Natural gradient descent with momentum

Nonlinear Approximation for High-Dimensional Problems 2025
Workshop in honor of Albert Cohen

Agustín Somacal

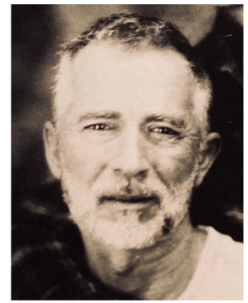
In collaboration with **Anthony Nouy**

École Centrale de Nantes

Laboratoire de Mathématiques Jean Leray


ARCTIC OCEAN
CANADA
RUSSIA
NORTH ATLANTIC OCEAN
UNITED STATES
MONGOLIA
PACIFIC OCEAN
CHINA
MEXICO
PACIFIC OCEAN
INDONESIA
BRAZIL
PERU
SOUTH ATLANTIC OCEAN
AUSTRALIA
INDIAN OCEAN
CHILE
SOUTHERN OCEAN

Albert Cohen



Welcome to the home page of Albert Cohen.

Address: Laboratoire Jacques-Louis Lions,
Sorbonne Université
4, Place Jussieu
75005 Paris, France

Phone: (33)(1)44277195
email : albert.cohen@sorbonne-universite.fr

Research interests

Abridged curriculum vitae

List of publications

Cours de M2 2023/2024

Activity group SMAI-SIGMA

The conference Curves and Surfaces 2022

The conference Foundations of Computational Mathematics 2023

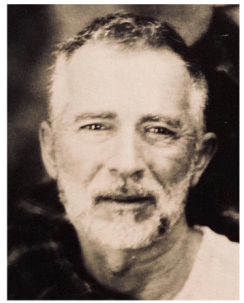
The FoCM society

Things I like to do ...

... below see level

... above see level

Albert Cohen



Welcome to the home page of Albert Cohen.

Address: Laboratoire Jacques-Louis Lions,
Sorbonne Université
4, Place Jussieu
75005 Paris, France

Phone: (33)(1)44277195
email : albert.cohen@sorbonne-universite.fr

Research interests

Abridged curriculum vitae

List of publications

Cours de M2 2023/2024

Activity group SMAI-SIGMA

The conference Curves and Surfaces 2022

The conference Foundations of Computational Mathematics 2023

The FoCM society

Things I like to do ...

.. below see level

.. above see level

Outline

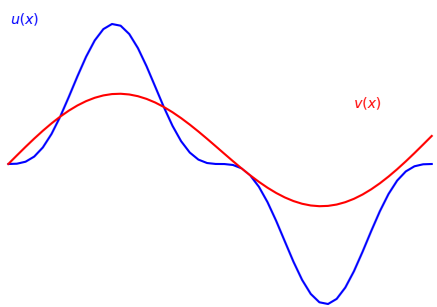
1. **What is natural gradient and why we may need it?**
   - Problem setting
   - From gradient descent to Newton's method.
   - From Newton's method to natural gradient.
   - Toy examples to gain intuition.
2. **What is momentum and when we may need it?**
3. **How to combine momentum and natural gradient.**
   - Two toy examples
   - Two less toy examples

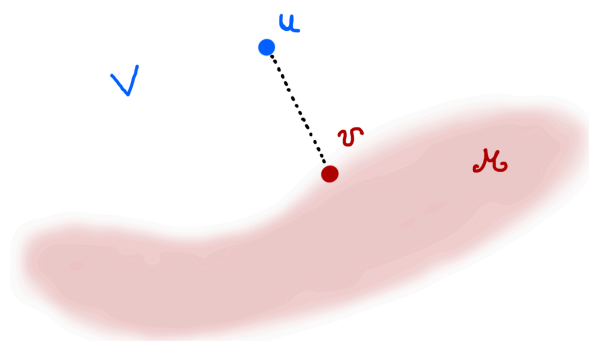
Problem formulation

Objective: approximate target function $u \in V$ by $v \in \mathcal{M} \subset V$

Target function $u : \mathbb{R}^d \to \mathbb{R} \in V$

Hilbert space $L^2(\Omega), H^1(\Omega), \ldots$





Problem formulation

Objective: approximate target function $u \in V$ by $v \in \mathcal{M} \subset V$

Target function

$$u : \mathbb{R}^d \to \mathbb{R} \in V$$

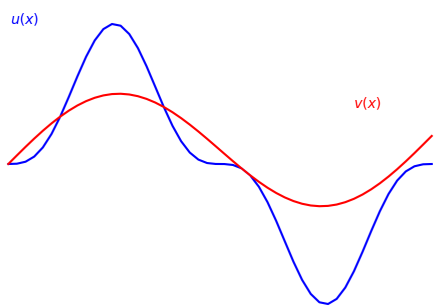
Hilbert space
$L^2(\Omega), H^1(\Omega), \ldots$

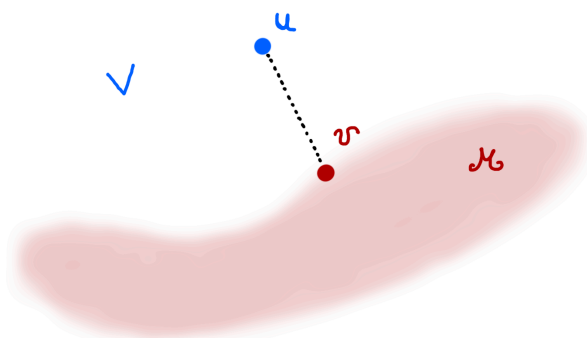
Approximation manifold

$$\mathcal{M} := \{v_\theta(x) = A(\theta)(x);\ \theta \in \mathbb{R}^p\}$$

Linear model

$$A(\theta)(x) = \theta_1 \phi_1(x) + \cdots + \theta_p \phi_p(x)$$
$$= \theta^T \mathbf{\Phi}(x)$$



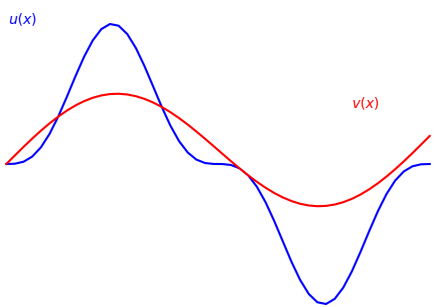


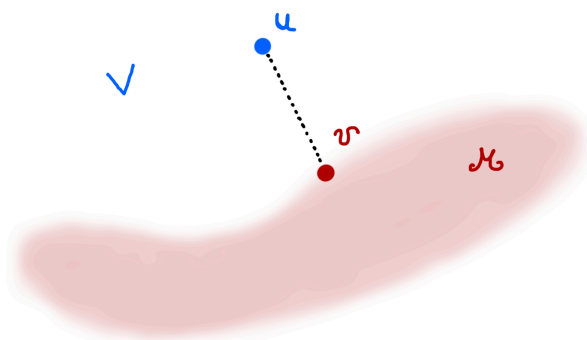
Problem formulation

Objective: approximate target function $u \in V$ by $v \in \mathcal{M} \subset V$

Target function $\quad u : \mathbb{R}^d \to \mathbb{R} \in V$ $\quad$ Hilbert space $L^2(\Omega), H^1(\Omega), \dots$

Approximation manifold $\quad \mathcal{M} := \{v_\theta(x) = A(\theta)(x);\ \theta \in \mathbb{R}^p\}$ $\quad$ Linear model, Neural network, ...





Minimization problem

Objective: approximate $u \in V$ by $v \in \mathcal{M}$.

Continuous problem

$$\begin{aligned}\mathcal{L}_u(v) &= \frac{1}{2}\|u - v\|_V^2 \\ &= \frac{1}{2}\langle u - v, u - v\rangle_V \\ &= \frac{1}{2}\int (u(x) - v(x))^2 \mathrm{d}\mu(x)\end{aligned}$$

Minimization problem

Objective: approximate $u \in V$ by $v \in \mathcal{M}$.

Continuous problem

$$\begin{aligned}\mathcal{L}_u(v) &= \frac{1}{2}\|u - v\|_V^2 \\ &= \frac{1}{2}\langle u - v, u - v\rangle_V \\ &= \frac{1}{2}\int (u(x) - v(x))^2 \mathrm{d}\mu(x)\end{aligned}$$

Discrete problem

$$\begin{aligned}\mathcal{L}_u(v) &= \frac{1}{2}\|u - v\|_m^2 \\ &= \frac{1}{2}\langle u - v, u - v\rangle_m \\ &= \frac{1}{2m}\sum_{i=1}^{m}(u(x_i) - v(x_i))^2\end{aligned}$$

Minimization problem

Objective: approximate $u \in V$ by $v \in \mathcal{M}$.

Continuous problem

$$\begin{aligned}\mathcal{L}_u(v) &= \frac{1}{2}\|u - v\|_V^2 \\ &= \frac{1}{2}\langle u - v, u - v\rangle_V \\ &= \frac{1}{2}\int (u(x) - v(x))^2 \mathrm{d}\mu(x)\end{aligned}$$

Discrete problem

$$\begin{aligned}\mathcal{L}_u(v) &= \frac{1}{2}\|u - v\|_m^2 \\ &= \frac{1}{2}\langle u - v, u - v\rangle_m \\ &= \frac{1}{2m}\sum_{i=1}^{m}(u(x_i) - v(x_i))^2\end{aligned}$$

Functional perspective

$$v^* = \underset{v \in \mathcal{M}}{\arg\min}\ \mathcal{L}_u(v)$$

Parameter perspective

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \mathcal{L}_u(A(\theta))$$

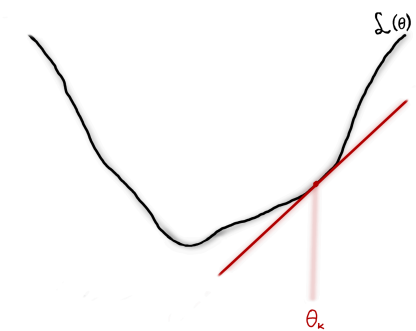
$$\mathcal{L}_u(A(\theta)) = (\mathcal{L}_u \circ A)(\theta) =: \mathcal{L}(\theta)$$

Gradient descent

Iteratively improve approximation by minimizing $\mathcal{L}(\theta_k)$.

Taylor expansion around current iterate θ_k.

$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta_k) + \langle \nabla_\theta \mathcal{L}(\theta_k), \theta - \theta_k \rangle_{\mathbb{R}^p}$$

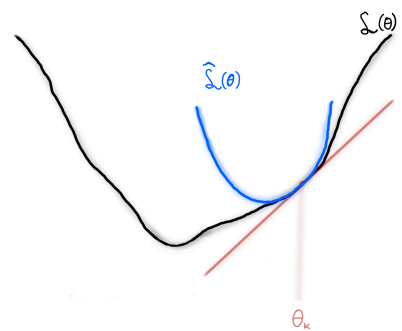


Gradient descent

Iteratively improve approximation by minimizing $\mathcal{L}(\theta_k)$.

Taylor expansion around current iterate θ_k plus penalization on the distance traveled on each step.

$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta_k) + \langle \nabla_\theta \mathcal{L}(\theta_k), \theta - \theta_k \rangle_{\mathbb{R}^p} + \frac{1}{2s} \rho(\theta, \theta_k)$$

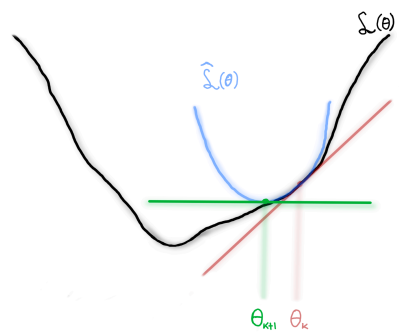


Gradient descent

Iteratively improve approximation by minimizing $\mathcal{L}(\theta_k)$.

