Derivative Free Optimization class
AMS & Optimization Masters
On the connection between affine-invariance, convergence and learning second order information

Anne Auger
Inria and CMAP, Ecole Polytechnique
France

Motivation

CMA-ES: a widely used randomized DFO algorithm [Hansen et al. 2001-2006]

for non-convex, non-smooth, difficult black-box problems

parameter-free

6.5 + 54 millions downloads for two main Python codes

does not even use function value

yet observed to learn "second-order" information in particular on $f(x) = g((x - x^\star)^\top H(x - x^\star))$, $H \succ 0$, $g : \mathbb{R} \to \mathbb{R}$ strict. increasing

sometimes presented as (randomized) quasi-Newton

How is that even possible??

Objectives

uncover the simple and nice mathematical arguments behind this learning

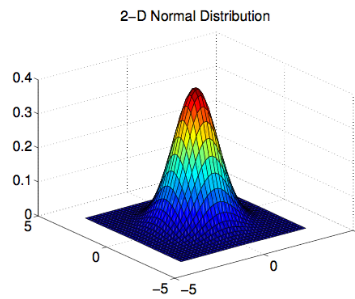
→ illustrate proofs on quasi-Newton algorithms (BFGS)

simpler and illustrates the generality of the ideas

Disclaimer: results on quasi-Newton presented are not new nor impressive *stronger results exists*

yet we show that they stem from simple fundamental properties

Adaptive Stochastic Optimization Algorithm

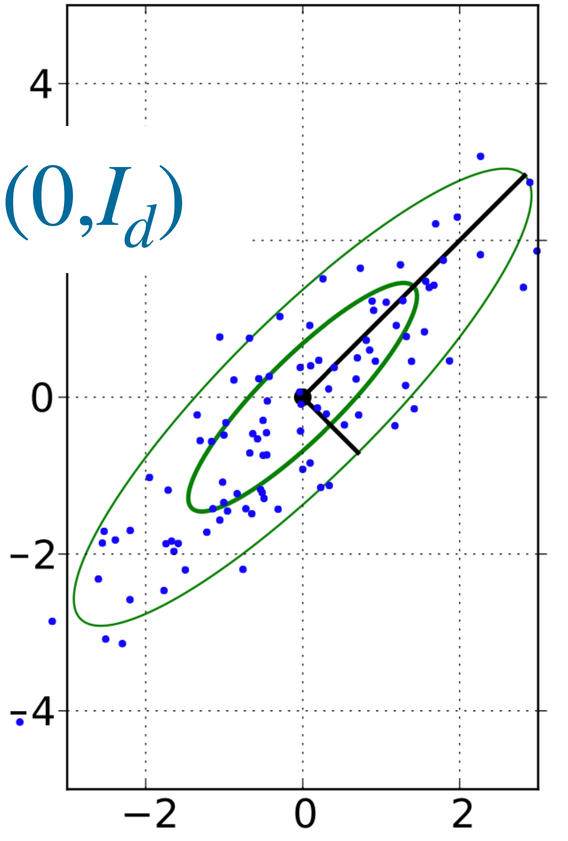


Given e.g. $\theta_t = (m_t, \sigma_t, C_t) \in \mathbb{R}^n \times \mathbb{R}_> \times \mathcal{S}^n_{++}$

❶ Sample candidate solutions $X^i_{t+1} \sim \mathcal{N}(m_t, \sigma_t^2 C_t)$, i.e.

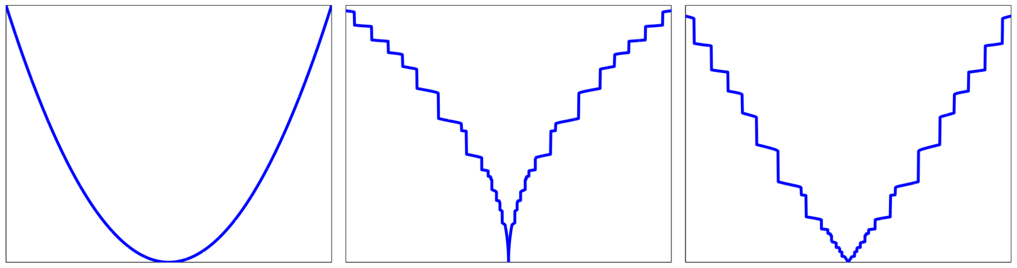
$$X^i_{t+1} = m_t + \sigma_t \sqrt{C_t} U^i_{t+1},\ i = 1, \ldots, \lambda$$

