Natural gradient descent with momentum

SMAI 2025

Agustín Somacal

In collaboration with Anthony Nouy

École Centrale de Nantes

Laboratoire de Mathématiques Jean Leray

Outline

1. **What is natural gradient and why we may need it?**
   - 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,

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, 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

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

Parameter perspective

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

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

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

Parameter perspective

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

Gradient descent

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

Taylor expansion around current iterate θ_k.

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

Gradient descent

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

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

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

Gradient descent

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

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

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

Gradient descent

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

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

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

Gradient descent

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

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

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

Gradient descent

$$\begin{aligned}\rho(\theta, \theta_k) &= \|\theta - \theta_k\|^2_{\mathbb{R}^p}\\ \nabla_\theta \rho(\theta, \theta_k) &= 2(\theta - \theta_k)\end{aligned}$$

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

Gradient descent

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

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

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

Gradient descent

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

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

Preconditioned gradient

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|^2_M$$
$$\nabla_\theta \rho(\theta, \theta_k) = 2M(\theta - \theta_k)$$

$$\theta = \theta_k - sM^{-1}\nabla \mathcal{L}_u(\theta_k)$$

Gradient descent

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

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

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

Gradient descent

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

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

Newton's method

$$\rho(\theta, \theta_k) = \|\theta - \theta_k\|^2_H$$
$$\nabla_\theta \rho(\theta, \theta_k) = 2H(\theta - \theta_k)$$

$$\theta = \theta_k - sH^{-1}\nabla \mathcal{L}_u(\theta_k)$$

Natural gradient

From Newton's method [Amari, Shun-ichi. 1998] [Martens, James 2020].

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

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

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

Natural gradient

From Newton's method [Amari, Shun-ichi. 1998] [Martens, James 2020].

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

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

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

Natural gradient

From Newton's method [Amari, Shun-ichi. 1998] [Martens, James 2020].

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

Natural gradient

From Newton's method [Amari, Shun-ichi. 1998] [Martens, James 2020].

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

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

Natural gradient

From Newton's method [Amari, Shun-ichi. 1998] [Martens, James 2020].

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

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

In the case of $\mathcal{L}_u(v) = \|u - v\|^2_{L^2(\Omega)}$ we have that $H_V\mathcal{L} = \delta(x,y)$ thus

Natural gradient

From Newton's method [Amari, Shun-ichi. 1998] [Martens, James 2020].

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

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

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

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

In the case of $\mathcal{L}_u(v) = \|u - v\|^2_{L^2(\Omega)}$ we have that $H_V\mathcal{L} = \delta(x,y)$ thus

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

Natural gradient

From Newton's method [Amari, Shun-ichi. 1998] [Martens, James 2020].

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

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

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

Model linearization

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

In the case of $\mathcal{L}_u(v) = \|u - v\|^2_{L^2(\Omega)}$ we have that $H_V\mathcal{L} = \delta(x,y)$ thus

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

Natural gradient

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

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

Natural gradient

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

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

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = -G^{-1}\nabla_\theta\mathcal{L}$$

Natural gradient

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

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

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = -G^{-1}\nabla_\theta\mathcal{L} \qquad\qquad \frac{\mathrm{d}v}{\mathrm{d}s} = -P_{\mathcal{T}_k}\nabla\mathcal{L}$$

Natural gradient

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

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

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = -G^{-1}\nabla_\theta\mathcal{L} \qquad\qquad \frac{\mathrm{d}v}{\mathrm{d}s} = -P_{\mathcal{T}_k}\nabla\mathcal{L}$$

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

Natural gradient

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

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

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = -G^{-1}\nabla_\theta\mathcal{L} \qquad\qquad \frac{\mathrm{d}v}{\mathrm{d}s} = -P_{\mathcal{T}_k}\nabla\mathcal{L}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = -\nabla_\theta\mathcal{L} \qquad\qquad \frac{\mathrm{d}v}{\mathrm{d}s} = -GP_{\mathcal{T}_k}\nabla\mathcal{L}$$

Toy example

Gradient descent trajectory.

