Ioannis Mitliagkas

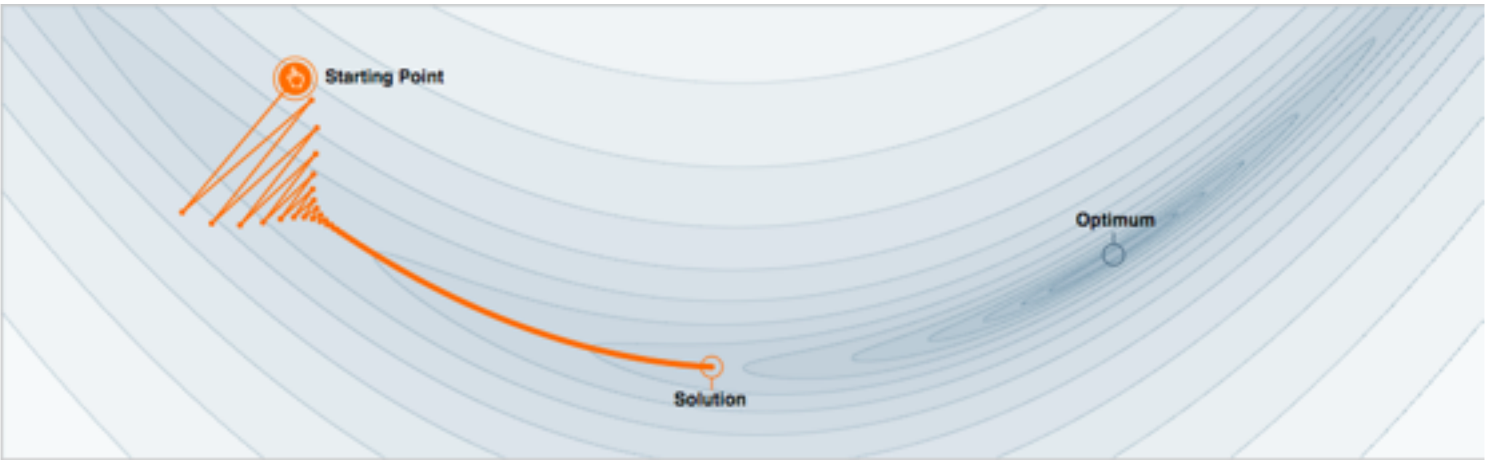
AN INTERESTING PROPERTY OF POLYAK'S MOMENTUM

MILA

GRADIENT DESCENT

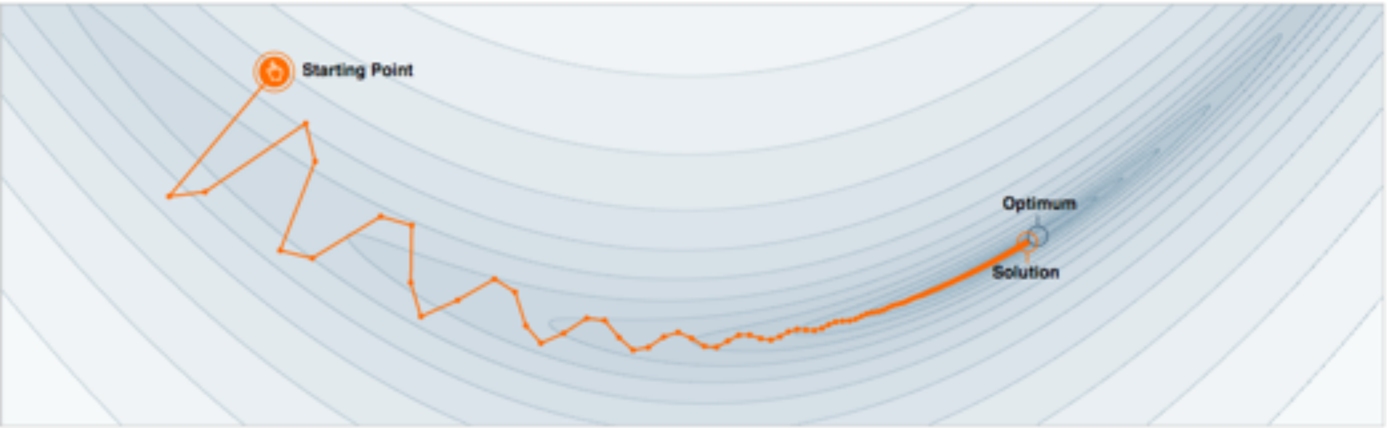
$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

Without momentum



With momentum [Polyak, 1964]

[Distill blog]



CONDITION NUMBER

Dynamic range of curvatures, κ

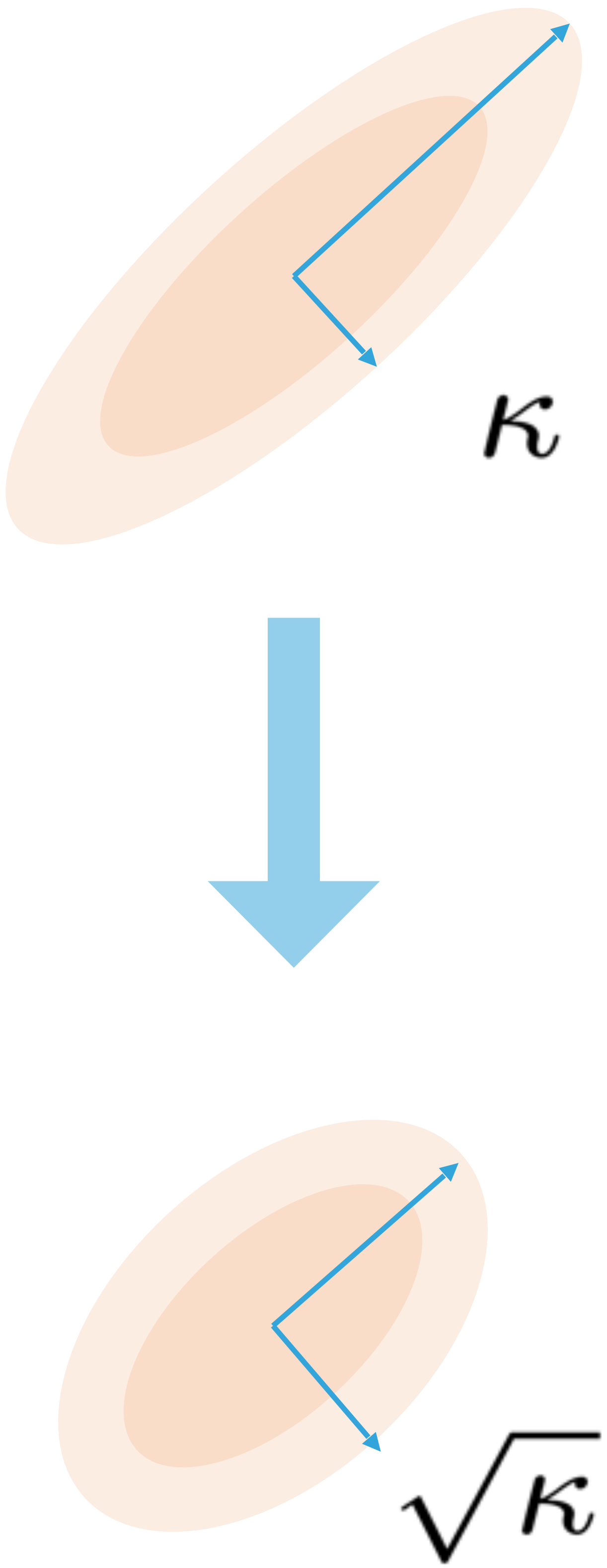
GRADIENT DESCENT ON STRONGLY CONVEX

Convergence rate $O(\frac{\kappa-1}{\kappa+1})$

GRADIENT DESCENT WITH MOMENTUM

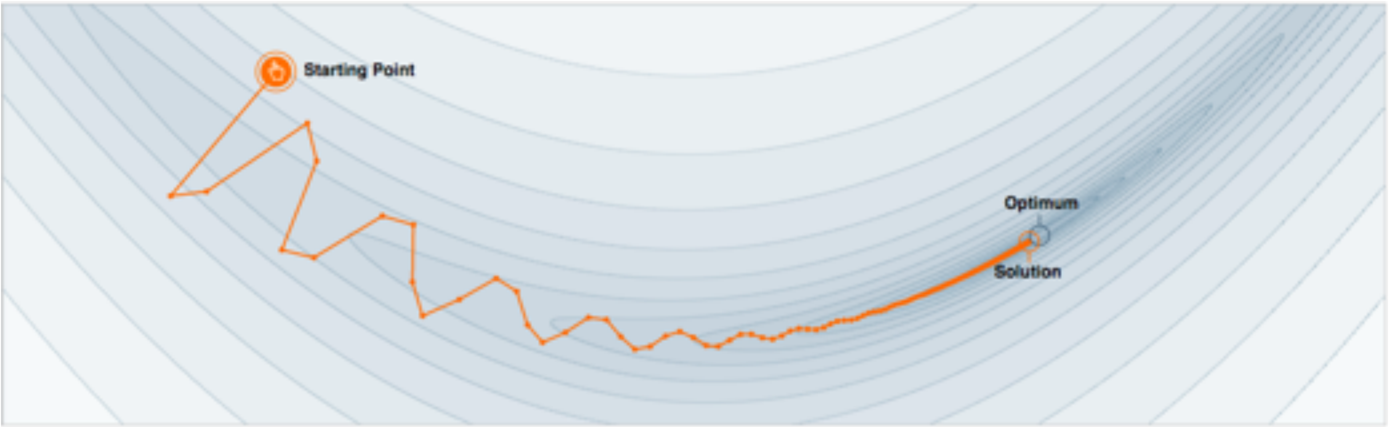
Dependence on κ changes

$$O(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^*$$



EFFECTIVELY IMPROVES THE CONDITION NUMBER

MOMENTUM ALGORITHM



[Distill blog]