$\{U_t, t \geq 1\}$ i.i.d.,$U^i_{t+1} \sim \mathcal{N}(0, I_d)$



❷ Evaluate and rank candidate solutions

$$f\left(X^{s_{t+1}(1)}_{t+1}\right) \leq \ldots \leq f\left(X^{s_{t+1}(\lambda)}_{t+1}\right)$$



❸ Update θ_t:

$$\theta_{t+1} = F\left(\theta_t, [U^{s_{t+1}(1)}_{t+1}, \ldots, U^{s_{t+1}(\lambda)}_{t+1}]\right)$$

should drive m_t towards the optimum

CMA-ES - simplified setting $\theta_t = (m_t, \sigma_t, C_t) \in \mathbb{R}^n \times \mathbb{R}_> \times \mathcal{S}^n_{++}$

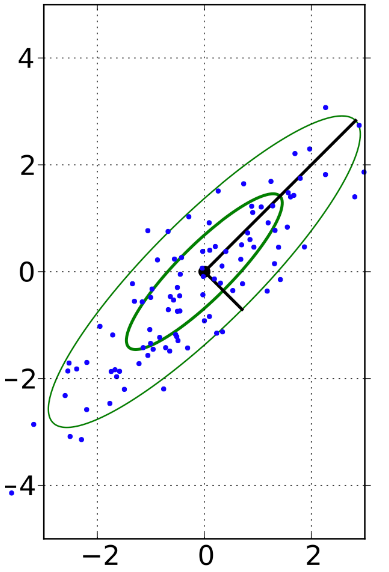
Sampling + ranking:

$y_i \sim \mathcal{N}(0, C_t)$

$$X^i_{t+1} = m_t + \sigma_t \sqrt{C_t} U^i_{t+1} \quad i = 1,\ldots,\lambda$$

$\{U_t, t \geq 1\}$ i.i.d $U^i_{t+1} \sim \mathcal{N}(0, I_d)$

$$f\left(X^{s_{t+1}(1)}_{t+1}\right) \leq \ldots \leq f\left(X^{s_{t+1}(\lambda)}_{t+1}\right)$$



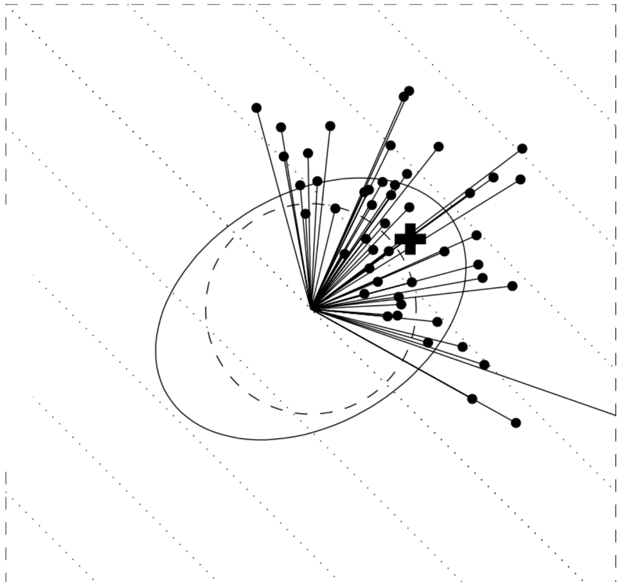
Update of θ_t:

$$m_{t+1} = \sum_{i=1}^{\mu} w_i X^{s_{t+1}(i)}_{t+1} = m_t + \sigma_t \sqrt{C_t} \sum_{i=1}^{\mu} w_i U^{s_{t+1}(i)}_{t+1}$$

$\sum_{i=1}^{\mu} w_i = 1, \mu_{\text{eff}} = 1 / \sum w_i^2$

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c_\sigma}{d_\sigma}\left[\frac{\sqrt{\mu_{\text{eff}}} \| \sum_{i=1}^{\mu} w_i U^{s_{t+1}(i)}_{t+1} \|}{E[\|\mathcal{N}(0, I_d)\|]} - 1\right]\right)$$

$$C_{t+1} = (1 - c_\mu \qquad) C_t + \underbrace{c_\mu \sqrt{C_t} \left(\sum_{i=1}^{\mu} w_i U^{s_{t+1}(i)}_{t+1} [U^{s_{t+1}(i)}_{t+1}]^\top\right) \sqrt{C_t}}_{\textit{rank } \mu \textit{ update}}$$

