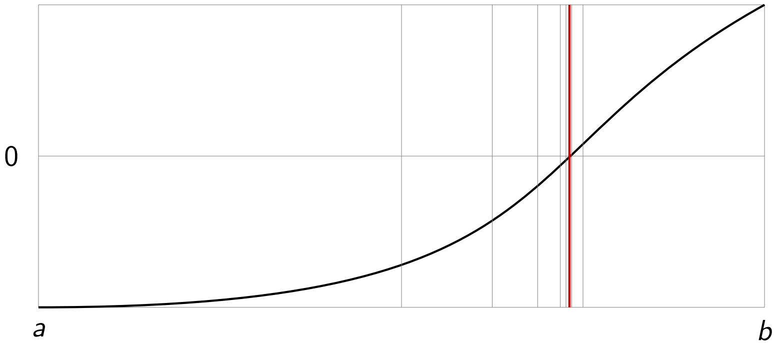
Iterative Algorithms in Optimization, Variational Analysis and Fixed Point Theory

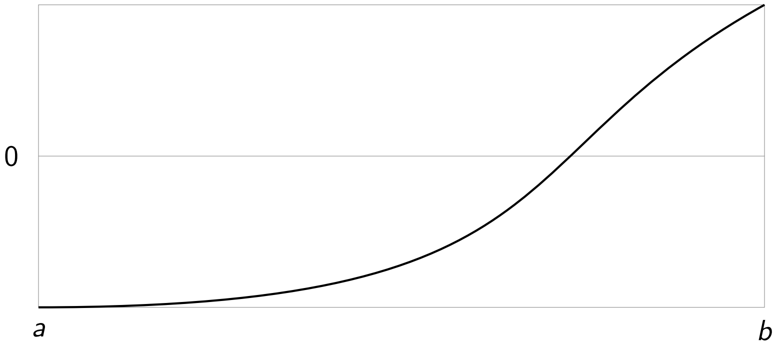
Unit 03: Second order in space and time.

university of groningen

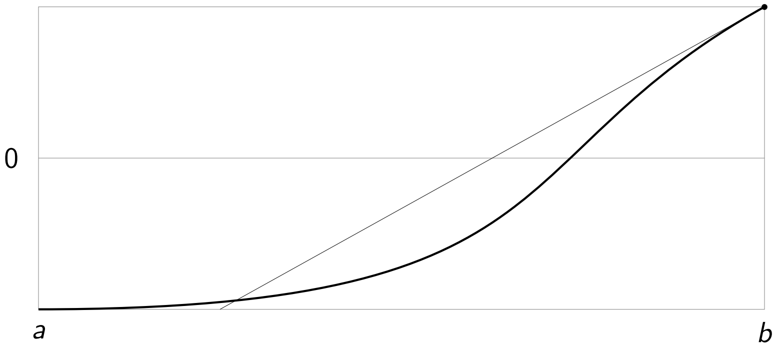
Recall the bisection method to solve $g(x) = 0$



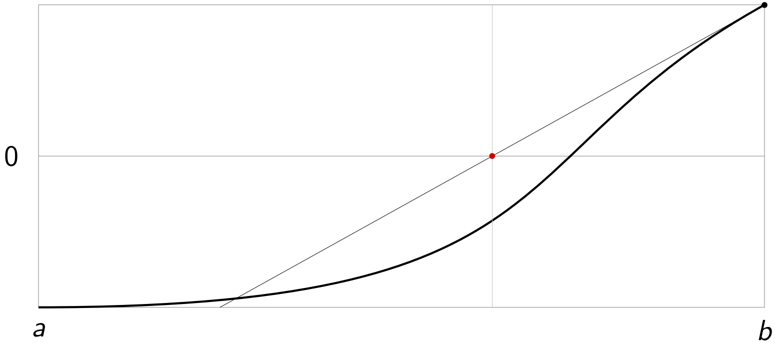
Let us use information about the function



