7. Symmetric Positive Definite Matrices

- Test for Positive Definite Matrix
  - All eigenvalues of S are positive (S: real symmetric matrix)
  - Energy $x^T S x$ is positive for $x \neq 0$ (best!)
  - $S = A^T A$ with independent columns of A
  - All the pivots are positive
  - All the leading determinants are positive
- Second derivative matrix is positive definite @ a minimum point
- Semidefinite allows zero eigenvalues/energy/pivots/determinants

$$\mathbf{S}=\mathbf{S}^T \rightarrow \begin{cases} \text{all } n \text{ eigenvalues } \boldsymbol{\lambda} \text{ are real numbers} \\ \text{all } n \text{ eigenvectors } \mathbf{q} \text{ can be chosen orthogonal (perpendicular each other)} \end{cases}$$

$$\mathbf{S}=\mathbf{I} \rightarrow \text{all } \lambda=1, \text{ every nonzero vector } \mathbf{x} \ (\mathbf{Ix}=\mathbf{1x})$$

$$\mathbf{q}(\text{unit length}) \rightarrow \mathbf{Q}=\begin{bmatrix} & & \\ q_1 & \cdots & q_n \\ & & \end{bmatrix} \xrightarrow{\text{orthonormal}} \begin{cases} \mathbf{Q}^T\mathbf{Q}=\mathbf{I} \\ \mathbf{Q}^T=\mathbf{Q}^{-1} \end{cases}$$

(Spectal Theorem) Every real symmetric matrix has the form $\mathbf{S}=\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$ $(\leftarrow \mathbf{A}=\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1})$ and every matrix of that form is symmetric.

(quick proofs for orthogonal eigenvectors and real eigenvalues)

$$\left\{ \begin{array}{l} \text{nonzero / zero eigenvalues: } \mathbf{Sx}=\lambda\mathbf{x} \text{ and } \mathbf{Sy}=0\mathbf{y} \\ \text{nonzero eigenvalues: } \mathbf{Sx}=\lambda\mathbf{x} \text{ and } \mathbf{Sy}=\alpha\mathbf{y} \end{array} \right\} \rightarrow \text{orthogonal eigenvector?}$$

$$\text{complex ?: multiply complex conjugate vector } \bar{\mathbf{x}}^T \rightarrow \bar{\mathbf{x}}^T\mathbf{Sx}=\lambda\bar{\mathbf{x}}^T\mathbf{x} \rightarrow \text{real eigenvalue?}$$

$$\mathbf{S}=\begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix}=\bar{\mathbf{S}}^T \rightarrow \lambda=8,-1 \text{ and } \mathbf{x}=\begin{bmatrix} 1 \\ 1+i \end{bmatrix}, \begin{bmatrix} 1-i \\ -1 \end{bmatrix}$$

Test 1: A positive definite matrix has all positive eigenvalues.

Test 2: $\mathbf{S}$ is positive definite if the energy $\mathbf{x}^T\mathbf{S}\mathbf{x}$ is positive for all vectors $\mathbf{x} \neq \mathbf{0}$

$\mathbf{x}^T\mathbf{S}\mathbf{x} > 0 \overset{?}{\longleftrightarrow} \lambda > 0$: $\mathbf{S}\mathbf{x} = \lambda\mathbf{x} \rightarrow \mathbf{x}^T\mathbf{S}\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x},\ \lambda > 0 \rightarrow \mathbf{x}^T\mathbf{S}\mathbf{x} > 0$

If $\mathbf{x}^T\mathbf{S}\mathbf{x} > 0$ for the eigenvectors of $\mathbf{S}$, then $\mathbf{x}^T\mathbf{S}\mathbf{x} > 0$ for every nonzero vector $\mathbf{x}$.

why? $\mathbf{x} = c_1\mathbf{x}_1 + \cdots + c_n\mathbf{x}_n \rightarrow \mathbf{x}^T\mathbf{S}\mathbf{x} = {c_1}^2\lambda_1{\mathbf{x}_1}^T\mathbf{x}_1 + \cdots + {c_n}^2\lambda_n{\mathbf{x}_n}^T\mathbf{x}_n > 0$ if every $\lambda_i > 0$

