pi_\theta^\top r$ $\quad r \in \mathbb{R}^{K}$

Softmax + low-dimensional feature (“log-linear policies”): $\pi_\theta = \mathrm{softmax}(X\theta)$

$$\pi_\theta(a) = \frac{\exp\{[X\theta](a)\}}{\sum_{a' \in [K]} \exp\{[X\theta](a')\}} \qquad X \in \mathbb{R}^{K \times d}$$

Problem and parameterization

Policy optimization: $\max_{\theta \in \mathbb{R}^d} \pi_\theta^\top r \qquad r \in \mathbb{R}^K$

Softmax + low-dimensional feature ("log-linear policies"): $\pi_\theta = \mathrm{softmax}(X\theta)$

$$\pi_\theta(a) = \frac{\exp\{[X\theta](a)\}}{\sum_{a' \in [K]} \exp\{[X\theta](a')\}} \qquad X \in \mathbb{R}^{K \times d}$$

Used in practice but **hard to analyze**

- non-concave maximization (softmax transform)
- not realizable if $d < K$ ($\pi_\theta = \mathrm{softmax}(X\theta)$, and $X\theta$ not equal $r \in \mathbb{R}^K$)

Algorithms

Problem: $$\max_{\theta \in \mathbb{R}^d} \pi_\theta^\top r$$

Algorithm 1 Softmax policy gradient (PG)

Input: Learning rate $\eta > 0$.
Output: Policies $\pi_{\theta_t} = \text{softmax}(X\theta_t)$.
Initialize parameter $\theta_1 \in \mathbb{R}^d$.
while $t \geq 1$ **do**
  $\theta_{t+1} \leftarrow \theta_t + \eta \cdot X^\top(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top)r$.
end while

Algorithm 2 Natural policy gradient (NPG)

Input: Learning rate $\eta > 0$.
Output: Policies $\pi_{\theta_t} = \text{softmax}(X\theta_t)$.
Initialize parameter $\theta_1 \in \mathbb{R}^d$.
while $t \geq 1$ **do**
  $\theta_{t+1} \leftarrow \theta_t + \eta \cdot (X^\top X)^{-1} X^\top r$.
end while

Softmax Policy Gradient (PG); Natural Policy Gradient (NPG)

Problem: $\max_{\theta \in \mathbb{R}^{d}} \pi_\theta^\top r$

Algorithm 1 Softmax policy gradient (PG)

Input: Learning rate $\eta > 0$.
Output: Policies $\pi_{\theta_t} = \text{softmax}(X\theta_t)$.
Initialize parameter $\theta_1 \in \mathbb{R}^d$.
while $t \geq 1$ **do**
  $\theta_{t+1} \leftarrow \theta_t + \eta \cdot X^\top (\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t} \pi_{\theta_t}^\top) r$.
end while

Algorithm 2 Natural policy gradient (NPG)

Input: Learning rate $\eta > 0$.
Output: Policies $\pi_{\theta_t} = \text{softmax}(X\theta_t)$.
Initialize parameter $\theta_1 \in \mathbb{R}^d$.
while $t \geq 1$ **do**
  $\theta_{t+1} \leftarrow \theta_t + \eta \cdot (X^\top X)^{-1} X^\top r$.
end while

$$\frac{d\,\pi_{\theta_t}^\top r}{d\theta_t} = \frac{d\,X\theta_t}{d\theta_t} \left(\frac{d\,\pi_{\theta_t}}{d\,X\theta_t} \right)^\top \frac{d\,\pi_{\theta_t}^\top r}{d\pi_{\theta_t}} = X^\top (\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t} \pi_{\theta_t}^\top)\, r$$

$$\left(X^\top X\right)^{-1} X^\top r = \arg\min_{w \in \mathbb{R}^d} \|Xw - r\|_2^2$$

Existing results

Problem: $\max_{\theta \in \mathbb{R}^d} \pi_\theta^\top r$ $\pi_\theta = \mathrm{softmax}(\theta)$

Softmax PG: asymptotic global convergence (Agarwal et al., 2019)

O(1/t) rate (Mei et al., 2020)

Poor constant dependence (Li et al., 2021)

$\pi_\theta = \mathrm{softmax}(X\theta)$: impossible to achieve global convergence, exponentially many bad local maxima (Chen et al., 2020).