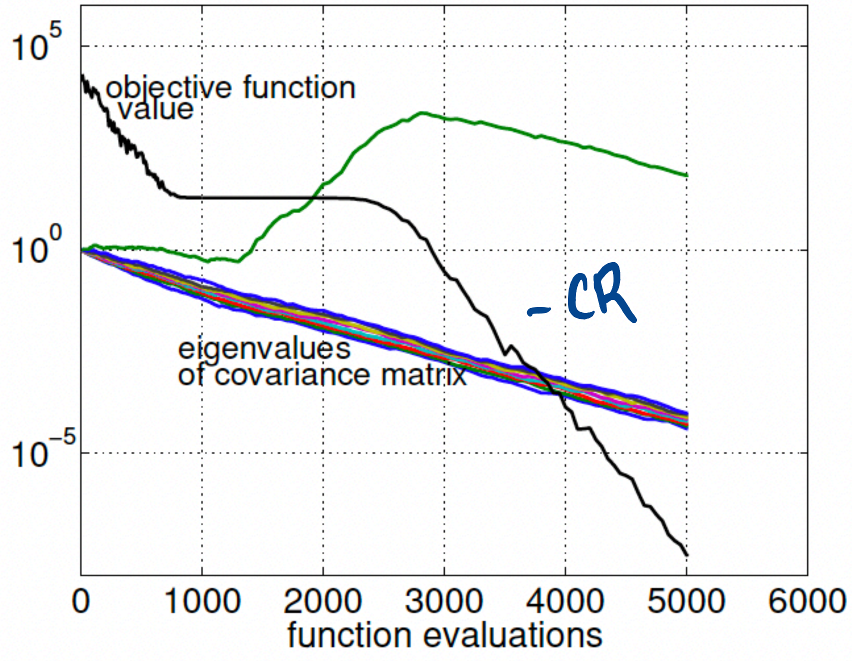


CMA-ES: Linear Convergence and Learning Inverse Hessian

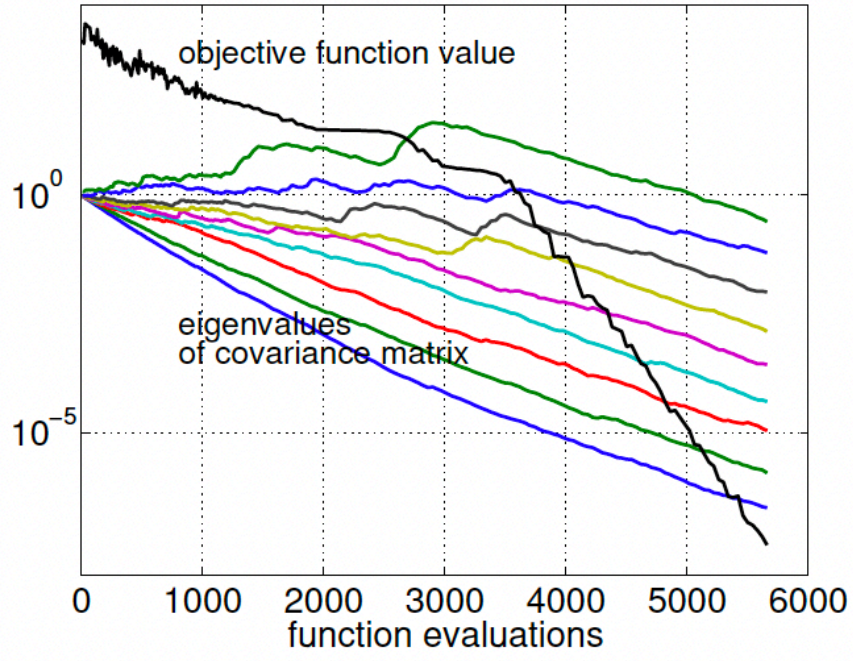
For all $f(x) = g\left(\frac{1}{2}(x - x^\star)^\top H (x - x^\star)\right)$, with $g : \mathrm{Im}(f) \to \mathbb{R}$ strict increasing, $H \succ 0$ (SDP)

$$\frac{1}{t} \ln \frac{\|m_t - x^\star\|}{\|m_0 - x^\star\|} \xrightarrow[t\to\infty]{} - \mathrm{CR} \qquad\qquad C_t \propto \alpha_t H^{-1} \;\text{ with } \alpha_t \to 0$$

Empirical observations:



$\mathrm{Eig}(H) = \left(10^{-6}, 1, \ldots, 1\right)$



$\mathrm{Eig}(H) = \left(1, \ldots, 10^{6\frac{i-1}{n-1}}, \ldots, 10^6\right)$

BFGS algorithm

$$\theta_t = (x_t, B_t)$$

incumbent

estimate of Hessian

```
initialize state $\theta_0 = (x_0, B_0) \in \mathbb{R}^n \times \mathscr{S}(n, \mathbb{R})$, $t = 0$
while stopping criterion not met do
    compute $p_t = -B_t^{-1} \nabla f(x_t)$
    compute step-size: $\alpha_t = \text{LineSearch}(x_t, p_t, f)$
    move in the direction of $p_t$: $x_{t+1} = x_t + \alpha_t p_t = x_t - \alpha_t B_t^{-1} \nabla f(x_t)$
    compute $s_t = \alpha_t p_t$
    compute $y_t = \nabla f(x_{t+1}) - \nabla f(x_t)$
    update estimate of Hessian: $B_{t+1} = B_t + \frac{y_t y_t^{\mathrm{T}}}{y_t^{\mathrm{T}} s_t} - \frac{B_t s_t s_t^{\mathrm{T}} B_t}{s_t^{\mathrm{T}} B_t s_t}$
    $t = t + 1$
end while
```

On affine-invariance

Affine-invariance

on $x \mapsto f(x)$ $\quad x_0 \quad x_1 \quad x_2 \quad \dots \quad x_t \quad \dots$

Affine-invariance

$A \in \mathrm{GL}n(\mathbb{R})$

on $x \mapsto f(x)$ $\quad x_0 \quad x_1 \quad x_2 \quad \ldots \quad x_t \quad \ldots$

on $x' \mapsto f(Ax')$ $\quad x'_0 \quad x'_1 \quad x'_2 \quad \ldots \quad x'_t \quad \ldots$

$f(Ax' \cancel{+b})$

↑ because of translation invariance

Affine-invariance

$A \in \mathrm{GL}n(\mathbb{R})$

$$\begin{tikzcd}
\text{on } x \mapsto f(x) & x_0 \arrow[d, leftrightarrow] & x_1 \arrow[d, leftrightarrow] & x_2 \arrow[d, leftrightarrow] & \dots & x_t \arrow[d, leftrightarrow] & \dots \\
\text{on } x' \mapsto f(Ax') & x'_0 \arrow[d, equal] & x'_1 \arrow[d, equal] & x'_2 \arrow[d, equal] & \dots & x'_t \arrow[d, equal] & \dots \\
 & A^{-1}x_0 & A^{-1}x_1 & A^{-1}x_2 & \dots & A^{-1}x_t & \dots
\end{tikzcd}$$

Affine-invariance

An algorithm is affine invariant if it produces the same trajectory when optimizing $f(x)$ or any $f(Ax')$ with $A \in \mathrm{GL}n(\mathbb{R})$

up to the change of variable: $x' = A^{-1}x$

$$\begin{tikzcd}
\text{on } x \mapsto f(x) & x_0 \arrow[d, leftrightarrow] & x_1 \arrow[d, leftrightarrow] & x_2 \arrow[d, leftrightarrow] & \dots & x_t \arrow[d, leftrightarrow] & \dots \\
\text{on } x' \mapsto f(Ax') & x'_0 \arrow[d, equal] & x'_1 \arrow[d, equal] & x'_2 \arrow[d, equal] & \dots & x'_t \arrow[d, equal] & \dots \\
& A^{-1}x_0 \arrow[d, equal] & A^{-1}x_1 & A^{-1}x_2 & \dots & A^{-1}x_t & \dots \\
& \Phi_A(x_0) & & & & &
\end{tikzcd}$$

Affine-invariance: commutative diagram

An algorithm is affine-invariant if for all $A \in \mathrm{GL}n(\mathbb{R})$, there exists Φ_A (a bijective change of state variables) s.t. the diagram commutes:

change of state variables

$$\begin{tikzcd}
x_t \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_A"'] & x_{t+1} \arrow[d, shift right, "\Phi_A"'] \\
x'_t \arrow[r, "x' \mapsto f(Ax')"] \arrow[u, shift right, "\Phi_{A^{-1}}"'] & x'_{t+1} \arrow[u, shift right, "\Phi_{A^{-1}}"']
\end{tikzcd}$$

$$x'_t = \Phi_A(x_t) = A^{-1}x_t$$

Affine-invariance: commutative diagram

often state not reduced to incumbent solutions

example: BFGS where $\theta_t = (x_t, B_t)$

State for CMA-ES $:= \theta_t \left(m_t, \sigma_t, C_t, p_t^\sigma, p_t^c\right)$