If $\mathbf{S}_1$ and $\mathbf{S}_2$ are symmetric positive definite, so is $\mathbf{S}_1 + \mathbf{S}_2$

Test 3: $\mathbf{S} = \mathbf{A}^T\mathbf{A}$ for a matrix $\mathbf{A}$ with independent columns

$$\mathbf{x}^T\mathbf{S}\mathbf{x} = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \left(\mathbf{A}\mathbf{x}\right)^T\mathbf{A}\mathbf{x} = \left\|\mathbf{A}\mathbf{x}\right\|^2 \geq 0,\ \mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \rightarrow \mathbf{S}\,?$$

Test 4: All the leading determinants $D_1, D_2, \ldots, D_n$ of $\mathbf{S}$ are positive

Test 5: All the pivots of $\mathbf{S}$ are positive (in elimination)

The kth pivot equals the ratio $\dfrac{D_k}{D_{k-1}}$ of the leading determinants $\left(\text{size } k \text{ and } k-1\right)$

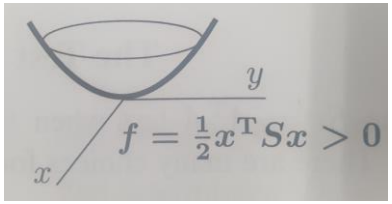
$\mathbf{S} = \mathbf{L}\mathbf{U} = \mathbf{L}\mathbf{D}\mathbf{L}^T = \mathbf{A}^T\mathbf{A} \rightarrow \mathbf{A} = \sqrt{\mathbf{D}}\mathbf{L}^T$ (Cholesky factorization)

$\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \rightarrow \mathbf{A} = \mathbf{Q}\sqrt{\mathbf{\Lambda}}\mathbf{Q}^T = \mathbf{A}^T$

$$\mathbf{S}=\begin{bmatrix}2&-1&0\\-1&2&-1\\0&-1&2\end{bmatrix}\Rightarrow\begin{cases}D_1=2,\ D_2=3,\ D_4=4\\ \xrightarrow{l_{21}=\frac{-1}{2},l_{31}=0}\begin{bmatrix}2&-1&0\\0&3/2&-1\\0&-1&2\end{bmatrix}\xrightarrow{l_{32}=\frac{-2}{3}}\begin{bmatrix}2&-1&0\\0&3/2&-1\\0&0&4/3\end{bmatrix}\end{cases}$$

$$\mathbf{S}=\begin{bmatrix}1&&\\-1/2&1&\\0&-2/3&1\end{bmatrix}\begin{bmatrix}2&-1&0\\&3/2&-1\\&&4/3\end{bmatrix}=\mathbf{LU}=\begin{bmatrix}1&&\\-1/2&1&\\0&-2/3&1\end{bmatrix}\begin{bmatrix}2&&\\&3/2&\\&&4/3\end{bmatrix}\begin{bmatrix}1&-1/2&0\\&1&-2/3\\&&1\end{bmatrix}=\mathbf{LDL}^T$$

$$=\begin{bmatrix}\sqrt{2}&&\\-\sqrt{1/2}&\sqrt{3/2}&\\0&-\sqrt{2/3}&\sqrt{4/3}\end{bmatrix}\begin{bmatrix}\sqrt{2}&-\sqrt{1/2}&0\\&\sqrt{3/2}&-\sqrt{2/3}\\&&\sqrt{4/3}\end{bmatrix}=\mathbf{A}^T\mathbf{A}$$