Taylor expansion around current iterate θ_k plus penalization on the distance traveled on each step.

$$0 = \nabla_\theta \left[\mathcal{L}(\theta_k) + \langle \nabla_\theta \mathcal{L}(\theta_k), \theta - \theta_k \rangle_{\mathbb{R}^p} + \frac{1}{2s} \rho(\theta, \theta_k) \right]$$



Gradient descent

Iteratively improve approximation by minimizing $\mathcal{L}(\theta_k)$.

Taylor expansion around current iterate θ_k plus penalization on the distance traveled on each step.

$$\frac{1}{2}\nabla_\theta \rho(\theta, \theta_k) = -s\nabla_\theta \mathcal{L}(\theta_k)$$

Gradient descent

Iteratively improve approximation by minimizing $\mathcal{L}(\theta_k)$.

Taylor expansion around current iterate θ_k plus penalization on the distance traveled on each step.

$$\frac{1}{2}\nabla_\theta \rho(\theta, \theta_k) = -s\nabla_\theta \mathcal{L}(\theta_k)$$

Gradient descent

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|^2_{\mathbb{R}^p}$$

$$\nabla_\theta \rho(\theta, \theta_k) = 2(\theta - \theta_k)$$

$$\theta = \theta_k - s\nabla_\theta \mathcal{L}(\theta_k)$$

Preconditioned gradient

Iteratively improve approximation by minimizing $\mathcal{L}(\theta_k)$.

Taylor expansion around current iterate θ_k plus penalization on the distance traveled on each step.

$$\frac{1}{2}\nabla_\theta \rho(\theta, \theta_k) = -s\nabla_\theta \mathcal{L}(\theta_k)$$

Gradient descent	Preconditioned gradient
$\rho(\theta, \theta_k) = \|\theta - \theta_k\|^2_{\mathbb{R}^p}$	$\rho(\theta, \theta_k) = \|\theta - \theta_k\|^2_M$
$\nabla_\theta \rho(\theta, \theta_k) = 2(\theta - \theta_k)$	$\nabla_\theta \rho(\theta, \theta_k) = 2M(\theta - \theta_k)$
$\theta = \theta_k - s\nabla_\theta \mathcal{L}(\theta_k)$	$\theta = \theta_k - sM^{-1}\nabla_\theta \mathcal{L}(\theta_k)$

Newton’s method

Iteratively improve approximation by minimizing $\mathcal{L}(\theta_k)$.

Taylor expansion around current iterate θ_k plus penalization on the distance traveled on each step.

$$\frac{1}{2}\nabla_\theta \rho(\theta, \theta_k) = -s\nabla_\theta \mathcal{L}(\theta_k)$$

Gradient descent	Newton's method
$\rho(\theta, \theta_k) = \|\theta - \theta_k\|^2_{\mathbb{R}^p}$	$\rho(\theta, \theta_k) = \|\theta - \theta_k\|^2_H$
$\nabla_\theta \rho(\theta, \theta_k) = 2(\theta - \theta_k)$	$\nabla_\theta \rho(\theta, \theta_k) = 2H(\theta - \theta_k)$
$\theta = \theta_k - s\nabla_\theta \mathcal{L}(\theta_k)$	$\theta = \theta_k - sH^{-1}\nabla_\theta \mathcal{L}(\theta_k)$

Functional gradient and geometric intuition

Some properties [Gruhlke, Robert, Anthony Nouy, and Philipp Trunschke. 2024].

$$H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] = \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right]$$

$$= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right]$$

Functional gradient and geometric intuition

Some properties [Gruhlke, Robert, Anthony Nouy, and Philipp Trunschke. 2024].

$$H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] = \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right]$$
$$= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right]$$

Functional gradient and geometric intuition

Some properties [Gruhlke, Robert, Anthony Nouy, and Philipp Trunschke. 2024].

$$H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] = \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right]$$
$$= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right]$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|^2_{L^2(\Omega)} \qquad \longrightarrow \qquad V \ni \nabla\mathcal{L}_u = u - v$$

Natural gradient from Newton's method

[Amari, Shun-ichi. 1998] [Martens, James 2020].

$$
\begin{aligned}
H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] &= \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right] \\
&= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right] \\
&= \left\langle H_V\mathcal{L}\frac{\partial A}{\partial\theta_i}, \frac{\partial A}{\partial\theta_j}\right\rangle_V + \left\langle \nabla\mathcal{L}_u, \frac{\partial^2 A}{\partial\theta_i\partial\theta_j}\right\rangle_V \\
&= G + \langle \nabla\mathcal{L}_u, H_A\rangle_V
\end{aligned}
$$

Natural gradient from Newton's method

[Amari, Shun-ichi. 1998] [Martens, James 2020].

$$
\begin{aligned}
H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] &= \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right] \\
&= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right] \\
&= \left\langle H_V\mathcal{L}\frac{\partial A}{\partial\theta_i}, \frac{\partial A}{\partial\theta_j}\right\rangle_V + \left\langle \nabla\mathcal{L}_u, \frac{\partial^2 A}{\partial\theta_i\partial\theta_j}\right\rangle_V \\
&= G + \langle \nabla\mathcal{L}_u, H_A\rangle_V
\end{aligned}
$$

Natural gradient from Newton's method

[Amari, Shun-ichi. 1998] [Martens, James 2020].

$$
\begin{aligned}
H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] = \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right] \\
&= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right] \\
&= \left\langle H_V\mathcal{L}\frac{\partial A}{\partial\theta_i}, \frac{\partial A}{\partial\theta_j}\right\rangle_V + \left\langle \nabla\mathcal{L}_u, \frac{\partial^2 A}{\partial\theta_i\partial\theta_j}\right\rangle_V \\
&= G + \langle \nabla\mathcal{L}_u, H_A\rangle_V
\end{aligned}
$$

$$
G_{ij} = \int \frac{\partial A}{\partial\theta_i}(x)[H_V\mathcal{L}(v)](x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x).
$$

Natural gradient from Newton's method

[Amari, Shun-ichi. 1998] [Martens, James 2020].

$$
\begin{aligned}
H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] = \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right] \\
&= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right] \\
&= \left\langle H_V\mathcal{L}\frac{\partial A}{\partial\theta_i}, \frac{\partial A}{\partial\theta_j}\right\rangle_V + \left\langle \nabla\mathcal{L}_u, \frac{\partial^2 A}{\partial\theta_i\partial\theta_j}\right\rangle_V \\
&= G + \langle \nabla\mathcal{L}_u, H_A\rangle_V
\end{aligned}
$$

$$
G_{ij} = \int \frac{\partial A}{\partial\theta_i}(x)[H_V\mathcal{L}(v)](x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x).
$$

$$
\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|^2_{L^2(\Omega)} \qquad \longrightarrow \qquad H_V\mathcal{L}(v)(x) = 1
$$

Natural gradient from Newton's method

[Amari, Shun-ichi. 1998] [Martens, James 2020].

$$
\begin{aligned}
H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] = \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right] \\
&= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right] \\
&= \left\langle H_V\mathcal{L}\frac{\partial A}{\partial\theta_i}, \frac{\partial A}{\partial\theta_j}\right\rangle_V + \left\langle \nabla\mathcal{L}_u, \frac{\partial^2 A}{\partial\theta_i\partial\theta_j}\right\rangle_V \\
&= G + \langle \nabla\mathcal{L}_u, H_A\rangle_V
\end{aligned}
$$

$$
G_{ij} = \int \frac{\partial A}{\partial\theta_i}(x)[H_V\mathcal{L}(v)](x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x).
$$

$$
\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|^2_{L^2(\Omega)} \qquad \longrightarrow \qquad H_V\mathcal{L}(v)(x) = 1
$$

$$
G_{ij} = \int \left[\frac{\partial A}{\partial\theta_i}\frac{\partial A}{\partial\theta_j}\right](x)\mathrm{d}x
$$

Natural gradient from Newton's method

[Amari, Shun-ichi. 1998] [Martens, James 2020].

$$H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} = \frac{\partial}{\partial\theta_i}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\right] = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}(\mathcal{L}_u \circ A)\right] = \frac{\partial}{\partial\theta_i}\left[\int_\Omega \nabla\mathcal{L}_u(v_\theta)(x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x)\right]$$

$$= \frac{\partial}{\partial\theta_i}\left[\left\langle \nabla\mathcal{L}_u, \frac{\partial A}{\partial\theta_j}\right\rangle_V\right]$$

$$= \left\langle H_V\mathcal{L}\frac{\partial A}{\partial\theta_i}, \frac{\partial A}{\partial\theta_j}\right\rangle_V + \left\langle \nabla\mathcal{L}_u, \frac{\partial^2 A}{\partial\theta_i\partial\theta_j}\right\rangle_V$$

$$= G + \cancel{\langle \nabla\mathcal{L}_u, H_A\rangle_V}$$

Model linearization

$$G_{ij} = \int \frac{\partial A}{\partial\theta_i}(x)[H_V\mathcal{L}(v)](x)\frac{\partial A}{\partial\theta_j}(x)\mathrm{d}\mu(x).$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|^2_{L^2(\Omega)} \qquad \longrightarrow \qquad H_V\mathcal{L}(v)(x) = 1$$

$$G_{ij} = \int \left[\frac{\partial A}{\partial\theta_i}\frac{\partial A}{\partial\theta_j}\right](x)\mathrm{d}x$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow
in **parameter space**

$$\frac{\partial\theta}{\partial s} = -M^{-1}\nabla_\theta\mathcal{L}$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow
in **parameter space**

$$\frac{\partial \theta}{\partial s} = -M^{-1}\nabla_\theta \mathcal{L}$$

$$\begin{aligned}
\frac{\partial v_\theta}{\partial s} = \frac{\partial v_\theta}{\partial \theta}\frac{\partial \theta}{\partial s} = \frac{\partial A}{\partial \theta}^T M^{-1}\nabla_\theta\mathcal{L} &= -\frac{\partial A}{\partial \theta}^T M^{-1}\left\langle \nabla\mathcal{L}, \frac{\partial A}{\partial \theta}\right\rangle_V \\
&= -\left(G^{-\frac{1}{2}}\frac{\partial A}{\partial \theta}\right)^T G^{\frac{1}{2}}M^{-1}G^{\frac{1}{2}}\left\langle \nabla\mathcal{L}, G^{-\frac{1}{2}}\frac{\partial A}{\partial \theta}\right\rangle_V \\
&= -E^T G^{\frac{1}{2}}M^{-1}G^{\frac{1}{2}}\left\langle \nabla\mathcal{L}, E\right\rangle_V
\end{aligned}$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow
in **parameter space**

$$\frac{\partial \theta}{\partial s} = -M^{-1}\nabla_\theta \mathcal{L}$$

$$\begin{aligned}
\frac{\partial v_\theta}{\partial s} = \frac{\partial v_\theta}{\partial \theta}\frac{\partial \theta}{\partial s} = \frac{\partial A}{\partial \theta}^T M^{-1}\nabla_\theta\mathcal{L} &= -\frac{\partial A}{\partial \theta}^T M^{-1}\left\langle \nabla\mathcal{L}, \frac{\partial A}{\partial \theta}\right\rangle_V \\
&= -\left(G^{-\frac{1}{2}}\frac{\partial A}{\partial \theta}\right)^T G^{\frac{1}{2}} M^{-1} G^{\frac{1}{2}} \left\langle \nabla\mathcal{L}, G^{-\frac{1}{2}}\frac{\partial A}{\partial \theta}\right\rangle_V \\
&= -E^T G^{\frac{1}{2}} M^{-1} G^{\frac{1}{2}} \left\langle \nabla\mathcal{L}, E\right\rangle_V
\end{aligned}$$

$$\langle E, E\rangle_V = \int_\Omega \left(G^{-\frac{1}{2}}\frac{\partial A}{\partial \theta}\right)^T \left(G^{-\frac{1}{2}}\frac{\partial A}{\partial \theta}\right) \mathrm{d}\mu(x) \quad = \int_\Omega \frac{\partial A}{\partial \theta}^T G^{-1}\frac{\partial A}{\partial \theta}\mathrm{d}\mu(x) = I$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow in **parameter space**

$$\frac{\partial \theta}{\partial s} = -M^{-1}\nabla_\theta \mathcal{L}$$

Preconditioned gradient flow in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G^{\frac{1}{2}} M^{-1} G^{\frac{1}{2}} \langle \nabla\mathcal{L}, E \rangle_V$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow
in **parameter space**

$$\frac{\partial \theta}{\partial s} = -M^{-1}\nabla_\theta \mathcal{L}$$

Preconditioned gradient flow
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G^{\frac{1}{2}} M^{-1} G^{\frac{1}{2}} \langle \nabla\mathcal{L}, E\rangle_V$$