Existing results

Problem: $\max_{\theta \in \mathbb{R}^d} \pi_\theta^\top r$ $\quad\quad \pi_\theta = \mathrm{softmax}(\theta)$

NPG: $O(1/t)$ global convergence (Agarwal et al., 2019)

$O(e^{-c \cdot t})$ rate (Khodadadia et al., 2021; Lan 2021; Xiao, 2022)

$\pi_\theta = \mathrm{softmax}(X\theta)$: additive approximation error (Agarwal et al., 2019)

$$(\pi^* - \pi_{\theta_t})^\top r \leq c_1/\sqrt{t} + c_2 \cdot \epsilon_{\mathrm{approx}}$$

Examples

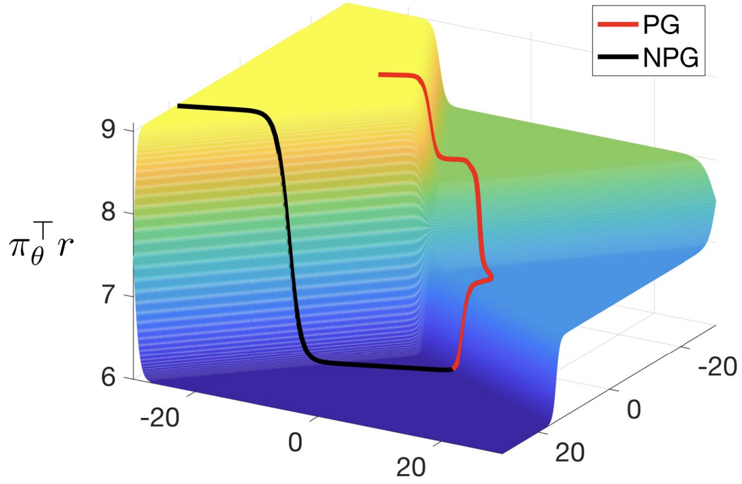
Example 1. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

$$\epsilon_{approx} = \min_{w \in \mathbb{R}^d} \|Xw - r\|_2 = \left\| X \left(X^\top X \right)^{-1} X^\top r - r \right\|_2 = \sqrt{202.6} \approx 14.2338$$

Examples

Example 1. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

$$\epsilon_{approx} = \min_{w \in \mathbb{R}^d} \|Xw - r\|_2 = \left\|X\left(X^\top X\right)^{-1} X^\top r - r\right\|_2 = \sqrt{202.6} \approx 14.2338$$

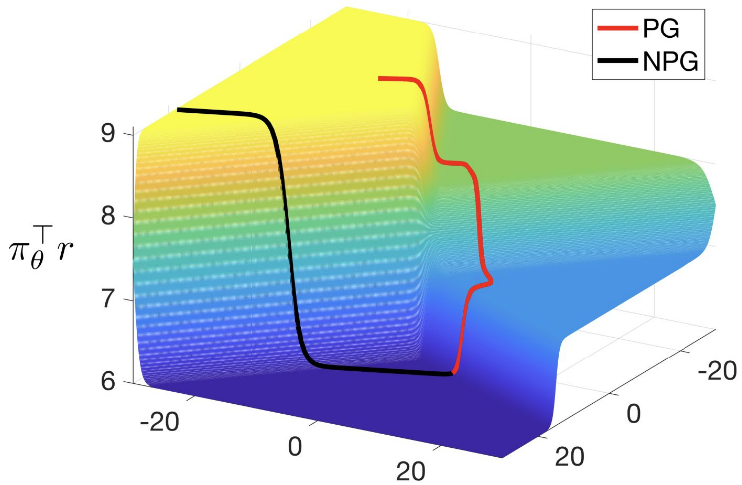


$$\max_{\theta \in \mathbb{R}^{d}} \pi_\theta^\top r \qquad \pi_\theta = \mathrm{softmax}(X\theta)$$

Examples

Example 1. $K = 4,\ d = 2,\ X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

$$\epsilon_{approx} = \min_{w \in \mathbb{R}^d} \|Xw - r\|_2 = \left\| X \left(X^\top X\right)^{-1} X^\top r - r \right\|_2 = \sqrt{202.6} \approx 14.2338$$



$$\max_{\theta \in \mathbb{R}^{d}} \pi_\theta^\top r \qquad \pi_\theta = \mathrm{softmax}(X\theta)$$

$$\text{Softmax PG: } \theta_{t+1} \leftarrow \theta_t + \eta \cdot X^\top (\mathrm{diag}(\pi_{\theta_t}) - \pi_{\theta_t} \pi_{\theta_t}^\top) r$$