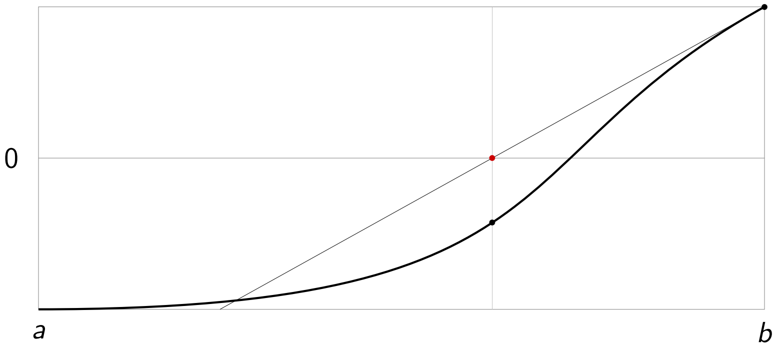
Let us use information about the function



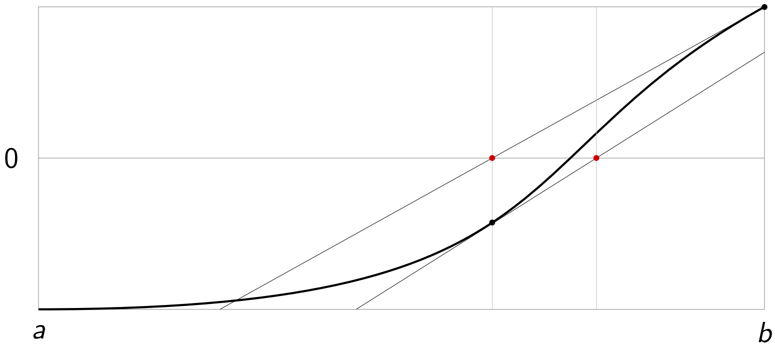
Let us use information about the function



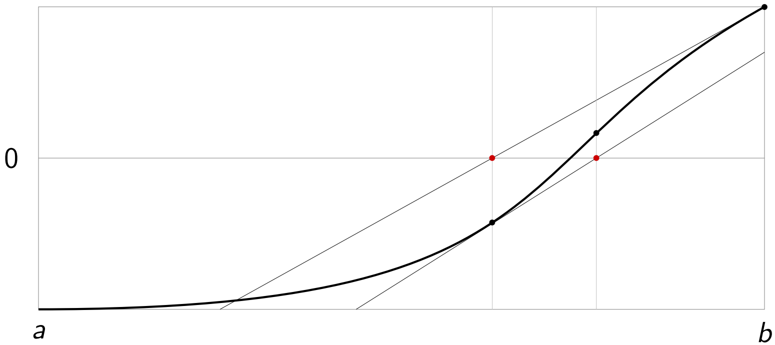
Let us use information about the function



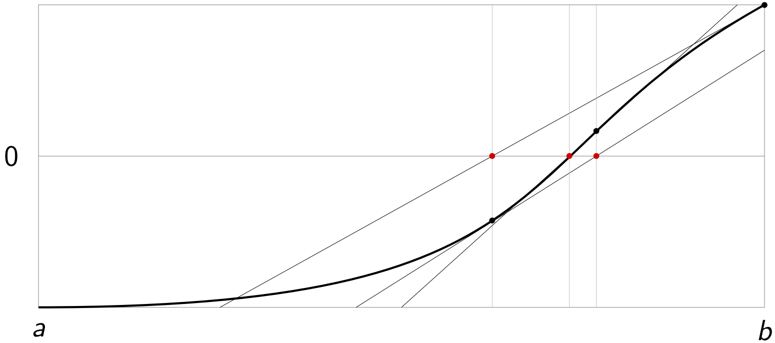
Let us use information about the function



Let us use information about the function



Let us use information about the function



Let us formalize the procedure

$g : \mathbb{R} \to \mathbb{R}$ differentiable

- Start with x_0 and compute $g(x_0)$ and $g'(x_0)$.

Let us formalize the procedure

$g : \mathbb{R} \to \mathbb{R}$ differentiable

- Start with x_0 and compute $g(x_0)$ and $g'(x_0)$.
- Draw the line $y = g(x_0) + g'(x_0)(x - x_0)$.