Gradient descent $M = I$
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G^{\frac{1}{2}} I^{-1} G^{\frac{1}{2}} \langle \nabla\mathcal{L}, E\rangle_V$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow
in **parameter space**

$$\frac{\partial \theta}{\partial s} = -M^{-1}\nabla_\theta \mathcal{L}$$

Preconditioned gradient flow
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G^{\frac{1}{2}} M^{-1} G^{\frac{1}{2}} \langle \nabla\mathcal{L}, E\rangle_V$$

Gradient descent $M = I$
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G \langle \nabla\mathcal{L}, E\rangle_V$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow
in **parameter space**

$$\frac{\partial \theta}{\partial s} = -M^{-1}\nabla_\theta \mathcal{L}$$

Preconditioned gradient flow
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G^{\frac{1}{2}} M^{-1} G^{\frac{1}{2}} \langle \nabla\mathcal{L}, E\rangle_V$$

Gradient descent $M = I$
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G \langle \nabla\mathcal{L}, E\rangle_V$$

Natural gradient $M = G$
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G^{\frac{1}{2}} G^{-1} G^{\frac{1}{2}} \langle \nabla\mathcal{L}, E\rangle_V$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow
in **parameter space**

$$\frac{\partial \theta}{\partial s} = -M^{-1}\nabla_\theta \mathcal{L}$$

Preconditioned gradient flow
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G^{\frac{1}{2}} M^{-1} G^{\frac{1}{2}} \langle \nabla\mathcal{L}, E\rangle_V$$

Gradient descent $M = I$
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G \langle \nabla\mathcal{L}, E\rangle_V$$

Natural gradient $M = G$
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T \langle \nabla\mathcal{L}, E\rangle_V$$

Natural gradient flow dynamics

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|_G^2 \qquad \longrightarrow \qquad \theta = \theta_k - sG^{-1}\nabla\mathcal{L}(\theta_k)$$

Preconditioned gradient flow
in **parameter space**

$$\frac{\partial \theta}{\partial s} = -M^{-1}\nabla_\theta \mathcal{L}$$

Preconditioned gradient flow
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G^{\frac{1}{2}} M^{-1} G^{\frac{1}{2}} \langle \nabla\mathcal{L}, E\rangle_V$$

Gradient descent $M = I$
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -E^T G \langle \nabla\mathcal{L}, E\rangle_V$$

Natural gradient $M = G$
in **functional space**

$$\frac{\partial v_\theta}{\partial s} = -P_{\mathcal{T}_k} \nabla\mathcal{L}$$

Toy example

Gradient descent trajectory.

$$u \in L^2([0,1])$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|^2$$

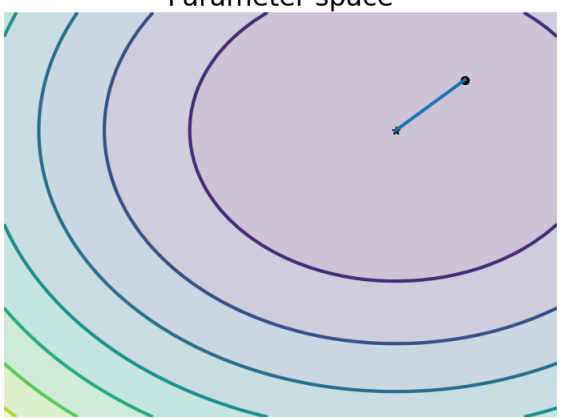
$$v_\theta(x) = A(\theta)(x) = \theta^T \Phi(x)$$

$$= \theta^T \begin{bmatrix} 1 \\ \sqrt{2}\sin(2\pi x) \end{bmatrix}$$

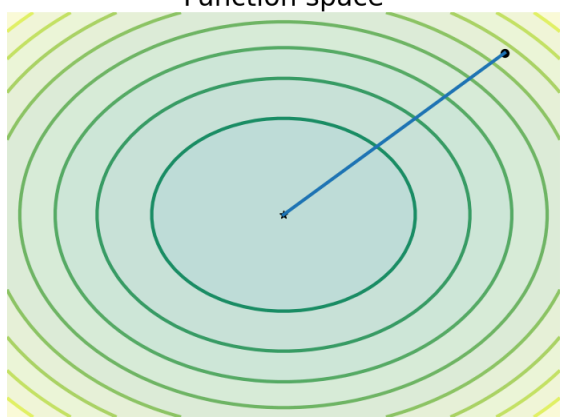
$$\frac{\partial A}{\partial \theta_i}(\theta)(x) = \Phi_i(x)$$

$$G_{ij} = \int \left[\frac{\partial A}{\partial \theta_i}\frac{\partial A}{\partial \theta_j}\right](x)\mathrm{d}x = \delta_{ij}$$

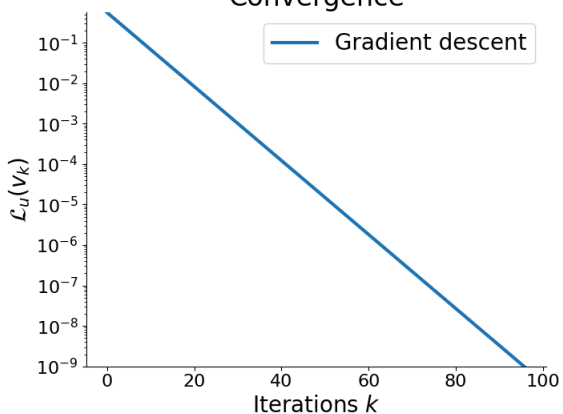
Parameter space



Function space



Convergence



Toy example

Gradient descent is biased in functional space.

$$u \in L^2([0,1])$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|^2$$

$$v_\theta(x) = A(\theta)(x) = \theta^T B \Phi(x)$$

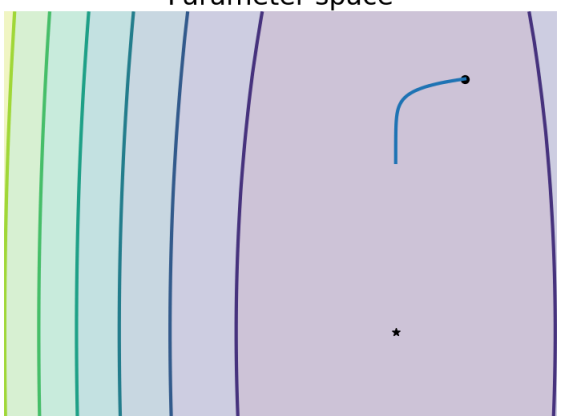
$$= \theta^T B \begin{bmatrix} 1 \\ \sqrt{2}\sin(2\pi x) \end{bmatrix}$$

$$\mathbb{R}^{p\times p} \ni B = \begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix}$$

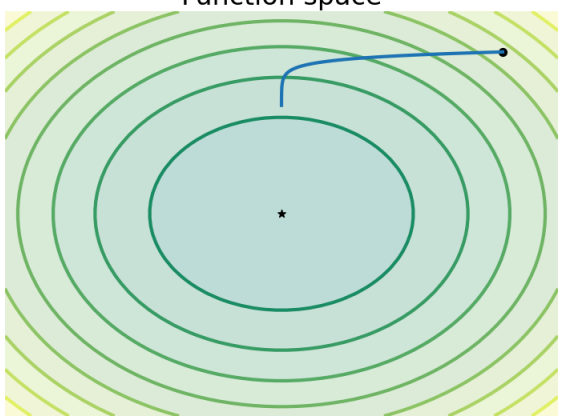
$$\frac{\partial A}{\partial \theta_i}(\theta)(x) = B_i \Phi_i(x)$$

$$G_{ij} = \int \left[\frac{\partial A}{\partial \theta_i}[B^T B]_{ij} \frac{\partial A}{\partial \theta_j}\right](x)\mathrm{d}x = [B^T B]_{ij}$$

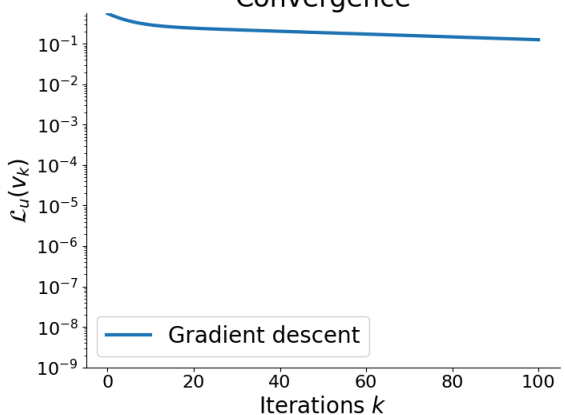
Parameter space



Function space



Convergence



Toy example

Natural gradient descent.

$$u \in L^2([0,1])$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|^2$$

$$v_\theta(x) = A(\theta)(x) = \theta^T B \Phi(x)$$

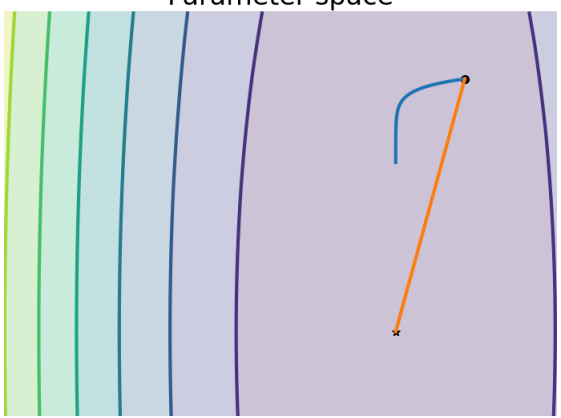
$$= \theta^T B \begin{bmatrix} 1 \\ \sqrt{2}\sin(2\pi x) \end{bmatrix}$$

$$\mathbb{R}^{p\times p} \ni B = \begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix}$$

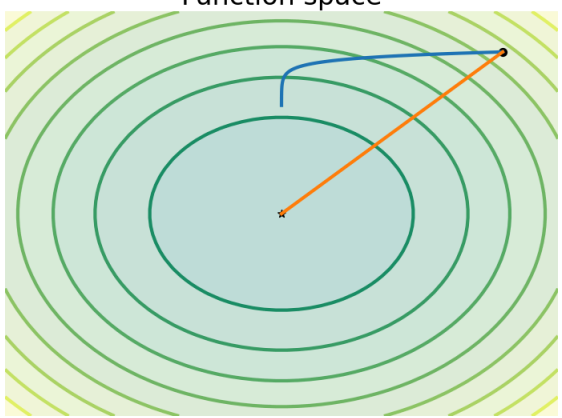
$$\frac{\partial A}{\partial \theta_i}(\theta)(x) = B_i \Phi_i(x)$$

$$G_{ij} = \int \left[\frac{\partial A}{\partial \theta_i}[B^T B]_{ij} \frac{\partial A}{\partial \theta_j}\right](x)\mathrm{d}x = [B^T B]_{ij}$$

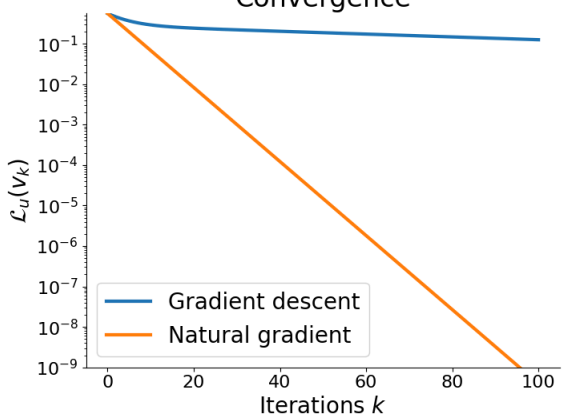
Parameter space



Function space



Convergence



Toy example

Non isotropic loss.