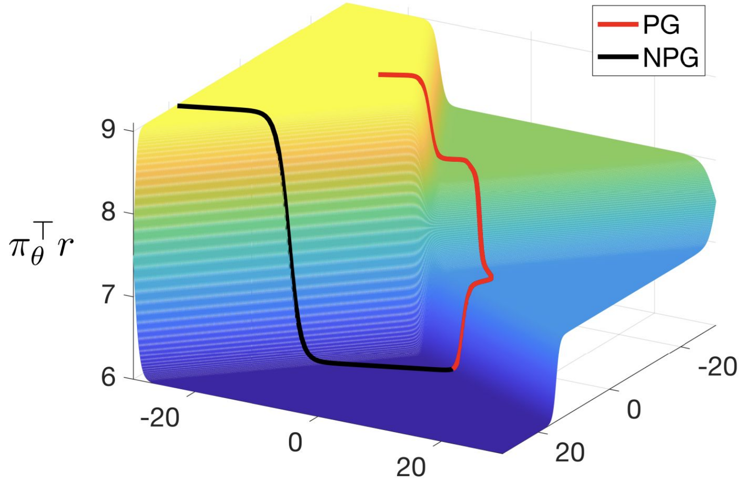
$$\text{NPG: } \theta_{t+1} \leftarrow \theta_t + \eta \cdot (X^\top X)^{-1} X^\top r$$

$$\theta_1 = (6, 8)^\top \in \mathbb{R}^2 \qquad \eta = 0.2$$

Examples

Example 1. $K = 4,\ d = 2,\ X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

$$\epsilon_{approx} = \min_{w \in \mathbb{R}^d} \|Xw - r\|_2 = \left\|X\left(X^\top X\right)^{-1} X^\top r - r\right\|_2 = \sqrt{202.6} \approx 14.2338$$



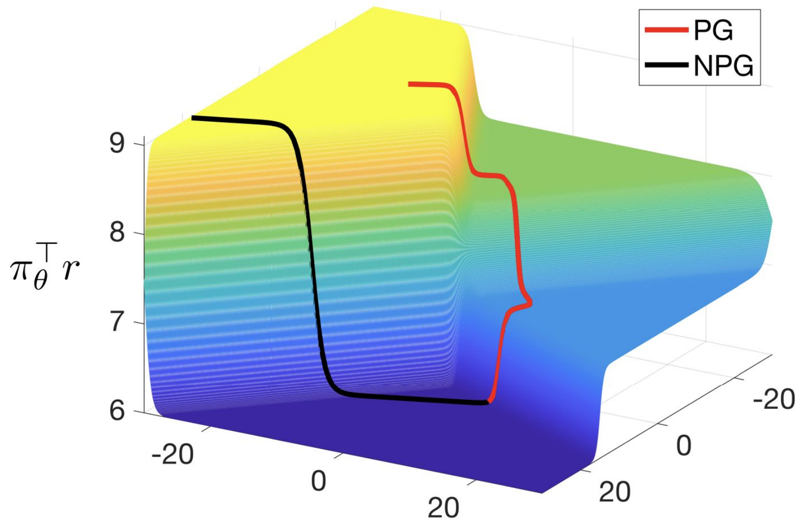
$$\max_{\theta \in \mathbb{R}^{d}} \pi_\theta^\top r \qquad \pi_\theta = \mathrm{softmax}(X\theta)$$

$$\theta_1 = (6, 8)^\top \in \mathbb{R}^2 \qquad \eta = 0.2$$

$$\pi_{\theta_t}^\top r \to 9 = r(a^*)$$

Findings

Finding 1: Softmax PG and NPG can achieve global convergence with non-zero approximation errors.



Example 1. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

Findings

Finding 1: Softmax PG and NPG can achieve global convergence with non-zero approximation errors.

Question: Is non-zero approximation error useful for characterizing global convergence?