Let us formalize the procedure

$g : \mathbb{R} \to \mathbb{R}$ differentiable

- Start with x_0 and compute $g(x_0)$ and $g'(x_0)$.
- Draw the line $y = g(x_0) + g'(x_0)(x - x_0)$.
- Intersect with $y = 0$ to find x_1: $g'(x_0)(x_1 - x_0) = -g(x_0)$.

Let us formalize the procedure

$g : \mathbb{R} \to \mathbb{R}$ differentiable

- Start with x_0 and compute $g(x_0)$ and $g'(x_0)$.
- Draw the line $y = g(x_0) + g'(x_0)(x - x_0)$.
- Intersect with $y = 0$ to find x_1: $g'(x_0)(x_1 - x_0) = -g(x_0)$.
- If $g'(x_0) \neq 0$, then $x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}$.

Let us formalize the procedure

$g : \mathbb{R} \to \mathbb{R}$ differentiable

- Start with x_0 and compute $g(x_0)$ and $g'(x_0)$.
- Draw the line $y = g(x_0) + g'(x_0)(x - x_0)$.
- Intersect with $y = 0$ to find x_1: $g'(x_0)(x_1 - x_0) = -g(x_0)$.
- If $g'(x_0) \neq 0$, then $x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}$.
- Repeat the procedure to find $x_2, x_3 \ldots$ by

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

if $g'(x_k) \neq 0$.

Newton's Method

Let $G : \mathbb{R}^N \to \mathbb{R}^N$, and consider the system

$$G(x) = 0,$$

which has N equations and N unknowns.

Newton's Method

Let $G : \mathbb{R}^N \to \mathbb{R}^N$, and consider the system

$$G(x) = 0,$$

which has N equations and N unknowns.

If G is differentiable, and $DG(x)$ is invertible for all $x \in \mathbb{R}^N$ (or in a convenient subset) we can iterate

$$x_{k+1} = x_k - \big[DG(x_k)\big]^{-1} G(x_k).$$

Newton's Method

Let $G : \mathbb{R}^N \to \mathbb{R}^N$, and consider the system

$$G(x) = 0,$$

which has N equations and N unknowns.

If G is differentiable, and $DG(x)$ is invertible for all $x \in \mathbb{R}^N$ (or in a convenient subset) we can iterate

$$x_{k+1} = x_k - \left[DG(x_k)\right]^{-1} G(x_k).$$

We expect x_k to converge to some $\hat{x}$ such that $G(\hat{x}) = 0$.

Newton's Method in optimization

If $G = \nabla f$, this is

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k),$$

and we expect $\hat{x}$ to be a critical point of f.

Newton's Method in optimization

If $G = \nabla f$, this is

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k),$$

and we expect $\hat{x}$ to be a critical point of f.

Example

If $f(x) = \frac{1}{2}\|Ax - b\|^2$, then $\nabla f(x) = A^T(Ax - b)$ and $\nabla^2 f(x) = A^T A$.

Newton's Method in optimization

If $G = \nabla f$, this is

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k),$$

and we expect $\hat{x}$ to be a critical point of f.

Example

If $f(x) = \frac{1}{2}\|Ax - b\|^2$, then $\nabla f(x) = A^T(Ax - b)$ and $\nabla^2 f(x) = A^T A$. Therefore,

$$x_1 = x_0 - \left[A^T A\right]^{-1} A^T (Ax_0 - b).$$

Newton's Method in optimization

If $G = \nabla f$, this is

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k),$$

and we expect $\hat{x}$ to be a critical point of f.

Example

If $f(x) = \frac{1}{2}\|Ax - b\|^2$, then $\nabla f(x) = A^T(Ax - b)$ and $\nabla^2 f(x) = A^T A$.
Therefore,

$$x_1 = x_0 - \left[A^T A\right]^{-1} A^T(Ax_0 - b).$$

Since the solution $\hat{x}$ satisfies $A^T(A\hat{x} - b) = 0$, we must have $x_1 = \hat{x}$.

Newton's Method in optimization

If $G = \nabla f$, this is

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k),$$

and we expect $\hat{x}$ to be a critical point of f.