Elegant, and very successful optimization method

Along with adaptive methods,
the **workhorse of modern machine learning**

BACKGROUND

GRADIENT DESCENT

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

SCALAR QUADRATIC

Curvature

Minimize $$f(x) = \frac{h}{2}x^2$$

$$\begin{aligned} x_{t+1} &= x_t - \alpha \nabla f(x_t) \\ &= x_t - \alpha h x_t \\ &= (1 - \alpha h) x_t \end{aligned}$$

LINEAR SYSTEM

$$x_{t+1} = (1 - \alpha h)^t x_1$$

Rate of convergence
$$\rho = |1 - \alpha h|$$

RELAXATION PROPERTY

CONVERGENCE RATE

Rate $$\rho = |1 - \alpha h|$$

Convergence $$\|x_t - x^*\| \leq C \cdot \rho^t$$

(there exists C)

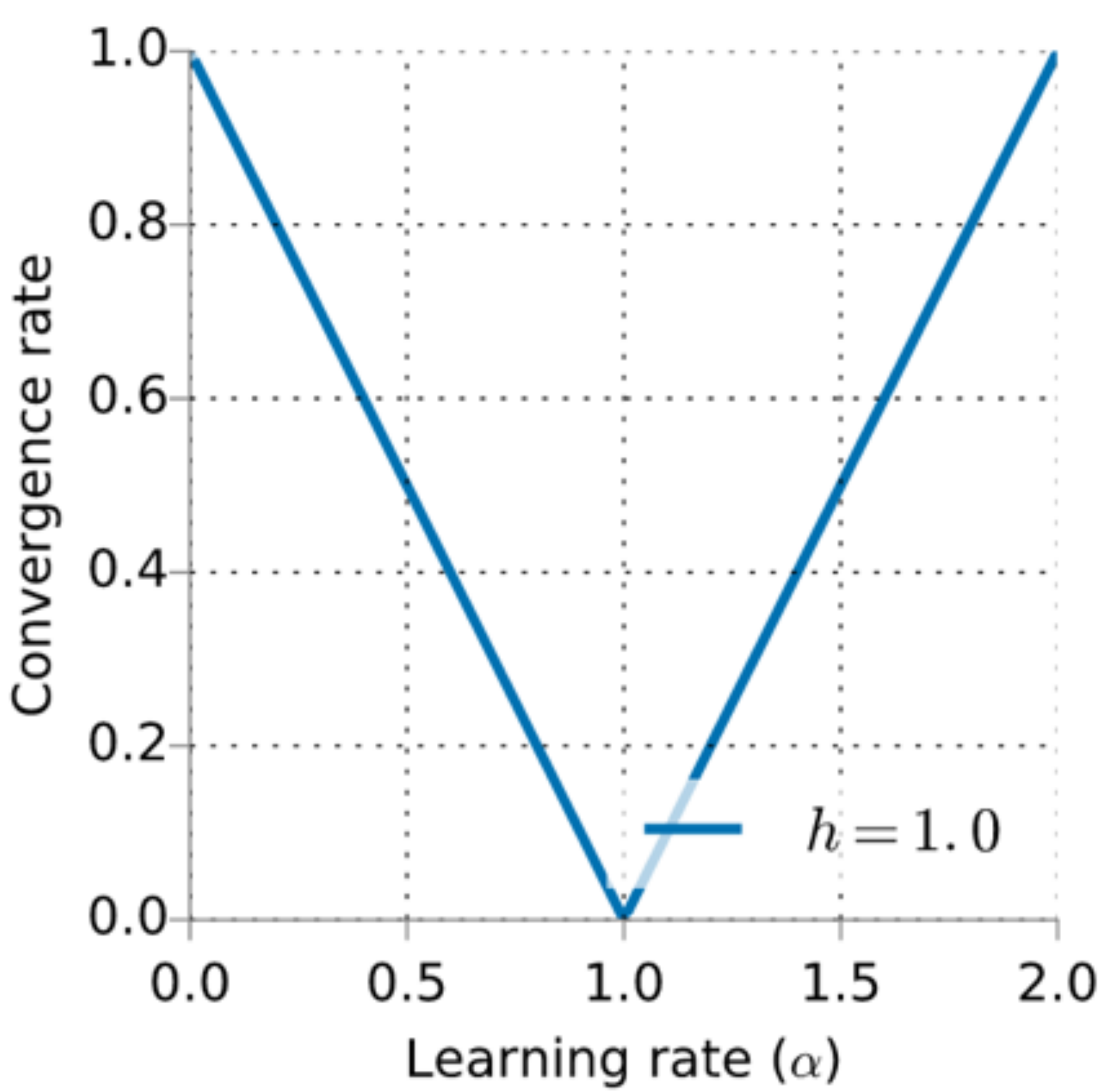
WE WANT THE RATE TO BE A SMALL NUMBER

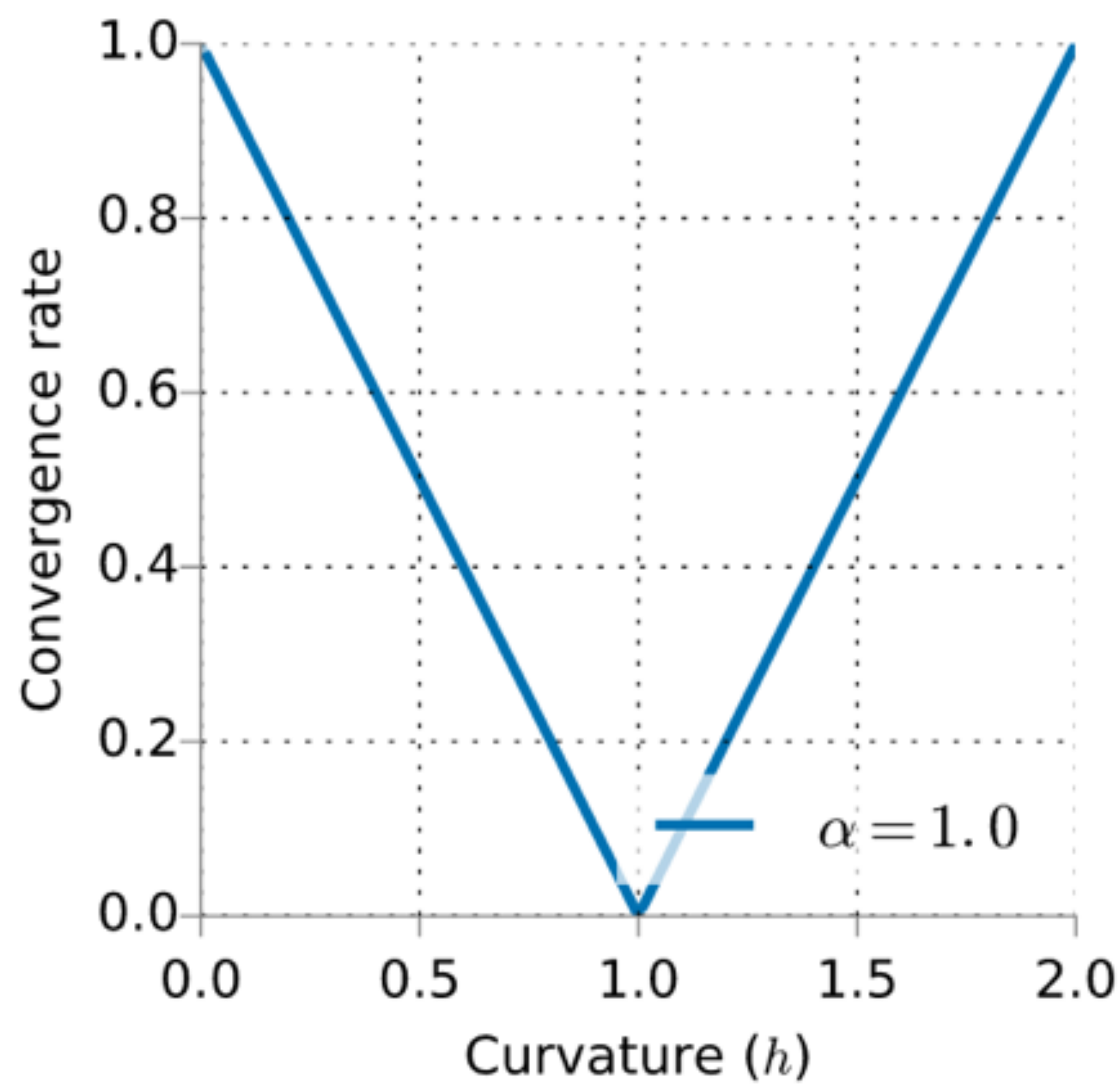
SCALAR QUADRATIC

$$f(x) = \frac{h}{2}x^2$$

Rate of convergence

$$\rho = |1 - \alpha h|$$





MULTIVARIATE QUADRATIC

WLOG, Hessian is diagonal
(why? separability along eigenbasis of H)