Examples

Example 1. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

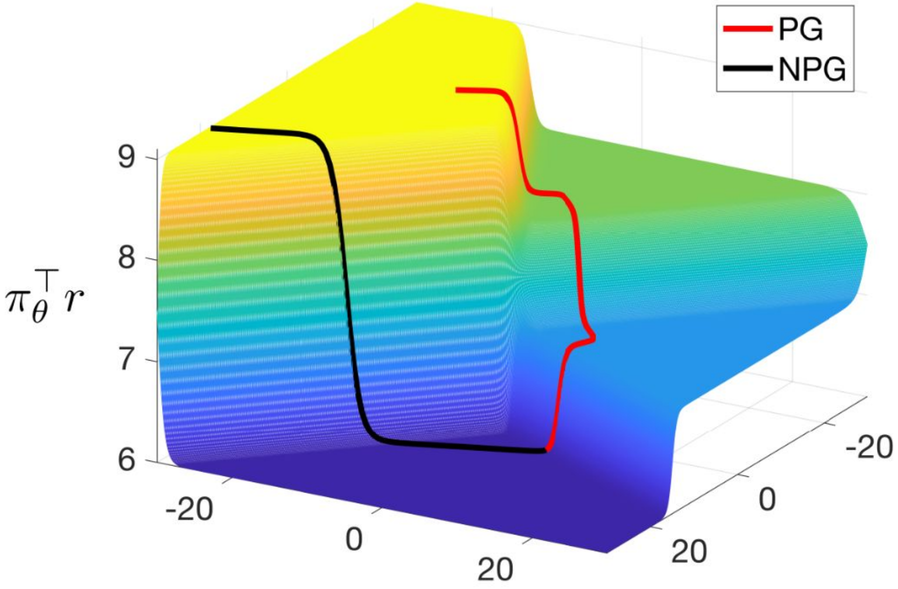
Example 2. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & 0 & -1 & 2 \\ -2 & 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{d \times K}$, *and* $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$

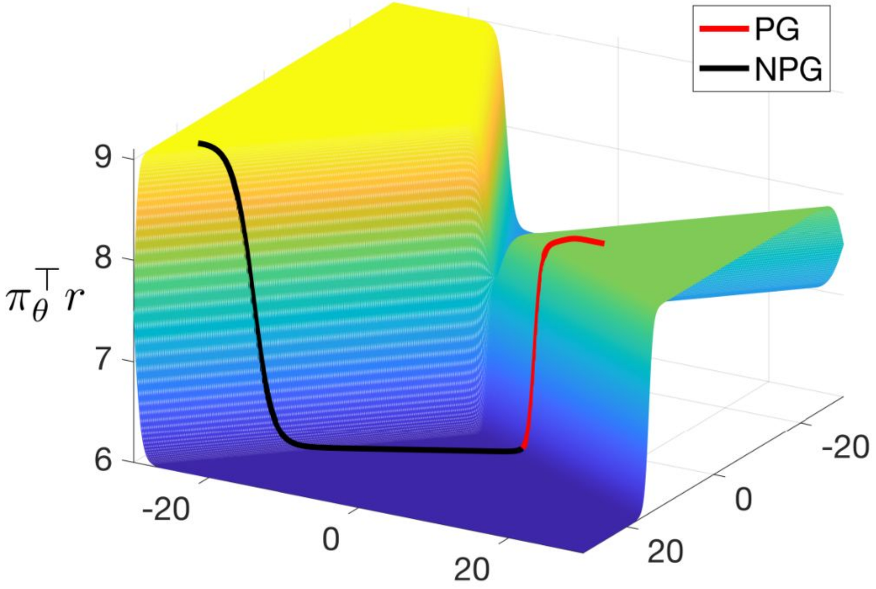
$$\left\| X \left(X^\top X\right)^{-1} X^\top r - r \right\|_2 = \sqrt{205} \approx 14.3178$$

Examples

Example 1. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

Example 2. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & 0 & -1 & 2 \\ -2 & 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{d \times K}$, *and* $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$





Examples

Example 1. $K = 4,\ d = 2,\ X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

Example 2. $K = 4,\ d = 2,\ X^\top = \begin{bmatrix} 0 & 0 & -1 & 2 \\ -2 & 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{d\times K}$, *and* $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$

Example 3. $K = 4,\ d = 2,\ X^\top = \begin{bmatrix} -1 & 0 & 0 & 2 \\ 0 & -2 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{d\times K}$, *and* $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$

