Introduction to Optimization

Improving convergence rates.

Lecture 07: Quadratic functions and finite termination.

Lecture 08: Inertial algorithms. Stochastic gradient.

university of groningen

Systems of linear equations

Let $\mathcal{A} \in \mathbb{R}^{N\times N}$ be invertible (to simplify), and let $\beta \in \mathbb{R}^N$. The problem

$$\mathcal{A}x = \beta$$

has a unique solution

Systems of linear equations

Let $\mathcal{A} \in \mathbb{R}^{N \times N}$ be invertible (to simplify), and let $\beta \in \mathbb{R}^N$. The problem

$$\mathcal{A}x = \beta$$

has a unique solution, which is the unique minimizer of

$$f(x) = \frac{1}{2}\|\mathcal{A}x - \beta\|^2$$

Systems of linear equations

Let $\mathcal{A} \in \mathbb{R}^{N\times N}$ be invertible (to simplify), and let $\beta \in \mathbb{R}^N$. The problem

$$\mathcal{A}x = \beta$$

has a unique solution, which is the unique minimizer of

$$f(x) = \frac{1}{2}\|\mathcal{A}x - \beta\|^2,$$

and also the unique minimizer of

$$\phi(x) = \frac{1}{2}x \cdot Ax - b \cdot x,$$

where $A = \mathcal{A}^T\mathcal{A}$ (symmetric and positive definite) and $b = \mathcal{A}^T\beta$.

Conjugate gradient methods

Conjugate gradient methods iterate

$$x_{n+1} = x_n + \alpha_n d_n$$

for convenient choices of d_n and α_n in such a way that

$$\{d_0, \ldots, d_{N-1}\}$$

is a basis of $\mathbb{R}^N$ and x_n minimizes ϕ on the affine subspace

$$x_0 + \mathrm{span}\{d_0, \ldots, d_{n-1}\}.$$

As a consequence, the solution is found in at most N steps for every initial point x_0.

Conjugate directions

Vectors $\{d_0, \dots, d_{N-1}\}$ are conjugate with respect to A if

$$d_i \cdot Ad_j = 0$$

for all i, j.

Conjugate directions

Vectors $\{d_0, \ldots, d_{N-1}\}$ are conjugate with respect to A if

$$d_i \cdot A d_j = 0$$

for all i, j.

Remark

The expression

$$\langle x, y \rangle_A := x \cdot Ay$$

is an inner product in $\mathbb{R}^N$.

Conjugate directions

Vectors $\{d_0, \ldots, d_{N-1}\}$ are conjugate with respect to A if

$$d_i \cdot Ad_j = 0$$

for all i, j.

Remark

The expression

$$\langle x, y \rangle_A := x \cdot Ay$$

is an inner product in $\mathbb{R}^N$. Being conjugate means being orthogonal with respect to this inner product.

Conjugate directions

Vectors $\{d_0, \ldots, d_{N-1}\}$ are conjugate with respect to A if

$$d_i \cdot Ad_j = 0$$

for all i, j.

Remark

The expression

$$\langle x, y \rangle_A := x \cdot Ay$$

is an inner product in $\mathbb{R}^N$. Being conjugate means being orthogonal with respect to this inner product. Conjugate vectors are linearly independent.

An abstract conjugate gradient method

Theorem

Let $\{d_0, \ldots, d_{N-1}\}$ *be conjugate, and let* $x_{n+1} = x_n + \alpha_n d_n$, *where*

$$\alpha_n = \operatorname{Argmin}_{\alpha>0} \phi(x_n + \alpha d_n) = -\frac{d_n \cdot \nabla \phi(x_n)}{\|d_n\|_A^2} = -\frac{d_n \cdot (Ax_n - b)}{\|d_n\|_A^2},$$

for $n = 0, \ldots, N-1$. *Then* x_N *minimizes* ϕ, *whence* $Ax_N = b$.

An abstract conjugate gradient method

Theorem

Let $\{d_0, \ldots, d_{N-1}\}$ be conjugate, and let $x_{n+1} = x_n + \alpha_n d_n$, where

$$\alpha_n = \operatorname{Argmin}_{\alpha>0} \phi(x_n + \alpha d_n) = -\frac{d_n \cdot \nabla\phi(x_n)}{\|d_n\|_A^2} = -\frac{d_n \cdot (Ax_n - b)}{\|d_n\|_A^2},$$

for $n = 0, \ldots, N-1$. Then x_N minimizes ϕ, whence $Ax_N = b$.