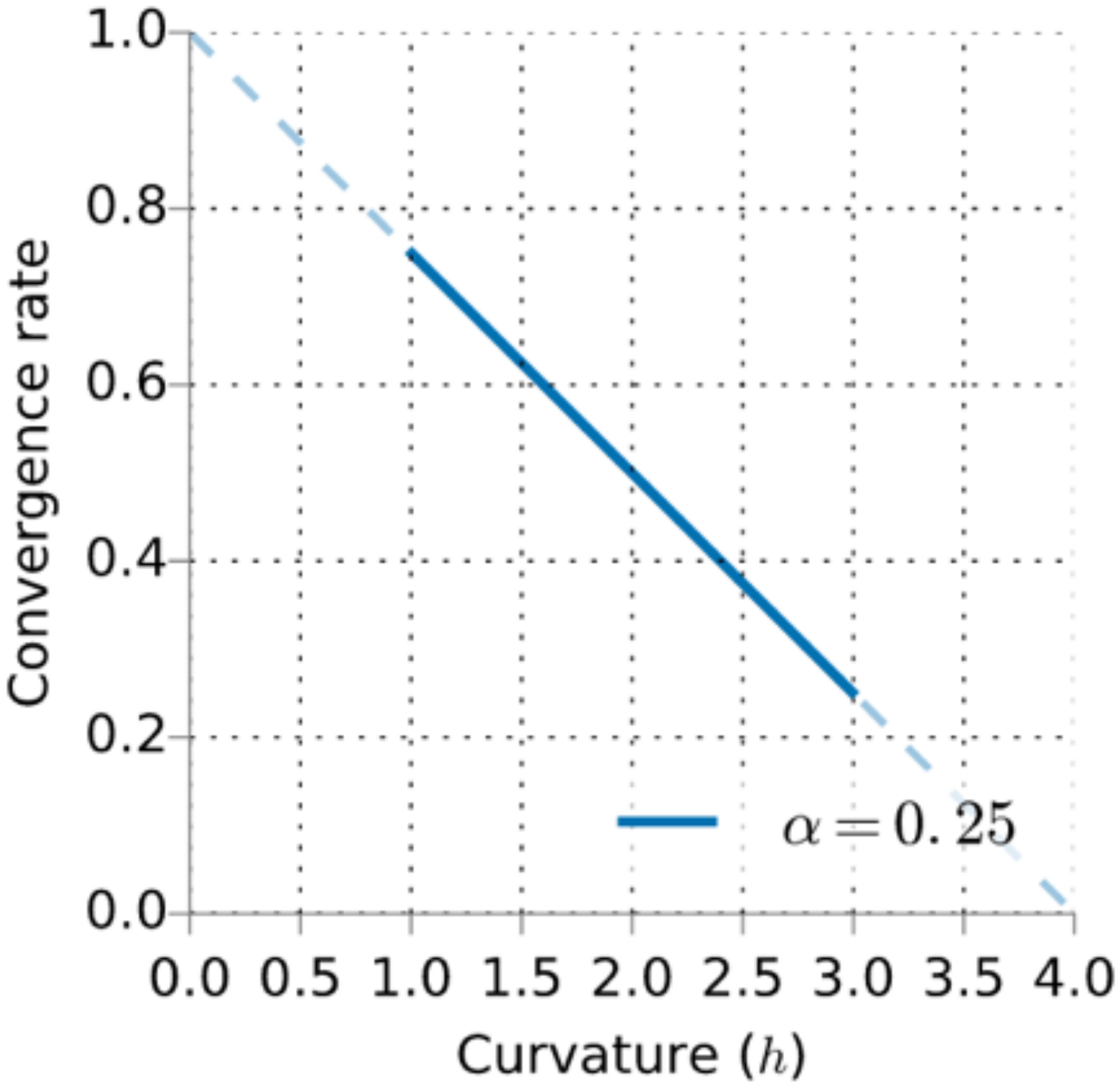
$$f(x) = \frac{1}{2} x^\top H x \qquad H = \begin{bmatrix} h_1 & 0 & 0 \\ 0 & h_2 & 0 \\ 0 & 0 & h_3 \end{bmatrix}$$

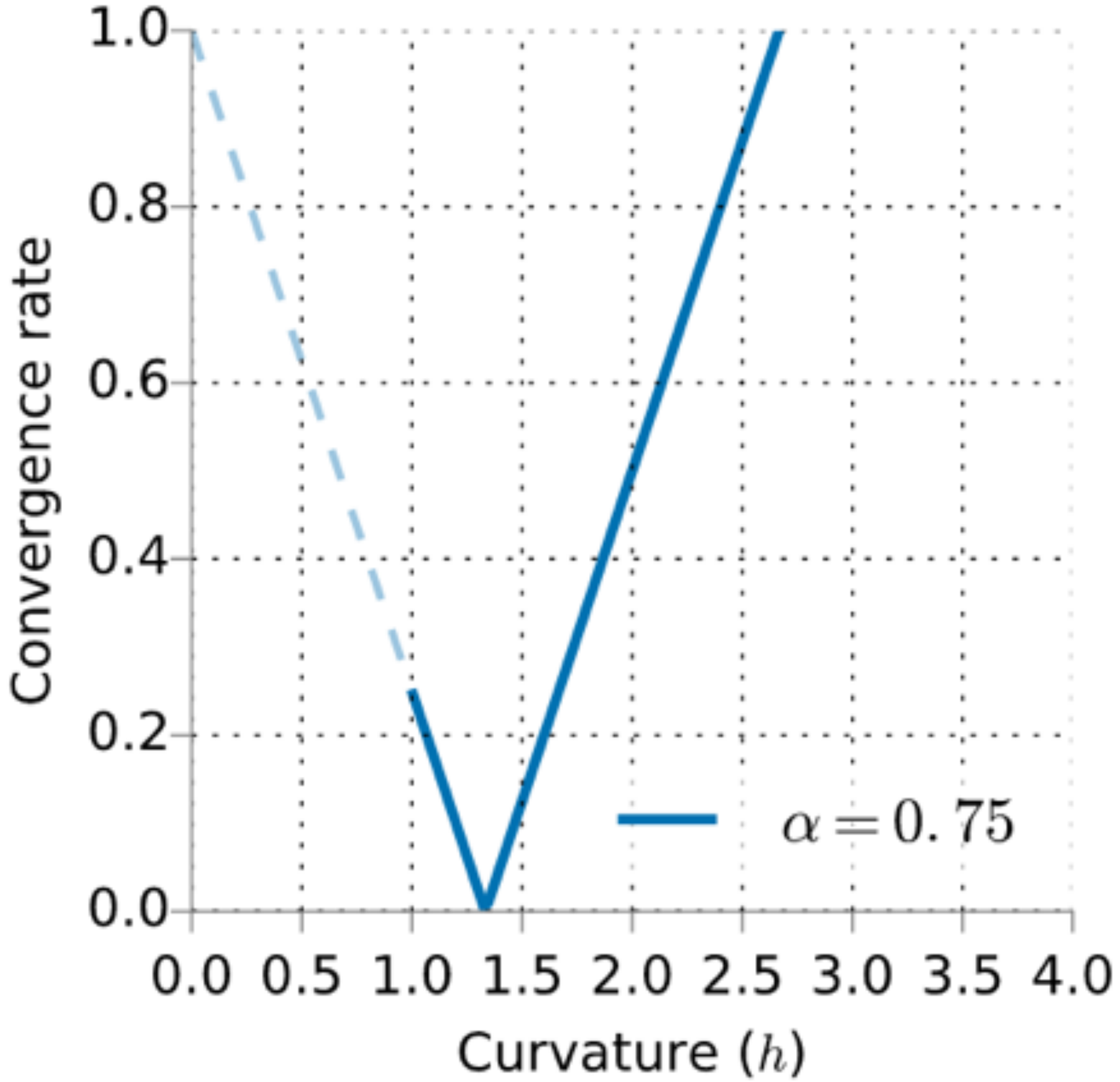
Analysis decomposes into scalar analyses of eigendirections

$$\begin{aligned} x_{t+1}(i) &= x_t(i) - \alpha h_i x_t(i) \\ &= (1 - \alpha h_i) x_t(i) \\ &= (1 - \alpha h_i)^t x_1(i) \end{aligned}$$