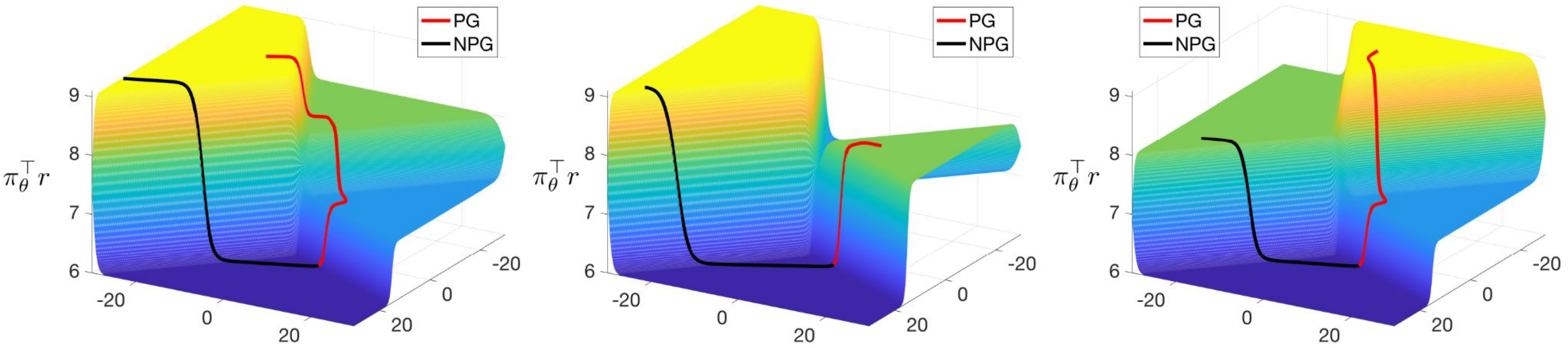
$$\left\| X \left(X^\top X\right)^{-1} X^\top r - r \right\|_2 = \sqrt{212} \approx 14.5602$$

Examples

Example 1. $K = 4,\, d = 2,\, X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

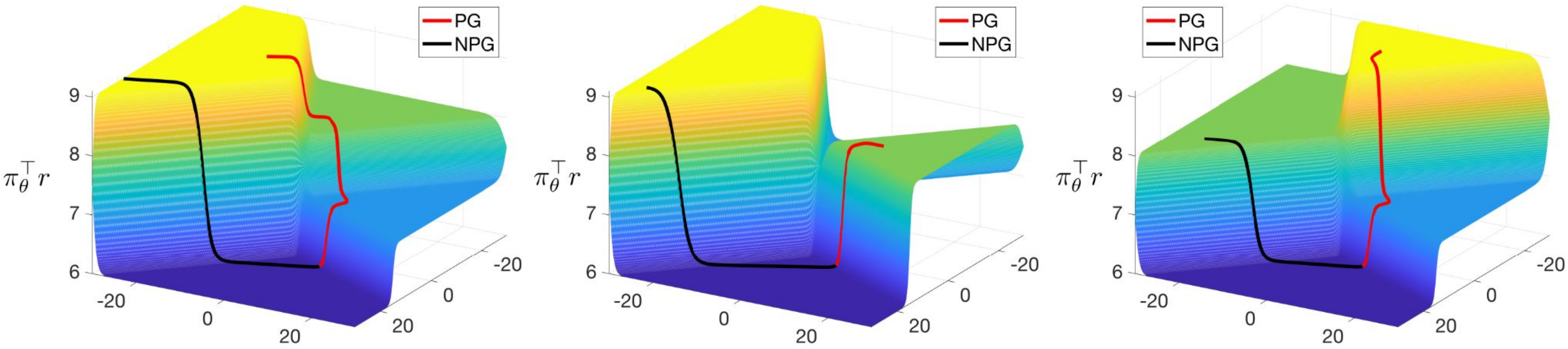
Example 2. $K = 4,\, d = 2,\, X^\top = \begin{bmatrix} 0 & 0 & -1 & 2 \\ -2 & 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{d \times K}$, *and* $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$

Example 3. $K = 4,\, d = 2,\, X^\top = \begin{bmatrix} -1 & 0 & 0 & 2 \\ 0 & -2 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{d \times K}$, *and* $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$



Findings

Finding 2: Non-zero approximation errors does not characterize global convergence for both algorithms.