Affine-invariance: commutative diagram

often state not reduced to incumbent solutions

example: BFGS where $\theta_t = (x_t, B_t)$

An algorithm is affine-invariant if for all $A \in \mathrm{GL}n(\mathbb{R})$, there exists a bijective change of state variables Φ_A s.t. the diagram commutes:

change of state variables

$$\begin{tikzcd}
\theta_t \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_A"'] & \theta_{t+1} \arrow[d, shift right, "\Phi_A"'] \\
\theta'_t \arrow[r, "x' \mapsto f(Ax')"] \arrow[u, shift right, "\Phi_{A^{-1}}"'] & \theta'_{t+1} \arrow[u, shift right, "\Phi_{A^{-1}}"']
\end{tikzcd}$$

$$\theta'_t = \Phi_A(\theta_t)$$

$$\theta_t = \Phi_{A^{-1}}(\theta'_t)$$

Affine-invariance: commutative diagram

Affine-invariance ⇒ rotational invariance

An algorithm is affine-invariant if for all $A \in \mathrm{GL}n(\mathbb{R})$, there exists a bijective change of state variables Φ_A s.t. the diagram commutes:

$$\begin{tikzcd}
\theta_t \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right=1, "\Phi_A"'] & \theta_{t+1} \arrow[d, shift right=1, "\Phi_A"'] \\
\theta'_t \arrow[r, "x' \mapsto f(Ax')"] \arrow[u, shift right=1, "\Phi_{A^{-1}}"'] & \theta'_{t+1} \arrow[u, shift right=1, "\Phi_{A^{-1}}"']
\end{tikzcd}$$

change of state variables

$$\theta'_t = \Phi_A(\theta_t)$$

$$\theta_t = \Phi_{A^{-1}}(\theta'_t)$$

Affine-invariance of BFGS

The BFGS algorithm (with affine-invariant step-size) satisfies for all $A \in GL(\mathbb{R}^n)$ the commutative diagram:

affine-invariant step-size: constant, exact line-search, …

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_A"'] & (x_{t+1}, B_{t+1}) \arrow[d, shift right, "\Phi_A"'] \\
(x'_t, B'_t) \arrow[r, "x' \mapsto f(Ax')"] \arrow[u, shift right, "\Phi_{A^{-1}}"'] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right, "\Phi_{A^{-1}}"']
\end{tikzcd}$$

with $(x'_t, B'_t) = \Phi_A(x_t, B_t) := (A^{-1}x_t, A^\top B_t A)$

Exercice: prove that the BFGS algorithm is affine-invariant

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Frechet differentiable objective function. Consider the BFGS algorithm defined as

```
initialize state $\theta_0 = (x_0, B_0) \in \mathbb{R}^n \times \mathcal{S}_{n,>}(\mathbb{R})$, $k = 0$
while stopping criterion not met do
    compute $d_k = -B_k^{-1} \nabla f(x_k)$
    compute step-size: $\alpha_k = \text{LineSearch}(x_k, d_k, f)$
    move in the direction of $d_k$: $x_{k+1} = x_k + \alpha_k d_k$
    compute $s_k = \alpha_k d_k$
    compute $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$
    update estimate of Hessian: $B_{k+1} = B_k + \frac{y_k y_k^{\mathrm{T}}}{y_k^{\mathrm{T}} s_k} - \frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k}$
    $k = k + 1$
end while
```

We will assume for the sake of simplicity that the step-size $\alpha_k = \alpha$ is constant.

Let $A \in \mathbb{R}^{n \times n}$ be an invertible matrix and let $x_0 \in \mathbb{R}^n$ and $B_0 \in \mathbb{R}^{n \times n}$ with $B_0 \succ 0$. Consider the sequence $(x_k, B_k)_{k \geq 1}$ generated by the BFGS algorithm optimizing $x \mapsto f(x)$. Let $(x'_0, B'_0) = (A^{-1} x_0, A^T B_0 A)$ and consider $(x'_k, B'_k)_{k \geq 1}$ the sequence of states of the BFGS algorithm optimizing $g(x') = f(Ax')$ and initialized in (x'_0, B'_0).

Prove that for all $k \geq 1$, $(x'_k, B'_k) = (A^{-1} x_k, A^T B_k A)$, i.e. that the BFGS algorithm is affine-invariant.