SLOWEST DIRECTION DOMINATES

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

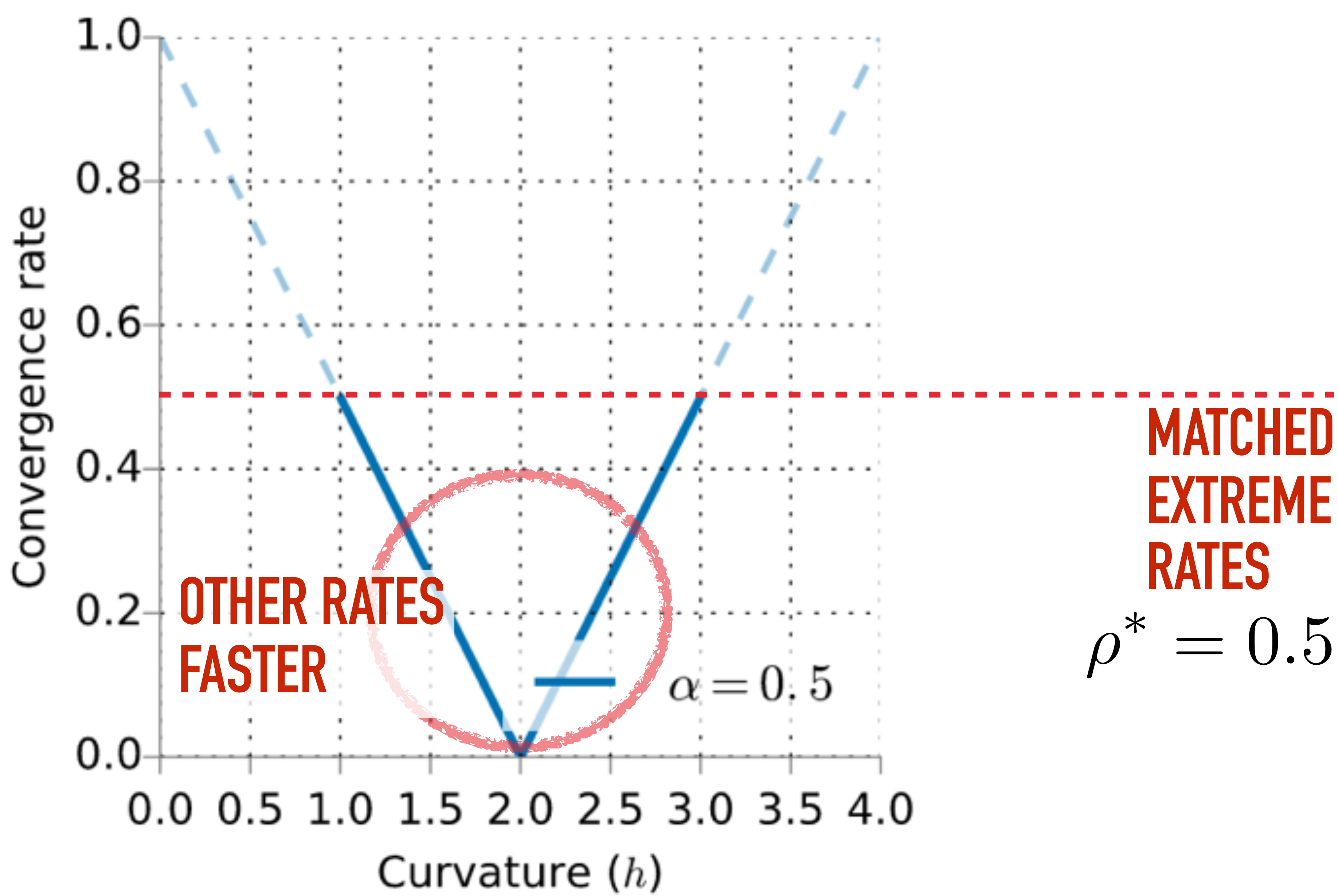




OPTIMAL LEARNING RATE

$$\alpha = 2/(h_{min} + h_{max}) = 0.5$$

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$



$$\rho^* = 0.5$$

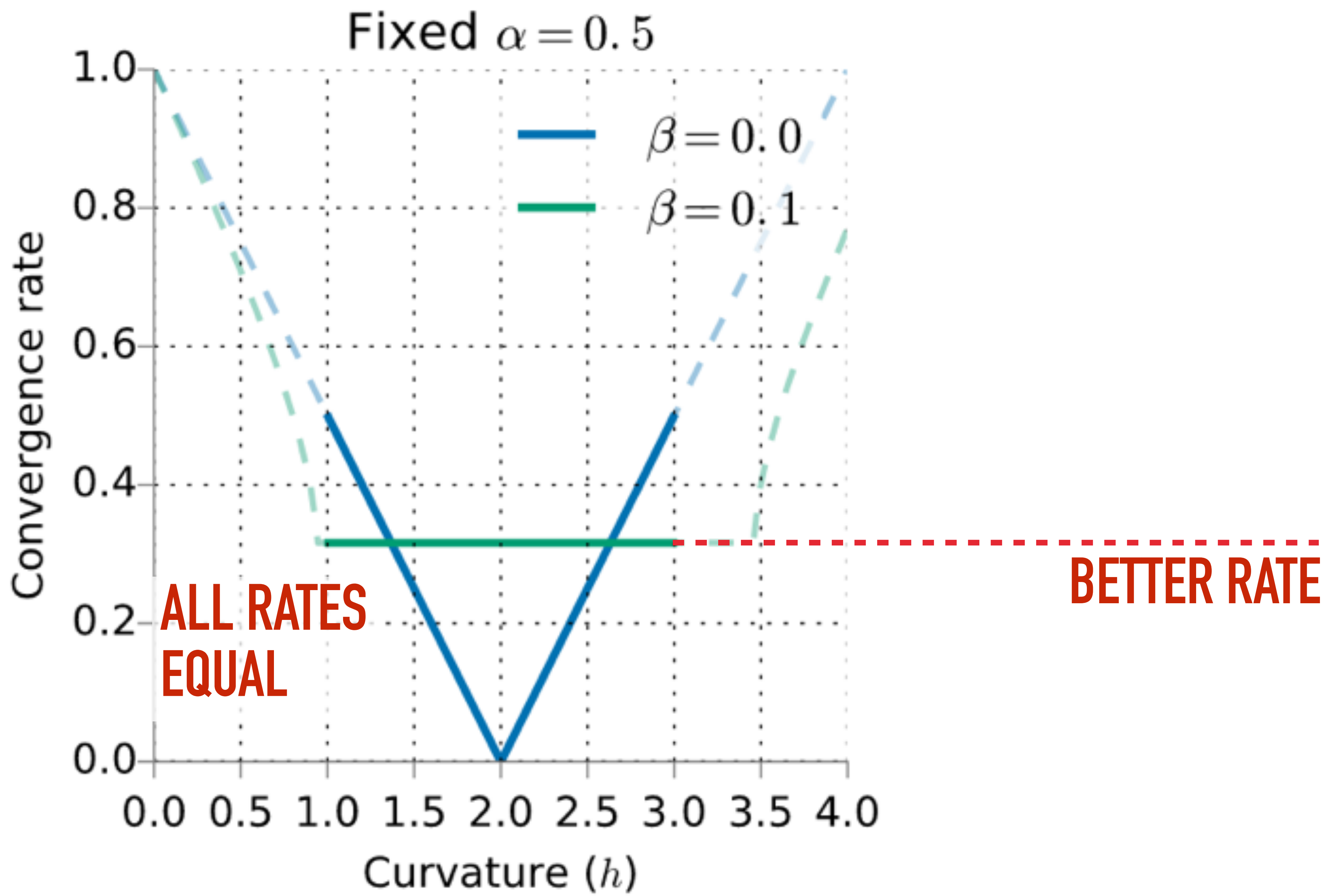
Polyak's momentum

(heavy ball method)

$$x_{t+1} = x_t - \alpha \nabla f(x_t) + \beta (x_t - x_{t-1})$$

PREVIEW

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$



WHAT'S HAPPENING?

POLYAK'S MOMENTUM

Curvature

Minimize $$f(x) = \frac{h}{2}x^2$$

$$\begin{aligned} x_{t+1} &= x_t - \alpha \nabla f(x_t) + \beta(x_t - x_{t-1}) \\ &= x_t - \alpha h x_t + \beta(x_t - x_{t-1}) \\ &= (1 + \beta - \alpha h)x_t - \beta x_{t-1} \end{aligned}$$

CAN WE WRITE AS A LINEAR SYSTEM?

STATE SPACE AUGMENTATION

$$f(x) = \frac{h}{2}x^2$$

$$x_{t+1} = (1 + \beta - \alpha h)x_t - \beta x_{t-1}$$

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} = \begin{bmatrix} 1 - \alpha h + \beta & -\beta \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}$$

LINEAR OPERATOR, A

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} = A^t \begin{bmatrix} x_1 \\ x_0 \end{bmatrix}$$

CONVERGENCE RATE?

CONVERGENCE RATE

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} = A^t \begin{bmatrix} x_1 \\ x_0 \end{bmatrix}$$

RELAXATION?

Asymptotically, spectral radius of A gives rate

$$\left\| \begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} \right\| = O(\rho(A)^t)$$