Examples

Proposition: $K = 3, \; d = 2$ $\quad X^\top = \begin{bmatrix} 0 & -10 & 0 \\ -2 & 4 & 1 \end{bmatrix} \in \mathbb{R}^{d \times K}$

$r = Xw = (4, 2, -2)^\top \quad w = (-1, -2)^\top \in \mathbb{R}^d$

Bad initialization: $\theta_1 = (-\ln 2, \ln 2)^\top$, using Softmax PG

$\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} < \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < \cdots < \frac{\pi_{\theta_1}(1)}{\pi_{\theta_1}(3)} < \frac{1}{2}$, implying that $\pi_{\theta_t}(1) \not\to 1$.

Findings

Finding 3: Linear realizability (zero approximation error) does not imply global convergence for Softmax PG.

Findings

Finding 1: Softmax PG and NPG can achieve global convergence with non-zero approximation errors.

Finding 2: Non-zero approximation errors does not characterize global convergence for both algorithms.

Finding 3: Linear realizability (zero approximation error) does not imply global convergence for Softmax PG.

Findings

Finding 1: Softmax PG and NPG can achieve global convergence with non-zero approximation errors.

Finding 2: Non-zero approximation errors does not characterize global convergence for both algorithms.

Finding 3: Linear realizability (zero approximation error) does not imply global convergence for Softmax PG.

Question: What conditions characterize global convergence of Softmax PG and NPG in unrealizable problem?

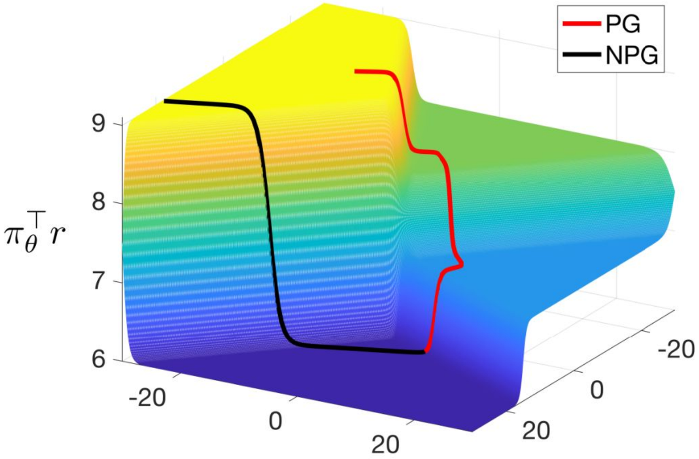
Main Results (Softmax PG)

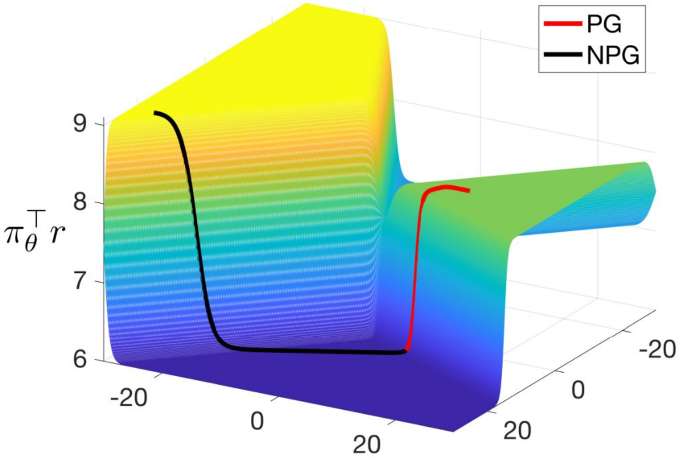
Softmax PG (sufficient, not necessary):

Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of feature matrix $X \in \mathbb{R}^{K \times d}$.
If (i) there exists at least one $w \in \mathbb{R}^d$, such that $r' := Xw \in \mathbb{R}^K$ preserves the ordering of $r \in \mathbb{R}^K$, i.e., $r(i) > r(j)$ if and only if $r'(i) > r'(j)$; (ii) $(x_i - x_j)^\top (x_{a^*} - x_j) \geq 0$ for all $r(a^*) > r(i) > r(j)$.

Main Results (Softmax PG)

Softmax PG: Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of feature matrix $X \in \mathbb{R}^{K \times d}$. If (i) there exists at least one $w \in \mathbb{R}^d$, such that $r' := Xw \in \mathbb{R}^K$ preserves the ordering of $r \in \mathbb{R}^K$, i.e., $r(i) > r(j)$ if and only if $r'(i) > r'(j)$; (ii) $(x_i - x_j)^\top (x_{a^*} - x_j) \geq 0$ for all $r(a^*) > r(i) > r(j)$.





