# Geometry and Convergence of Natural Policy Gradients


Guido Montúfar

UCLA & MPI MiS

With Johannes Müller MPI MiS


Online Machine Learning Seminar, University of Nottingham, Feb 2023


UCLA

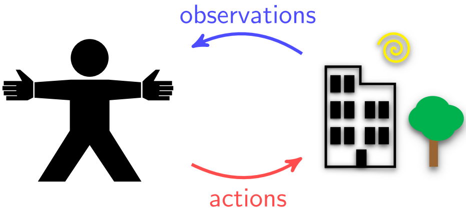
Max Planck Institute for Mathematics in the Sciences

European Research Council. Established by the European Commission

DFG Deutsche Forschungsgemeinschaft German Research Foundation

NSF

$$
\begin{tikzcd}
\text{Parameter space } \Theta \subseteq \mathbb{R}^p \arrow[r, "P"] \arrow[dr, "L"'] & \text{Model space } \mathcal{M} \arrow[d, "\ell"] \\
& \text{Reals } \mathbb{R}
\end{tikzcd}
$$




[Müller and Montúfar, 2022]

# Parameter space and function space

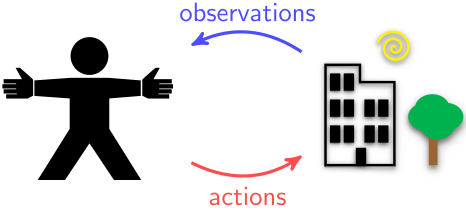
- Often we have a parametrized set of hypotheses $\{P_\theta \colon \theta \in \Theta\} \subseteq \mathcal{M}$
- Seek to optimize an objective function of the form

$$L(\theta) = \ell(P_\theta),$$

  interested in $P_{\theta^*}$ rather than $\theta^*$
- We can use the steepest direction in $\mathcal{M}$ rather than $\Theta$
- We still need to decide how to define the geometry of $\mathcal{M}$

| | Unregularized | | Regularized | |
|---|---|---|---|---|
| | Discr. time | Cts. time | Discr. time | Cts. time |
| Vanilla | $O(t^{-1})$ | – | linear | – |
| Kakade | linear | **linear** | **quadratic**<br>linear | **linear** |
| Morimura | – | **linear** | **quadratic** | **linear** |
| $\sigma > 1$ | – | $\mathbf{O(t^{-\frac{1}{\sigma-1}})}$ | **quadratic** | **linear** |

Table 1: Our work covers the bold results; previously shown were results for vanilla [Mei et al., 2020, Mei et al., 2021], Kakade discrete time – regularized [Cen et al., 2021] and unregularized [Khodadadian et al., 2021]




Want to optimize the action selection mechanism (policy)

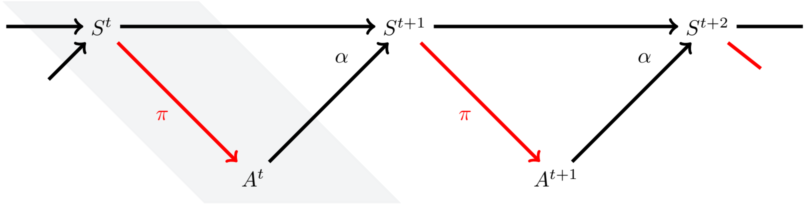
# Markov Decision Process

- $\mathcal{S}$ states
- $\mathcal{A}$ actions
- $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ transition probabilities
- $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ instantaneous reward
- $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ memoryless stochastic policy - the search variable

In this talk we focus on fully observable case; for POMDPs $\pi = \pi' \circ \beta$

A policy $\pi$ induces transition kernels $P_\pi \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}\times\mathcal{A}}$ and $p_\pi \in \Delta_{\mathcal{S}}^{\mathcal{S}}$

$$P_\pi(s', a'|s, a) = \alpha(s'|s, a)\pi(a'|s') \qquad p_\pi(s'|s) = \sum_{a\in\mathcal{A}} \alpha(s'|s, a)\pi(a|s)$$




In this talk we focus on fully observable case; for POMDPs $\pi = \pi' \circ \beta$

At each time step, the agent receives an instantaneous reward $r(s, a)$ for taking action $a$ at state $s$. Want to optimize long-term reward:

## Expected discounted reward

$$R_\gamma^\mu(\pi) := \mathbb{E}_{\mathbb{P}^{\pi,\mu}}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right]$$

Properties: Non-convex, rational function of $\pi$

In this talk we focus on discounted reward; for mean reward $\gamma \to 1$

The reward can be written as

$$R^{\mu}_{\gamma}(\pi) = \sum_{s,a} r(s,a)\eta^{\pi,\mu}_{\gamma}(s,a) = \langle r, \eta^{\pi,\mu}_{\gamma}\rangle_{\mathcal{S}\times\mathcal{A}},$$

where the expected discounted state-action frequency is

$$\eta^{\pi,\mu}_{\gamma}(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^{\pi,\mu}(s_t = s, a_t = a),$$

which can be interpreted as a discounted stationary distribution.

Observation: The optimization problem is linear in $\eta$

Idea: Study the problem over $\eta$ and the factorization

$$\begin{matrix} \text{parameter} & & \text{policy} & & \text{state-action frequency} & & \text{reward} \\ \theta & \mapsto & \pi & \mapsto & \eta & \mapsto & R \end{matrix}$$

For MDPs the feasible values of $\eta$ form a polytope:

## Proposition 1 (State-action polytope of MDPs, [Derman, 1970])

*The set $\mathcal{N}$ of state-action frequencies is a polytope given by $\mathcal{N} = \mathcal{L} \cap \Delta_{\mathcal{S}\times\mathcal{A}}$, where*

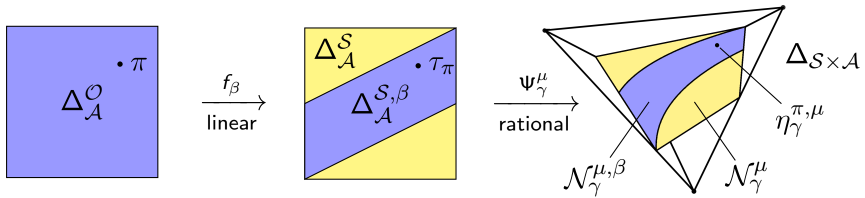
$$\mathcal{L} = \left\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \ell_s(\eta) = 0 \text{ for all } s \in \mathcal{S}, \eta \geq 0\right\}, \tag{1}$$