Moreover, $\nabla\phi(x_n) \cdot d_j = 0$ for $j = 0, \ldots, n-1$, and x_n minimizes ϕ over

$$x_0 + \operatorname{span}\{d_0, \ldots, d_{n-1}\}.$$

This is known as expanding subspace minimization.

Implementation of a conjugate gradient method

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla\phi(x_0) = b - Ax_0$.

Implementation of a conjugate gradient method

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla\phi(x_0) = b - Ax_0$.

At step n, we know x_n, $g_n = \nabla\phi(x_n)$ and d_n.

Implementation of a conjugate gradient method

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla\phi(x_0) = b - Ax_0$.

At step n, we know x_n, $g_n = \nabla\phi(x_n)$ and d_n.

If $g_n = 0$, we stop; otherwise, we compute $\alpha_n = \dfrac{\|g_n\|^2}{\|d_n\|_A^2}$, and then update

$$\begin{array}{rcl} x_{n+1} & = & x_n + \alpha_n d_n \\ g_{n+1} & = & g_n + \alpha_n A d_n \\ d_{n+1} & = & -g_{n+1} + \beta_n d_n, \quad \text{with} \quad \dfrac{\|g_{n+1}\|^2}{\|g_n\|^2}. \end{array}$$

Convergence

Theorem

The procedure described above produces a conjugate set $\{d_0, \ldots, d_n\}$, *with* $n \leq N$, *and terminates at* x_n, *where* $Ax_n = b$.

Convergence

Theorem

The procedure described above produces a conjugate set $\{d_0, \ldots, d_n\}$*, with* $n \leq N$*, and terminates at* x_n*, where* $Ax_n = b$*.*

Remark

If A *has exactly* k *distinct eigenvalues, the algorithm terminates in at most* k *steps.*

Nonlinear extensions

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.

Nonlinear extensions

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla f(x_0)$.

Nonlinear extensions

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla f(x_0)$.

At step n, we know x_n, $g_n = \nabla f(x_n)$ and d_n.

Nonlinear extensions

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla f(x_0)$.

At step n, we know x_n, $g_n = \nabla f(x_n)$ and d_n.

If $g_n = 0$, we stop; otherwise, we compute (backtracking) α_n satisfying

$$\begin{aligned} f(x_n + \alpha_n d_n) &\leq f(x_n) + c_1 \alpha_n \, g_n \cdot d_n \\ |\nabla f(x_n + \alpha_n d_n) \cdot d_n| &\leq -c_2 \, g_n \cdot d_n, \end{aligned}$$

with $0 < c_1 < c_2 < 1/2$ (strong Wolfe conditions).

Nonlinear extensions, continued

Then, we update

$$\begin{aligned} x_{n+1} &= x_n + \alpha_n d_n \\ d_{n+1} &= -g_{n+1} + \beta_n d_n, \end{aligned}$$

where several choices for β_n are possible, such as:

- Fletcher-Reeves: $\frac{\|g_{n+1}\|^2}{\|g_n\|^2}$
- Polak-Ribière: $\frac{g_{n+1}\cdot(g_{n+1}-g_n)}{\|g_n\|^2}$
- Hestenes-Stiefel: $\frac{g_{n+1}\cdot(g_{n+1}-g_n)}{d_n\cdot(g_{n+1}-g_n)}$
- Dai-Yuan: $\frac{\|g_{n+1}\|^2}{d_n\cdot(g_{n+1}-g_n)}$

Break

Newton's method

Quadratic model of f given by Taylor's expansion at x_n:

$$f(x) \simeq f(x_n) + \nabla f(x_n) \cdot (x - x_n) + \frac{1}{2}(x - x_n) \cdot \nabla^2 f(x_n)(x - x_n).$$

Newton's method

Quadratic model of f given by Taylor's expansion at x_n:

$$f(x) \simeq f(x_n) + \nabla f(x_n) \cdot (x - x_n) + \frac{1}{2}(x - x_n) \cdot \nabla^2 f(x_n)(x - x_n).$$

Set x_{n+1} as the minimizer: $\nabla f(x_n) + \nabla^2 f(x_n)(x_{n+1} - x_n) = 0$

Newton's method

Quadratic model of f given by Taylor's expansion at x_n:

$$f(x) \simeq f(x_n) + \nabla f(x_n) \cdot (x - x_n) + \frac{1}{2}(x - x_n) \cdot \nabla^2 f(x_n)(x - x_n).$$

Set x_{n+1} as the minimizer: $\nabla f(x_n) + \nabla^2 f(x_n)(x_{n+1} - x_n) = 0$, or

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

Newton's method

Quadratic model of f given by Taylor's expansion at x_n:

$$f(x) \simeq f(x_n) + \nabla f(x_n) \cdot (x - x_n) + \frac{1}{2}(x - x_n) \cdot \nabla^2 f(x_n)(x - x_n).$$

Set x_{n+1} as the minimizer: $\nabla f(x_n) + \nabla^2 f(x_n)(x_{n+1} - x_n) = 0$, or

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

if $\nabla^2 f(x_n)$ is invertible.