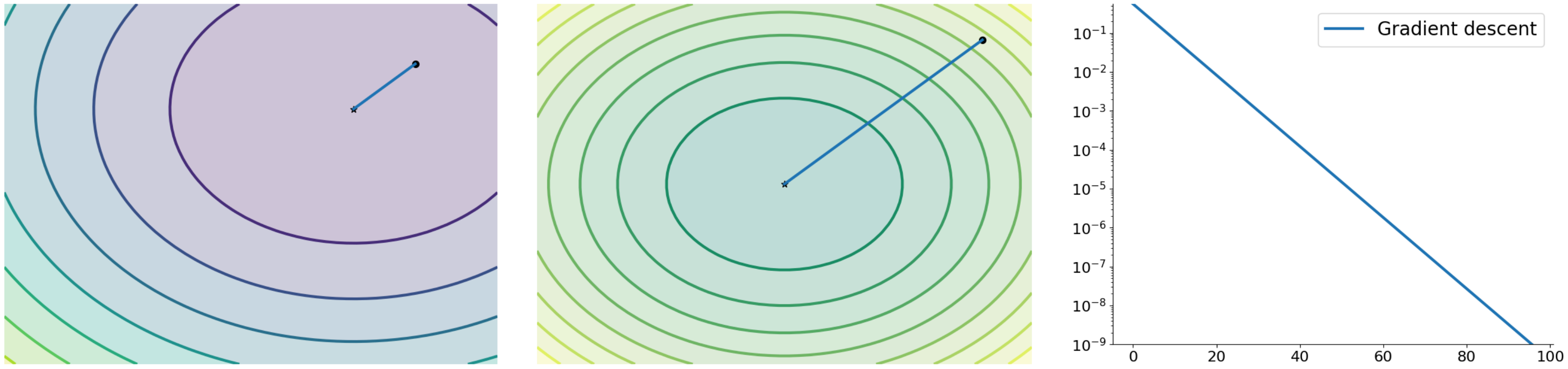
$$u \in L^2([0,1])$$

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

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

$$\begin{aligned} v_\theta(x) = A(\theta)(x) &= \theta^T \Phi(x) \\ &= \theta^T \begin{bmatrix} 1 \\ \sqrt{2}\sin(2\pi x) \end{bmatrix} \end{aligned}$$

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



Toy example

Gradient descent is biased in functional space.

$$B \in \mathbb{R}^{p\times p}$$

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

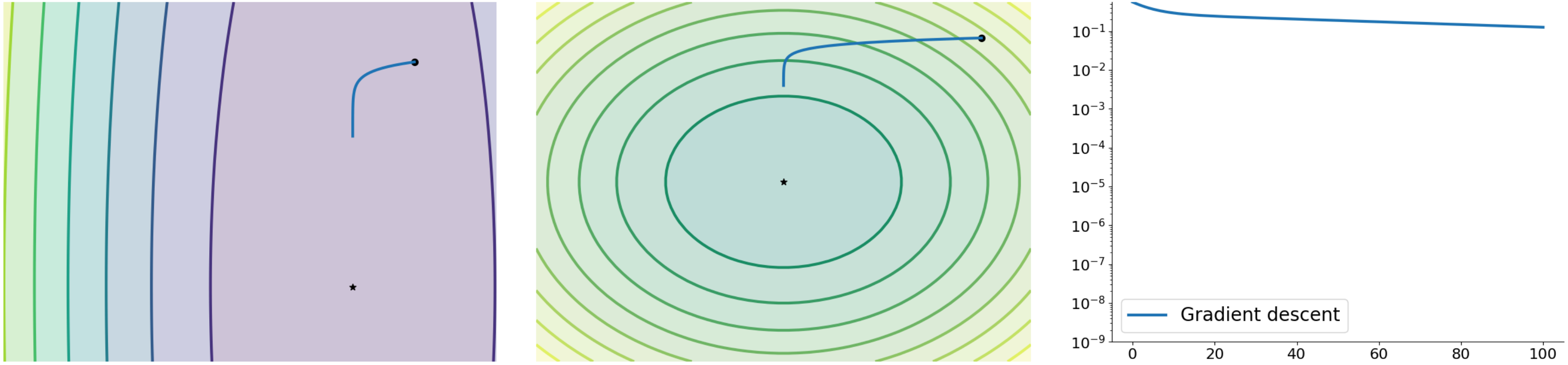
$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|$$

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

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

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

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



Toy example

Natural gradient descent.