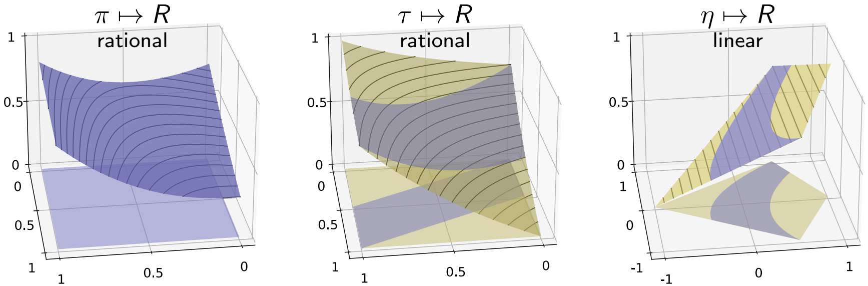
*and* $\ell_s(\eta) \coloneqq \sum_a \eta_{sa} - \gamma \sum_{s',a'} \eta_{s'a'} \alpha(s|s',a') - (1-\gamma)\mu_s$.

## Corollary 2

*The MDP problem is a linear program over $\eta$.*

Observation policies State policies State-action frequencies



## Assumption 1 (Positivity)

For every $s \in \mathcal{S}$ and $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$, we assume that $\sum_a \eta_{sa}^{\pi} > 0$.

Note: This positivity assumption is satisfied e.g. if $\mu > 0$, and is required for global convergence of PG methods [Mei et al., 2020].

We will use this to have a diffeomorphism between $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mathcal{N}$:

## Proposition 3 ([Müller and Montúfar, 2022])

*Under Assumption 1, the mapping $\Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathcal{N}, \omega \mapsto \eta$ is rational and bijective with rational inverse given by conditioning $\mathcal{N} \to \Delta_{\mathcal{A}}^{\mathcal{S}}, \eta \mapsto \omega$, where $\omega_{as} = \frac{\eta_{sa}}{\sum_{a'} \eta_{sa'}}$.*

# Natural Gradients

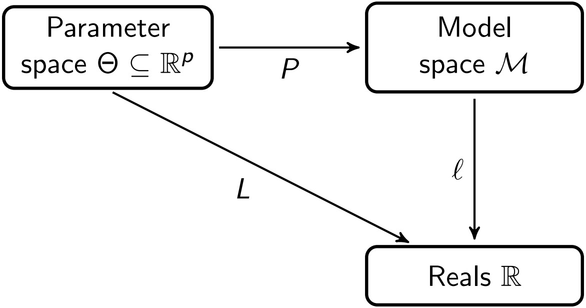



Figure 1: Parametric model and factorizing objective.

# Riemannian gradients

- Steepest direction of $L(\theta)$ at $\theta$

$$\begin{aligned} \min_{d\theta} \quad & L(\theta + d\theta) \\ \text{s.t.} \quad & |d\theta|^2 = \epsilon^2 \end{aligned}$$

- In a Riemannian manifold with metric $G(\theta) = (g_{ij}(\theta))$,

$$|d\theta|^2 = \sum_{ij} g_{ij}(\theta) d\theta_i d\theta_j$$

- Leads to $d\theta \propto G(\theta)^{-1} \nabla L(\theta)$

# Natural Gradients

For an objective function $R$, Natural Gradients take the form

$$\theta_{k+1} = \theta_k + \Delta t \, G(\theta_k)^+ \, \nabla R(\theta_k),$$

where

- $G(\theta)_{ij} = g(dP_\theta e_i, dP_\theta e_j)$ is a Gram matrix
- $G(\theta)^+$ pseudo inverse
- $g$ Riemannian metric
- $P(\theta)$ representation of the parameter

## Example 4 (Fisher Natural Gradient)

- $P(\theta) \in \Delta_{\mathcal{X}}$ a probability distribution parametrized by $\theta$
- $g$ Fisher information metric

$$g_P(u,v) = \sum_x \frac{u_x v_x}{P_x}, \quad \text{for all } u, v \in T_P\Delta_{\mathcal{X}}$$

- $G(\theta)_{ij} = \sum_x \frac{\partial_i P_x(\theta) \partial_j P_x(\theta)}{P_x(\theta)}$

## Definition 5 (General natural gradient)

Consider an objective $L\colon \Theta \to \mathbb{R}$, where the *parameter space* $\Theta \subseteq \mathbb{R}^p$ an open subset. Further, assume that the objective factorizes as $L = \ell \circ P$, where $P\colon \Theta \to \mathcal{M}$ is a *model parametrization* with $\mathcal{M}$ a Riemannian manifold with Riemannian metric $g$, and $\ell\colon \mathcal{M} \to \mathbb{R}$ is a *loss in model space*, as shown in Figure 1. For $\theta \in \Theta$ we define the Gram matrix

$$G(\theta)_{ij} := g_{P(\theta)}(dP_\theta e_i, dP_\theta e_j)$$

and call

$$\nabla^N L(\theta) := G(\theta)^+ \nabla L(\theta)$$

the natural gradient (NG) of $L$ at $\theta$ with respect to the factorization $L = \ell \circ P$ and the metric $g$.

## Best improvement direction

### Theorem 6 (NG leads to steepest descent in model space)

*Consider the settings of Definition 5, where $\mathcal{M}$ is a Riemannian manifold with metric $g$. Let $\nabla^N L(\theta) := G(\theta)^+ \nabla_\theta L(\theta)$ denote the natural gradient with respect to this factorization. Then it holds that*

$$dP_\theta(\nabla^N L(\theta)) = \Pi_{T_\theta \mathcal{M}_\Theta}(\nabla^g \ell(P(\theta))).$$

# Choice of the geometry in model space

# Invariance axiomatic

## Definition 7 (Invariance)

Given $(\mathcal{E}, g)$, $(\mathcal{E}', g')$ and an embedding $f\colon \mathcal{E} \to \mathcal{E}'$, the metric is said to be invariant if the embedding is an isometry, meaning that

$$g_p(u, v) = g'_{f(p)}(f_* u, f_* v), \quad \text{for all } p \in \mathcal{E} \text{ and } u, v \in T_p\mathcal{E},$$

where $f_*\colon T_p\mathcal{E} \to T_{f(p)}\mathcal{E}'$ is the pushforward of $f$.