(See separate correction) Assume optimal step-size

$\alpha_k = \underset{\alpha}{\text{argmin}}\, f(x_k + \alpha d_k)$

Affine-invariance of CMA-ES

The CMA-ES algorithm is affine-invariant: for all $A \in GL(\mathbb{R}^n)$ the following commutative diagram holds:

$$\begin{tikzcd}
(m_t, C_t, \sigma_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_A"'] & (m_{t+1}, C_{t+1}, \sigma_{t+1}) \arrow[d, shift right, "\Phi_A"'] \\
(m'_t, C'_t, \sigma'_t) \arrow[r, "x' \mapsto f(Ax')"] \arrow[u, shift right, "\Phi_{A^{-1}}"'] & (m'_{t+1}, C'_{t+1}, \sigma'_{t+1}) \arrow[u, shift right, "\Phi_{A^{-1}}"']
\end{tikzcd}$$

$$\Phi_A(m_t, C_t, \sigma_t) = (A^{-1}m_t, A^{-1}C_t(A^{-1})^\top, \sigma_t)$$

How affine-invariance and stability imply learning a matrix proportional to Hessian on convex-quadratic functions

Consequence of rotational invariance + stability

Lemma: Consider a rotational invariant function $f: \mathbb{R}^n \to \mathbb{R}$ [such that $f(Rx) = f(x)$ for all $R \in O_n(\mathbb{R})$]

f is a radial function: $f(x) = g(\|x\|)$

example: $f(x) = \gamma x^\top x = \gamma \|x\|^2$

Consequence of rotational invariance + stability

Lemma: Consider a rotational invariant function $f : \mathbb{R}^n \to \mathbb{R}$ [such that $f(Rx) = f(x)$ for all $R \in O_n(\mathbb{R})$]

f is a radial function: $f(x) = g(\|x\|)$

example: $f(x) = \gamma x^\top x = \gamma \|x\|^2$

[stability] Suppose that $\exists !\ (x^*, B^*) \in \mathbb{R}^n \times \mathcal{S}_>$ such that BFGS converges globally to (x^*, B^*).

for all (x_0, B_0), $\lim_{t \to \infty} (x_t, B_t) = (x^*, B^*)$

Consequence of rotational invariance + stability

Lemma: Consider a rotational invariant function $f : \mathbb{R}^n \to \mathbb{R}$ [such that $f(Rx) = f(x)$ for all $R \in O_n(\mathbb{R})$]

f is a radial function: $f(x) = g(\|x\|)$
example: $f(x) = \gamma x^\top x = \gamma \|x\|^2$

[stability] Suppose that $\exists!\ (x^*, B^*) \in \mathbb{R}^n \times \mathcal{S}_>$ such that BFGS converges globally to (x^*, B^*).

for all (x_0, B_0), $\lim_{t\to\infty} (x_t, B_t) = (x^*, B^*)$

Then for all R

$$Rx^* = x^* \qquad \Rightarrow x^* = 0$$

$$R^\top B^* R = B^* \qquad \Rightarrow B^* = \alpha I_d$$

Corollary: If $f(x) = \dfrac{1}{2}\|x\|^2$, $\lim_{t\to\infty} B_t = \alpha \nabla^2 f$.

Proof: uses only rotational invariance and stability

Proof: uses only rotational invariance and stability

Let $(x_t, B_t) \to (x^*, B^*)$. Let $R \in O_n(\mathbb{R})$

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] & (x_{t+1}, B_{t+1}) \arrow[r, "t \to \infty"'] & (x^*, B^*)
\end{tikzcd}$$

Proof: uses only rotational invariance and stability

Let $(x_t, B_t) \to (x^*, B^*)$. Let $R \in O_n(\mathbb{R})$

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_R"'] & (x_{t+1}, B_{t+1}) \arrow[r, "t \to \infty"'] \arrow[d, shift right] & (x^*, B^*) \\
(x'_t, B'_t) \arrow[r, "x' \mapsto f(Rx')"] \arrow[u, shift right] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right] &
\end{tikzcd}$$

The diagram commutes with: $(x'_t, B'_t) = \Phi_R(x_t, B_t) = (R^\top x_t, R^\top B_t R)$

Proof: uses only rotational invariance and stability

Let $(x_t, B_t) \to (x^*, B^*)$. Let $R \in O_n(\mathbb{R})$

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_R"'] & (x_{t+1}, B_{t+1}) \arrow[r, "t \to \infty"'] \arrow[d, shift right] & (x^*, B^*) \\
(x'_t, B'_t) \arrow[r, "x' \mapsto f(Rx')"] \arrow[u, shift right] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right]
\end{tikzcd}$$

The diagram commutes with: $(x'_t, B'_t) = \Phi_R(x_t, B_t) = (R^\top x_t, R^\top B_t R)$

Since f is rotational invariant $f(Rx') = f(x')$ so (x'_t, B'_t) optimizes f also.