$$B \in \mathbb{R}^{p\times p}$$

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

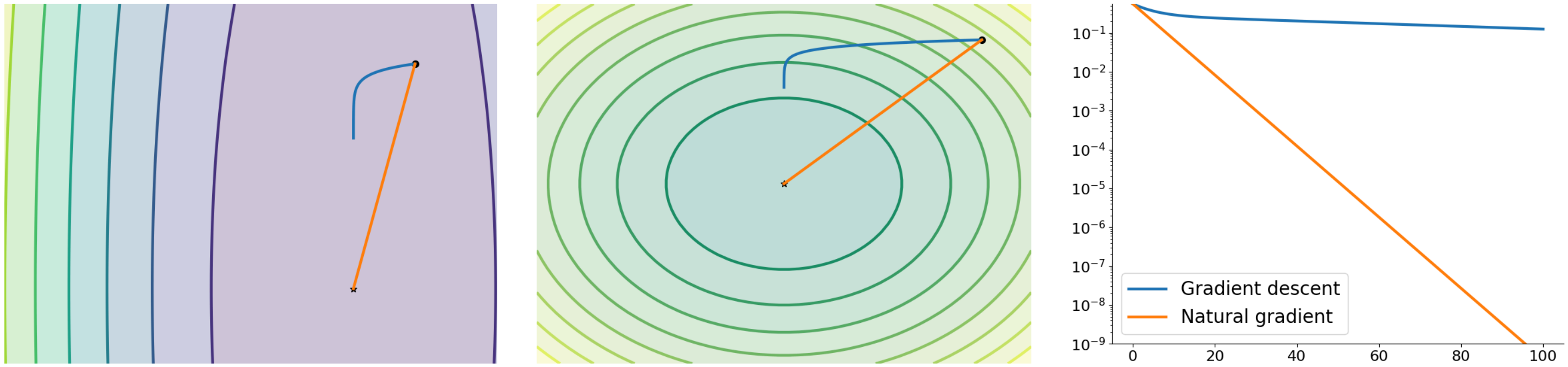
$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|$$

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

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

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

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



Toy example

Non isotropic loss.

$$B \in \mathbb{R}^{p\times p}$$

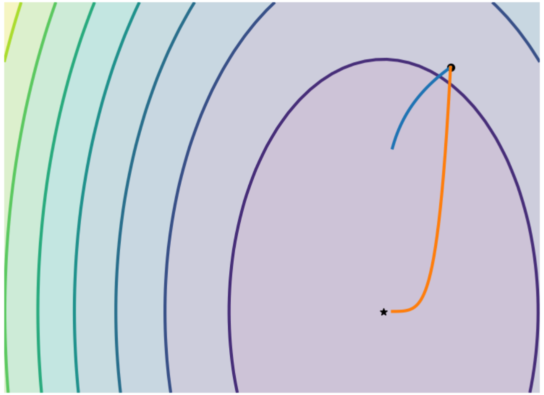
$$u \in L^2([0,1])$$

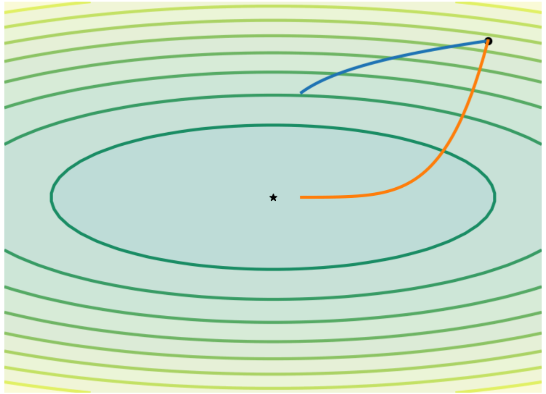
$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K$$

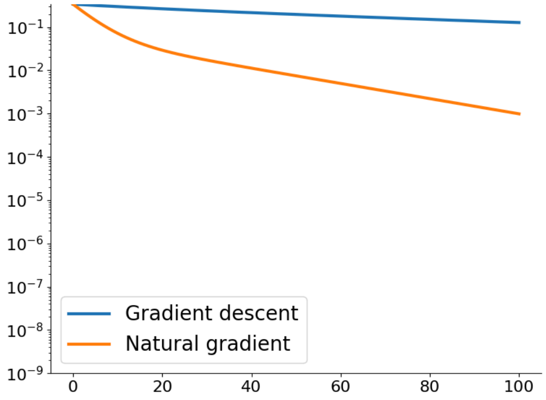
$$G_{ij} = \int \left[\frac{\partial A}{\partial \theta_i}[B^T K B]_{ij} \frac{\partial A}{\partial \theta_j}\right](x)\mathrm{d}x = [B^T K B]_{ij}$$

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

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







Toy example

Natural gradient descent with loss hessian.

$$B \in \mathbb{R}^{p\times p}$$

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

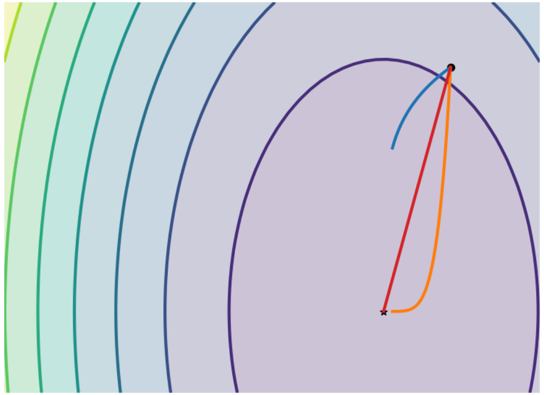
$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K$$