Main Results (Softmax PG)

Softmax PG: Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of feature matrix $X \in \mathbb{R}^{K \times d}$. If (i) there exists at least one $w \in \mathbb{R}^d$, such that $r' := Xw \in \mathbb{R}^K$ preserves the ordering of $r \in \mathbb{R}^K$, i.e., $r(i) > r(j)$ if and only if $r'(i) > r'(j)$; (ii) $(x_i - x_j)^\top (x_{a^*} - x_j) \geq 0$ for all $r(a^*) > r(i) > r(j)$.

Example 1. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

$w = (-1, -1)^\top \in \mathbb{R}^d$ $\quad$ $r' := Xw = (2, 1, -1, -2)^\top \in \mathbb{R}^K$

Main Results (Softmax PG)

Softmax PG: Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of feature matrix $X \in \mathbb{R}^{K \times d}$.
If (i) there exists at least one $w \in \mathbb{R}^d$, such that $r' := Xw \in \mathbb{R}^K$ preserves the ordering of $r \in \mathbb{R}^K$, i.e., $r(i) > r(j)$ if and only if $r'(i) > r'(j)$; (ii) $(x_i - x_j)^\top (x_{a^*} - x_j) \geq 0$ for all $r(a^*) > r(i) > r(j)$.

Example 1. $K = 4,\ d = 2,\ X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ *and* $r = (9, 8, 7, 6)^\top$

$w = (-1, -1)^\top \in \mathbb{R}^d$ $\quad\quad r' := Xw = (2, 1, -1, -2)^\top \in \mathbb{R}^K$

Example 2. $K = 4,\ d = 2,\ X^\top = \begin{bmatrix} 0 & 0 & -1 & 2 \\ -2 & 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{d \times K}$, *and* $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$

$w = (w(1), w(2))^\top$ $\quad\quad r' := Xw = (-2 \cdot w(2), w(2), -w(1), 2 \cdot w(1))^\top$

Main Results (Softmax PG)

Softmax PG: Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of feature matrix $X \in \mathbb{R}^{K \times d}$. If (i) there exists at least one $w \in \mathbb{R}^d$, such that $r' := Xw \in \mathbb{R}^K$ preserves the ordering of $r \in \mathbb{R}^K$, i.e., $r(i) > r(j)$ if and only if $r'(i) > r'(j)$; (ii) $(x_i - x_j)^\top (x_{a^*} - x_j) \geq 0$ for all $r(a^*) > r(i) > r(j)$.

(i) ensures “no finite stationary points”, implying that $\|\theta_t\|_2 \to \infty$.

$$w^\top X^\top \left(\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top\right) r = {r'}^\top \left(\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top\right) r$$

$$= \sum_{i=1}^{K-1} \pi_\theta(i) \cdot \sum_{j=i+1}^{K} \pi_\theta(j) \cdot (r'(i) - r'(j)) \cdot (r(i) - r(j))$$

(ii) avoids existence of bad local maxima on sub-optimal plateaus.

Main Results (NPG)

NPG (sufficient and necessary):

Denote $\hat{r} := X\left(X^\top X\right)^{-1} X^\top r$, a sufficient and necessary condition for NPG to achieve $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ (from any initialization $\theta_1 \in \mathbb{R}^d$) is that $\hat{r}(a^*) > \hat{r}(a)$ for all $a \neq a^*$ such that $a^* := \mathrm{argmax}_{a \in [K]} r(a)$.

If the condition is satisfied, then the rate is $(\pi^* - \pi_{\theta_t})^\top r \in O(e^{-c \cdot t})$.

Main Results (NPG)

NPG: Denote $\hat{r} := X\left(X^\top X\right)^{-1} X^\top r$, a sufficient and necessary condition for NPG to achieve $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ (from any initialization $\theta_1 \in \mathbb{R}^d$) is that $\hat{r}(a^*) > \hat{r}(a)$ for all $a \neq a^*$ such that $a^* := \operatorname{argmax}_{a \in [K]} r(a)$.