Probability distributions: [Čencov, 1982, Campbell, 1986, Ay et al., 2017] characterize Fisher as the unique metric (up to scaling) that is invariant with respect to congruent embeddings by Markov mappings.

Conditional probability distributions: Product of Fisher metric satisfies invariance properties [Lebanon, 2005, Montúfar et al., 2014]; nevertheless, choice less clear than on the simplex.

# Hessian geometries

Idea: Select a metric based on the optimization problem at hand.
If the objective $\ell \colon \mathcal{M} \to \mathbb{R}$ has a positive definite Hessian at every point, it induces a Riemannian metric via

$$g_p(v, w) = v^\top \nabla^2 \ell(p) w,$$

in local coordinates, that we call the *Hessian geometry*;
see [Amari and Cichocki, 2010, Shima, 2007].

## Example 8 (Hessian geometries)

The following Riemannian geometries are induced by strictly convex functions.

1. *Euclidean geometry:* The Euclidean geometry on $\mathbb{R}^d$ is induced by the convex function $x \mapsto \frac{1}{2}\sum_i x_i^2$.
2. *Fisher geometry:* The Fisher metric on $\mathbb{R}^d_{>0}$ is induced by the negative entropy $x \mapsto \sum_i x_i \log(x_i)$.
3. *Itakura-Saito:* The logarithmic barrier function $x \mapsto \sum_i \log(x_i)$ of the positive cone $\mathbb{R}^d_{>0}$ yields the Itakura-Saito metric (see the next item).

4. *$\sigma$-geometries:* All of the above examples can be interpreted as special cases of a parametric family of Hessian metrics. Let

$$\phi_\sigma(x) := \begin{cases} \sum_i x_i \log(x_i) & \text{if } \sigma = 1 \\ -\sum_i \log(x_i) & \text{if } \sigma = 2 \\ \frac{1}{(2-\sigma)(1-\sigma)} \sum x_i^{2-\sigma} & \text{otherwise} \end{cases} \tag{2}$$

Then the resulting Riemannian metric on $\mathbb{R}^d$ for $\sigma \in (-\infty, 0]$ and on $\mathbb{R}^d_{>0}$ for $\sigma \in (0, \infty)$ is given by

$$g_x^\sigma(v, w) = \sum_i \frac{v_i w_i}{x_i^\sigma}. \tag{3}$$

This recovers the Euclidean geometry for $\sigma = 0$, the Fisher metric for $\sigma = 1$, and the Itakura-Saito metric for $\sigma = 2$.

5. *Conditional entropy:* Consider the conditional entropy

$$\phi_C(\mu) := H(\mu|\mu_X) = H(\mu) - H(\mu_X), \tag{4}$$

which is convex on $\Delta_{\mathcal{X}\times\mathcal{Y}}$.
The Hessian of the conditional entropy is given by

$$\partial_{(s,a)}\partial_{(s',a')}\phi_C(\mu) = \delta_{xx'}\left(\delta_{yy'}\mu(x,y)^{-1} - \mu_X(x)^{-1}\right) \tag{5}$$

This is a Riemannian metric on the interior of $\{\mu \in \Delta_{\mathcal{X}\times\mathcal{X}} : \mu_X = \nu(\mu_{Y|X})\}$, for a smooth $\nu\colon \operatorname{int}(\Delta_{\mathcal{Y}}^{\mathcal{X}}) \to \operatorname{int}(\Delta_{\mathcal{X}})$.
Indeed, it is the pull back of the Riemannian metric

$$g\colon T\Delta_{\mathcal{Y}}^{\mathcal{X}} \times T\Delta_{\mathcal{Y}}^{\mathcal{X}} \to \mathbb{R}, \quad g_{\mu(\cdot|\cdot)}(v,w) := \sum_x \nu(x) \sum_y \frac{v(x,y)w(x,y)}{\mu(y|x)}.$$

# Natural Policy Gradients

## Softmax policy parametrization

The tabular softmax parametrization is given by

$$\pi_\theta(a|s) := \frac{e^{\theta_{sa}}}{\sum_{a'} e^{\theta_{sa'}}} \quad \text{for all } a \in \mathcal{A}, s \in \mathcal{S}, \quad \text{for } \theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}. \tag{6}$$

## Definition 9 (Regular policy parametrization)

We call a policy parametrization $\mathbb{R}^p \to \operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$; $\theta \mapsto \pi_\theta$ *regular* if it is differentiable and satisfies

$$\operatorname{span}\{\partial_{\theta_i}\pi_\theta : i = 1, \ldots, p\} = T_{\pi_\theta}\Delta_{\mathcal{A}}^{\mathcal{S}} \quad \text{for every } \theta \in \mathbb{R}^p.$$

This assumes an unconstrained parameter, can be overparametrized.

# Policy Gradient Theorem

## Theorem 10 (Policy gradient theorem)

*Consider an MDP* $(\mathcal{S}, \mathcal{A}, \alpha, r), \gamma \in [0,1)$ *and a parametrized policy class. It holds that*

$$\begin{aligned}\partial_{\theta_i} R(\theta) &= \sum_s \rho_\theta(s) \sum_a \partial_{\theta_i} \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \\ &= \sum_{s,a} \eta_\theta(s,a) \partial_{\theta_i} \log(\pi_\theta(a|s)) Q^{\pi_\theta}(s,a),\end{aligned}$$

*where* $Q^\pi := (I - \gamma P_\pi)^{-1} r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ *is the state-action value function.*