Newton's method

Quadratic model of f given by Taylor's expansion at x_n:

$$f(x) \simeq f(x_n) + \nabla f(x_n) \cdot (x - x_n) + \frac{1}{2}(x - x_n) \cdot \nabla^2 f(x_n)(x - x_n).$$

Set x_{n+1} as the minimizer: $\nabla f(x_n) + \nabla^2 f(x_n)(x_{n+1} - x_n) = 0$, or

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

if $\nabla^2 f(x_n)$ is invertible.

Example: $f(x) = \frac{1}{2}\|Ax - b\|^2$.

Newton's method

$x_{n+1} = x_n - \nabla^2 f(x_n)^{-1} \nabla f(x_n)$

Theorem

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

Suppose $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$ *and* $\|\nabla^2 f(x)^{-1}\| \leq M$ *for all* x, y *in a neighborhood of* $\hat{x}$.

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

$$\|x_n - \hat{x}\| \leq r^{2^n}$$

for some $r \in (0, 1)$ *and all* $n \geq 0$.

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

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

$$D_n \sim \nabla^2 f(x_n)^{-1}$$

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

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

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

$$D_n \sim \nabla^2 f(x_n)^{-1}$$

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

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

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

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

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

$$D_n \sim \nabla^2 f(x_n)^{-1}$$

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

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

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

Another idea is to define D_{n+1} as a function of D_n (not $\nabla^2 f(x_n)$ or $\nabla^2 f(x_{n+1})$) while keeping the essence of Newton's method.

Quasi-Newton methods

The key equality defining Newton's method is the secant condition:

$$\nabla^2 f(x_n)^{-1}\left[\nabla f(x_{n+1}) - \nabla f(x_n)\right] = x_{n+1} - x_n.$$

Quasi-Newton methods

The key equality defining Newton's method is the secant condition:

$$\nabla^2 f(x_n)^{-1}\left[\nabla f(x_{n+1}) - \nabla f(x_n)\right] = x_{n+1} - x_n.$$

Let us construct D_{n+1} so that it is symmetric (like the Hessian)

Quasi-Newton methods

The key equality defining Newton's method is the secant condition:

$$\nabla^2 f(x_n)^{-1}\left[\nabla f(x_{n+1}) - \nabla f(x_n)\right] = x_{n+1} - x_n.$$

Let us construct D_{n+1} so that it is symmetric (like the Hessian), no too different from D_n

Quasi-Newton methods

The key equality defining Newton's method is the secant condition:

$$\nabla^2 f(x_n)^{-1}\left[\nabla f(x_{n+1}) - \nabla f(x_n)\right] = x_{n+1} - x_n.$$

Let us construct D_{n+1} so that it is symmetric (like the Hessian), no too different from D_n, and satisfies

$$D_{n+1} g_n = s_n,$$

where $g_n = \nabla f(x_{n+1}) - \nabla f(x_n)$ and $s_n = x_{n+1} - x_n$.

Popular instances

- DFP: Davidon (1959), Fletcher and Powell (1987)

$$D_{n+1} = D_n - \frac{(D_n g_n)(D_n g_n)^T}{g_n \cdot D_n g_n} + \rho_n (s_n s_n^T), \qquad \rho_n = \frac{1}{g_n \cdot s_n}.$$

Popular instances

- DFP: Davidon (1959), Fletcher and Powell (1987)

$$D_{n+1} = D_n - \frac{(D_n g_n)(D_n g_n)^T}{g_n \cdot D_n g_n} + \rho_n (s_n s_n^T), \qquad \rho_n = \frac{1}{g_n \cdot s_n}.$$

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

$$D_{n+1} = (I - \rho_n s_n g_n^T) D_n (I - \rho_n s_n g_n^T)^T + \rho_n (s_n s_n^T).$$

Popular instances

- DFP: Davidon (1959), Fletcher and Powell (1987)

$$D_{n+1} = D_n - \frac{(D_n g_n)(D_n g_n)^T}{g_n \cdot D_n g_n} + \rho_n (s_n s_n^T), \qquad \rho_n = \frac{1}{g_n \cdot s_n}.$$

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

$$D_{n+1} = (I - \rho_n s_n g_n^T) D_n (I - \rho_n s_n g_n^T)^T + \rho_n (s_n s_n^T).$$

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.