MORE ATTENTION NEEDED FOR FINITE STEPS

$$\rho(A) = \max\{|\lambda_1(A)|, |\lambda_2(A)|\}$$

SPECTRUM OF MOMENTUM OPERATOR

$$A = \begin{bmatrix} 1 - \alpha h + \beta & -\beta \\ 1 & 0 \end{bmatrix}$$

First sign of 'robustness': $\lambda_1 \lambda_2 = \det(A) = \beta$

Two regions, depending on discriminant

$$\Delta = tr(A)^2 - 4det(A)$$

$\Delta \geq 0$

Two real eigenvalues

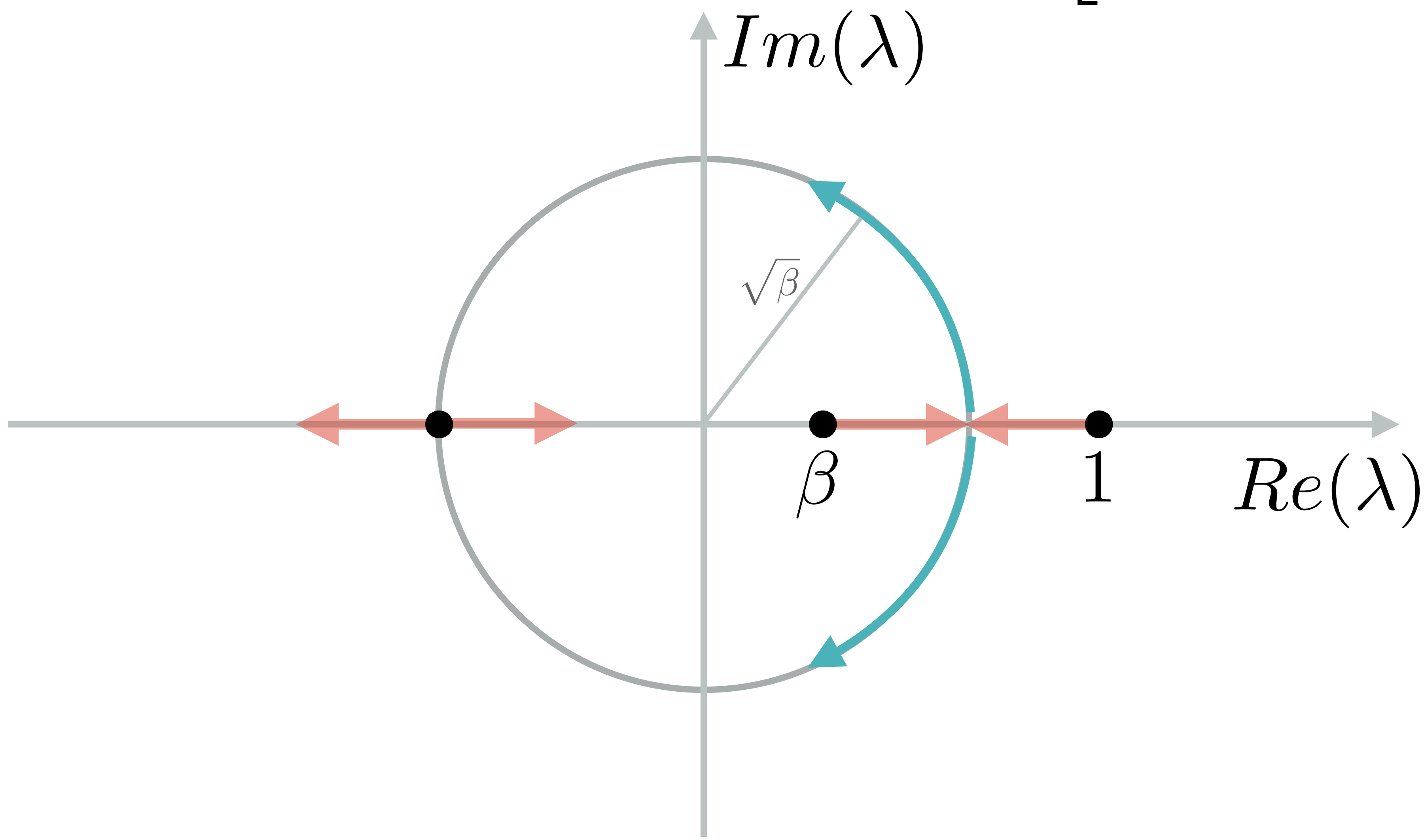
$\Delta < 0$

Two complex conjugate eigenvalues

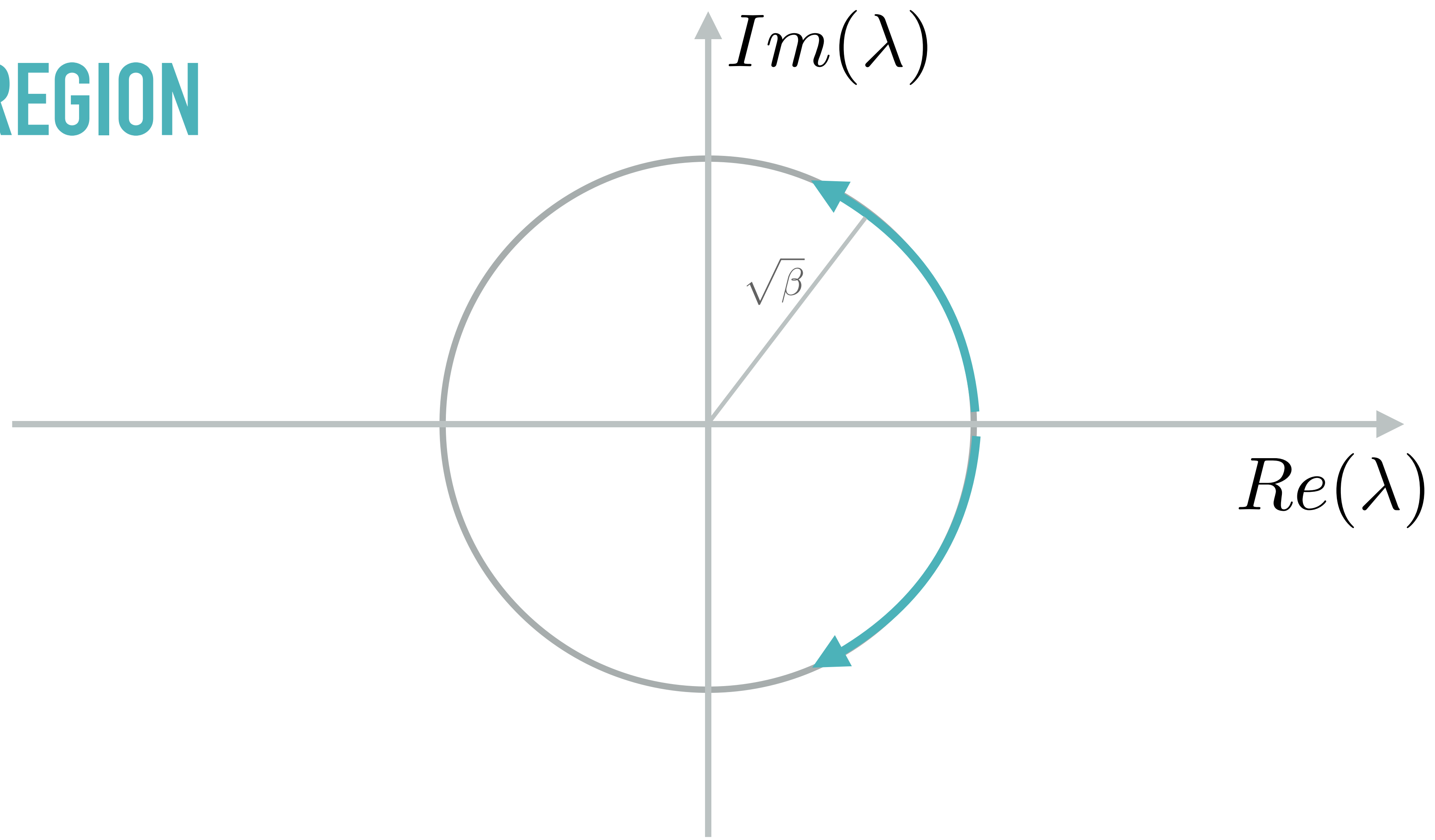
$$\frac{(1-\sqrt{\beta})^2}{h} \leq \alpha \leq \frac{(1+\sqrt{\beta})^2}{h}$$

SPECTRUM OF MOMENTUM OPERATOR

$$A = \begin{bmatrix} 1 - \alpha h + \beta & -\beta \\ 1 & 0 \end{bmatrix}$$



ROBUST REGION

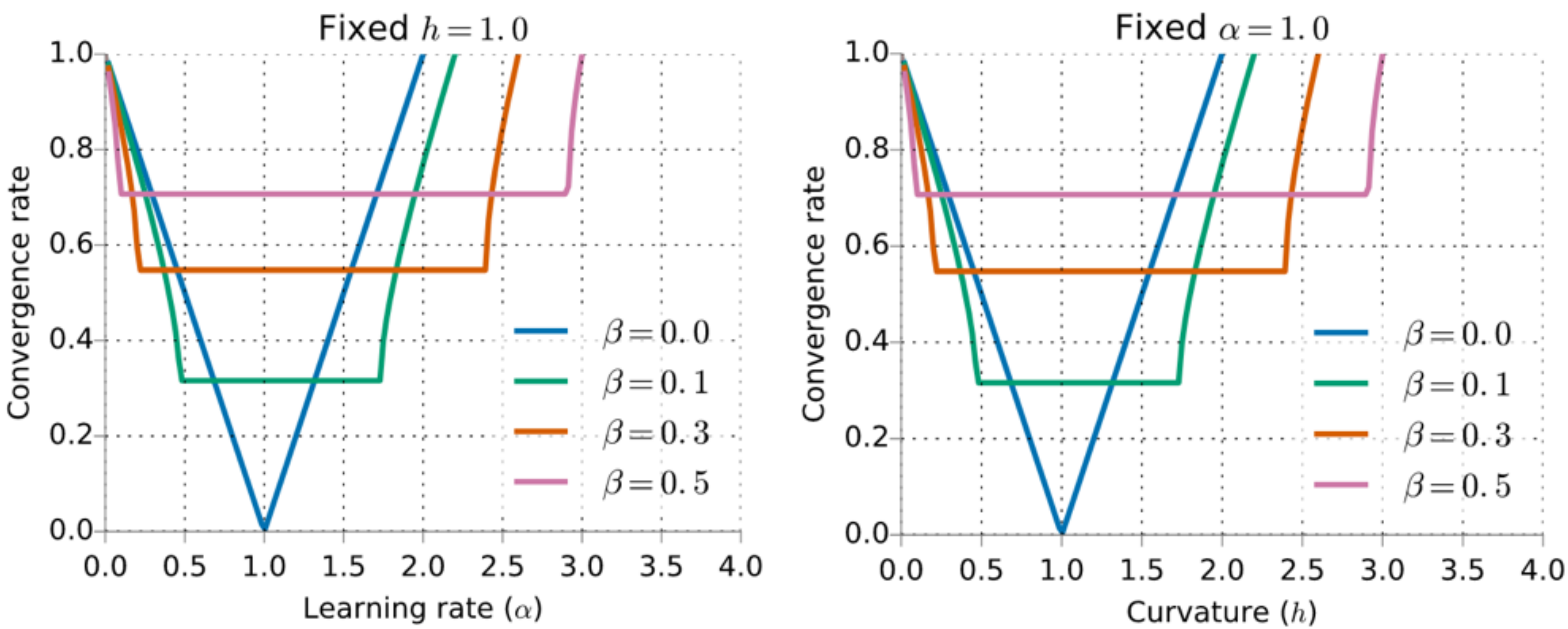


$$\frac{(1-\sqrt{\beta})^2}{h} \leq \alpha \leq \frac{(1+\sqrt{\beta})^2}{h} \qquad \rho(A) = \sqrt{\beta}$$