# Kakade's NPG

### Definition 11 (Kakade's NPG and geometry in policy space)

We refer to the natural gradient $\nabla^K R(\theta) := G_K(\theta)^+ \nabla_\theta R(\pi_\theta)$ as *Kakade's natural policy gradient (K-NPG)*, where $G_K$ is defined by

$$G_K(\theta)_{ij} = \sum_s \rho_\theta(s) \sum_a \frac{\partial_{\theta_i} \pi_\theta(a|s) \partial_{\theta_j} \pi_\theta(a|s)}{\pi_\theta(a|s)}. \tag{7}$$

Hence, Kakade's NPG is the NPG induced by the factorization $\theta \mapsto \pi_\theta \mapsto R(\theta)$ and the Riemannian metric on $\mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{S}})$ given by

$$g_\pi^K(v, w) := \sum_s \rho^\pi(s) \sum_a \frac{v(s,a) w(s,a)}{\pi(a|s)} \quad \text{for all } v, w \in T_\pi \Delta_{\mathcal{A}}^{\mathcal{S}}. \tag{8}$$

[Kakade, 2001]

## Theorem 12 (Kakade's geometry as cond. entropy Hessian geometry)

*Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha)$ and fix $\mu \in \Delta_{\mathcal{S}}$ and $\gamma \in (0, 1)$ such that Assumption 1 holds. Then, Kakade's geometry on $\Delta_{\mathcal{A}}^{\mathcal{S}}$ is the pull back of the Hessian geometry induced by the conditional entropy on the state-action polytope $\mathcal{N} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$ along $\pi \mapsto \eta^{\pi}$.*

*In particular, K-NPG is the NPG induced by factorization $\theta \mapsto \eta_\theta \mapsto R(\theta)$ with respect to the conditional entropy Hessian geometry, i.e.,*

$$G_K(\theta)_{ij} = \sum_{s,a} \frac{\partial_{\theta_i}\eta_\theta(s,a)\partial_{\theta_j}\eta_\theta(s,a)}{\eta_\theta(s,a)} - \sum_{s} \frac{\partial_{\theta_i}\rho_\theta(s)\partial_{\theta_j}\rho_\theta(s)}{\rho_\theta(s)}. \tag{9}$$

K-NPG is known to converge at a locally quadratic rate under conditional entropy regularization [Cen et al., 2021], which in policy space is

$$\psi(\pi) = \sum_s \rho^\pi(s) \sum_a \pi(a|s) \log(\pi(a|s)) = \sum_s \rho^\pi(s) H(\pi(\cdot|s)).$$

However Kakade's geometry in policy space $g^K$ is not the Hessian geometry induced by $\psi$ in policy space, which would take the form

$$\begin{aligned}\nabla^2\psi(\pi) = &\sum_s \rho^\pi(s)\nabla^2 H(\pi(\cdot|s)) + \sum_s H(\pi(\cdot|s))\nabla^2\rho^\pi(s) \\ &+ \sum_s (\nabla H(\cdot|s)^\top \nabla\rho^\pi(s) + \nabla H(\cdot|s)\nabla\rho^\pi(s)^\top).\end{aligned}$$

Kakade's metric only considers the first term; see (8).

# Morimura's NPG

## Definition 13 (Morimura's NPG)

We refer to the natural gradient $\nabla^M R(\theta) := G_M(\theta)^+ \nabla_\theta R(\pi_\theta)$ as *Morimura's natural policy gradient (M-NPG)*, where $G_M$ is given by

$$G_M(\theta)_{ij} = \sum_{s,a} \partial_{\theta_i} \log(\eta_\theta(s,a)) \partial_{\theta_j} \log(\eta_\theta(s,a)) \eta_\theta(s,a). \tag{10}$$

Hence, Morimura's NPG is the NPG induced by the factorization $\theta \mapsto \eta_\theta \mapsto R(\theta)$ and the Fisher metric on $\mathrm{int}(\Delta_{\mathcal{S}\times\mathcal{A}})$.

[Morimura et al., 2008]

## Comparison of Kakade and Morimura

By (9) the Gram matrix proposed by Morimura and co-authors and the Gram matrix proposed by Kakade are related to each other by

$$G_K(\theta) = G_M(\theta) - F_\rho(\theta),$$

where $F_\rho(\theta)_{ij} = \sum_s \rho_\theta(s)\partial_{\theta_i}\log(\rho_\theta(s))\partial_{\theta_j}\log(\rho_\theta(s))$ denotes the Fisher information matrix of the state distributions.

# General Hessian NPG

## Definition 14 (Hessian NPG)

We refer to the natural gradient $\nabla^{\phi} R(\theta) := G_{\phi}(\theta)^{+} \nabla_{\theta} R(\pi_{\theta})$ as *Hessian NPG with respect to* $\phi$ or $\phi$*-natural policy gradient (*$\phi$*-NPG)*.

In particular:

## Definition 15 ($\sigma$-NPG)

We refer to the natural gradient $\nabla^{\sigma} R(\theta) := G_{\sigma}(\theta)^{+} \nabla_{\theta} R(\pi_{\theta})$ as the $\sigma$*-natural policy gradient (*$\sigma$*-NPG)*. Hence $\sigma$-NPG is the NPG induced by factorization $\theta \mapsto \eta_{\theta} \mapsto R(\theta)$ and metric $g^{\sigma}$ on $\operatorname{int}(\Delta_{\mathcal{S}\times\mathcal{A}})$ defined in (3).

For $\sigma = 1$ we recover the Fisher geometry and hence M-NPG; for $\sigma = 2$ the Itakura-Saito metric; and for $\sigma = 0$ the Euclidean geometry.

Later, we show that the Hessian gradient flows exist globally for $\sigma \in [1, \infty)$ and provide convergence rates depending on $\sigma$.