Example

If $f(x) = \frac{1}{2}\|Ax - b\|^2$, then $\nabla f(x) = A^T(Ax - b)$ and $\nabla^2 f(x) = A^T A$. Therefore,

$$x_1 = x_0 - \left[A^T A\right]^{-1} A^T(Ax_0 - b).$$

Since the solution $\hat{x}$ satisfies $A^T(A\hat{x} - b) = 0$, we must have $x_1 = \hat{x}$. One iteration of Newton's Method is equivalent to solving the problem.

Newton's Method in optimization

$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$

- If f is $L-$smooth, $\|\nabla^2 f(x)d\| \leq L\|d\|$ for all $x, d \in \mathbb{R}^N$.

Newton's Method in optimization

$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$

- If f is $L-$smooth, $\|\nabla^2 f(x)d\| \leq L\|d\|$ for all $x, d \in \mathbb{R}^N$.
- If f is convex, then $\nabla^2 f(x)$ is positive-semidefinite: $d^T \nabla^2 f(x) d \geq 0$ for all $d \in \mathbb{R}^N$.

Newton's Method in optimization

$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$

- If f is $L-$smooth, $\|\nabla^2 f(x)d\| \leq L\|d\|$ for all $x, d \in \mathbb{R}^N$.
- If f is convex, then $\nabla^2 f(x)$ is positive-semidefinite: $d^T \nabla^2 f(x) d \geq 0$ for all $d \in \mathbb{R}^N$.
- If f is $\mu-$strongly convex, $\nabla f(x)$ is positive definite, with $d^T \nabla^2 f(x) d \geq \mu\|d\|^2$ for all $x, d \in \mathbb{R}^N$.

Newton's Method in optimization

$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$

- If f is $L-$smooth, $\|\nabla^2 f(x) d\| \leq L\|d\|$ for all $x, d \in \mathbb{R}^N$.
- If f is convex, then $\nabla^2 f(x)$ is positive-semidefinite: $d^T \nabla^2 f(x) d \geq 0$ for all $d \in \mathbb{R}^N$.
- If f is $\mu-$strongly convex, $\nabla f(x)$ is positive definite, with $d^T \nabla^2 f(x) d \geq \mu\|d\|^2$ for all $x, d \in \mathbb{R}^N$. In particular, $\nabla^2 f(x)$ is invertible and $\left\|\left[\nabla^2 f(x)\right]^{-1} h\right\| \leq \mu^{-1}\|h\|$ for all $x, h \in \mathbb{R}^N$.

Newton's Method in optimization

$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$

- If f is $L-$smooth, $\|\nabla^2 f(x)d\| \leq L\|d\|$ for all $x, d \in \mathbb{R}^N$.

- If f is convex, then $\nabla^2 f(x)$ is positive-semidefinite: $d^T \nabla^2 f(x) d \geq 0$ for all $d \in \mathbb{R}^N$.

- If f is $\mu-$strongly convex, $\nabla f(x)$ is positive definite, with $d^T \nabla^2 f(x) d \geq \mu\|d\|^2$ for all $x, d \in \mathbb{R}^N$. In particular, $\nabla^2 f(x)$ is invertible and $\left\|\left[\nabla^2 f(x)\right]^{-1} h\right\| \leq \mu^{-1}\|h\|$ for all $x, h \in \mathbb{R}^N$.

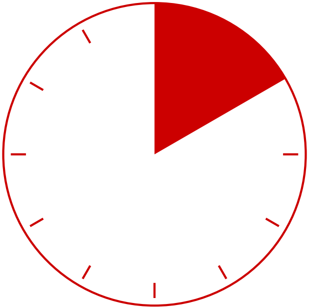
- If f is strongly convex, then $-\left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$ is a descent direction for f at x.

Newton's Method in optimization

$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$

- If f is $L-$smooth, $\|\nabla^2 f(x)d\| \leq L\|d\|$ for all $x, d \in \mathbb{R}^N$.
- If f is convex, then $\nabla^2 f(x)$ is positive-semidefinite: $d^T \nabla^2 f(x) d \geq 0$ for all $d \in \mathbb{R}^N$.
- If f is $\mu-$strongly convex, $\nabla f(x)$ is positive definite, with $d^T \nabla^2 f(x) d \geq \mu\|d\|^2$ for all $x, d \in \mathbb{R}^N$. In particular, $\nabla^2 f(x)$ is invertible and $\left\| \left[\nabla^2 f(x)\right]^{-1} h \right\| \leq \mu^{-1}\|h\|$ for all $x, h \in \mathbb{R}^N$.
- If f is strongly convex, then $-\left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$ is a descent direction for f at x. Which numbers are descent step sizes?