Proof: uses only rotational invariance and stability

Let $(x_t, B_t) \to (x^*, B^*)$. Let $R \in O_n(\mathbb{R})$

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_R"'] & (x_{t+1}, B_{t+1}) \arrow[r, "t \to \infty"'] \arrow[d, shift right] & (x^*, B^*) \\
(x'_t, B'_t) \arrow[r, "x' \mapsto f(Rx')"] \arrow[u, shift right] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right] &
\end{tikzcd}$$

The diagram commutes with: $(x'_t, B'_t) = \Phi_R(x_t, B_t) = (R^\top x_t, R^\top B_t R)$

Since f is rotational invariant $f(Rx') = f(x')$ so (x'_t, B'_t) optimizes f also.

Hence by stability $(x'_t, B'_t) = (R^\top x_t, R^\top B_t R) \to (x^*, B^*)$

Proof: uses only rotational invariance and stability

Let $(x_t, B_t) \to (x^*, B^*)$. Let $R \in O_n(\mathbb{R})$

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_R"'] & (x_{t+1}, B_{t+1}) \arrow[r, "t \to \infty"'] \arrow[d, shift right] & (x^*, B^*) \\
(x'_t, B'_t) \arrow[r, "x' \mapsto f(Rx')"] \arrow[u, shift right] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right] &
\end{tikzcd}$$

The diagram commutes with: $(x'_t, B'_t) = \Phi_R(x_t, B_t) = (R^\top x_t, R^\top B_t R)$

Since f is rotational invariant $f(Rx') = f(x')$ so (x'_t, B'_t) optimizes f also.

Hence by stability

$$\begin{tikzcd}
(x'_t, B'_t) = (R^\top x_t, R^\top B_t R) \arrow[r] \arrow[d] & (x^*, B^*) \arrow[dl, equal, "\text{by unicity}"] \\
(R^\top x^*, R^\top B^* R) &
\end{tikzcd}$$

Proof: uses only rotational invariance and stability

Let $(x_t, B_t) \to (x^*, B^*)$. Let $R \in O_n(\mathbb{R})$

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_R"'] & (x_{t+1}, B_{t+1}) \arrow[r, "t \to \infty"'] \arrow[d, shift right] & (x^*, B^*) \\
(x'_t, B'_t) \arrow[r, "x' \mapsto f(Rx')"] \arrow[u, shift right] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right] &
\end{tikzcd}$$

The diagram commutes with: $(x'_t, B'_t) = \Phi_R(x_t, B_t) = (R^\top x_t, R^\top B_t R)$

Since f is rotational invariant $f(Rx') = f(x')$ so (x'_t, B'_t) optimizes f also.

Hence by stability

$$\begin{tikzcd}
(x'_t, B'_t) = (R^\top x_t, R^\top B_t R) \arrow[r] \arrow[d] & (x^*, B^*) \arrow[dl, phantom, sloped, "=", "\text{by unicity}"'] \\
(R^\top x^*, R^\top B^* R) &
\end{tikzcd}$$

Then for all R

$$Rx^* = x^*$$
$$R^\top B^* R = B^*$$

Proof: uses only rotational invariance and stability

Let $(x_t, B_t) \to (x^*, B^*)$. Let $R \in O_n(\mathbb{R})$

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_R"'] & (x_{t+1}, B_{t+1}) \arrow[r, "t \to \infty"'] \arrow[d, shift right] & (x^*, B^*) \\
(x'_t, B'_t) \arrow[r, "x' \mapsto f(Rx')"] \arrow[u, shift right] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right] &
\end{tikzcd}$$

The diagram commutes with: $(x'_t, B'_t) = \Phi_R(x_t, B_t) = (R^\top x_t, R^\top B_t R)$

Since f is rotational invariant $f(Rx') = f(x')$ so (x'_t, B'_t) optimizes f also.

$$\begin{tikzcd}
\text{Hence by stability } (x'_t, B'_t) = (R^\top x_t, R^\top B_t R) \arrow[r] \arrow[d] & (x^*, B^*) \arrow[dl, equal, "\text{by unicity}"] \\
(R^\top x^*, R^\top B^* R) &
\end{tikzcd}$$

Then for all R

$$Rx^* = x^* \qquad \Rightarrow x^* = 0$$

$$R^\top B^* R = B^* \qquad \Rightarrow B^* = \alpha I_d$$

Consequence of affine-invariance + stability

Lemma: Consider $h(x') = f(H^{1/2}x')$ with f rotational invariant, $H \succ 0$.

example: $h(x) = \frac{1}{2}\gamma x^\top H x, H \succ 0$

Consequence of affine-invariance + stability

Lemma: Consider $h(x') = f(H^{1/2}x')$ with f rotational invariant, $H \succ 0.$

example: $h(x) = \frac{1}{2}\gamma x^\top Hx, H \succ 0$

[stability on f] Suppose that $\exists!\ (x^*, B^*) \in \mathbb{R}^n \times \mathcal{S}_>$ s.t. BFGS optimizing f converges globally to (x^*, B^*).