MOMENTUM, β, CONTROLS WIDTH OF ROBUST REGION

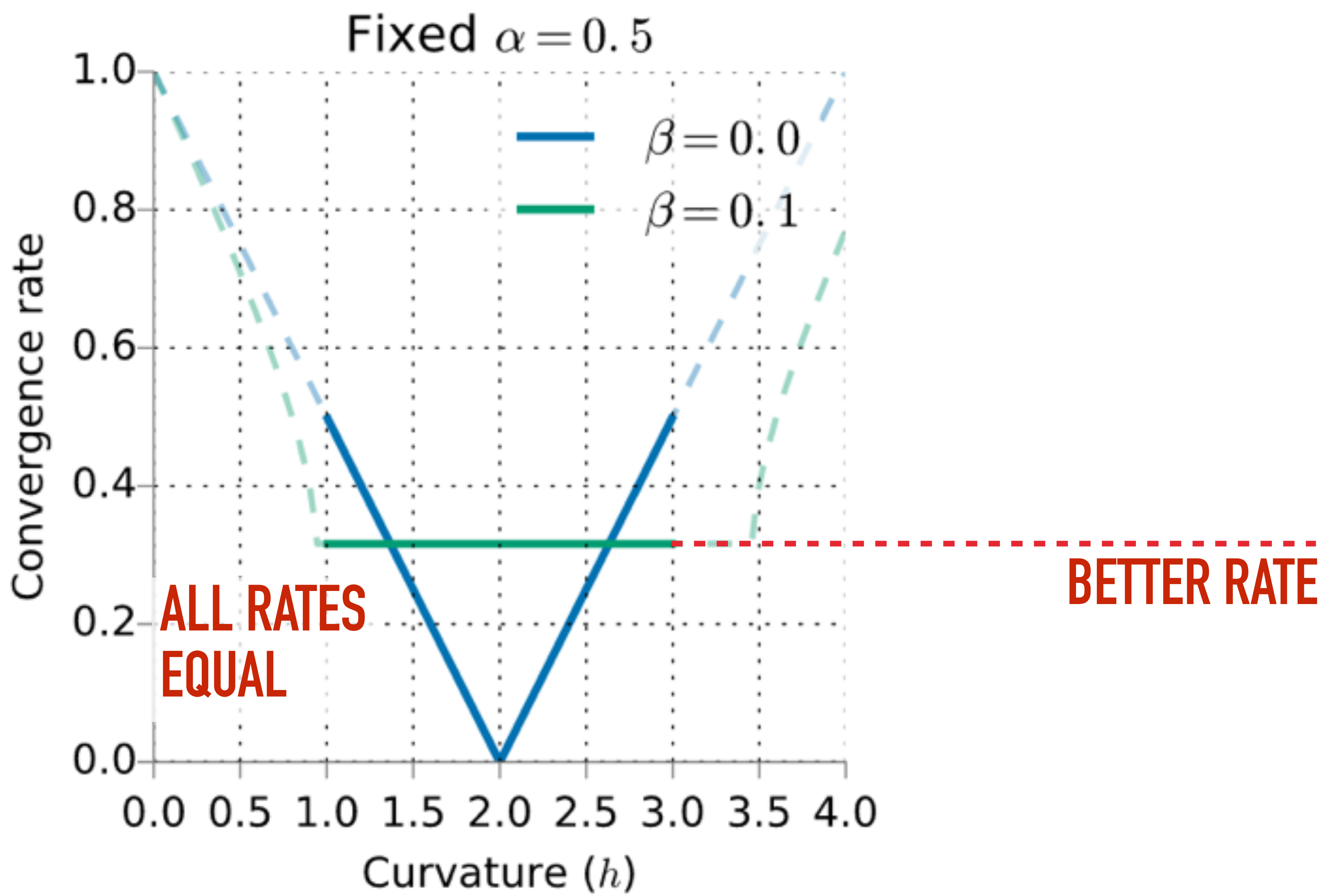
ROBUST REGION

$$\frac{(1-\sqrt{\beta})^2}{h} \leq \alpha \leq \frac{(1+\sqrt{\beta})^2}{h} \qquad \rho(A) = \sqrt{\beta}$$