Break



Convergence of Newton's method

$x_{k+1} = x_k - [DG(x_k)]^{-1}\, G(x_k)$

Theorem

Consider $G : \mathbb{R}^N \to \mathbb{R}^N$ *and* $\hat{x} \in \mathbb{R}^N$ *such that* $G(\hat{x}) = 0$.

Suppose $\|DG(x) - DG(y)\| \le L\|x - y\|$ *and* $\|DG(x)^{-1}\| \le M$ *for all* x, y *in a neighborhood of* $\hat{x}$.

Then, there is $\delta > 0$ *such that if* $\|x_0 - \hat{x}\| < \delta$, *then*

$$\|x_{k+1} - \hat{x}\| \le \frac{LM}{2}\|x_k - \hat{x}\|^2$$

for all $k \ge 0$, *and so*

$$\|x_k - \hat{x}\| \le Cr^{2^k}$$

for some $C > 0$, $r \in (0,1)$ *and all* $k \ge 0$.

Avoid the cost of $\nabla^2 f(x_k)^{-1}$

Define $x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k)$, where

$$D_k \sim \nabla^2 f(x_k)^{-1}$$

in some sense, and is not as costly to compute.

Avoid the cost of $\nabla^2 f(x_k)^{-1}$

Define $x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k)$, where

$$D_k \sim \nabla^2 f(x_k)^{-1}$$

in some sense, and is not as costly to compute.

One simple heuristic is periodic evaluation: Choose $p \in \mathbb{N}$ and define

$$D_k = \nabla^2 f(x_{jp})^{-1} \qquad \text{for} \quad k = jp, jp+1, \ldots, (j+1)p - 1.$$

Avoid the cost of $\nabla^2 f(x_k)^{-1}$

Define $x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k)$, where

$$D_k \sim \nabla^2 f(x_k)^{-1}$$

in some sense, and is not as costly to compute.

One simple heuristic is periodic evaluation: Choose $p \in \mathbb{N}$ and define

$$D_k = \nabla^2 f(x_{jp})^{-1} \qquad \text{for} \quad k = jp, jp+1, \ldots, (j+1)p - 1.$$

Another idea is to update D_k to D_{k+1} keeping the essence of Newton's method, but using first order information only.

Quasi-Newton methods

Newton's method is based on the fact that

$$\nabla f(x_{k+1}) \sim \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k),$$

which implies

$$\nabla^2 f(x_k)^{-1}\left[\nabla f(x_{k+1}) - \nabla f(x_k)\right] \sim x_{k+1} - x_k.$$

Quasi-Newton methods

Newton's method is based on the fact that

$$\nabla f(x_{k+1}) \sim \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k),$$

which implies

$$\nabla^2 f(x_k)^{-1}\left[\nabla f(x_{k+1}) - \nabla f(x_k)\right] \sim x_{k+1} - x_k.$$

Now, approximating backwards, $\nabla^2 f(x_{k+1})$ should also satisfy

$$\nabla^2 f(x_{k+1})^{-1}\left[\nabla f(x_{k+1}) - \nabla f(x_k)\right] \sim x_{k+1} - x_k.$$

Quasi-Newton methods

Newton's method is based on the fact that

$$\nabla f(x_{k+1}) \sim \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k),$$

which implies

$$\nabla^2 f(x_k)^{-1} \left[\nabla f(x_{k+1}) - \nabla f(x_k) \right] \sim x_{k+1} - x_k.$$

Now, approximating backwards, $\nabla^2 f(x_{k+1})$ should also satisfy

$$\nabla^2 f(x_{k+1})^{-1} \left[\nabla f(x_{k+1}) - \nabla f(x_k) \right] \sim x_{k+1} - x_k.$$

D_{k+1} will be symmetric, close to D_k, and satisfy the secant condition:

$$D_{k+1} \left[\nabla f(x_{k+1}) - \nabla f(x_k) \right] = x_{k+1} - x_k.$$

BFGS: Broyden, Fletcher, Goldfarb and Shanno (1970)