$$u \in L^2([0,1])$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K^2$$

$$v_\theta(x) = A(\theta)(x) = \theta^T B \Phi(x)$$

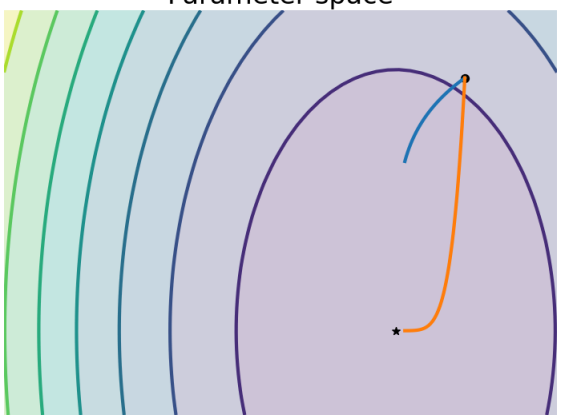
$$= \theta^T B \begin{bmatrix} 1 \\ \sqrt{2}\sin(2\pi x) \end{bmatrix}$$

$$\mathbb{R}^{p\times p} \ni B = \begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix}$$

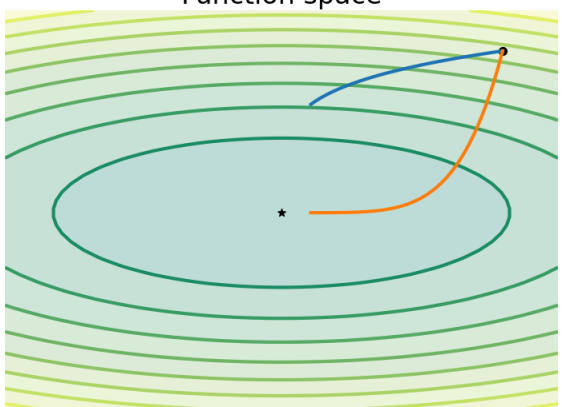
$$\frac{\partial A}{\partial \theta_i}(\theta)(x) = B_i \Phi_i(x)$$

$$G_{ij} = \int \left[\frac{\partial A}{\partial \theta_i}[B^T K B]_{ij} \frac{\partial A}{\partial \theta_j}\right](x)\mathrm{d}x = [B^T K B]_{ij}$$

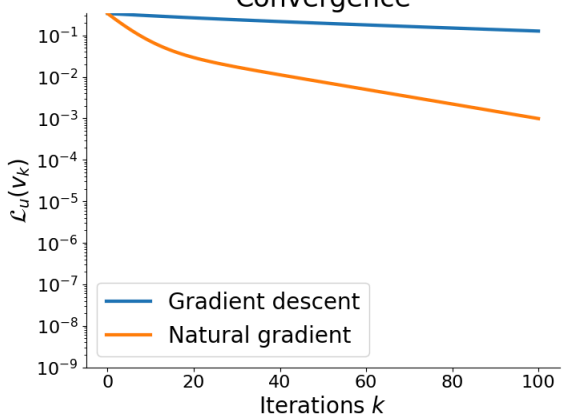
Parameter space



Function space



Convergence



Toy example

Natural gradient descent with loss hessian.

$$u \in L^2([0,1])$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K^2$$

$$v_\theta(x) = A(\theta)(x) = \theta^T B \Phi(x)$$

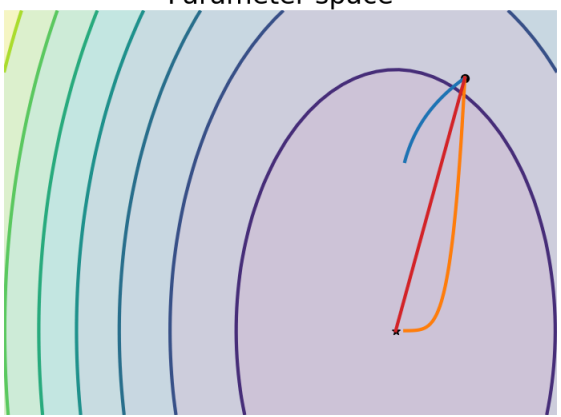
$$= \theta^T B \begin{bmatrix} 1 \\ \sqrt{2}\sin(2\pi x) \end{bmatrix}$$

$$\mathbb{R}^{p\times p} \ni B = \begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix}$$

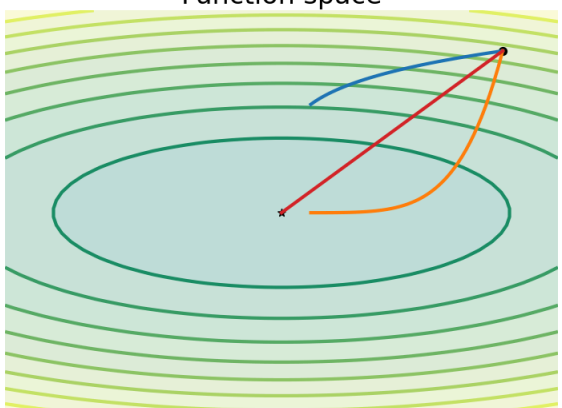
$$\frac{\partial A}{\partial \theta_i}(\theta)(x) = B_i \Phi_i(x)$$

$$G_{ij} = \int \left[\frac{\partial A}{\partial \theta_i}[B^T K B]_{ij} \frac{\partial A}{\partial \theta_j}\right](x)\mathrm{d}x = [B^T K B]_{ij}$$

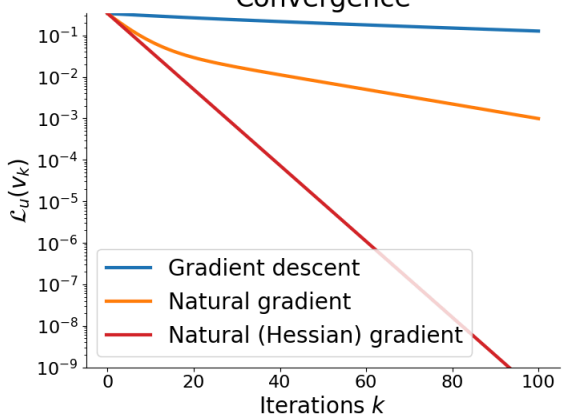
Parameter space



Function space



Convergence



Toy example

Natural gradient and Newton method are equivalent for linear models.

$$u \in L^2([0,1])$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K^2$$

$$v_\theta(x) = A(\theta)(x) = \theta^T B \Phi(x)$$

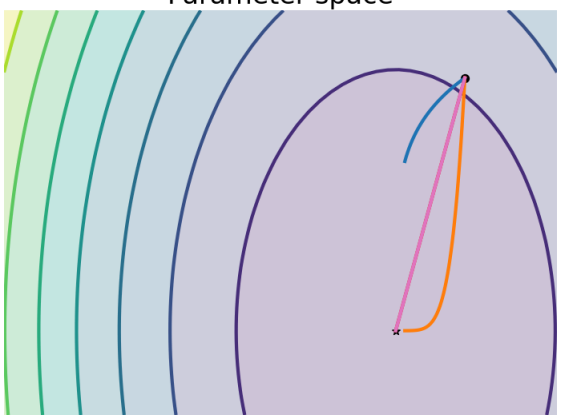
$$= \theta^T B \begin{bmatrix} 1 \\ \sqrt{2}\sin(2\pi x) \end{bmatrix}$$

$$\mathbb{R}^{p\times p} \ni B = \begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix}$$

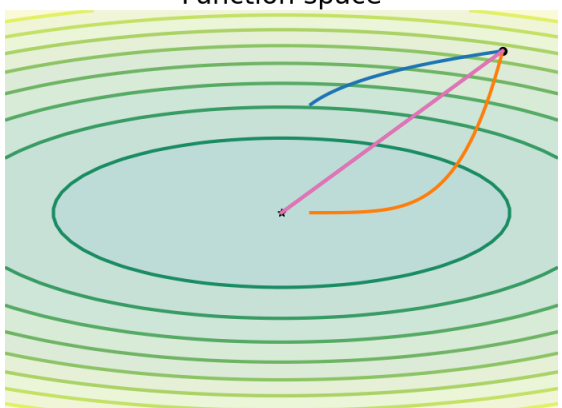
$$\frac{\partial A}{\partial \theta_i}(\theta)(x) = B_i \Phi_i(x)$$

$$G_{ij} = \int \left[\frac{\partial A}{\partial \theta_i}[B^T K B]_{ij} \frac{\partial A}{\partial \theta_j}\right](x)\mathrm{d}x = [B^T K B]_{ij}$$

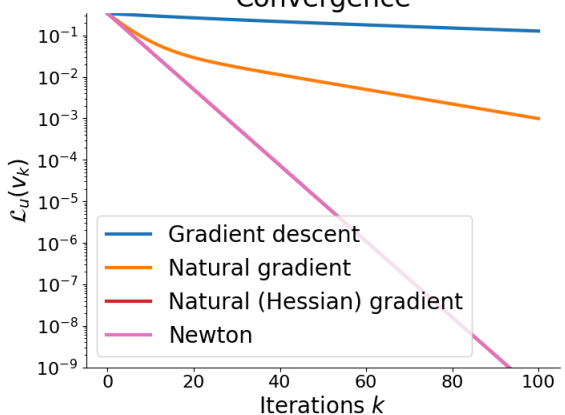
Parameter space



Function space



Convergence



Toy example

Nonlinear manifold.

$$u \in L^2([0,1])$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K^2$$

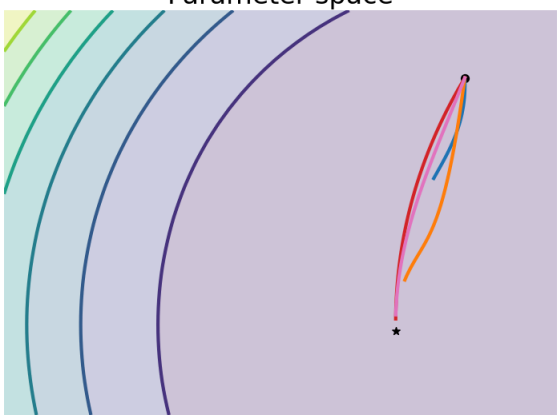
$$v_\theta(x) = A(\theta)(x) = \theta^T(B + Q\theta)\Phi(x)$$

$$\mathbb{R}^{p\times p} \ni B = \begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix}, Q \in \mathbb{R}^{p\times p\times p}$$

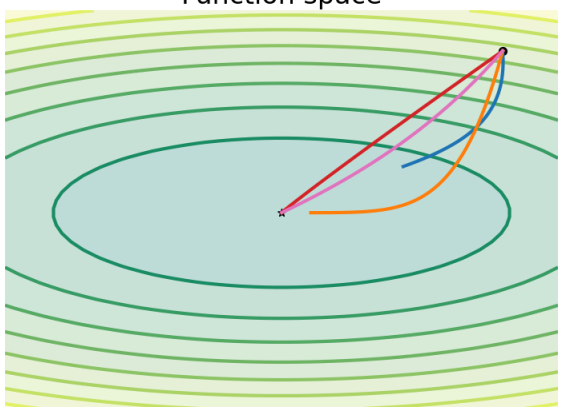
$$\frac{\partial A}{\partial \theta_i}(\theta)(x) = (B_i + Q_i\theta)\Phi_i(x)$$

$$G_{ij}(\theta) = [(B + Q_i\theta)^T K (B + Q_i\theta)]_{ij}$$

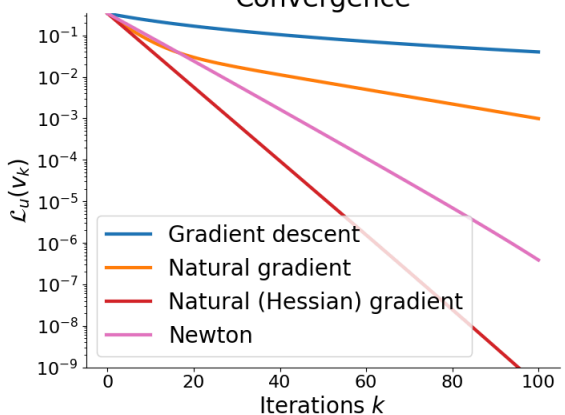
Parameter space



Function space



Convergence



Why we need momentum

Using momentum to **scape local minima** and **correct the biased step** of NGD.