Intuition (using Example 1):

- $$\frac{\pi_{\theta_{t+1}}(a^*)}{\pi_{\theta_{t+1}}(a)} = \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} \cdot \exp\left\{\eta \cdot (\hat{r}(a^*) - \hat{r}(a))\right\} = \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} \cdot \exp\left\{\eta \cdot \frac{26}{5}\right\}$$

Softmax PG condition is sufficient but not necessary

Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of feature matrix $X \in \mathbb{R}^{K \times d}$.
If (i) there exists at least one $w \in \mathbb{R}^d$, such that $r' := Xw \in \mathbb{R}^K$ preserves the ordering of $r \in \mathbb{R}^K$, i.e., $r(i) > r(j)$ if and only if $r'(i) > r'(j)$; (ii) $(x_i - x_j)^\top (x_{a^*} - x_j) \geq 0$ for all $r(a^*) > r(i) > r(j)$.

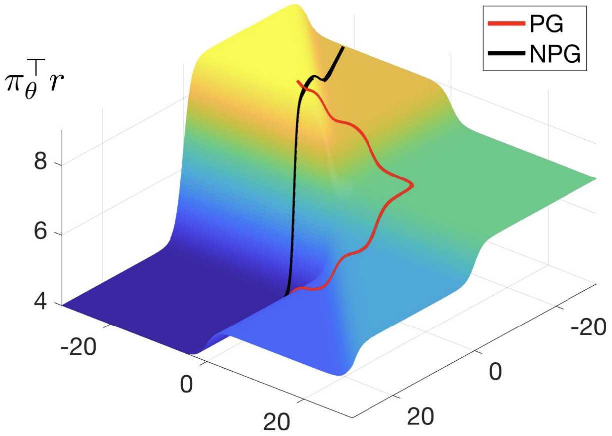
Example 5. $K = 6$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & -1 & 0 & 1 & 1 \\ -1 & 0 & 1 & 1 & 0 & -1 \end{bmatrix}$, *and* $r = (9, 8, 7, 6, 5, 4)^\top$.

Softmax PG condition is sufficient but not necessary

Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of feature matrix $X \in \mathbb{R}^{K \times d}$. If (i) there exists at least one $w \in \mathbb{R}^d$, such that $r' := Xw \in \mathbb{R}^K$ preserves the ordering of $r \in \mathbb{R}^K$, i.e., $r(i) > r(j)$ if and only if $r'(i) > r'(j)$; (ii) $(x_i - x_j)^\top (x_{a^*} - x_j) \geq 0$ for all $r(a^*) > r(i) > r(j)$.

Example 5. $K = 6$, $d = 2$,

$$X^\top = \begin{bmatrix} 0 & -1 & -1 & 0 & 1 & 1 \\ -1 & 0 & 1 & 1 & 0 & -1 \end{bmatrix}, \textit{ and } r = (9, 8, 7, 6, 5, 4)^\top.$$



Softmax PG condition is sufficient but not necessary

Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of feature matrix $X \in \mathbb{R}^{K \times d}$.
If (i) there exists at least one $w \in \mathbb{R}^d$, such that $r' := Xw \in \mathbb{R}^K$ preserves the ordering of $r \in \mathbb{R}^K$, i.e., $r(i) > r(j)$ if and only if $r'(i) > r'(j)$; (ii) $(x_i - x_j)^\top (x_{a^*} - x_j) \geq 0$ for all $r(a^*) > r(i) > r(j)$.

Example 5. $K = 6$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & -1 & 0 & 1 & 1 \\ -1 & 0 & 1 & 1 & 0 & -1 \end{bmatrix}$, *and* $r = (9, 8, 7, 6, 5, 4)^\top$.

Speculation: for all $r(a^*) > r(i)$, there exists $r(k) > r(i)$, such that for all $r(j) < r(i)$, it holds that $(x_i - x_j)^\top (x_k - x_j) \geq 0$.

Summary (ordering-based conditions)

NPG (sufficient and necessary): weaker than zero approximation error

Softmax PG (sufficient, not necessary): approximation error irrelevant

Future directions

General MDPs

Stochastic updates

Sufficient and necessary conditions

Representation learning

Transformers (softmax attention)

RLHF (preference-based data vs. ordering-based conditions)

References

Agarwal et al., On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2019.

Lan, Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes, 2021.

Xiao, On the Convergence Rates of Policy Gradient Methods, 2022.

Mei et al., On the global convergence rates of softmax policy gradient methods, 2020.

Chen et al., Surrogate objectives for batch policy optimization in one-step decision making, 2019.

Khodadadian et al., On the Linear convergence of Natural Policy Gradient Algorithm, 2021.

Li et al., Softmax policy gradient methods can take exponential time to converge, 2021.

End

Thanks! Questions?