$$\mathbf{S}=\begin{bmatrix}a&b\\b&c\end{bmatrix}\rightarrow\begin{cases}\text{determinants: } a>0,\ ac-b^2>0\\ \text{pivots: } a>0,\ c-b\dfrac{b}{a}=\dfrac{ac-b^2}{a}>0\\ \text{eigenvalues: } \lambda_1>0,\ \lambda_2>0\\ \text{energy: } ax^2+2bxy+cy^2>0\end{cases}\xrightarrow[b=4]{a=c=5}\begin{cases}E=\mathbf{x}^T\mathbf{S}\mathbf{x}=\begin{bmatrix}x&y\end{bmatrix}\begin{bmatrix}5&4\\4&5\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}=5x^2+8xy+5y^2>0\\ \rightarrow \text{bowl opening upwards, strictly convex}\\ \text{Test: matrix of second derivatives is positive definite}\\ \text{at all points}\end{cases}$$

Test for a minimum $\left(\text{maximum, saddle point}\right)$

$f\left(x\right)$: $\dfrac{df}{dx}=0$ and $\dfrac{d^2f}{dx^2}>0$ at $x=x_0$

$f\left(x,y\right)$: $\left(\dfrac{df}{dx}=0,\ \dfrac{df}{dy}=0\right)$ and $\begin{bmatrix}\partial^2 f/\partial x^2 & \partial^2 f/\partial x\partial y\\ \partial^2 f/\partial x\partial y & \partial^2 f/\partial y^2\end{bmatrix}$ is positive definite at x_0, y_0

Optimization and Machine Learning

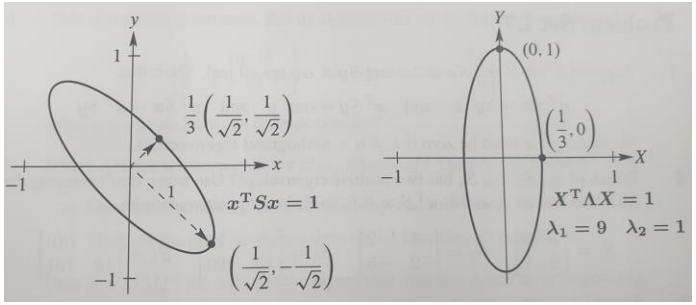
- Calculus: partial derivatives of f are all zero at $\mathbf{x}^* \rightarrow \dfrac{\partial f}{\partial x_i}=0 \rightarrow$ direction to move
- Linear Algebra: matrix $\mathbf{S}$ of second derivatives is positive definite

The graph (cut through the bowl) of $\mathbf{x}^T\mathbf{S}\mathbf{x}=1$ is an ellipse , with its axes pointing along the eigenvectors of $\mathbf{S}$

$$\mathbf{x}^T\mathbf{S}\mathbf{x}=\mathbf{x}^T\left(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\right)\mathbf{x}=\left(\mathbf{x}^T\mathbf{Q}\right)\mathbf{\Lambda}\left(\mathbf{Q}^T\mathbf{x}\right)=\begin{bmatrix}x & y\end{bmatrix}\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}X & Y\end{bmatrix}\mathbf{\Lambda}\begin{bmatrix}X\\ Y\end{bmatrix}$$

$$\mathbf{S}=\begin{bmatrix}5 & 4\\ 4 & 5\end{bmatrix}=\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}9 & \\ & 1\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$$

$$5x^2+8xy+5y^2=1\rightarrow 9\left(\frac{x+y}{\sqrt{2}}\right)^2+1\left(\frac{x-y}{\sqrt{2}}\right)^2=1\rightarrow 9X^2+Y^2=1$$