# Reduction to state-action space

## Proposition 16 (Evolution in state-action space)

*Consider an MDP $(\mathcal{S}, \mathcal{A}, \alpha)$, a Riemannian metric $g$ on $\operatorname{int}(\mathcal{N}) = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and an differentiable objective function $\mathfrak{R}\colon \operatorname{int}(\Delta_{\mathcal{S}\times\mathcal{A}}) \to \mathbb{R}$. Consider a regular policy parametrization and the objective $R(\theta) := \mathfrak{R}(\eta_\theta)$ and a solution $\theta\colon [0, T] \to \Theta = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ of the NPG flow*

$$\partial_t \theta(t) = \nabla^N R(\theta(t)) = G(\theta(t))^+ \nabla R(\theta(t)), \tag{11}$$

*where $G(\theta)_{ij} = g_\eta(\partial_{\theta_i}\eta_\theta, \partial_{\theta_j}\eta_\theta)$ and $G(\theta)^+$ denotes a pseudo inverse of $G(\theta)$. Setting $\eta(t) := \eta_{\theta(t)}$ we have that $\eta\colon [0, T] \to \Delta_{\mathcal{S}\times\mathcal{A}}$ is the gradient flow with respect to the metric $g|_{\mathcal{N}}$ and the objective $\mathfrak{R}$, i.e., solves*

$$\partial_t \eta(t) = \nabla^{g|_{\mathcal{N}}} \mathfrak{R}(\eta(t)) = \Pi^g_{T\mathcal{L}}(\nabla^g \mathfrak{R}(\eta(t))), \tag{12}$$

*where $\Pi^g_{T\mathcal{L}}$ is the $g$-orthogonal projection onto $T\mathcal{L}$ with $\mathcal{L}$ defined in (1).*

# Convergence of unregularized Hessian NPG flows

## Setting 17

- Objective $\mathfrak{R}\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}\cup\{-\infty\}$ that is finite, differentiable and concave on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and cts on $\mathrm{dom}(\mathfrak{R}) = \{\eta\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \mathfrak{R}(\eta)\in\mathbb{R}\}$.
- Function $\phi\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}\cup\{+\infty\}$, finite and $C^2$ on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$, with $\nabla^2\phi(\eta)$ positive definite on $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ for $\eta\in\mathrm{int}(\mathcal{N})$.
- Solution $\eta\colon [0,T)\to\mathcal{N}$ of the Hessian gradient flow

$$\partial_t\eta(t) = \Pi_{T\mathcal{L}}(\nabla^2\phi(\eta(t))^{-1}\nabla\mathfrak{R}(\eta(t))). \tag{13}$$

- We denote[1] $R^* := \sup_{\eta\in\mathcal{N}}\mathfrak{R}(\eta) < \infty$ and by $\eta^*\in\mathcal{N}$, we denote a maximizer – if one exists – of $\mathfrak{R}$ over $\mathcal{N}$.
- We denote the policies corresponding to $\eta_0$ and $\eta^*$ by $\pi_0$ and $\pi^*$.

[1] Note that $\mathfrak{R}$ is bounded over the bounded set $\mathcal{N}$ as a concave function.

# Sublinear rates for general case

## Lemma 18 (Convergence of Hessian NPG flows)

*Consider Setting 17 and assume there exists a solution $\eta\colon [0, T) \to \operatorname{int}(\mathcal{N})$ of the NPG flow (13) with initial condition $\eta(0) = \eta_0$. Then for any $\eta' \in \mathcal{N}$ and $t \in [0, T)$ it holds that*

$$\mathfrak{R}(\eta') - \mathfrak{R}(\eta(t)) \leq D_\phi(\eta', \eta_0) t^{-1}, \tag{14}$$

*where $D_\phi$ denotes the Bregman divergence of $\phi$.*

*In particular, $\mathfrak{R}(\eta(t)) \to R^*$ as $T \to \infty$. Further, convergence happens at a rate $O(t^{-1})$ if there is a maximizer $\eta^* \in \mathcal{N}$ of $\mathfrak{R}$ with $\phi(\eta^*) < \infty$.*

Similar to [Alvarez et al., 2004, Prop. 4.4]

Thus proving convergence of NPG reduces to ensuring well-posedness

To induce Hessian geometries that prevent finite-time hitting boundary:

## Definition 19 (Legendre type functions)

We call $\phi\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}\cup\{+\infty\}$ a *Legendre type function* if:

1. *Domain:* It holds that $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0} \subseteq \mathrm{dom}(\phi) \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}$, where $\mathrm{dom}(\phi) = \{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \phi(\eta) < \infty\}$.
2. *Smoothness and convexity:* We assume $\phi$ to be continuous on $\mathrm{dom}(\phi)$ and twice continuous differentiable on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ and such that $\nabla^2\phi(\eta)$ is positive definite on $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ for every $\eta \in \mathrm{int}(\mathcal{N})$.
3. *Gradient blowup at boundary:* For any $(\eta_k) \subseteq \mathrm{int}(\mathcal{N})$ with $\eta_k \to \eta \in \partial\mathcal{N}$ we have $\|\nabla\phi(\eta_k)\| \to \infty$.

Slight generalization of [Alvarez et al., 2004] important for our analysis

## Example 20

Legendre-type functions cover the functions inducing K-NPG and M-NPG via their Hessian geometries.

1. The functions $\phi_\sigma$ in (2) that define the $\sigma$-NPG are of Legendre-type for $\sigma \in [1, \infty)$. This includes the Fisher geometry (M-NPG) for $\sigma = 1$, but excludes the Euclidean geometry, which corresponds to $\sigma = 0$.
2. The conditional entropy $\phi_C$ in (4) is a Legendre-type function. Its Hessian geometry induces the K-NPG.

   In this case the gradient blowup holds on the boundary of $\mathcal{N}$ but not on the boundary of $\Delta_{\mathcal{S}\times\mathcal{A}}$ or even $\mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}$.