Beyond L^2 loss

Classification problem: KL-div

$$\mathcal{L}_u(v) = \int v(x) \log \frac{v(x)}{u(x)} \mathrm{d}x$$

Stochastic setting empirical loss

$$\mathcal{L}_u(v_k) = \|u - v_k\|_m^2$$
$$\frac{1}{2m} \sum_{i=1}^{m} (u(x_{I_i^k}) - v(x_{I_i^k}))^2$$

PDE residual

$$\mathcal{L}(v) = \|R(v)\|^2$$
$$= \| - \epsilon \partial_{xx} v + \partial_x v - 1\|^2$$

Escape local minima

$$\frac{\partial v_\theta}{\partial s} = -P_{\mathcal{T}_k} \nabla \mathcal{L}$$

Non-linear model

$$\theta_{k+1} = \theta_k - s G_k^{-1} \nabla_\theta \mathcal{L}(\theta_k)$$

$$\theta_s = \theta_k - \int_0^s G_k^{-1}(\theta_\ell) \nabla_\theta \mathcal{L}(\theta_\ell) \mathrm{d}\ell$$

Momentum dynamics

From gradient flow to momentum [Polyak, B.T. 1964] [Nesterov, Yurii. 1983].

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = -\nabla_\theta \mathcal{L}$$

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}s^2} = -\gamma\frac{\mathrm{d}\theta}{\mathrm{d}s} - \nabla_\theta \mathcal{L}$$

Heavy-ball

$$\begin{aligned}\theta_{k+1} &= \theta_k + p_k \\ p_k &= \beta p_{k-1} - \alpha \nabla_\theta \mathcal{L}_u(\theta_k)\end{aligned}$$

Nestorov

$$\begin{aligned}y_k &= \theta_k + \beta(\theta_k - \theta_{k-1}) \\ \theta_{k+1} &= y_k - \alpha \nabla_\theta \mathcal{L}_u(y_k)\end{aligned}$$

$$\theta_{k+1} = \theta_k + \beta p_{k-1} - \alpha \nabla_\theta \mathcal{L}_u(\theta_k)$$

$$\theta_{k+1} = \theta_k + \beta(\theta_k - \theta_{k-1}) - \alpha \nabla_\theta \mathcal{L}_u(y_k)$$

Momentum dynamics in functional space

From momentum in parameter space to functional space.

Heavy-ball

$$\theta_{k+1} - \theta_k = \beta p_{k-1} - \alpha \nabla_\theta \mathcal{L}_u(\theta_k)$$

$$\mathcal{T}_k \ni \mathrm{d}v_k = P_{\mathcal{T}_k}[\beta p_{k-1} - \alpha \nabla \mathcal{L}(v_k)]$$

$$P_{\mathcal{T}_k} \nabla \mathcal{L}(v_k)$$

$$P_{\mathcal{T}_k} p_{k-1}$$

Nestorov

$$\theta_{k+1} - \theta_k = \beta(\theta_k - \theta_{k-1}) - \alpha \nabla_\theta \mathcal{L}_u(y_k)$$

$$\mathcal{T}_k \ni \mathrm{d}v_k = P_{\mathcal{T}_k}[\beta(v_k - v_{k-1}) - \alpha \nabla \mathcal{L}(v_k)]$$

$$P_{\mathcal{T}_k}(v_k - v_{k-1})$$

Momentum dynamics in functional space

From momentum in parameter space to functional space.

Heavy-ball

$$\theta_{k+1} - \theta_k = \beta p_{k-1} - \alpha \nabla_\theta \mathcal{L}_u(\theta_k)$$

$$\mathcal{T}_k \ni \mathrm{d}v_k^{HB} = P_{\mathcal{T}_k}[\beta p_{k-1} - \alpha \nabla \mathcal{L}(v_k)]$$

$$P_{\mathcal{T}_k} \nabla \mathcal{L}(v_k)$$

$$G_k^{-1} \nabla_\theta \mathcal{L}(\theta_k)$$

$$P_{\mathcal{T}_k} p_{k-1}$$

$$G_k^{-1} G_{k,k-1} p_{k-1}$$

$$G_{k,k-1} = \int \left[\left. \frac{\partial A}{\partial \theta} \right|_{\theta_k} \left. \frac{\partial A}{\partial \theta} \right|_{\theta_{k-1}} \right] (x) \mathrm{d}x$$

Nestorov

$$\theta_{k+1} - \theta_k = \beta(\theta_k - \theta_{k-1}) - \alpha \nabla_\theta \mathcal{L}_u(y_k)$$

$$\mathcal{T}_k \ni \mathrm{d}v_k^{N} = P_{\mathcal{T}_k}[\beta(v_k - v_{k-1}) - \alpha \nabla \mathcal{L}(v_k)]$$

$$P_{\mathcal{T}_k}(v_k - v_{k-1})$$

$$G_k^{-1} \int \left[\left. \frac{\partial A}{\partial \theta} \right|_{\theta_k} (v_k - v_{k-1}) \right] (x) \mathrm{d}x$$

Momentum dynamics in functional space

From momentum in parameter space to functional space.

Heavy-ball

$$\theta_{k+1} - \theta_k = \beta p_{k-1} - \alpha \nabla_\theta \mathcal{L}_u(\theta_k)$$

$$\mathcal{T}_k \ni \mathrm{d}v_k^{HB} = \beta P_{\mathcal{T}_k} p_{k-1} - \alpha P_{\mathcal{T}_k}^{H_\mathcal{L}} \nabla \mathcal{L}(v_k)$$

$$P_{\mathcal{T}_k}^{H_\mathcal{L}} \nabla \mathcal{L}(v_k)$$

$$G_k^{-1} \nabla_\theta \mathcal{L}(\theta_k)$$

$$G_k = \int \left[\frac{\partial A}{\partial \theta} H_\mathcal{L} \frac{\partial A}{\partial \theta} \right] (x) \mathrm{d}x$$

$$P_{\mathcal{T}_k} p_{k-1}$$

$$G_k^{-1} G_{k,k-1} p_{k-1}$$

$$G_{k,k-1} = \int \left[\frac{\partial A}{\partial \theta}\bigg|_{\theta_k} \frac{\partial A}{\partial \theta}\bigg|_{\theta_{k-1}} \right] (x) \mathrm{d}x$$

Nestorov

$$\theta_{k+1} - \theta_k = \beta(\theta_k - \theta_{k-1}) - \alpha \nabla_\theta \mathcal{L}_u(y_k)$$

$$\mathcal{T}_k \ni \mathrm{d}v_k^{N} = \beta P_{\mathcal{T}_k} (v_k - v_{k-1}) - \alpha P_{\mathcal{T}_k}^{H_\mathcal{L}} \nabla \mathcal{L}(v_k)$$

$$P_{\mathcal{T}_k} (v_k - v_{k-1})$$

$$G_k^{-1} \int \left[\frac{\partial A}{\partial \theta}\bigg|_{\theta_k} (v_k - v_{k-1}) \right] (x) \mathrm{d}x$$

Toy example

Escaping local minima.

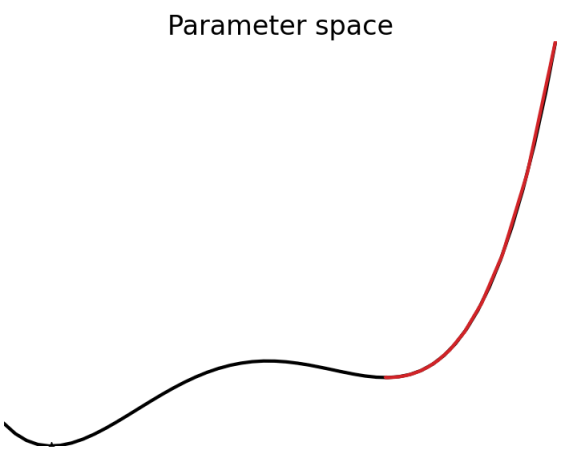
$$u \in L^2([0,1])$$

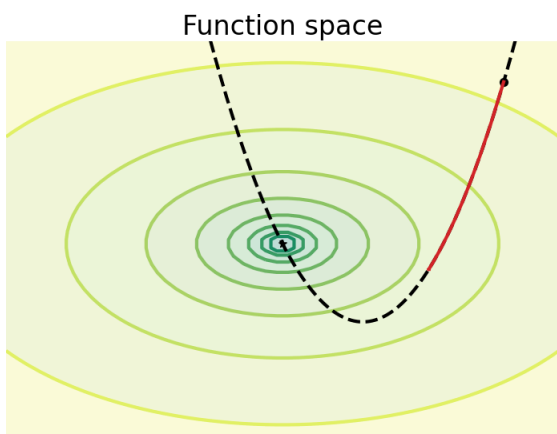
$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K^2$$

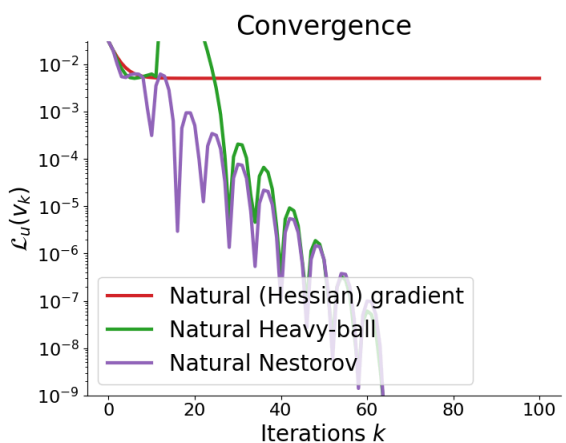
$$v_\theta(x) = \theta_1 b^T \Phi(x) + \theta_1^2 b^{\perp T} \Phi(x)$$

$$\mathrm{d}v_k^{HB} = P_{\mathcal{T}_k}[\beta p_{k-1} - \alpha \nabla \mathcal{L}(v_k)]$$

$$\mathrm{d}v_k^{N} = P_{\mathcal{T}_k}[\beta (v_k - v_{k-1}) - \alpha \nabla \mathcal{L}(v_k)]$$







Toy example

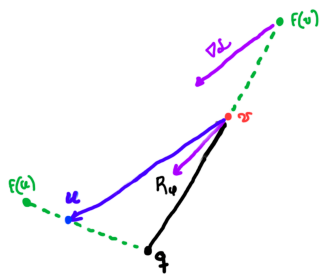
Not L^2 loss.

$$u \in L^2([0,1])$$

$$\mathcal{L}_u(v) = \frac{1}{2}\|f(u) - f(v)\|_K^2$$

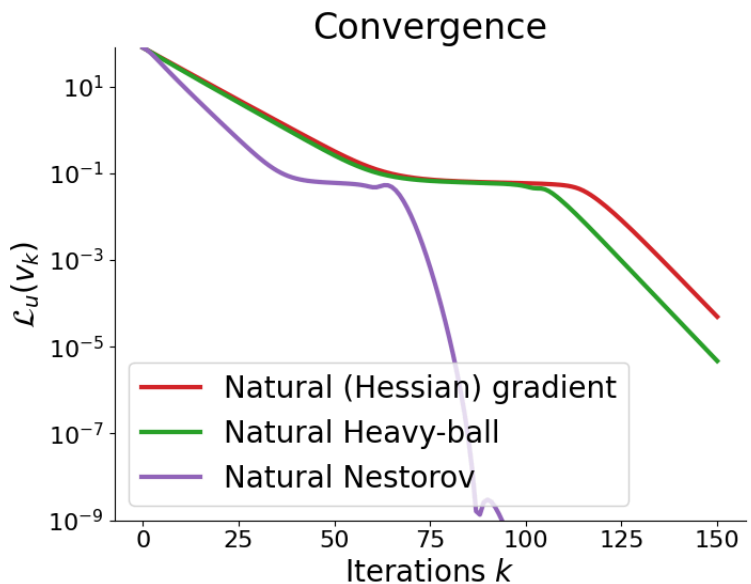
$$f(v) = (1 + \omega\|v - q\|^2)(v - q) + q$$

$$q = R_\varphi(u - v) + v$$



$$\mathrm{d}v_k^{HB} = P_{\mathcal{T}_k}[\beta p_{k-1} - \alpha\nabla\mathcal{L}(v_k)]$$

$$\mathrm{d}v_k^{N} = P_{\mathcal{T}_k}[\beta(v_k - v_{k-1}) - \alpha\nabla\mathcal{L}(v_k)]$$