Multivariate Calculus

one function / one variable: $F\left(x+\Delta x\right) \approx F\left(x\right)+\left(\Delta x\right)\frac{dF}{dx}+\frac{1}{2}\left(\Delta x\right)^2\frac{d^2F}{dx^2}$

one function / n variables: $F\left(\mathbf{x}+\Delta\mathbf{x}\right) \approx F\left(\mathbf{x}\right)+\left(\Delta\mathbf{x}\right)^T\nabla F\left(\mathbf{x}\right)+\frac{1}{2}\left(\Delta\mathbf{x}\right)^T\mathbf{H}\left(\mathbf{x}\right)\left(\Delta\mathbf{x}\right)$

$$\mathbf{x}=\left(x_1,\ldots,x_n\right),\ \underbrace{\nabla F\left(\mathbf{x}\right)=\begin{bmatrix}\partial F/\partial x_1\\ \vdots\\ \partial F/\partial x_n\end{bmatrix}}_{n \text{ derivatives}},\quad \underbrace{H_{jk}=\frac{\partial^2 F}{\partial x_j\partial x_k}=H_{kj}}_{\text{symmetric: } n^2\rightarrow\left(\frac{1}{2}n^2+\frac{1}{2}n\right) \text{ derivatives}}$$

m functions / n variables: $\mathbf{f}=\left(f_1\left(\mathbf{x}\right),\ldots,f_m\left(\mathbf{x}\right)\right)$

$$\mathbf{f}\left(\mathbf{x}+\Delta\mathbf{x}\right)\approx\mathbf{f}\left(\mathbf{x}\right)+\mathbf{J}\left(\mathbf{x}\right)\left(\Delta\mathbf{x}\right)$$

$$\mathbf{J}\left(\mathbf{x}\right)=\begin{bmatrix}\left(\nabla f_1\right)^T\\ \vdots\\ \left(\nabla f_m\right)^T\end{bmatrix}=\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \vdots & & \vdots\\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n}\end{bmatrix},\quad \underbrace{J_{jk}=\frac{\partial f_j}{\partial x_k}}_{\text{Jacobian matrix}}\leftarrow \text{multivariable calculus}$$

Optimization (1)

	unknowns	constraints	condition
Applied Mathematics	vectors	matrices	first derivative = zero
Calculus of Variation	functions	integrals	first variation = zero

$\arg\min F(\mathbf{x}) = \text{value(s) of } \mathbf{x} \text{ where } F \text{ reaches its minimum} = \mathbf{x}^*$

$$\text{minimize } F(\mathbf{x}) \rightarrow \begin{cases} \text{solving } \nabla F(\mathbf{x}) = 0 \\ \text{solving } \mathbf{f}(\mathbf{x}) \left[= \nabla F(\mathbf{x}) \right] = 0 \rightarrow n \text{ equations and } n \text{ unknowns} \end{cases}$$

1) steepest descent: $\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla F(\mathbf{x}_k)$

$s_k = \text{step size (learning rate)} \rightarrow \text{exact line search for best } s_k$

2) Newon's method: $\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}(\mathbf{x}_k)^{-1} \nabla F(\mathbf{x}_k)$

$\mathbf{H}(\mathbf{x}_k) = \text{Hessian, Jacobian of the gradient}$

$$\nabla F(\mathbf{x}_{k+1}) \approx \nabla F(\mathbf{x}_k) + \mathbf{H}(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k) = 0$$

Optimization (2)

$$\text{convergence}\begin{cases}1)\ \text{steepest descent: linear, } \left\|\mathbf{x}_{k+1}-\mathbf{x}^*\right\| \le C\left\|\mathbf{x}_k-\mathbf{x}^*\right\| \ \ (C<1) \\ 2)\ \text{Newton's method: quadratic, } \left\|\mathbf{x}_{k+1}-\mathbf{x}^*\right\| \le C\left\|\mathbf{x}_k-\mathbf{x}^*\right\|^2\end{cases}$$

$$\mathbf{f}\left(\mathbf{x}_{k+1}\right)=0 \rightarrow \mathbf{f}\left(\mathbf{x}_k\right)+J\left(\mathbf{x}_k\right)\left(\mathbf{x}_{k+1}-\mathbf{x}_k\right)=0 \rightarrow \mathbf{x}_{k+1}=\mathbf{x}_k-J\left(\mathbf{x}_k\right)^{-1}\mathbf{f}\left(\mathbf{x}_k\right)$$