## Theorem 21 (Conv. of K-NPG flow for unregularized reward)

*Consider Setting 17 with* $\phi = \phi_C$ *the conditional entropy, let* $\mathfrak{R}(\eta) = \langle r, \eta \rangle$ *denote the unregularized reward, and fix an* $\eta_0 \in \mathrm{int}(\mathcal{N})$. *Then there exists a unique global solution* $\eta\colon [0, \infty) \to \mathrm{int}(\mathcal{N})$ *of K-NPG flow with initial condition* $\eta(0) = \eta_0$ *and it holds that*

$$R^* - \mathfrak{R}(\eta(t)) \leq t^{-1} D_{\phi_C}(\eta^*, \eta_0) = t^{-1} \sum_s \rho^*(s) D_{KL}(\pi^*(\cdot|s), \pi_0(\cdot|s)),$$

*where* $D_{\phi_C}$ *denotes the conditional relative entropy. In particular, we have* $\mathrm{dist}(\eta(t), S) \in O(t^{-1})$, *where* $S = \{\eta \in \mathcal{N} : \langle r, \eta \rangle = R^*\}$ *denotes the solution set and* dist *denotes the Euclidean distance.*

## Theorem 22 (Convergence of $\sigma$-NPG flow for unregularized reward)

*Consider Setting 17 with $\phi = \phi_\sigma$ for some $\sigma \in [1, \infty)$ being defined in (2). Denote the unregularized reward by $\mathfrak{R}(\eta) = \langle r, \eta \rangle$ and fix an element $\eta_0 \in \operatorname{int}(\mathcal{N})$. Then there exists a unique global solution $\eta\colon [0, \infty) \to \operatorname{int}(\mathcal{N})$ of the Hessian NPG flow (13) with inital condition $\eta(0) = \eta_0$ and it holds that $R^* - \mathfrak{R}(\eta(t)) = O(f_\sigma(t))$ as $t \to \infty$, where*

$$f_\sigma(t) := \begin{cases} t^{-1} & \text{for } \sigma \in [1, 2) \\ \log(t) t^{-1} & \text{for } \sigma = 2 \\ t^{\sigma - 3} & \text{for } \sigma \in (2, \infty). \end{cases}$$

*In particular, we have $\operatorname{dist}(\eta(t), S) \in O(f_\sigma(t))$, where $S = \{\eta \in \mathcal{N} : \langle r, \eta \rangle = R^*\}$ denotes the solution set and dist denotes the Euclidean distance. This result covers M-NPG flow as special case $\sigma = 1$.*

## Remark 23

- Theorem 22 and Theorem 21 show global convergence of $\sigma$-NPG and K-NPG flows to a maximizer of the unregularized problem.

  This is possible because one works not with a regularized objective but rather with geometry from regularization and original objective.
- For $\sigma < 1$ the flow may reach a face of the feasible set in finite time; see Figure 3. For $\sigma \geq 3$ Theorem 22 is uninformative.
- One can show that the trajectory converges to the maximizer that is closest to $\eta_0$ wrt the Bregman divergence [Alvarez et al., 2004].

# Faster rates for $\sigma \in [1, 2)$ and K-NPG

## Lemma 24 (Convergence rates for gradient flow trajectories)

*Consider Setting 17 and assume that there is a global solution $\eta \colon [0, \infty) \to \operatorname{int}(\mathcal{N})$ of the Hessian gradient flow (13). Assume that there is $\eta^* \in \mathcal{N}$ such that $\phi(\eta^*) < +\infty$ as well as a neighborhood $N$ of $\eta^*$ in $\mathcal{N}$ and $\omega \in (0, \infty)$ and $\tau \in [1, \infty)$ such that*

$$\mathfrak{R}(\eta^*) - \mathfrak{R}(\eta) \geq \omega D_\phi(\eta^*, \eta)^\tau \quad \text{for all } \eta \in N. \tag{15}$$

*Then there is a constant $c > 0$ such that*

1. *if $\tau = 1$, then $D_\phi(\eta^*, \eta(t)) \leq ce^{-\omega t}$,*
2. *if $\tau > 1$, then $D_\phi(\eta^*, \eta(t)) \leq ct^{-1/(\tau-1)}$.*

Similar to [Alvarez et al., 2004, Prop. 4.9] but relaxing assumptions.

Thus can get faster NPG rates by ensuring (15); a form of strong convexity.

## Theorem 25 (Linear convergence of unregularized K-NPG flow)

*Consider Setting 17, where* $\phi = \phi_C$ *is the conditional entropy defined in* (4) *and assume that there is a unique maximizer* $\eta^*$ *of the unregularized reward* $\mathfrak{R}$*. Then* $R^* - \mathfrak{R}(\eta(t)) = O(e^{-ct})$ *for some* $c > 0$*.*

### Theorem 26 (Linear convergence of unregularized M-NPG flow / improved rates for $\sigma$-NPG flow)

*Consider Setting 17, where $\phi = \phi_\sigma$ for some $\sigma \in [1, 2)$ as defined in (2), and assume that there is a unique maximizer $\eta^*$ of the unregularized reward $\mathfrak{R}$. Denote $\eta\colon [0, \infty) \to \operatorname{int}(\mathcal{N})$ the solution of the $\sigma$-NPG flow. Then $R^* - \mathfrak{R}(\eta(t)) \in O(g_\sigma(t))$, where*

$$g_\sigma(t) = \begin{cases} e^{-ct} & \text{if } \sigma = 1 \\ t^{-1/(\sigma-1)} & \text{if } \sigma \in (1, 2), \end{cases}$$

*for some $c > 0$.*

# Numerical examples I

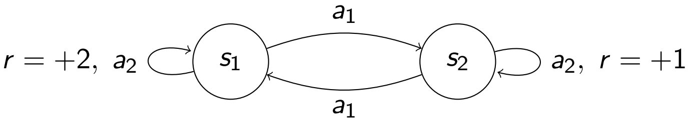



Figure 2: MDP example transition graph and reward.

# Numerical examples II

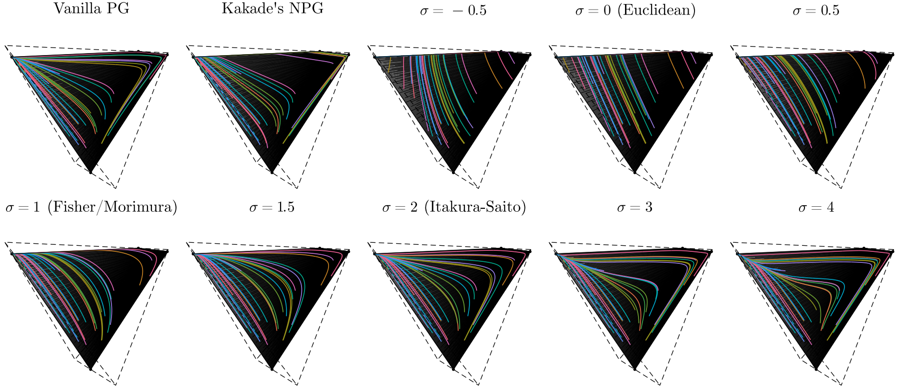



Figure 3: State-action trajectories for different PG methods: vanilla PG, K-NPG and $\sigma$-NPG, where M-NPG corresponds to $\sigma = 1$;