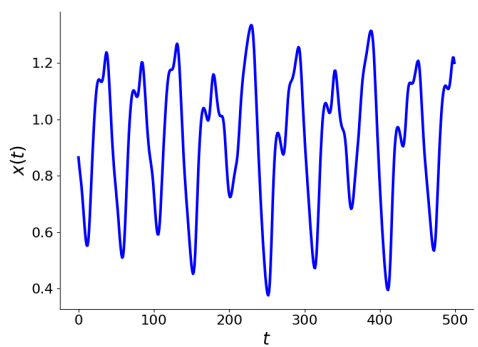
Mackey Glass

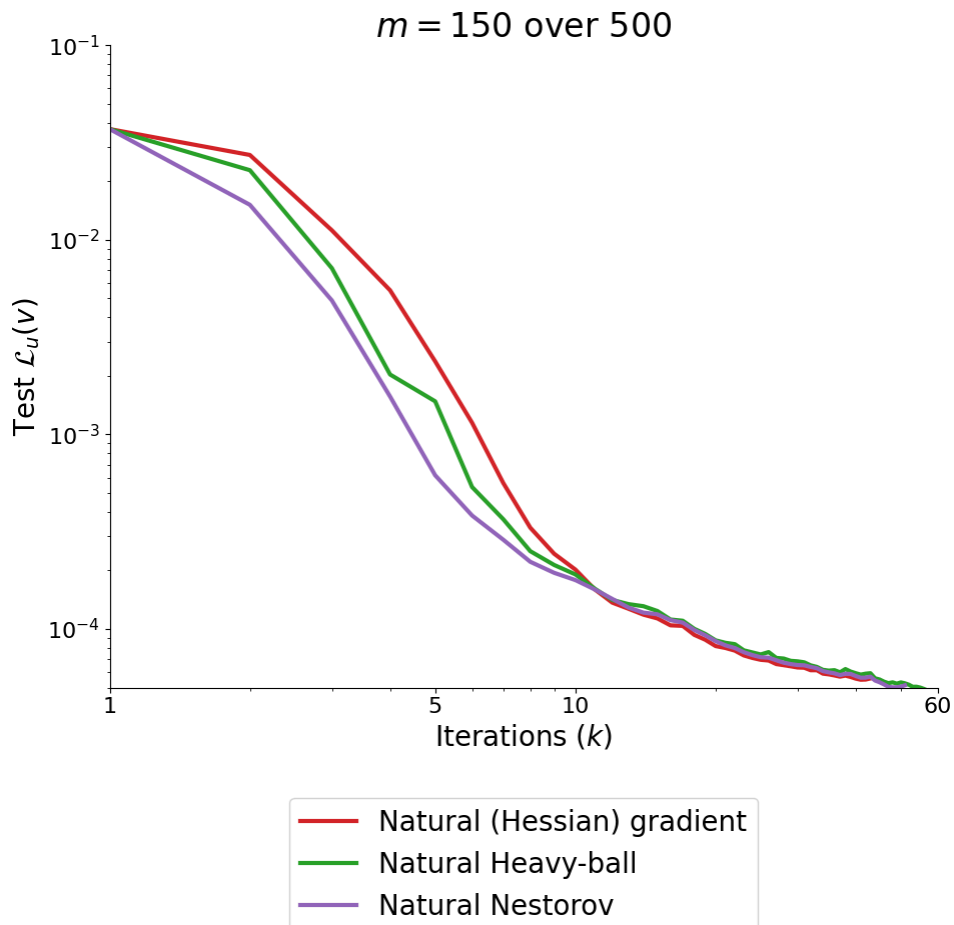
A less toy example [Park, H, S.-I Amari, and K Fukum (2000)].

Mackey Glass caotic time series:

- $z(t+1) = (1-b)z(t) + a\frac{z(t-\tau)}{1+z(t-\tau)^{10}}$
- Input: $z(t), z(t-6), z(t-12), z(t-18)$
- Output: $z(t+6)$

Model (shallow NN 10 neurons): $v_\theta : \mathbb{R}^4 \to \mathbb{R}$





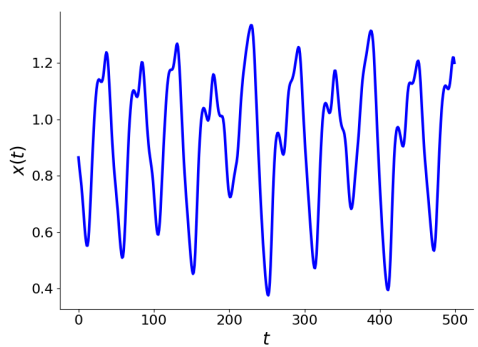
Mackey Glass

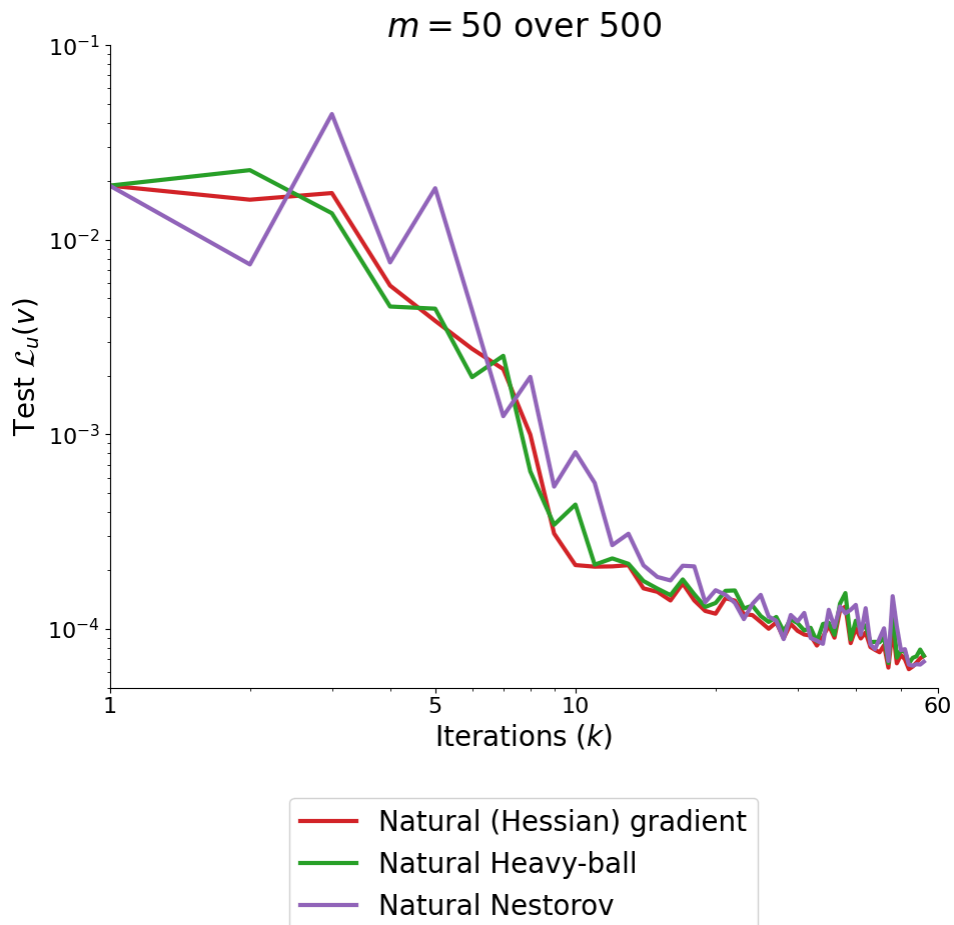
A less toy example [Park, H, S.-I Amari, and K Fukum (2000)].

Mackey Glass caotic time series:

- $z(t+1) = (1-b)z(t) + a\frac{z(t-\tau)}{1+z(t-\tau)^{10}}$
- Input: $z(t), z(t-6), z(t-12), z(t-18)$
- Output: $z(t+6)$

Model (shallow NN 10 neurons): $v_\theta : \mathbb{R}^4 \to \mathbb{R}$





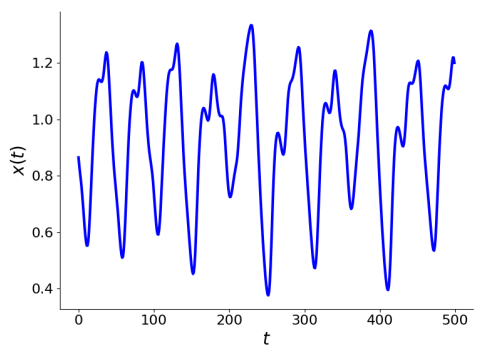
Mackey Glass

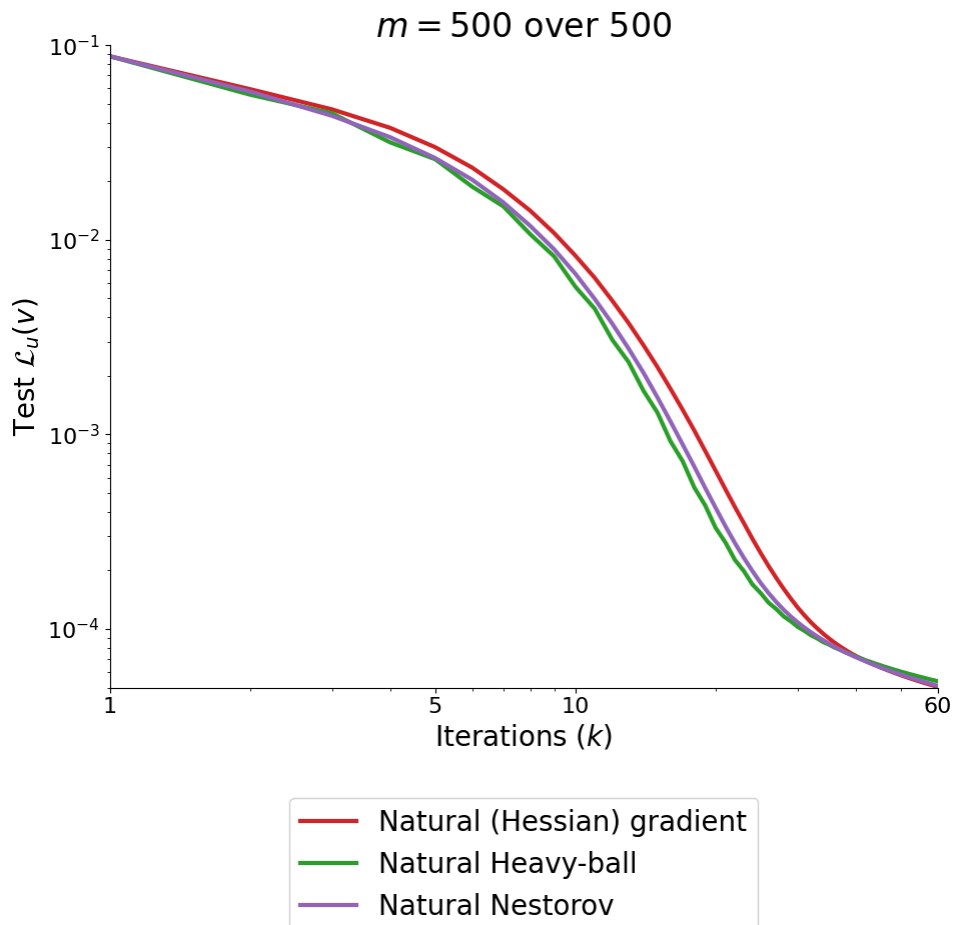
A less toy example [Park, H, S.-I Amari, and K Fukum (2000)].