example: $f\left(x\right)=x^2-9=0$

Gradient Descent: Downhill to a Minimum

"gradient descent" uses the derivatives $\frac{\partial f}{\partial x_i}$ to find a direction that reduces $f(\mathbf{x})$

$$z = f(x,y) \rightarrow \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right), \|\nabla f\| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2} = (\text{steepest slope of } f)$$

$$\begin{cases} f = \text{const} \ (\text{level set, surface}) \\ \nabla f = 0 \text{ on the surface} \\ \mathbf{H} = \nabla^2 f \rightarrow \begin{cases} \text{positive semi-definite: convexity} \\ \text{positive definite: strict convexity} \end{cases} \end{cases}$$

pure quadratic, unknown $x,\ y$

$$f = \frac{1}{2}\mathbf{x}^T\mathbf{S}\mathbf{x} = \frac{1}{2}\left(x^2 + by^2\right) \rightarrow \mathbf{S} = \begin{bmatrix} 1 & 0 \\ 0 & b \end{bmatrix} \xrightarrow{b<1} \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{1}{b} = (\text{condition number})$$

$$f(x,y) = 2x + 5y \rightarrow \nabla f = \begin{bmatrix} 2 \\ 5 \end{bmatrix} \rightarrow \mathbf{H} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Example

$$\left.\begin{aligned} f(\mathbf{x}) &= \frac{1}{2}\mathbf{x}^T\mathbf{S}\mathbf{x} - \mathbf{a}^T\mathbf{x}(-b) \\ \nabla f(\mathbf{x}) &= \mathbf{S}\mathbf{x} - \mathbf{a} \\ \mathbf{H} &= \mathbf{S} \end{aligned}\right\} \rightarrow \begin{cases} \text{minimum of } f(\mathbf{x}) \text{ is at } \mathbf{x}^* = \mathbf{S}^{-1}\mathbf{a} = \arg\min f(\mathbf{x}) \\ f^* = \frac{1}{2}\left(\mathbf{S}^{-1}\mathbf{a}\right)^T \mathbf{S}\left(\mathbf{S}^{-1}\mathbf{a}\right) - \mathbf{a}^T\left(\mathbf{S}^{-1}\mathbf{a}\right) \\ = \frac{1}{2}\mathbf{a}^T\left(\mathbf{S}^{-1}\mathbf{S}\right)\mathbf{S}^{-1}\mathbf{a} - \mathbf{a}^T\mathbf{S}^{-1}\mathbf{a} = -\frac{1}{2}\mathbf{a}^T\mathbf{S}^{-1}\mathbf{a} \end{cases}$$

$$f(\mathbf{x}) = \log(\det \mathbf{X}) \rightarrow \frac{\partial f}{\partial(\det \mathbf{X})}\frac{\partial(\det \mathbf{X})}{\partial x_{ij}} = \frac{C_{ij}}{\det \mathbf{X}} \rightarrow \nabla f(\mathbf{x}) = \text{entries of } \mathbf{X}^{-1} \rightarrow \nabla^2 f(\mathbf{x})?$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \rightarrow \begin{cases} \det \mathbf{X} = x_{11}\left|\text{minor}\right| - x_{12}\left|\text{minor}\right| + \cdots + x_{1n}\left|\text{minor}\right| \\ = x_{11}C_{11} - x_{12}C_{12} + \cdots + x_{1n}C_{1n} \end{cases}$$

$$\frac{\partial(\det \mathbf{X})}{\partial x_{11}} = C_{11} \rightarrow {x_{11}}^{-1} = \frac{C_{11}}{\det \mathbf{X}}$$