# Numerical examples III

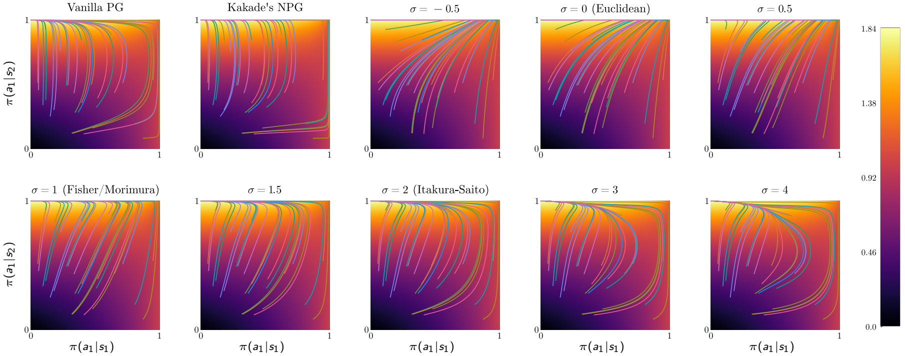



Figure 4: Heatmap of $\pi \mapsto R(\pi)$ and trajectories of individual methods over $\Delta_{\mathcal{A}}^{\mathcal{S}} \cong [0,1]^2$; maximizer $\pi^*$ is at the upper left corner.

# Numerical examples IV

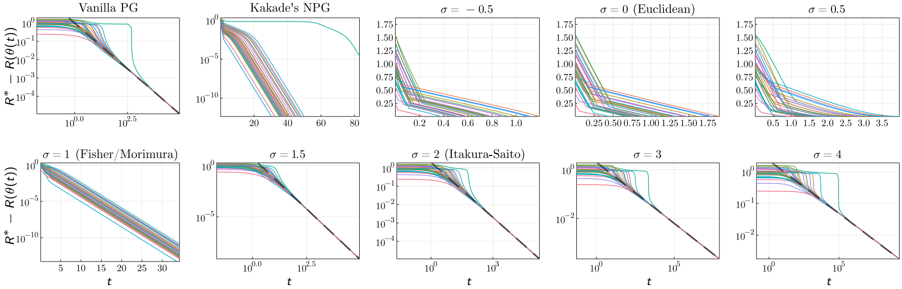



Figure 5: Optimality gap $R^* - R(\theta(t))$; vanilla PG and $\sigma > 1$ in log-log as we expect decay $t^{-1}$ and $t^{-1/(\sigma-1)}$ (shown dashed); K-NPG and M-NPG in log-y as we expect linear convergence; for $\sigma < 1$ we observe finite time convergence.

# Linear convergence of regularized Hessian NPG flows

## Theorem 27 (Linear convergence for regularized objective)

*Consider Setting 17, let $\phi$ be a Legendre-type function, denote the regularized reward by $\mathfrak{R}_\lambda(\eta) = \langle r, \eta \rangle - \lambda\phi(\eta)$ for some $\lambda > 0$, and assume that the global maximizer $\eta^*_\lambda$ of $\mathfrak{R}_\lambda$ over $\mathcal{N}$ lies in the interior $\mathrm{int}(\mathcal{N})$. Fix an $\eta_0 \in \mathrm{int}(\mathcal{N})$ and assume $\eta\colon [0,\infty) \to \mathrm{int}(\mathcal{N})$ solves the NPG flow wrt $\mathfrak{R}_\lambda$ and the Hessian geometry induced by $\phi$.*

*Then, for any $c \in (0,\lambda)$ there exists $K > 0$ st $D_\phi(\eta^*_\lambda, \eta(t)) \leq Ke^{-ct}$.*

*In particular, for any $\kappa \in (\kappa_c, \infty)$ this implies $R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) \leq \kappa\lambda Ke^{-ct}$, where $\kappa_c$ denotes the condition number of $\nabla^2\phi(\eta^*)$.*

Using Lemma 24 and Lemma 31.

Condition $\eta^*_\lambda \in \mathrm{int}(\mathcal{N})$ is satisfied if gradient blow-up in Definition 19 is slightly strengthened; Remark 32.

## Corollary 28 (Linear convergence of regularized K-NPG flow)

*Assume that* $\eta \colon [0, \infty) \to \operatorname{int}(\mathcal{N})$ *solves the NPG flow with respect to the regularized reward* $\mathfrak{R}_\lambda$ *and the Hessian geometry induced by* $\phi$*. For any* $\omega \in (0, \lambda)$ *there exists a constant* $K > 0$ *such that* $D_\phi(\eta^*, \eta(t)) \leq K e^{-\omega t}$*. In particular, for any* $\kappa \in (\kappa_c, \infty)$ *this implies* $R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) \leq \kappa K e^{-\omega t}$*, where* $\kappa_c$ *denotes the condition number of* $\nabla^2 \phi_C(\eta^*)$*.*

### Corollary 29 (Linear convergence for regularized $\sigma$-NPG flow)

*Consider Setting 17 with $\phi = \phi_\sigma$ for some $\sigma \in [1, \infty)$ and denote the regularized reward by $\mathfrak{R}_\lambda(\eta) = \langle r, \eta \rangle - \lambda\phi(\eta)$ and fix an element $\eta_0 \in \operatorname{int}(\mathcal{N})$. Assume that $\eta \colon [0, \infty) \to \operatorname{int}(\mathcal{N})$ solves the natural policy gradient flow with respect to the regularized reward $\mathfrak{R}_\lambda$ and the Hessian geometry induced by $\phi$. For any $\omega \in (0, \lambda)$ there exists a constant $K > 0$ such that $D_\phi(\eta^*, \eta(t)) \leq Ke^{-\omega t}$. In particular, for any $\kappa \in (\kappa(\eta^*)^\sigma, \infty)$ this implies $R^*_\lambda - \mathfrak{R}_\lambda(\eta(t)) \leq \kappa K e^{-\omega t}$, where $\kappa(\eta^*) = \frac{\max \eta^*}{\min \eta^*}$*