Mackey Glass caotic time series:

- $z(t+1) = (1-b)z(t) + a\frac{z(t-\tau)}{1+z(t-\tau)^{10}}$
- Input: $z(t), z(t-6), z(t-12), z(t-18)$
- Output: $z(t+6)$

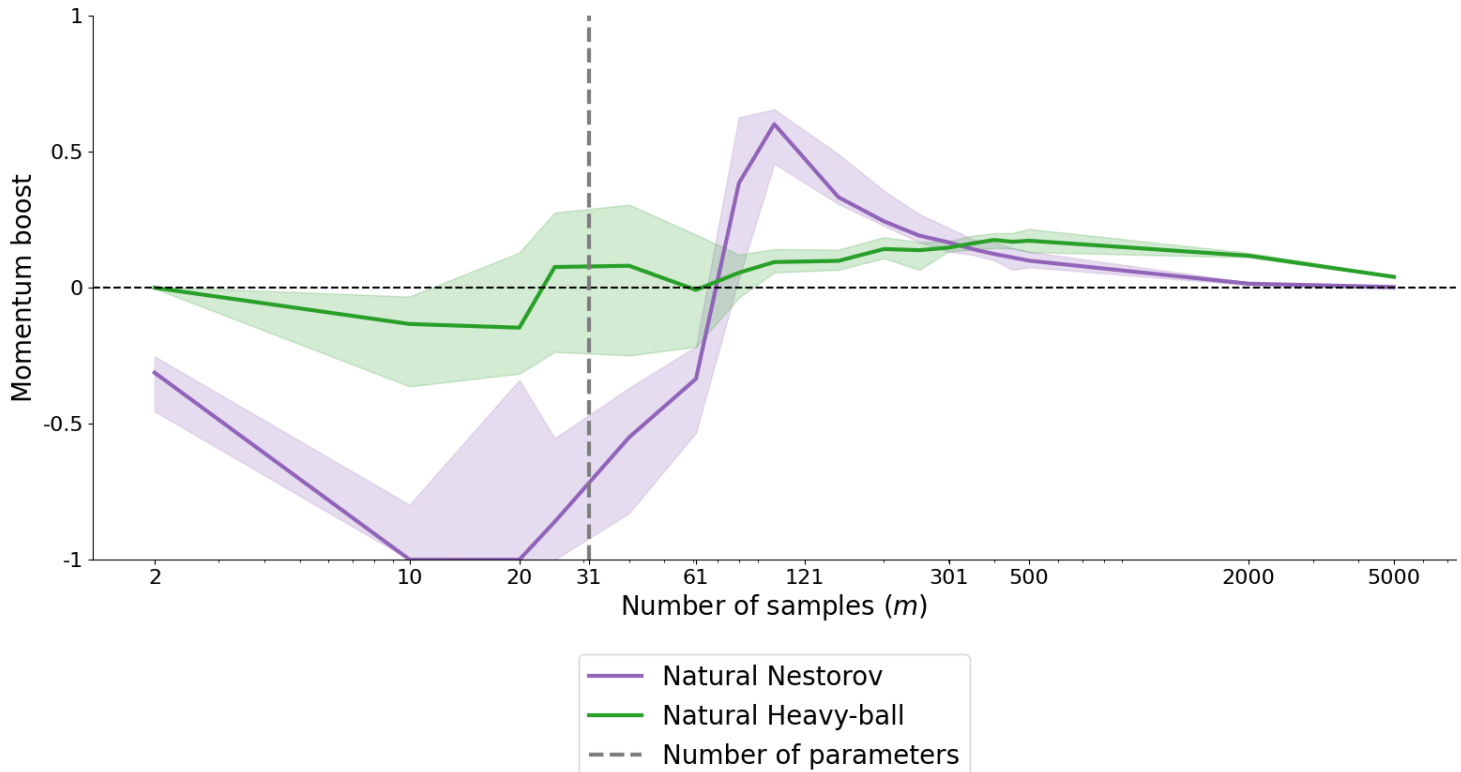
Model (shallow NN 10 neurons): $v_\theta : \mathbb{R}^4 \to \mathbb{R}$





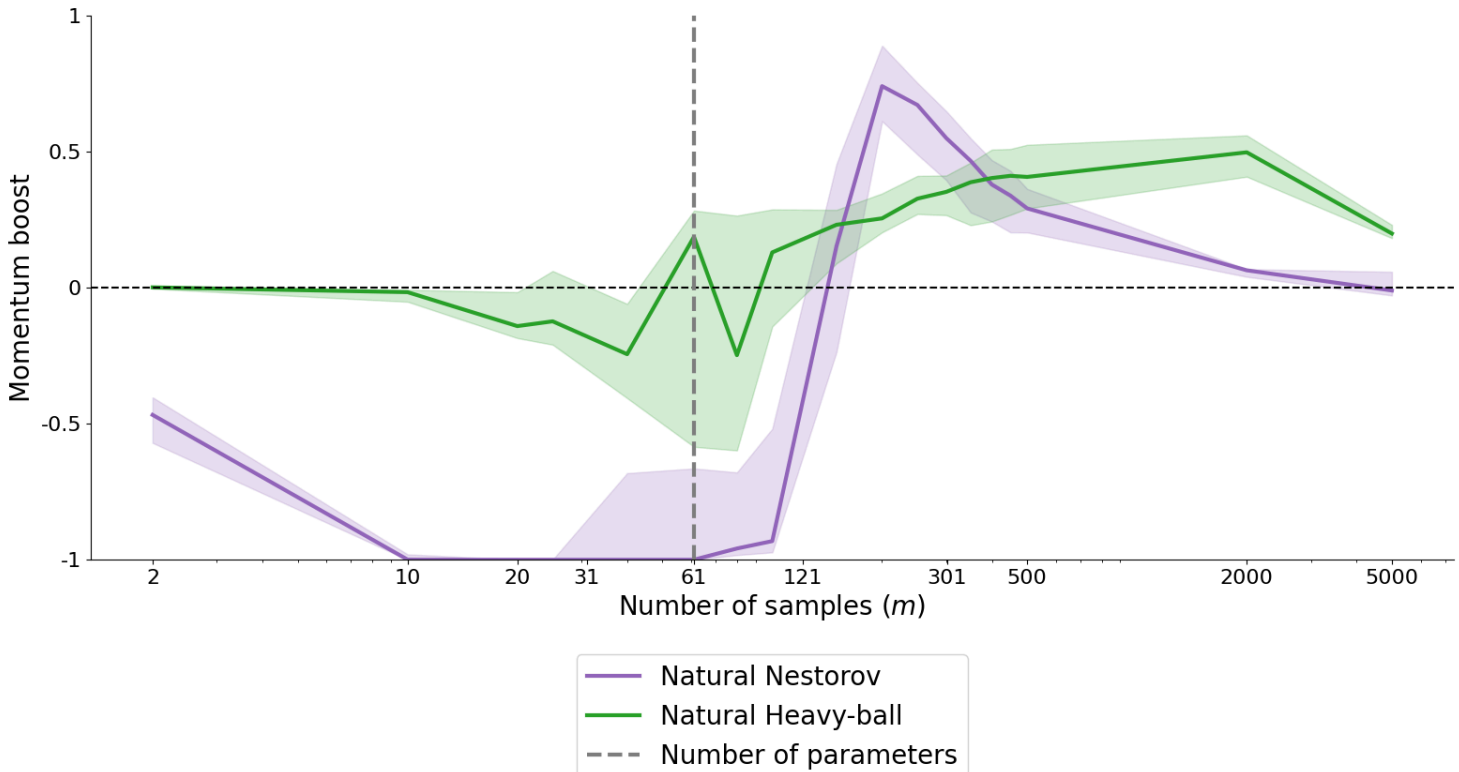
Mackey Glass: sampling

Effect of natural gradient direction estimation with limited number of samples.



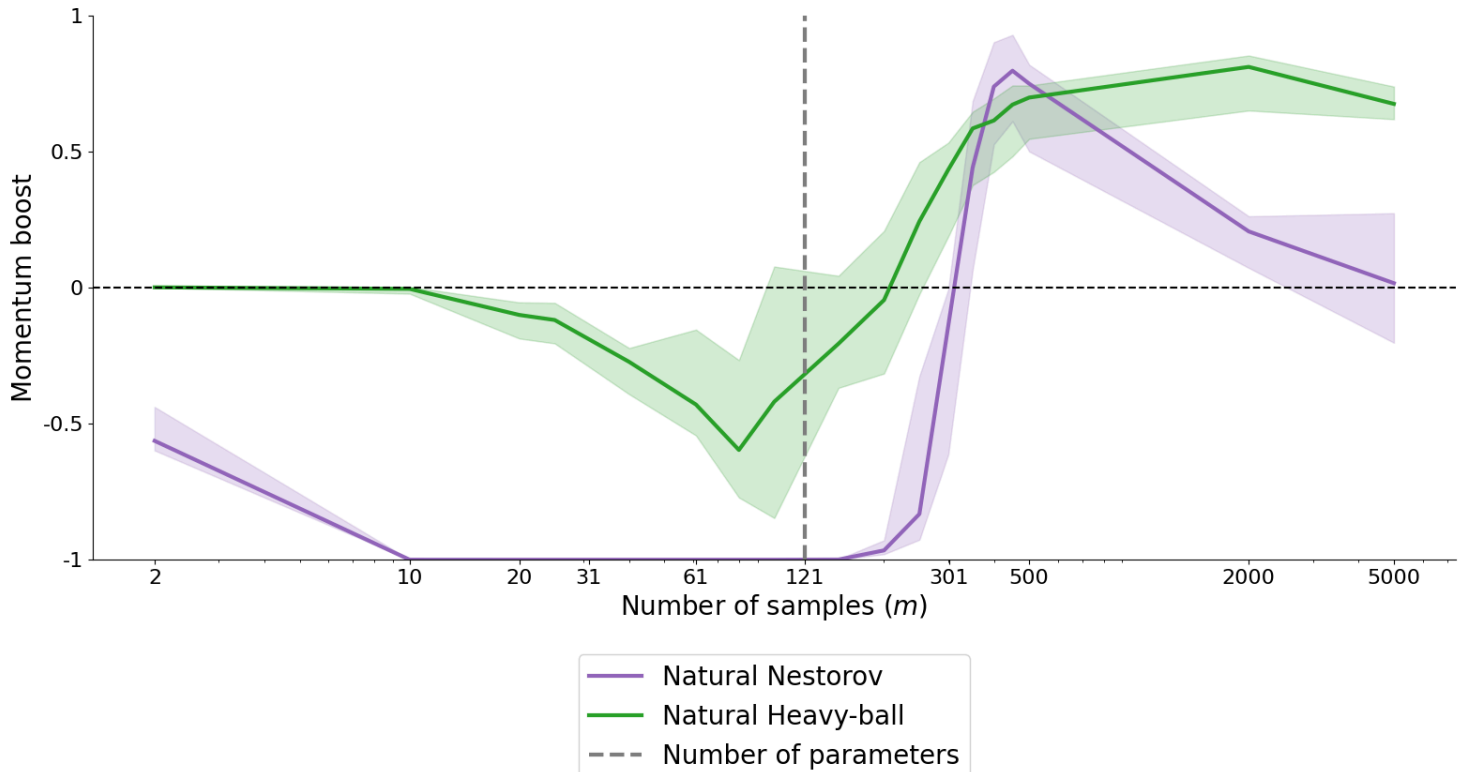
Mackey Glass: sampling

Effect of natural gradient direction estimation with limited number of samples.