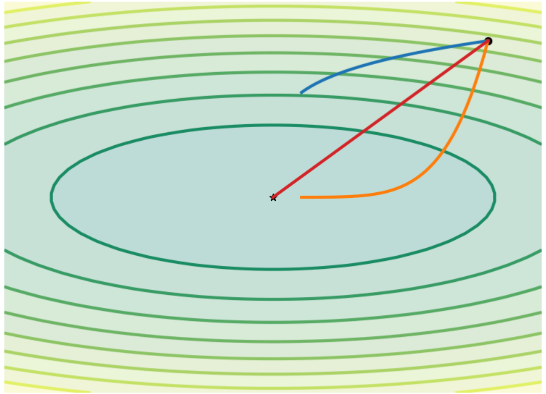
$$G_{ij} = \int \left[\frac{\partial A}{\partial \theta_i}[B^T K B]_{ij}\frac{\partial A}{\partial \theta_j}\right](x)\mathrm{d}x = [B^T K B]_{ij}$$

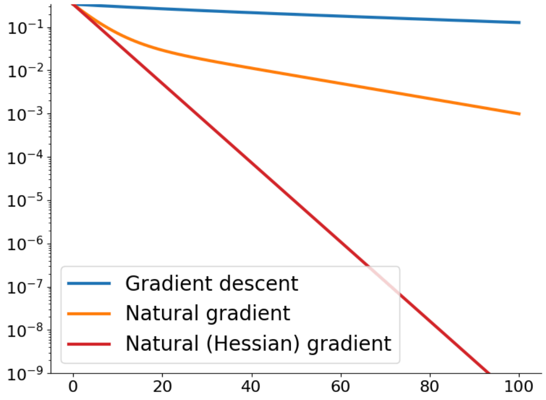
$$v_\theta(x) = A(\theta)(x) = \theta^T B\Phi(x)$$

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

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







Toy example

Natural gradient and Newton method are equivalent for linear models.

$$\color{red}{B} \in \mathbb{R}^{p\times p}$$

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

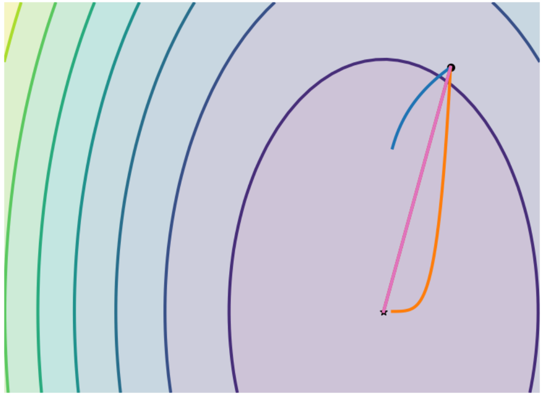
$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_{\color{blue}{K}}$$

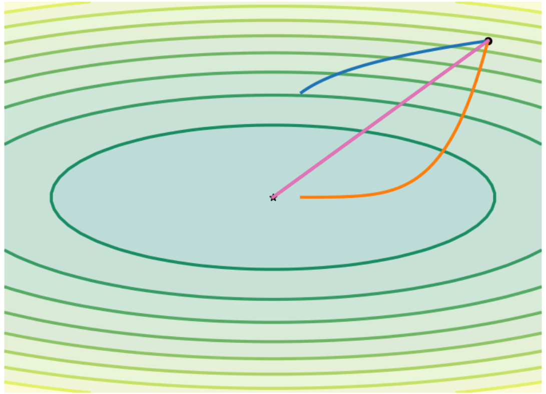
$$G_{ij} = \int \left[\color{green}{\frac{\partial A}{\partial \theta_i}}[\color{red}{B}^T \color{blue}{K} \color{red}{B}]_{ij} \color{green}{\frac{\partial A}{\partial \theta_j}}\right](x)\mathrm{d}x = [\color{red}{B}^T \color{blue}{K} \color{red}{B}]_{ij}$$