Gradient Descent: Convergence

gradient $\left(\text{steepest}\right)$ descent: $\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla f\left(\mathbf{x}_k\right)$

s_k : step size $\rightarrow$
- too big: oscillate all over the place
- too small: take too long
- exact line search: choose s_k to make $f\left(\mathbf{x}_{k+1}\right)$ a minimum in search direction
- bracketing: $s_0, \alpha s_0, \alpha^2 s_0$

convergence analysis

convexity: $\mathbf{H} = \nabla^2 f$ has eigenvalues between $0 < m \le \lambda \le M$ at all $\mathbf{x}$

$$f\left(\mathbf{x}_{k+1}\right) \le f\left(\mathbf{x}_k\right) + \left(\nabla f\right)^T \left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) + \frac{M}{2}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2$$

$$= f\left(\mathbf{x}_k\right) - s\left\|\nabla f\right\|^2 + \frac{Ms^2}{2}\left\|\nabla f\right\|^2 \rightarrow \text{best: } s = \frac{1}{M}$$

$$\left.\begin{aligned} f\left(\mathbf{x}_{k+1}\right) &\le f\left(\mathbf{x}_k\right) - \frac{1}{2M}\left\|\nabla f\left(\mathbf{x}_k\right)\right\|^2 \\ f\left(\mathbf{x}^*\right) &\ge f\left(\mathbf{x}_k\right) - \frac{1}{2m}\left\|\nabla f\left(\mathbf{x}_k\right)\right\|^2 \end{aligned}\right\} \rightarrow f\left(\mathbf{x}_{k+1}\right) - f\left(\mathbf{x}^*\right) \le \left(1 - \frac{m}{M}\right)\left[f\left(\mathbf{x}_k\right) - f\left(\mathbf{x}^*\right)\right]$$

Example with Zig-Zag

$$f=\frac{1}{2}\left(x^2+by^2\right)\rightarrow\begin{bmatrix}x\\y\end{bmatrix}_{k+1}=\begin{bmatrix}x\\y\end{bmatrix}_k-s_k\begin{bmatrix}x\\by\end{bmatrix}_k \text{ with } \left(x_0,y_0\right)=\left(b,1\right),\ f_0=\frac{1}{2}\left(b^2+b\right)$$

$$\left(x_k,y_k\right)=\left(b\left(\frac{b-1}{b+1}\right)^k,\left(\frac{1-b}{b+1}\right)^k\right),\ f_k=\left(\frac{1-b}{b+1}\right)^{2k}f_0$$

$$\text{small } b\,?\left(b=0.1\right)\begin{bmatrix}0.1\\1\end{bmatrix}\rightarrow\begin{bmatrix}0.1\left(\frac{-0.9}{1.1}\right)\\\frac{0.9}{1.1}\end{bmatrix}\rightarrow\begin{bmatrix}0.1\left(\frac{-0.9}{1.1}\right)^2\\\left(\frac{0.9}{1.1}\right)^2\end{bmatrix},\ \text{zig-zag (bouncing back and forth)}$$

$\left(1-b\right)$ is the critical quantity for convergence

accelerated gradient descent
1. momentum add (heavy ball)
2. Nesterov formula

$\Leftrightarrow$

stochastic gradient descent
mini-batch of samples each step
not one nor millions

Descent with Momentum

* momentum: means of dampening oscillations and speeding up the iteration leading to faster convergence, it allows a large range of step size

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla f\left(\mathbf{x}_k\right) \xrightarrow{\text{descent with momentum}} \begin{cases} \mathbf{x}_{k+1} = \mathbf{x}_k - s\mathbf{z}_k \\ \mathbf{z}_k = \nabla f\left(\mathbf{x}_k\right) + \beta \underbrace{\mathbf{z}_{k-1}}_{\substack{\text{memory} \\ \text{momentum} \\ \text{heavy ball}}} \end{cases} \rightarrow \begin{cases} \mathbf{x}_{k+1} = \mathbf{x}_k - s_k\mathbf{z}_k \\ \mathbf{z}_{k+1} = \nabla f\left(\mathbf{x}_{k+1}\right) + \beta\mathbf{z}_k \end{cases}$$

$$\frac{d^2y}{dt^2} + b\frac{dy}{dt} + ky = 0 \rightarrow \begin{cases} \frac{d}{dt}(y) = 0y + 1\frac{dy}{dt} \\ \frac{d}{dt}\left(\frac{dy}{dt}\right) = -ky - b\frac{dy}{dt} \end{cases} \rightarrow \frac{d}{dt}\begin{bmatrix} y \\ \frac{dy}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -k & -b \end{bmatrix}\begin{bmatrix} y \\ \frac{dy}{dt} \end{bmatrix}$$