At iteration k, use x_k, $\nabla f(x_k)$ and D_k to compute

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k), \quad \text{with } \alpha_k \text{ given by line search.}$$

BFGS: Broyden, Fletcher, Goldfarb and Shanno (1970)

At iteration k, use x_k, $\nabla f(x_k)$ and D_k to compute

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k), \quad \text{with } \alpha_k \text{ given by line search.}$$

Then compute $\nabla f(x_{k+1})$, $g_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, $s_k = x_{k+1} - x_k$, and

$$D_{k+1} = D_k + \left(\frac{g_k^T s_k + g_k^T D_k g_k}{(g_k^T s_k)^2} \right) s_k s_k^T + \frac{1}{g_k^T s_k} \left(D_k g_k s_k^T + \left(D_k g_k s_k^T \right)^T \right),$$

which minimizes the Frobenius distance to D_k, subject to the constraints.

BFGS: Broyden, Fletcher, Goldfarb and Shanno (1970)

At iteration k, use x_k, $\nabla f(x_k)$ and D_k to compute

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k), \quad \text{with } \alpha_k \text{ given by line search.}$$

Then compute $\nabla f(x_{k+1})$, $g_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, $s_k = x_{k+1} - x_k$, and

$$D_{k+1} = D_k + \left(\frac{g_k^T s_k + g_k^T D_k g_k}{(g_k^T s_k)^2} \right) s_k s_k^T + \frac{1}{g_k^T s_k} \left(D_k g_k s_k^T + \left(D_k g_k s_k^T \right)^T \right),$$

which minimizes the Frobenius distance to D_k, subject to the constraints.

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be μ-strongly convex and L-smooth, and let D_0 be positive definite. Then, x_n converges to the minimizer of f.
It does so in at most N steps if f is quadratic.

Iterative Algorithms in Optimization, Variational Analysis and Fixed Point Theory

Unit 03: Second order in space and time.

university of groningen

Momentum, inertia, acceleration

$x_{n+1} = x_n - \alpha_n \nabla f(x_n)$ is equivalent to

$$-\frac{x_{n+1} - x_n}{\alpha_n} = \nabla f(x_n),$$

which is an approximation of the steepest descent evolution equation

$$-\dot{x}(t) = \nabla f\big(x(t)\big).$$

Momentum, inertia, acceleration

$x_{n+1} = x_n - \alpha_n \nabla f(x_n)$ is equivalent to

$$-\frac{x_{n+1} - x_n}{\alpha_n} = \nabla f(x_n),$$

which is an approximation of the steepest descent evolution equation

$$-\dot{x}(t) = \nabla f\big(x(t)\big).$$

Other dynamics are related to minimization of potentials. For example,

$$m\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f\big(x(t)\big) = 0.$$

Discretization

We discretize

$$m\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f\big(x(t)\big) = 0$$

to obtain

$$m\,\frac{x_{k+1} - 2x_k + x_{k-1}}{h_k^2} + \gamma_k\,\frac{x_k - x_{k-1}}{h_k} + \nabla f(y_k) = 0.$$

Discretization

We discretize

$$m\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f\big(x(t)\big) = 0$$

to obtain

$$m\,\frac{x_{k+1} - 2x_k + x_{k-1}}{h_k^2} + \gamma_k\,\frac{x_k - x_{k-1}}{h_k} + \nabla f(y_k) = 0.$$

Equivalently,

$$x_{k+1} = x_k + \beta_k\,(x_k - x_{k-1}) - \alpha_k\,\nabla f(y_k),$$

with $\alpha_k = \frac{h_k^2}{m}$ and $\beta_k = 1 - \frac{\gamma_k h_k}{m}$.