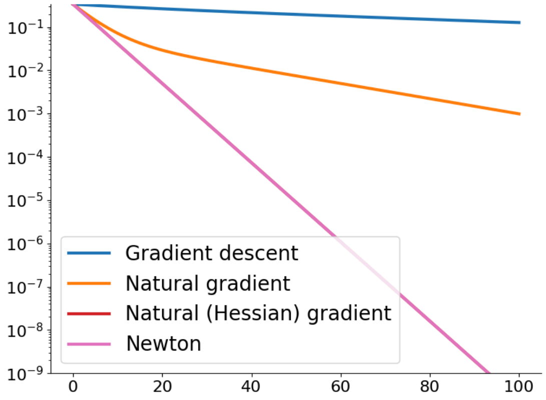
$$v_\theta(x) = A(\theta)(x) = \theta^T \color{red}{B}\Phi(x)$$

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

$$\color{green}{\frac{\partial A}{\partial \theta_i}}(\theta)(x) = \color{red}{B_i}\Phi(x)$$







Toy example

Nonlinear manifold.

$$B \in \mathbb{R}^{p\times p}, Q \in \mathbb{R}^{p\times p\times p}$$

$$u \in L^2([0,1]) \quad \mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K$$

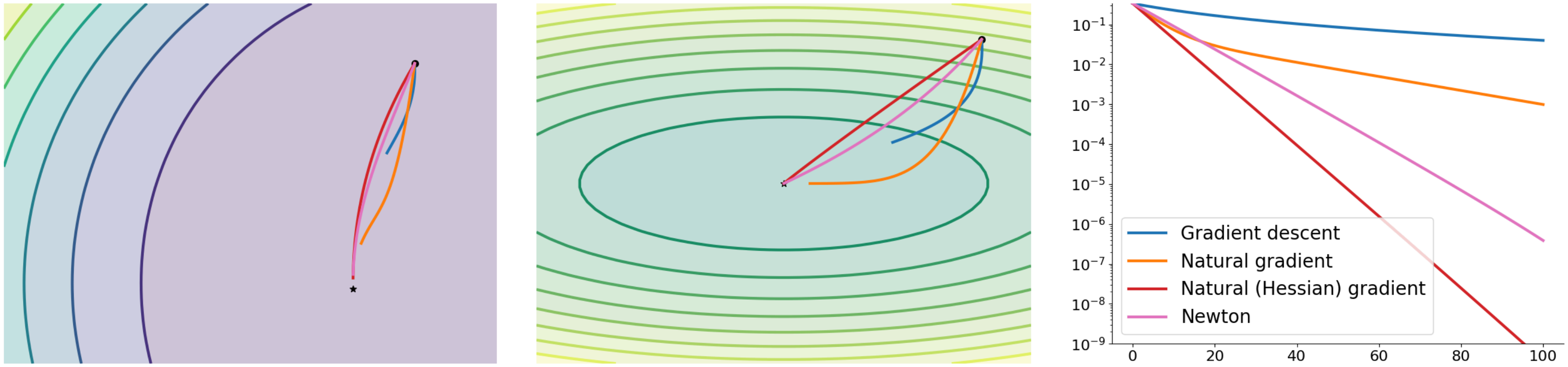
$$v_\theta(x) = A(\theta)(x) = (\theta^T B + \frac{1}{2}\theta^T Q\theta)\Phi(x)$$

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

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

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

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



Why we need momentum

Beyond L^2 loss.

Natural gradient will be biased if $\mathcal{L}_u(v) \neq \|u - v\|_K^2$

KL-divergence

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

Stochastic setting

$$\mathcal{L}_u(v_k) = \|u - v_k\|_m^2$$

$$\frac{1}{2m} \sum_{i=1}^{m} (u(x_{I_i^k}) - v(x_{I_i^k}))^2$$

PDE residual

$$\mathcal{L}(v) = \|R(v)\|^2$$

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

Escape local minima

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

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

Nestorov

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

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

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

Momentum dynamics in functional space

From momentum in parameter space to functional space.

Heavy-ball

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

$$v_{k+1} = R[v_k + p_k]$$
$$p_k = P_{\mathcal{T}_k}[\beta p_{k-1} - \alpha \nabla \mathcal{L}_u(v_k)]$$

Nestorov

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

$$w_k = R[v_k + \beta P_{\mathcal{T}_k}(v_k - v_{k-1})]$$
$$v_{k+1} = R[w_k - \alpha P_{\mathcal{T}_k} \nabla \mathcal{L}_u(w_k)]$$

$$P_{\mathcal{T}_k} \nabla \mathcal{L}_u(v_k) \qquad P_{\mathcal{T}_k} p_{k-1} \qquad P_{\mathcal{T}_k}(v_k - v_{k-1})$$

Momentum dynamics in functional space

From momentum in parameter space to functional space.