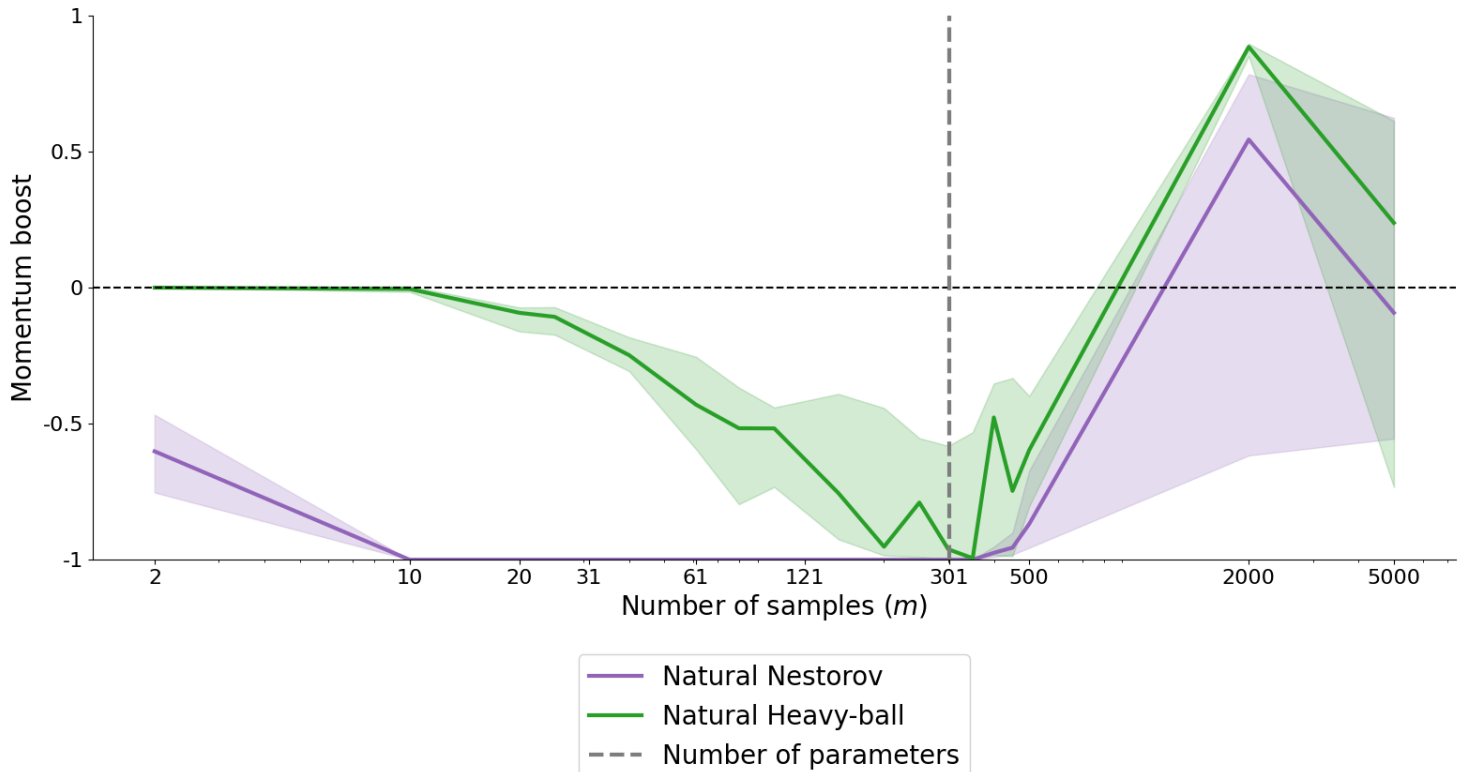
Mackey Glass: sampling

Effect of natural gradient direction estimation with limited number of samples.



Mackey Glass: sampling

Effect of natural gradient direction estimation with limited number of samples.



Classification task

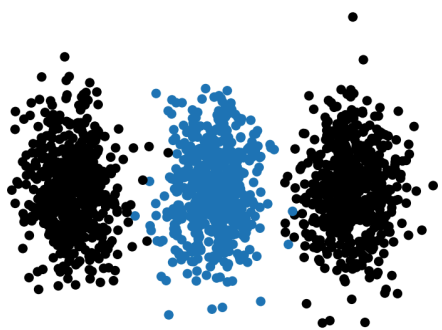
Another less toy example from [Park, H, S.-I Amari, a K Fukumizu (2000)].

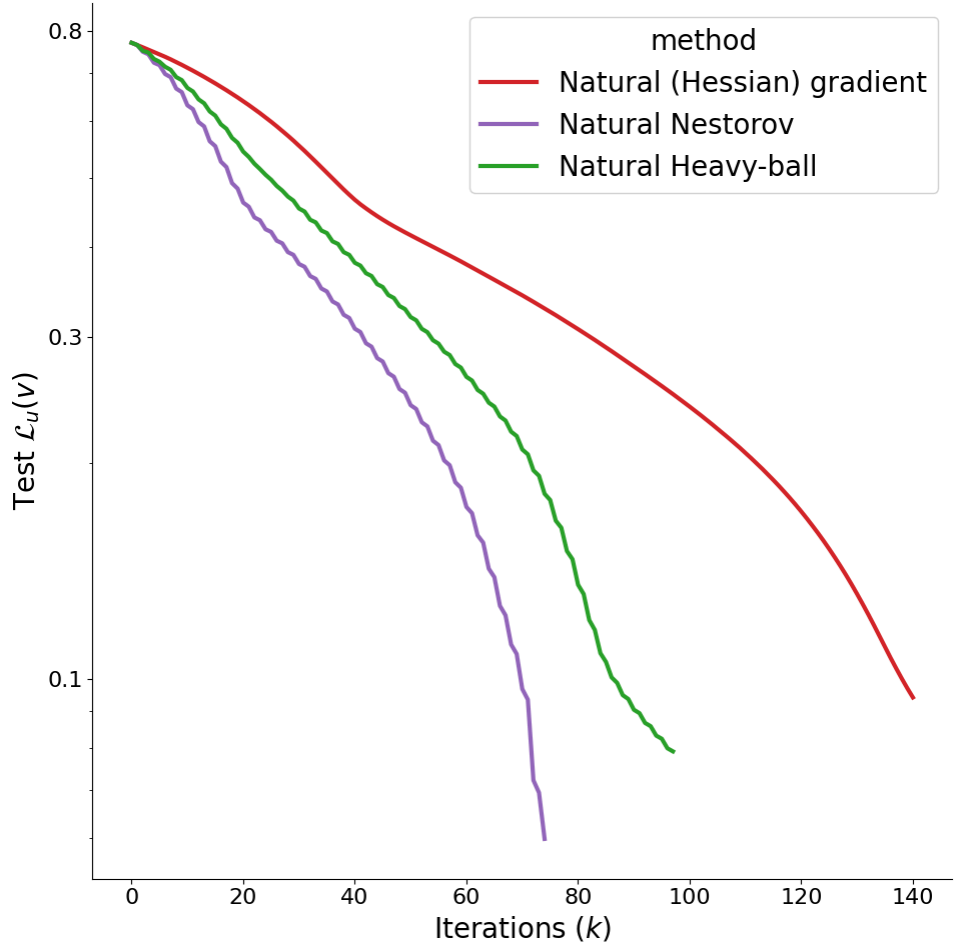
Mackey Glass caotic time series:

- Input: $z \in \mathbb{R}^2$
- Output: $c \in \{0, 1\}$

$$\mathcal{L}_u(v) = \int -u \log(v) - (1-u) \log(1-v)$$

Model (Shallow NN): $v_\theta : \mathbb{R}^2 \to [0, 1]$





Classification task

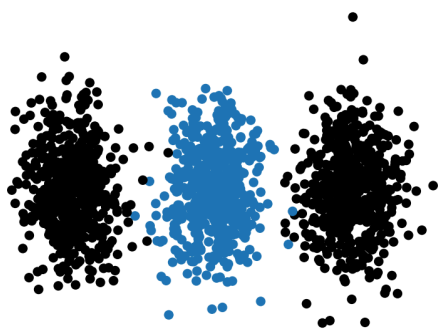
Another less toy example from [Park, H, S.-I Amari, a
K Fukumizu (2000)].

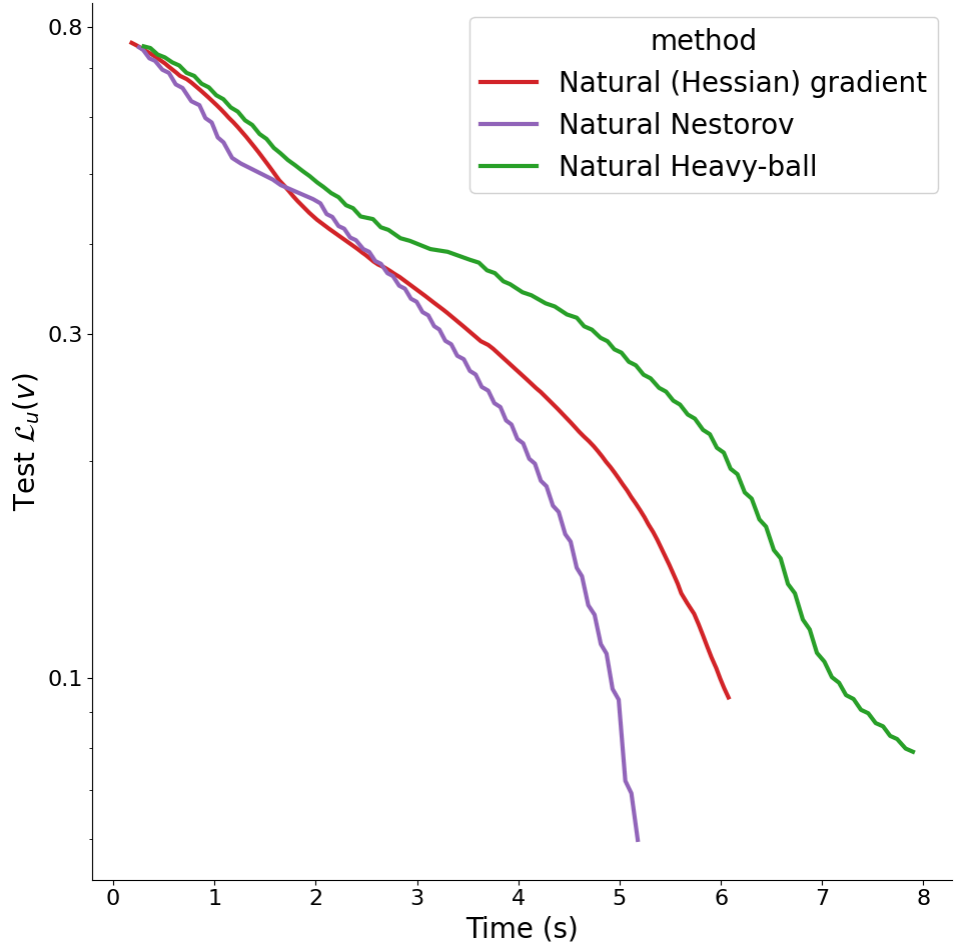
Mackey Glass caotic time series:

- Input: $z \in \mathbb{R}^2$
- Output: $c \in \{0, 1\}$

$$\mathcal{L}_u(v) = \int -u \log(v) - (1-u) \log(1-v)$$

Model (Shallow NN): $v_\theta : \mathbb{R}^2 \to [0, 1]$






CHEROKEE

Thank you!

(2024) R. Gruhlke, A. Nouy, P. Trunschke. NGD and optimal sampling.

(2020) J. Martens. NGD Review.

(2000) H. Park, S. Amari, K. Fukumizu. NGD Experiments.

(1998) S. Amari. NGD Initial paper.

(1983) Y. Nesterov. Nestorov acceleration.

(1964) B.T. Polyak. Heavy Ball acceleration.

Powered by Slidev