Descent with Momentum: Quadratic Model

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{S}\mathbf{x} \rightarrow \nabla f = \mathbf{S}\mathbf{x} \rightarrow \begin{pmatrix} \mathbf{S}\mathbf{q} = \lambda\mathbf{q} \\ \mathbf{x}_k = c_k\mathbf{q} \\ \mathbf{z}_k = d_k\mathbf{q} \end{pmatrix} \rightarrow \nabla f(\mathbf{x}_k) = \mathbf{S}\mathbf{x}_k = \lambda\mathbf{x}_k = \lambda c_k\mathbf{q}$$

$$\left.\begin{aligned} c_{k+1} &= c_k - sd_k \\ d_{k+1} &= \lambda c_{k+1} + \beta d_k \end{aligned}\right\} \rightarrow \begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix}\begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix}\begin{bmatrix} c_k \\ d_k \end{bmatrix} \rightarrow \begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & -s \\ \lambda & \beta - \lambda s \end{bmatrix}\begin{bmatrix} c_k \\ d_k \end{bmatrix} = \mathbf{R}\begin{bmatrix} c_k \\ d_k \end{bmatrix}$$

both eigenvalues e_1 and e_2 of $\mathbf{R}$ to be as small as possible

choose s and β to minimize $\max\left[\left|e_1(\lambda)\right|, \left|e_2(\lambda)\right|\right]$ for $\lambda_{\min}(\mathbf{S}) \le \lambda \le \lambda_{\max}(\mathbf{S})$

$$\rightarrow s = \left(\frac{2}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}}\right)^2, \beta = \left(\frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}}\right)^2$$

accelerated descent factor: $\left(\frac{1-b}{1+b}\right)^2 \rightarrow \left(\frac{1-\sqrt{b}}{1+\sqrt{b}}\right)^2$ when $b = \frac{1}{100}$, $\left(\frac{0.99}{1.01}\right)^2 = 0.96 \rightarrow \left(\frac{0.9}{1.1}\right)^2 = 0.67$

Nesterov Acceleration

evaluating the gradient ∇f at $\mathbf{x}_k \rightarrow \mathbf{x}_k + \gamma_k \left(\mathbf{x}_k - \mathbf{x}_{k-1} \right)$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla f\left(\mathbf{x}_k\right) \xrightarrow[\text{Nesterov}]{\gamma_k = \beta} \begin{cases} \mathbf{x}_{k+1} = \mathbf{y}_k - s\nabla f\left(\mathbf{y}_k\right) \\ \mathbf{y}_k = \mathbf{x}_k + \beta\left(\mathbf{x}_k - \mathbf{x}_{k-1}\right) \end{cases} \rightarrow \begin{cases} \mathbf{x}_{k+1} = \mathbf{y}_k - s\nabla f\left(\mathbf{y}_k\right) \\ \mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \beta\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) \end{cases}$$

$$f\left(\mathbf{x}\right) = \frac{1}{2}\mathbf{x}^T\mathbf{S}\mathbf{x} \rightarrow \nabla f = \mathbf{S}\mathbf{x} \rightarrow \begin{pmatrix} \mathbf{S}\mathbf{q} = \lambda\mathbf{q} \\ \mathbf{x}_k = c_k\mathbf{q} \\ \mathbf{y}_k = d_k\mathbf{q} \end{pmatrix} \rightarrow \nabla f\left(\mathbf{y}_k\right) = \mathbf{S}\mathbf{y}_k = \lambda\mathbf{y}_k = \lambda d_k\mathbf{q}$$

$$\left.\begin{aligned} c_{k+1} &= d_k - s\lambda d_k \\ d_{k+1} &= c_{k+1} + \beta\left(c_{k+1} - c_k\right) \end{aligned}\right\} \rightarrow \begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 0 & 1-s\lambda \\ -\beta & \left(1+\beta\right)\left(1-s\lambda\right) \end{bmatrix} \begin{bmatrix} c_k \\ d_k \end{bmatrix} = \mathbf{R}\begin{bmatrix} c_k \\ d_k \end{bmatrix}$$