Heavy-ball

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

Nestorov

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

$$v_{k+1} = R[v_k + p_k]$$
$$p_k = P_{\mathcal{T}_k}[\beta p_{k-1} - \alpha \nabla \mathcal{L}_u(v_k)]$$

$$w_k = R[v_k + \beta P_{\mathcal{T}_k}(v_k - v_{k-1})]$$
$$v_{k+1} = R[w_k - \alpha P_{\mathcal{T}_k} \nabla \mathcal{L}_u(w_k)]$$

$$P_{\mathcal{T}_k} \nabla \mathcal{L}_u(v_k) \qquad P_{\mathcal{T}_k} p_{k-1} \qquad P_{\mathcal{T}_k}(v_k - v_{k-1})$$

$$G_k^{-1} \nabla_\theta \mathcal{L}_u(\theta_k) \qquad G_k^{-1} G_{k,k-1} p_{k-1} \qquad G_k^{-1} \int \left[\frac{\partial A}{\partial \theta}(v_k - v_{k-1})\right](x) \mathrm{d}x$$

Toy example

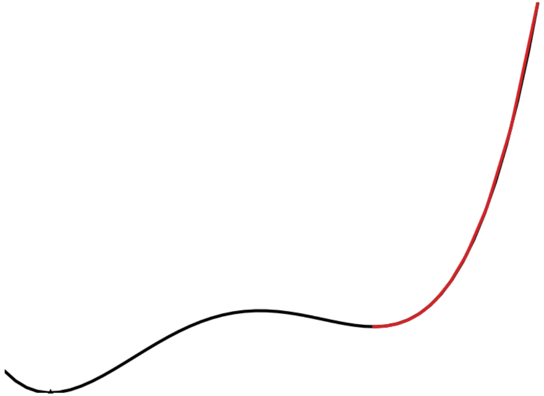
Escaping local minima.

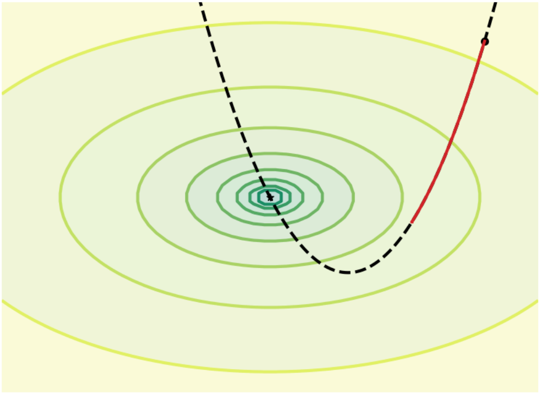
$$u \in L^2([0,1])$$

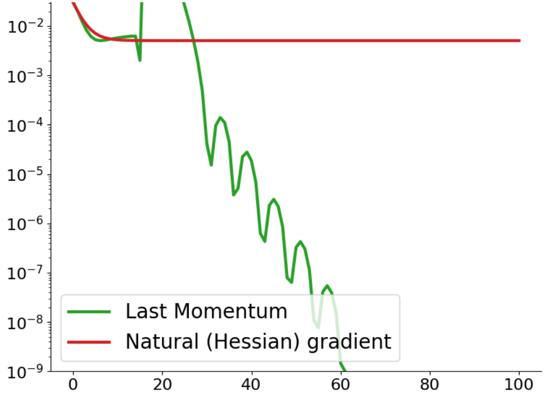
$$\mathcal{L}_u(v) = \frac{1}{2}\|u - v\|_K$$

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

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





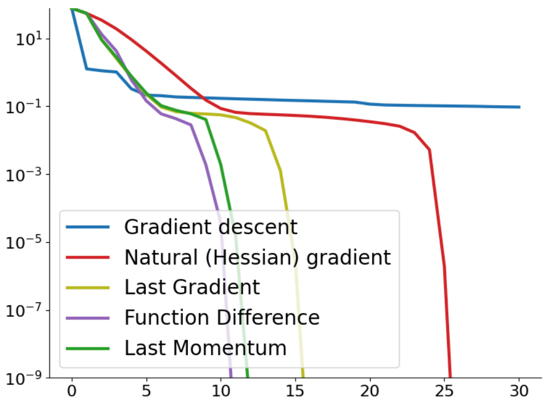


Toy example

Not L^2 loss.