## Theorem 30 (Locally quadratic convergence of reg. NPGs)

*Consider a real-valued function* $\phi\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}\cup\{+\infty\}$*, which we assume to be finite and twice continuously differentiable on* $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{>0}$ *and such that* $\nabla^2\phi(\eta)$ *is pos. def. on* $T_\eta\mathcal{N} = T\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ *for every* $\eta \in \operatorname{int}(\mathcal{N})$*. Further, consider a regular policy parametrization and the regularized reward* $R_\lambda(\theta) := R(\theta) + \lambda\phi(\eta_\theta)$ *and assume that* $\eta^* \in \operatorname{int}(\mathcal{N})$*, i.e., the maximizer lies in the interior of the state-action polytope. Consider the NPG induced by the Hessian geometry of* $\phi$*, i.e.,*

$$\theta_{k+1} = \theta_k + \Delta t G(\theta_k)^+ \nabla R_\lambda(\theta_k),$$

*with step size* $\Delta t = \lambda$*, where* $G(\theta_k)^+$ *denotes the Moore-Penrose inverse. Assume that* $R_\lambda(\theta_k) \to R^*_\lambda$ *for* $k \to \infty$*. Then* $\theta_k \to \theta^*$ *at a (locally) quadratic rate and hence* $R_\lambda(\theta_k) \to R^*_\lambda$ *at a (locally) quadratic rate.*

Using inexact Newton method Theorem 33 and a corresponding description of reg. NPG by Lemma 34 and 35.

| | Unregularized | | Regularized | |
|---|---|---|---|---|
| | Discr. time | Cts. time | Discr. time | Cts. time |
| Vanilla | $O(t^{-1})$ | – | linear | – |
| Kakade | linear | **linear** | **quadratic** $(\Delta_t = \lambda)$ linear $(\Delta_t \leq \lambda)$ | **linear** |
| Morimura | – | **linear** | **quadratic** $(\Delta_t = \lambda)$ | **linear** |
| $\sigma > 1$ | – | $\mathbf{O}(\mathbf{t}^{-\frac{1}{\sigma-1}})$ | **quadratic** $(\Delta_t = \lambda)$ | **linear** |

Table 2: Our work covers the bold results; previously shown were results for vanilla [Mei et al., 2020, Mei et al., 2021], Kakade discrete time – regularized [Cen et al., 2021] and unregularized [Khodadadian et al., 2021]

## Why is the analysis easier in state-action space?

- Problem is strongly convex in state-action space, whereas in policy and parameter space it is non-convex.
- Further, in policy space the corresponding Riemannian metric might not be the Hessian metric of the regularizer.
- In the parameter $\theta$, the NPG algorithm can be perceived as a generalized Gauss-Newton method; however, the reward function is non-convex in parameter space.
- For overparametrized models, $\dim(\Theta) > \dim(\Delta_{\mathcal{A}}^{\mathcal{S}})$, Hessian $\nabla^2 R(\theta^*)$ not positive definite, which complicates analysis in parameter space.

## Conclusion

- Study of a general class of natural policy gradient methods arising from Hessian geometries in state-action space.
- Linear convergence for Kakade's and Morimura's NPG for unregularized reward.
- Locally quadratic convergence for regularized NPG with respect to the Hessian geometry of the regularizer.

## Outlook

- General parametric policy classes and partially observable MDPs.
- Develop NPG methods without plateaus.
- Study NPG methods in state-action space with estimation.

# References I


Alvarez, F., Bolte, J., and Brahic, O. (2004).
Hessian Riemannian gradient flows in convex programming.
*SIAM journal on control and optimization*, 43(2):477–501.

Amari, S. and Cichocki, A. (2010).
Information geometry of divergence functions.
*Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195.

Ay, N., Jost, J., Vân Lê, H., and Schwachhöfer, L. (2017).
*Information geometry*.
Springer, Cham.

Campbell, L. (1986).
An extended Čencov characterization of the information metric.
*Proceedings of the American Mathematical Society*, 98:135–141.

# References II


Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2021).
Fast global convergence of natural policy gradient methods with entropy regularization.
*Operations Research*.

Čencov, N. N. (1982).
*Statistical decision rules and optimal inference*, volume 53 of *Translations of Mathematical Monographs*.
American Mathematical Society, Providence, R.I.
Translation from the Russian edited by Lev J. Leifman.

Dembo, R. S., Eisenstat, S. C., and Steihaug, T. (1982).
Inexact Newton methods.
*SIAM Journal on Numerical analysis*, 19(2):400–408.

# References III


Derman, C. (1970).
*Finite state Markovian decision processes*.
Academic Press, New York.

Kakade, S. M. (2001).
A natural policy gradient.
*Advances in Neural Information Processing Systems*, 14.

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. (2021).
On the linear convergence of natural policy gradient algorithm.
In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE.

Lebanon, G. (2005).
Axiomatic geometry of conditional models.
*IEEE Transactions on Information Theory*, 51:1283–1294.

# References IV


Mei, J., Gao, Y., Dai, B., Szepesvari, C., and Schuurmans, D. (2021).
Leveraging non-uniformity in first-order non-convex optimization.
In *International Conference on Machine Learning*, pages 7555–7564. PMLR.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020).
On the global convergence rates of softmax policy gradient methods.
In *International Conference on Machine Learning*, pages 6820–6829. PMLR.

Montúfar, G., Rauh, J., and Ay, N. (2014).
On the Fisher metric of conditional probability polytopes.
*Entropy*, 16(6):3207–3233.

# References V


Morimura, T., Uchibe, E., Yoshimoto, J., and Doya, K. (2008).
A new natural policy gradient by stationary distribution metric.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 82–97. Springer.

Müller, J. and Montúfar, G. (2022).
Geometry and convergence of natural policy gradients.
*MPI MiS Preprint 31/2022*.

Müller, J. and Montúfar, G. (2022).
The geometry of memoryless stochastic policy optimization in infinite-horizon POMDPs.
In *International Conference on Learning Representations*.

Shima, H. (2007).
*The geometry of Hessian structures*.
World Scientific, Singapore.