Two popular choices

Polyak's heavy ball (1964)

$$x_{k+1} = x_k + \beta_k (x_k - x_{k-1}) - \alpha_k \nabla f(x_k).$$

Two popular choices

Polyak's heavy ball (1964)

$$x_{k+1} = x_k + \beta_k\,(x_k - x_{k-1}) - \alpha_k\,\nabla f(x_k).$$

Nesterov's extrapolation (1983)

$$\left\{ \begin{array}{rcl} y_k & = & x_k + \beta_k\,(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - \alpha_k\,\nabla f(y_k). \end{array} \right.$$

Quadratic case:

$$f(x) = f^* + \frac{1}{2}(x - x^*)^T H(x - x^*).$$

Remark

f is L-smooth and μ-strongly convex if, and only if, $\sigma(H) \subset [\mu, L]$.

The Heavy Ball method

$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(x_k)$

Quadratic case:

$$f(x) = f^* + \frac{1}{2}(x - x^*)^T H(x - x^*).$$

Remark

f is L-smooth and μ-strongly convex if, and only if, $\sigma(H) \subset [\mu, L]$.

The Heavy Ball method is equivalent to

$$\begin{pmatrix} x_{k+1} \\ x_k \end{pmatrix} = \begin{pmatrix} 1 + \beta - \alpha H & -\beta \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix}.$$

The convergence rate depends on the eigenvalues of that matrix

$$\begin{pmatrix} 1+\beta-\alpha\lambda & -\beta \\ 1 & 0 \end{pmatrix}, \qquad \text{for} \qquad \lambda \in [\mu, L].$$

Proposition

If $\alpha L \geq 2(1+\beta)$, there is a matrix H, with $\sigma(H) \subset [\mu, L]$, for which the Heavy Ball does not converge for any initial condition.

The Heavy Ball method

$x_{k+1} = x_k + \beta\,(x_k - x_{k-1}) - \alpha\nabla f(x_k)$

The convergence rate depends on the eigenvalues of that matrix

$$\begin{pmatrix} 1+\beta-\alpha\lambda & -\beta \\ 1 & 0 \end{pmatrix}, \qquad \text{for} \qquad \lambda \in [\mu, L].$$

Proposition

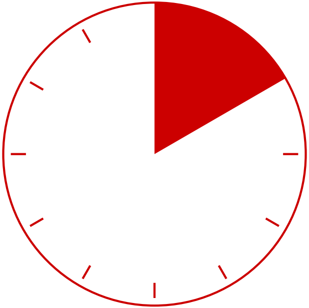
If $\alpha L \geq 2(1+\beta)$, there is a matrix H, with $\sigma(H) \subset [\mu, L]$, for which the Heavy Ball does not converge for any initial condition.

Proposition

The best uniform convergence rate is

$$\rho^* = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}, \quad \textit{obtained when} \quad \beta^* = (\rho^*)^2 \textit{ and } \alpha^* = \frac{2(1+\beta^*)}{L+\mu}.$$

Break



Nesterov's extrapolation

Standard formulation

$$\begin{cases} y_k &= x_k + \beta_k (x_k - x_{k-1}) \\ x_{k+1} &= y_k - \alpha_k \nabla f(y_k). \end{cases}$$

Nesterov's extrapolation

Standard formulation

$$\left\{\begin{array}{rcl} y_k & = & x_k + \beta_k\,(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - \alpha_k\,\nabla f(y_k). \end{array}\right.$$

Alternative description

$$y_{k+1} = y_k - \alpha_k \nabla f(y_k) + \beta_k(y_k - y_{k-1}) - \alpha_k\beta_k\big(\nabla f(y_k) - \nabla f(y_{k-1})\big).$$

Nesterov's extrapolation

Standard formulation

$$\left\{\begin{array}{rcl} y_k & = & x_k + \beta_k\,(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - \alpha_k\,\nabla f(y_k). \end{array}\right.$$

Alternative description

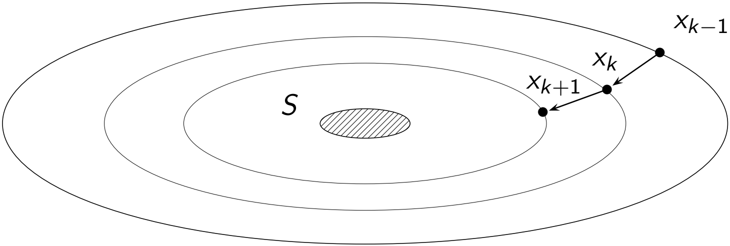
$$y_{k+1} = y_k - \alpha_k \nabla f(y_k) + \beta_k(y_k - y_{k-1}) - \alpha_k\beta_k\big(\nabla f(y_k) - \nabla f(y_{k-1})\big).$$