$$\begin{aligned} u &\in L^2([0,1]) \\ \mathcal{L}_u(v) &= \frac{1}{2}\|f(u) - f(v)\|_K \\ f(v) &= (1 + \omega\|v - q\|^2)(v - q) + q \\ q &= R(u - v) + v \end{aligned}$$

$$\begin{aligned} \mathrm{d}v_k{}^{LM} &= P_{\mathcal{T}_k}[\beta p_{k-1} - \alpha \nabla \mathcal{L}_u(v_k)] \\ \mathrm{d}v_k{}^{FD} &= P_{\mathcal{T}_k}[\beta(v_k - v_{k-1}) - \alpha \nabla \mathcal{L}_u(v_k)] \\ \mathrm{d}v_k{}^{LG} &= P_{\mathcal{T}_k}[\beta \nabla \mathcal{L}_u(v_{k-1}) - \alpha \nabla \mathcal{L}_u(v_k)] \end{aligned}$$



Mackey Glass

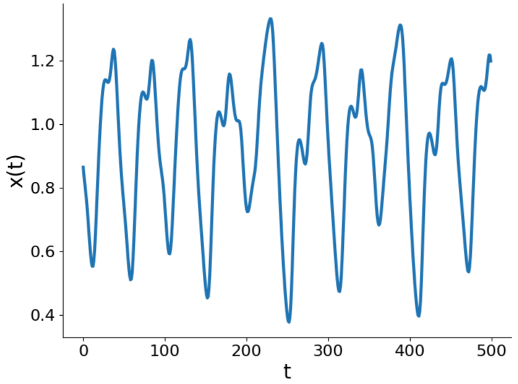
A less toy example [Park, H, S.-I Amari, and K Fukumizu (2000)].

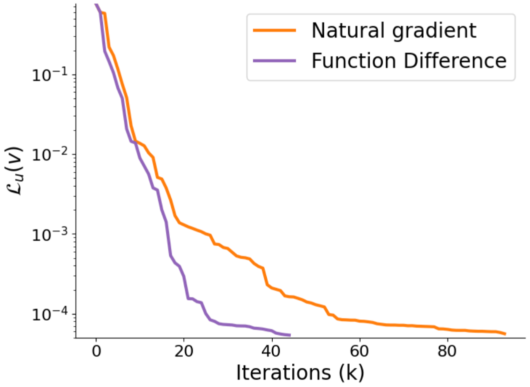
Mackey Glass caotic time series:

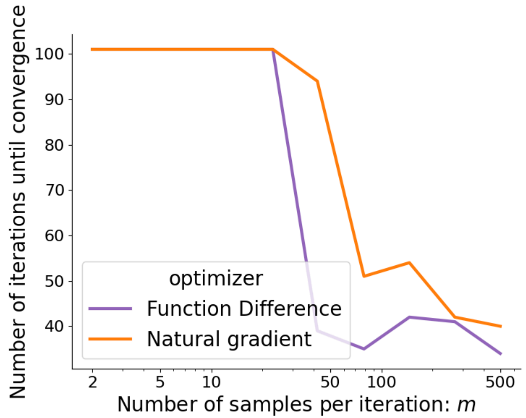
- $x(t+1) = (1-b)x(t) + a\frac{x(t-\tau)}{1+x(t-\tau)^{10}}$
- Input: $x(t), x(t-6), x(t-12), x(t-18)$
- Output: $x(t+6)$

Model: $v_\theta : \mathbb{R}^4 \to \mathbb{R}$

- Shallow neural network with 10 neurons.
- Total number of parameters: 61







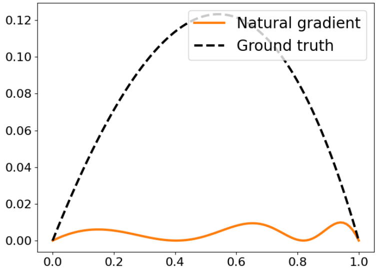
Physics informed learning.

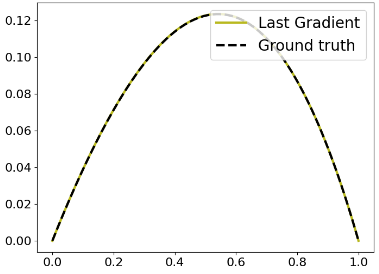
Physics informed neural networks (PINNs)
[Schwencke N., Furtlehner C. (2024)]
[Müller J., Zeinhofer M. (2024)].