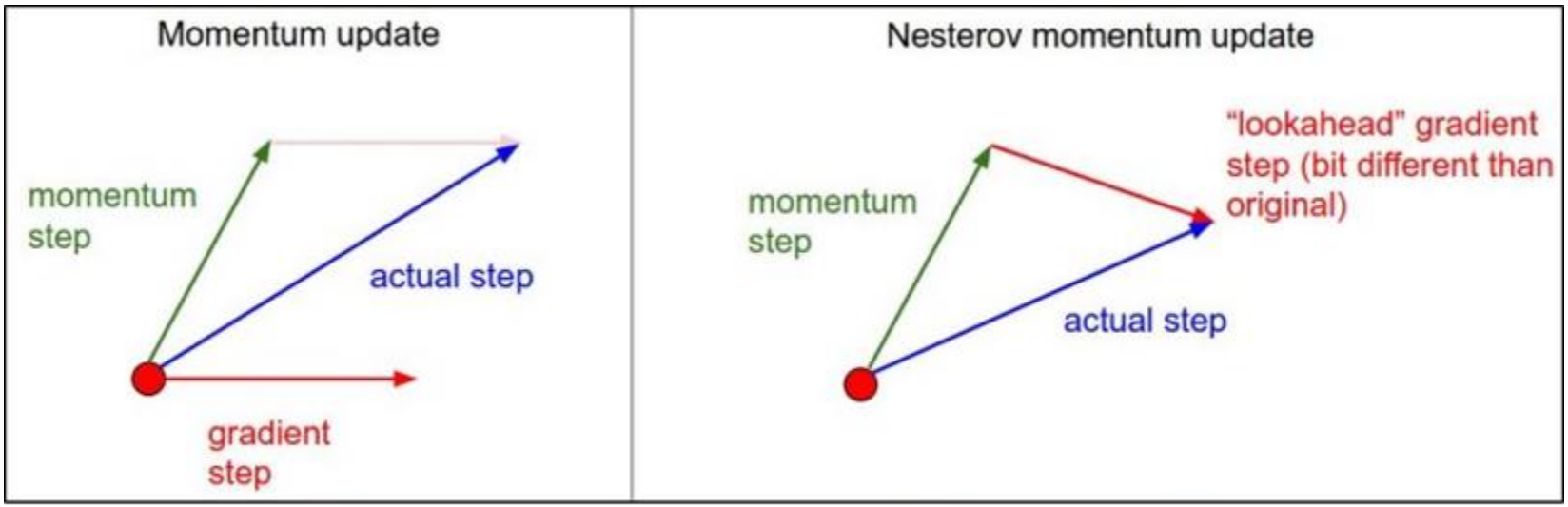
both eigenvalues e_1 and e_2 of $\mathbf{R}$ to be as small as possible

choose s and β to minimize $\max\left[\left|e_1\left(\lambda\right)\right|, \left|e_2\left(\lambda\right)\right|\right]$ for $\lambda_{\min}\left(\mathbf{S}\right) \le \lambda \le \lambda_{\max}\left(\mathbf{S}\right)$

$$\rightarrow s = \frac{1}{\lambda_{\max}}, \beta = \frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}} \rightarrow \max\left(\left|e_1\left(\lambda\right)\right|, \left|e_2\left(\lambda\right)\right|\right) = \frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}}$$

$\left(\text{example}\right)$ convergence factor: $1 - \sqrt{b}$

Momentum vs. Nesterov



Difference between Momentum and NAG. Picture from CS231.