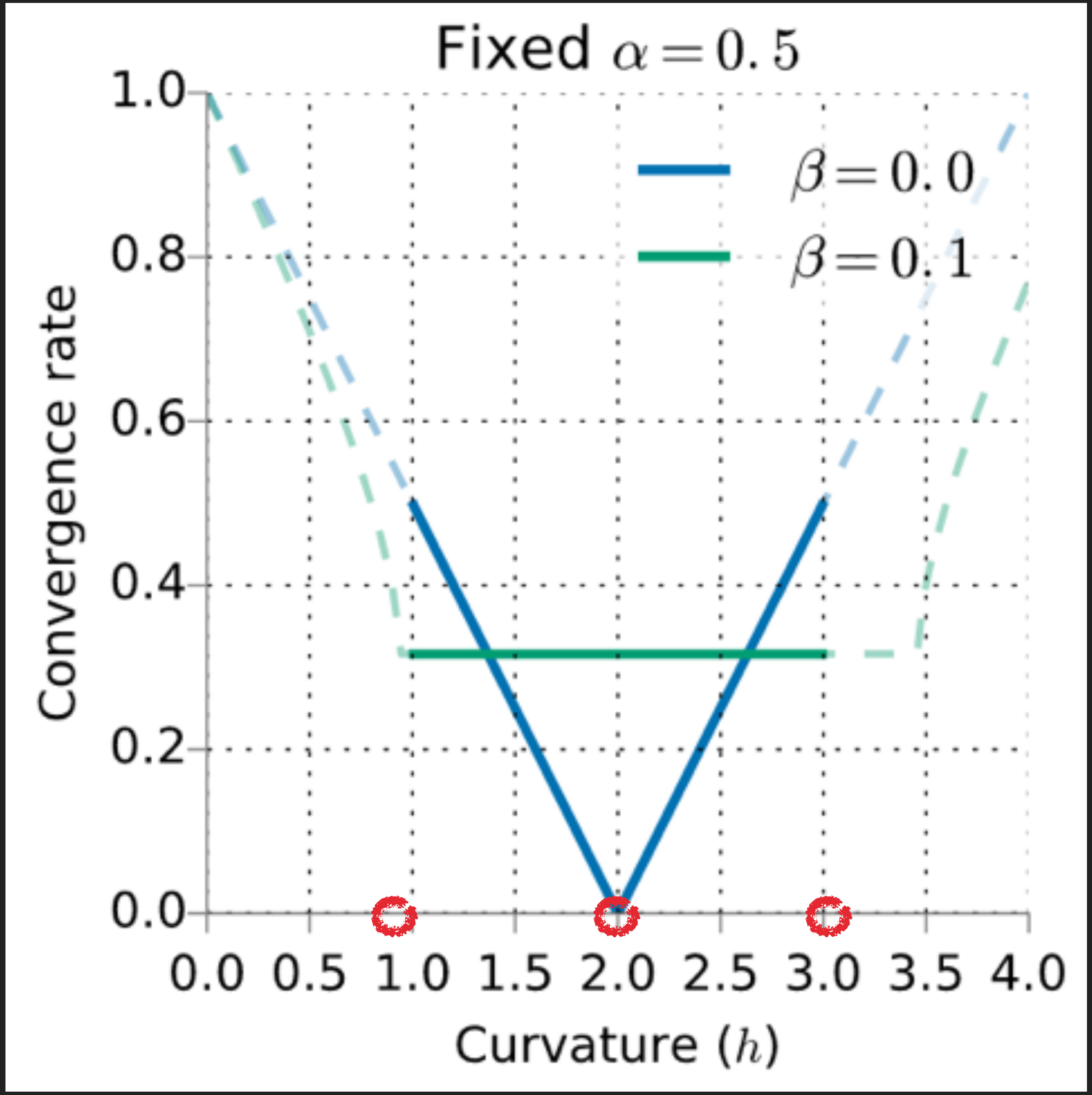


BACK TO EXAMPLE

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

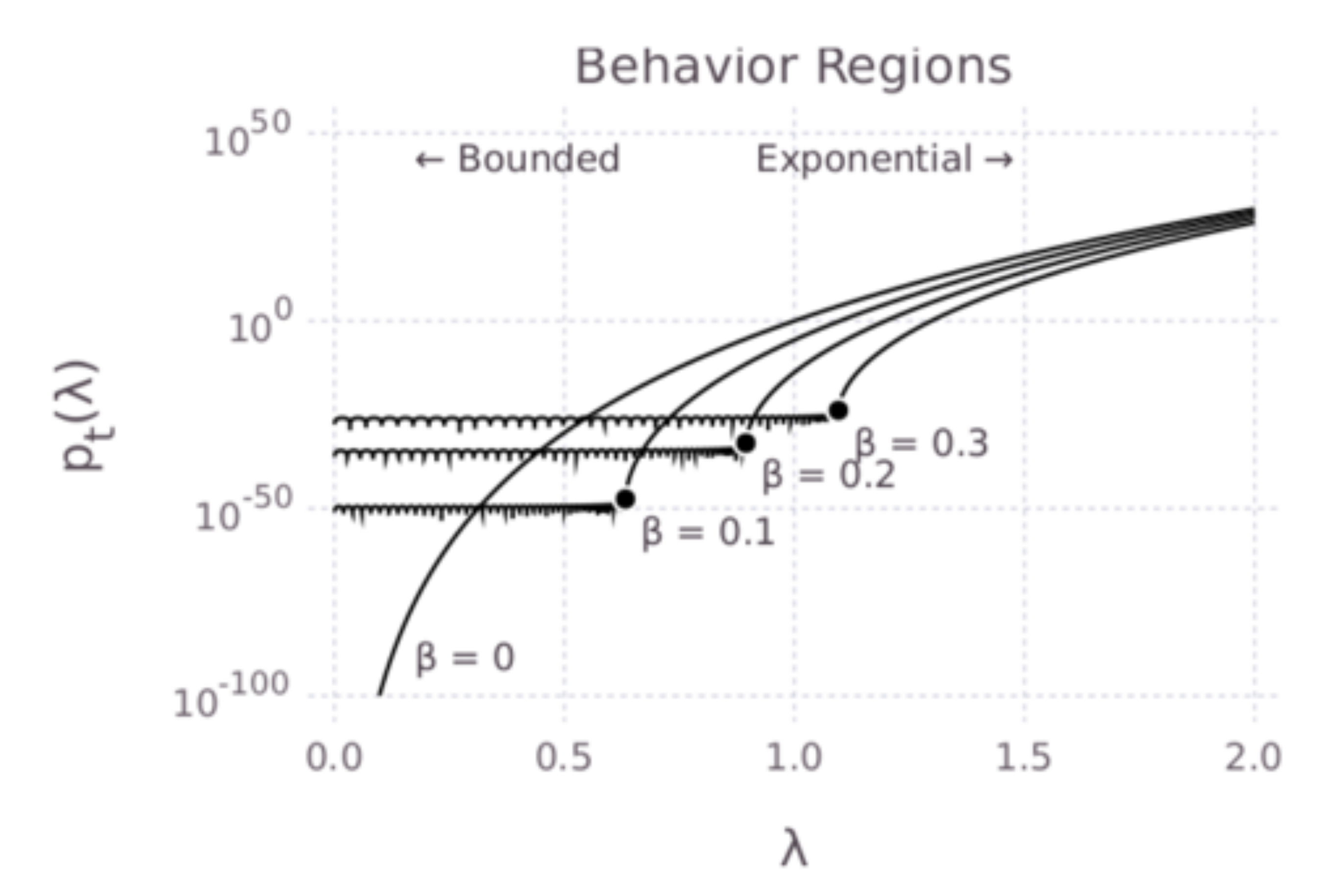


THEME: TUNE LINEAR OPERATOR TO MANIPULATE EIGENVALUES

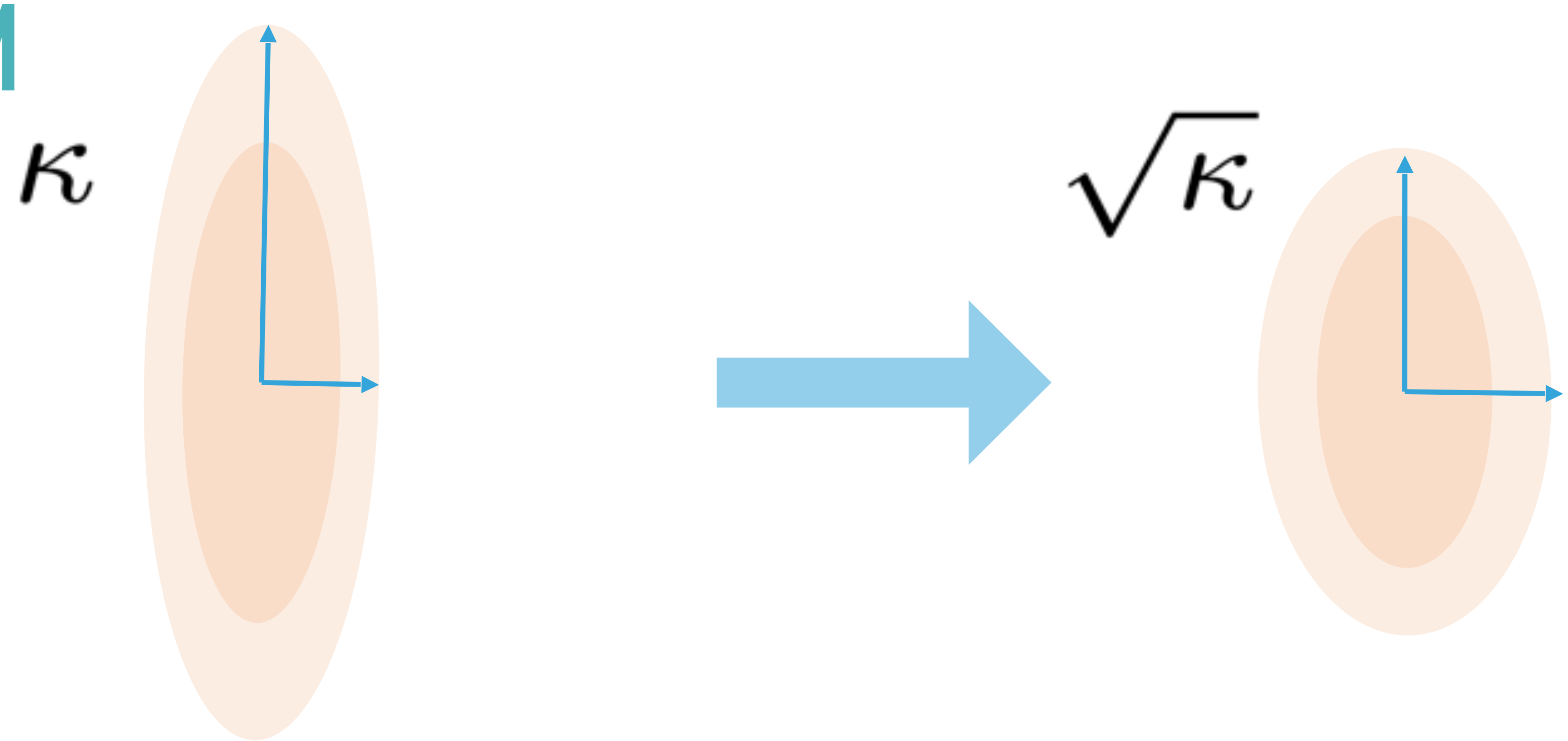


CHRISTOPHER DE SA, BRYAN HE, IOANNIS MITLIAGKAS, CHRISTOPHER RE, PENG XU

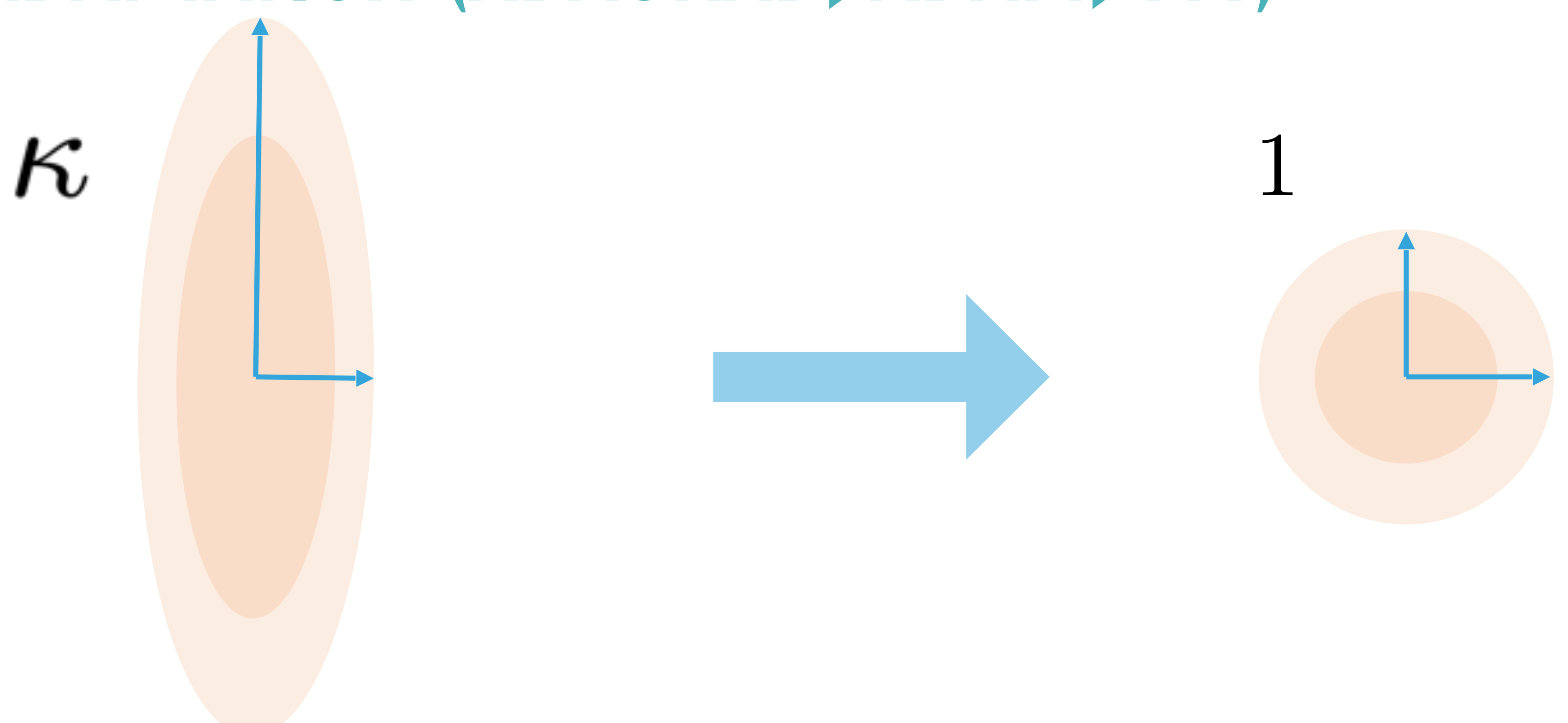
ACCELERATED STOCHASTIC POWER ITERATION



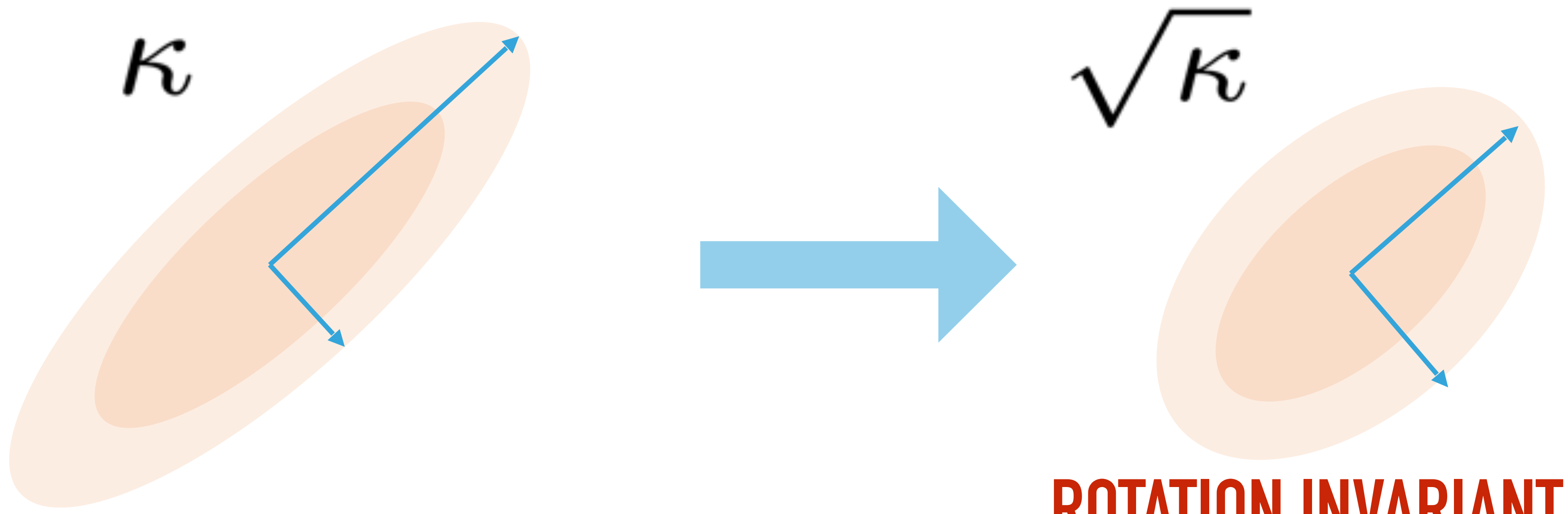
MOMENTUM



DIAGONAL ADAPTATION (ADAGRAD, ADAM, …)