Consequence of affine-invariance + stability

Lemma: Consider $h(x') = f(H^{1/2}x')$ with f rotational invariant, $H \succ 0$.

example: $h(x) = \frac{1}{2}\gamma x^\top Hx, H \succ 0$

[stability on f] Suppose that $\exists! \ (x^*, B^*) \in \mathbb{R}^n \times \mathcal{S}_>$ s.t. BFGS optimizing f converges globally to (x^*, B^*).

Then on h:
$$\lim_{t\to\infty} x'_t = 0$$
$$\lim_{t\to\infty} B'_t = \alpha H.$$

Consequence of affine-invariance + stability

Lemma: Consider $h(x') = f(H^{1/2}x')$ with f rotational invariant, $H \succ 0$.

example: $h(x) = \frac{1}{2}\gamma x^\top Hx, H \succ 0$

[stability on f] Suppose that $\exists!\ (x^*, B^*) \in \mathbb{R}^n \times \mathcal{S}_>$ s.t. BFGS optimizing f converges globally to (x^*, B^*).

Then on h:

$$\lim_{t\to\infty} x'_t = 0$$
$$\lim_{t\to\infty} B'_t = \alpha H.$$

Corollary: If $h(x') = \frac{1}{2}x'^\top Hx', H \succ 0$, $\lim_{t\to\infty} B'_t = \alpha\nabla^2 f(x)$.

Proof: uses only affine-invariance and stability

Consider (x'_t, B'_t) optimizing h.

$$\begin{tikzcd}(x'_t, B'_t) \arrow[r, "x' \mapsto h(x') = f(H^{1/2}x')"] & (x'_{t+1}, B'_{t+1})\end{tikzcd}$$

Proof: uses only affine-invariance and stability

Consider (x'_t, B'_t) optimizing h.

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_{H^{1/2}}"'] & (x_{t+1}, B_{t+1}) \arrow[d, shift right] \\
(x'_t, B'_t) \arrow[r, "x' \mapsto h(x') = f(H^{1/2}x')"] \arrow[u, shift right, "\Phi_{H^{-1/2}}"'] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right, "\Phi_{H^{-1/2}}"']
\end{tikzcd}$$

By affine-invariance $(x_t, B_t) = \Phi_{H^{-1/2}}(x'_t, B'_t) = (H^{1/2}x'_t, H^{-1/2}B'_tH^{-1/2})$ optimizes f.

Proof: uses only affine-invariance and stability

Consider (x'_t, B'_t) optimizing h.

$$\begin{tikzcd}
(x_t, B_t) \arrow[r, "x \mapsto f(x)"] \arrow[d, shift right, "\Phi_{H^{1/2}}"'] & (x_{t+1}, B_{t+1}) \arrow[r, "t \to \infty"'] \arrow[d, shift right] & (0,\alpha I_d) \\
(x'_t, B'_t) \arrow[r, "x' \mapsto h(x') = f(H^{1/2}x')"] \arrow[u, shift right, "\Phi_{H^{-1/2}}"'] & (x'_{t+1}, B'_{t+1}) \arrow[u, shift right, "\Phi_{H^{-1/2}}"'] &
\end{tikzcd}$$

By affine-invariance $(x_t, B_t) = \Phi_{H^{-1/2}}(x'_t, B'_t) = (H^{1/2}x'_t, H^{-1/2}B'_tH^{-1/2})$ optimizes f.

By stability on f, then $(x_t, B_t) \to (0,\alpha I_d)$, such that:

$$\lim_{t\to\infty} x'_t = H^{-1/2}0 = 0$$
$$\lim_{t\to\infty} B'_t = \alpha H^{1/2} I_d H^{1/2} = \alpha H.$$

Recap

learning Hessian and convergence to the optimum on convex-quadratic implied from:

❶ affine-invariance

❷ stability (convergence to unique point from any starting point)

The same two ingredients and proof ideas applies to CMA-ES to imply:

learning of inverse-Hessian by the covariance matrix on

$$g((x - x^\star)^\top H(x - x^\star)), H \succ 0$$

quite tricky to prove stability in the CMA-ES case

[see PhD thesis Armand Gissler]

Thank you !