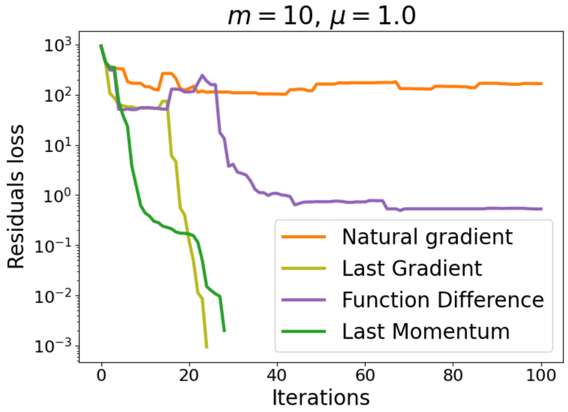
$$\mathcal{L}(v) = \|R(v)\|^2$$

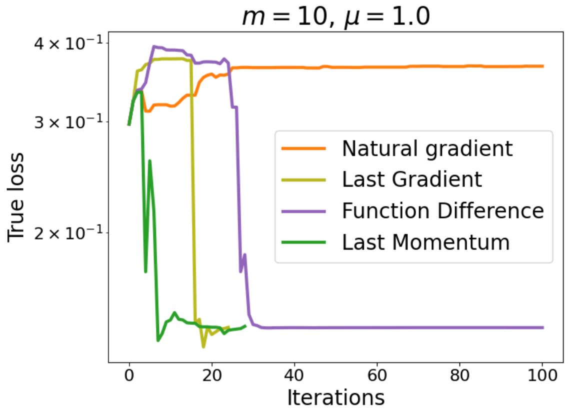
$$\mathcal{L}(v) = \| -\epsilon\partial_{xx} v + \partial_x v - 1\|^2$$

$$\mathcal{L}(v_k) = \frac{1}{2m}\sum_i^m \left(-\epsilon\partial_{xx} v(x_{I_i^k}) + \partial_x v(x_{I_i^k}) - 1\right)^2$$









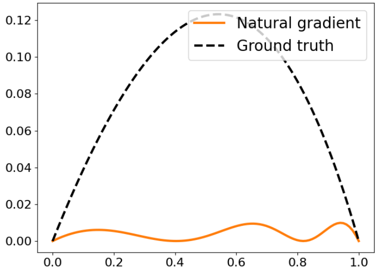
Physics informed learning.

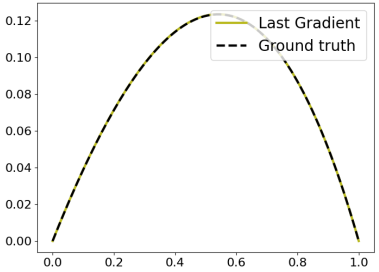
Physics informed neural networks (PINNs)
[Schwencke N., Furtlehner C. (2024)]
[Müller J., Zeinhofer M. (2024)].

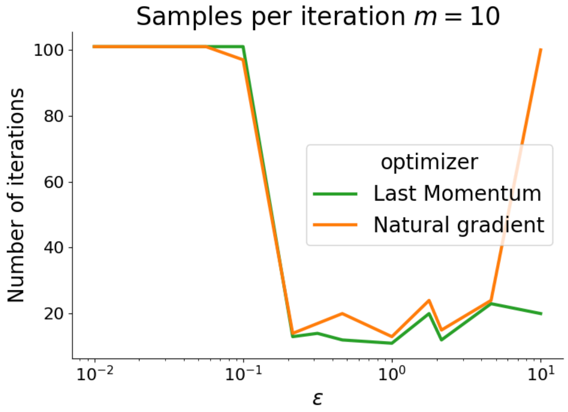
$$\mathcal{L}(v) = \|R(v)\|^2$$

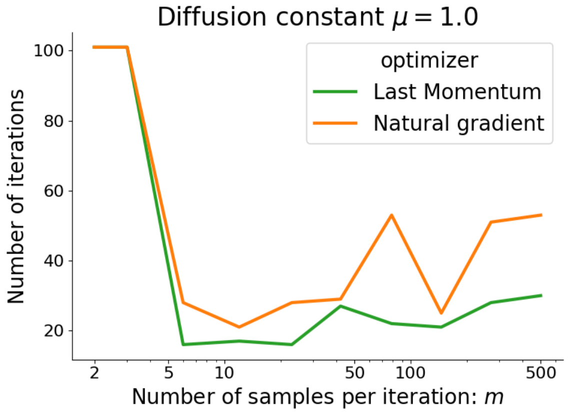
$$\mathcal{L}(v) = \| -\epsilon\partial_{xx}v + \partial_x v - 1\|^2$$

$$\mathcal{L}(v_k) = \frac{1}{2m}\sum_i^m (-\epsilon\partial_{xx}v(x_{I_i^k}) + \partial_x v(x_{I_i^k}) - 1)^2$$









Thanks!

Powered by Slidev