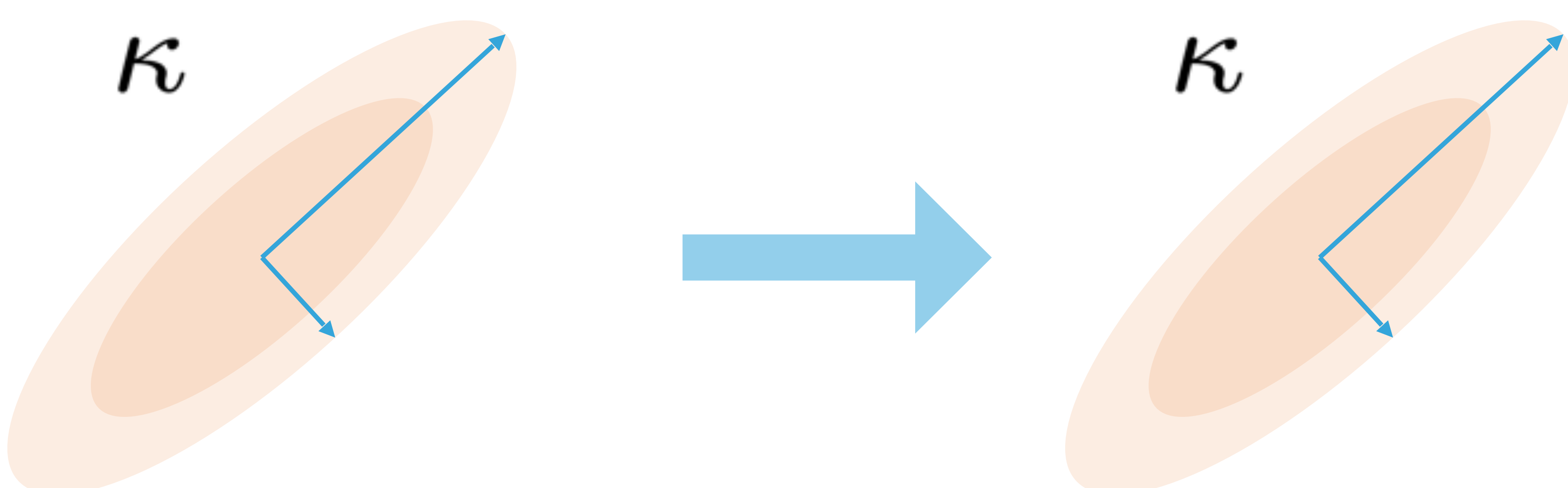


MOMENTUM



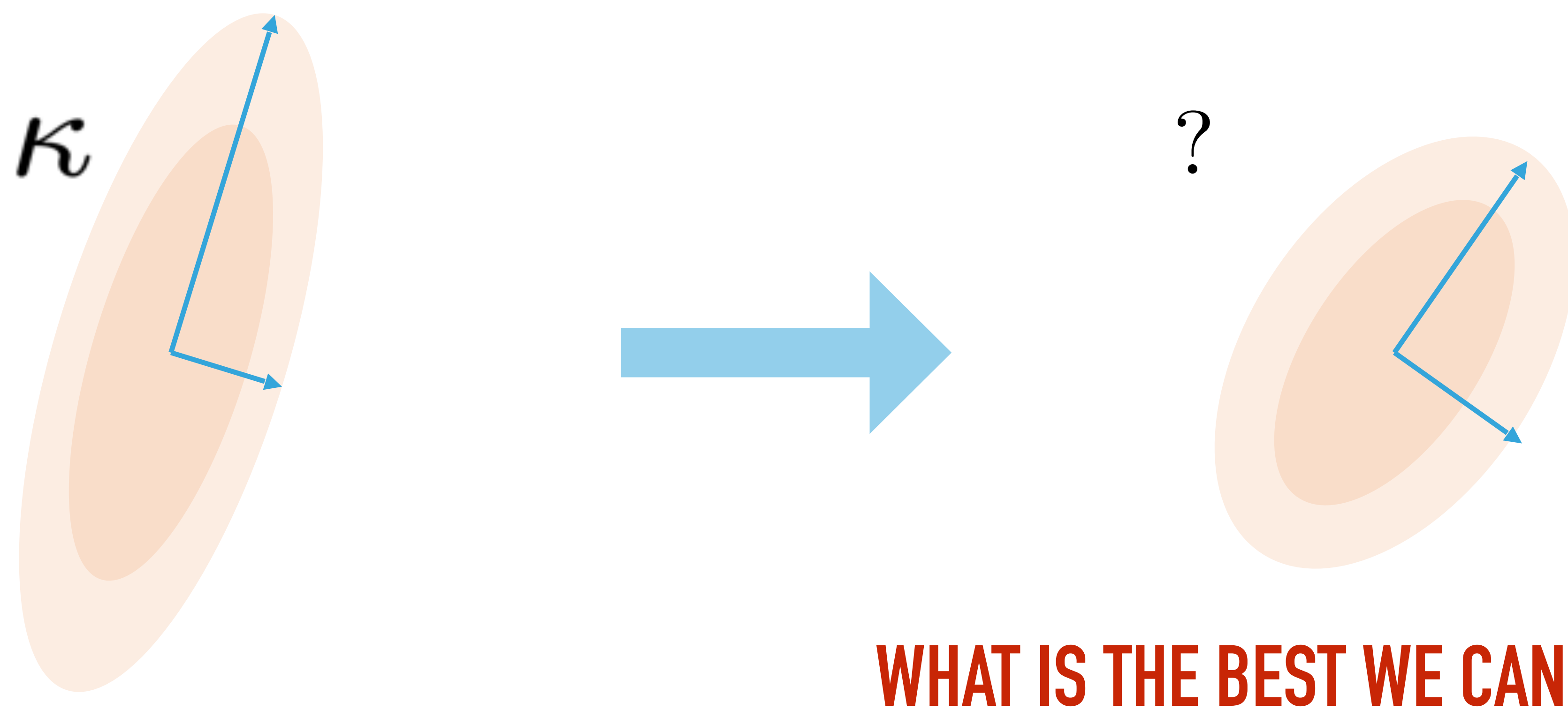
ROTATION INVARIANT

DIAGONAL ADAPTATION (ADAGRAD, ADAM, . . .)



NOT ROTATION INVARIANT

DIAGONAL ADAPTATION (ADAGRAD, ADAM, . . .)



WHAT IS THE BEST WE CAN DO WITH PER VARIABLE STEP SIZES?

$$D^* = \mathrm{argmin}_{D \text{ diagonal}} \frac{\lambda_{max}(D^{1/2} H D^{1/2})}{\lambda_{min}(D^{1/2} H D^{1/2})}$$

EXPLORATION:
OPTIMAL COMBINATION OF DIAGONAL ADAPTATION AND MOMENTUM