High-resolution differential equation

$$\ddot{y}(t) + A(t)\dot{x}(t) + B(t)\nabla^2 f\big(y(t)\big)\dot{y}(t) + C(t)\nabla f\big(y(t)\big) = 0.$$

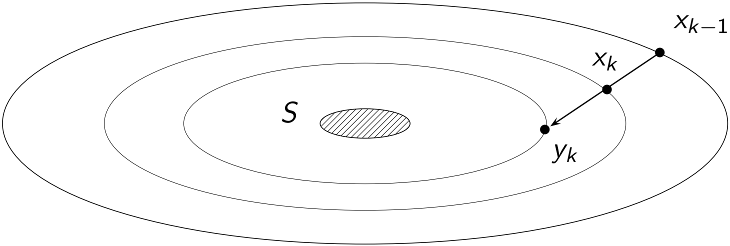
Nesterov's extrapolation

The main idea is the following: Instead of doing this



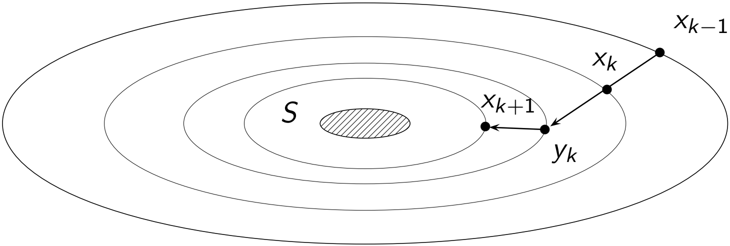
Nesterov's extrapolation

Better try this



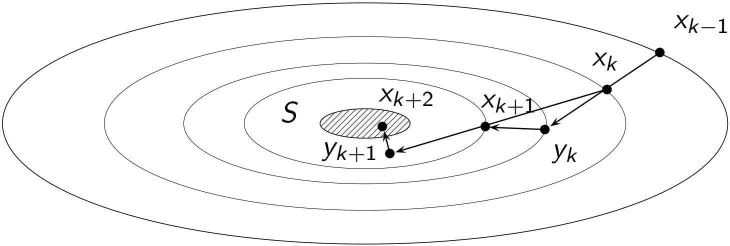
Nesterov's extrapolation

Better try this



Nesterov's extrapolation

Better try this



Convergence of Nesterov's method

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be an L-smooth convex function with minimizers, and let (x_k, y_k) be generated by Nesterov's method with convenient α_k, β_k.

Convergence of Nesterov's method

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be an L-smooth convex function with minimizers, and let (x_k, y_k) be generated by Nesterov's method with convenient α_k, β_k.

- *Then,* $f(x_k) - \min(f) \leq \dfrac{L\,\mathrm{dist}(x_0, S)^2}{(k+1)^2}$ *for all* $k \geq 1$. *In addition,* $\lim_{k\to\infty} k^2\big(f(x_k) - \min(f)\big) = 0.$

$\alpha_k \equiv \frac{1}{L}, \quad \beta_k \sim 1 - \frac{r}{k}$

Convergence of Nesterov's method

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be an L-smooth convex function with minimizers, and let (x_k, y_k) be generated by Nesterov's method with convenient α_k, β_k.

- *Then,* $f(x_k) - \min(f) \leq \dfrac{L\,\mathrm{dist}(x_0, S)^2}{(k+1)^2}$ *for all* $k \geq 1$. *In addition,*
  $\lim\limits_{k\to\infty} k^2\big(f(x_k) - \min(f)\big) = 0.$

$$\alpha_k \equiv \frac{1}{L}, \quad \beta_k \sim 1 - \frac{r}{k}$$

- *If, moreover,* f *is* μ*-strongly convex, then*
  $$f(x_k) - \min(f) \leq L\,\mathrm{dist}(x_0, S)^2 \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \text{ for all } k \geq 1.$$

$$\alpha_k \equiv \frac{1}{L}, \quad \beta_k \equiv \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$$