# Optimization Methods I
# First and Second Order Methods

Anders Hansson and Martin Andersen

Linköping University and Technical Unversity of Denmark

March 26, 2024

# Gradient Descent Method

Consider

$$\text{minimize} \quad f(x) \tag{1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable, *i.e.*, gradient of $f(x)$

$$\nabla f(x) = \begin{bmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{bmatrix}$$

exists and is a continuous function.

Assume $f$ bounded below and that minimum attained. Given initial point $x_0$, *gradient descent method* updates

$$x_{k+1} = x_k - t_k \nabla f(x_k) \tag{2}$$

where $t_k > 0$ is the step size.

# $\beta$-smoothness

The function $f$ is said to be $\beta$-smooth if $\exists\ \beta > 0$:

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq \beta \|y - x\|_2, \quad \forall x, y \in \mathbf{R}^n. \tag{3}$$

$$\Longrightarrow$$

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathbf{R}^n. \tag{4}$$

For twice differentiable $f$, $\beta$-smoothness implies $\|\nabla^2 f(x)\|_2 \leq \beta,\ \forall x$.

# Majorization Minimization Principle

Let $\bar{f}(\cdot \,|\, \cdot) : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ be given by

$$\bar{f}(y \,|\, x) = f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|y - x\|_2^2$$

Then $f(y) \leq \bar{f}(y \,|\, x)$ and $\bar{f}(x \,|\, x) = f(x)$ which implies that $\bar{f}$ majorizes $f$ at $x$.

The *majorization minimization algorithm* is

$$x_{k+1} = \operatorname*{argmin}_y \bar{f}(y \,|\, x_k), \quad k = 0, 1, 2, \ldots, \tag{5}$$

where $x_0 \in \mathbf{R}^n$ given.

Descent method, i.e. $f(x_{k+1}) \leq f(x_k)$.

# Interpretation as Gradient Descent

We have

$$\operatorname*{argmin}_y \bar{f}(y \,|\, x) = x - \frac{1}{\beta}\nabla f(x),$$

and hence (5) is gradient descent

$$x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k), \quad k = 0, 1, 2, \ldots \tag{6}$$

with step size $t_k = 1/\beta$.

# Convergence

Substituting $x_{k+1}$ for $y$ and $x_k$ for $x$ in (4), we obtain

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2$$

which shows that (6) always reduces the objective value unless $x_k$ is a stationary point.

If $f$ is bounded below

$$\lim_{k\to\infty} \|\nabla f(x_k)\|_2 = 0$$

If $f$ is convex, then convergence to global optimum.

# Convergence ctd.

If $f$ is differentiable, convex, and $\beta$-smooth, then

$$f(x^{(k)}) - p^\star = O(1/k)$$

where $p^\star$ is optimal value, i.e. *sublinear* convergence rate.

If $f$ is also $\alpha$-strongly convex, then

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbf{R}^n,$$

for some $\alpha > 0$. With constant $t_k = t$:

$$\|x_k - x^\star\|_2^2 \leq \left(1 - \frac{2t\alpha\beta}{\alpha + \beta}\right)^k \|x_0 - x^\star\|_2^2$$

where $0 < t \leq 2/(\alpha + \beta)$. Best upper bound for $t = 2/(\alpha + \beta)$:

$$1 - \frac{2t\alpha\beta}{\alpha + \beta} = \left(\frac{\beta - \alpha}{\beta + \alpha}\right)^2 = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2$$

where $\kappa = \beta/\alpha$, If $f$ in addition twice continuously differentiable, then the eigenvalues $\lambda\left(\nabla^2 f(x)\right) \in [\alpha, \beta]$.

# Proximal Gradient Method

Consider

$$\text{minimize} \quad g(x) + h(x) \tag{7}$$

where $g$ is continuously differentiable and $\beta$-smooth, and $h$ is proper, closed, and convex but not necessarily differentiable. Then

$$\bar{f}(y \mid x) = g(x) + \nabla g(x)^T (y - x) + \frac{\beta}{2} \|y - x\|_2^2 + h(y)$$

is a majorization of $f = g + h$ at $x$.

Majorization minimization algorithm results in *proximal gradient method*:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_y \bar{f}(y \mid x_k) \\ &= \operatorname*{argmin}_y \left\{ h(y) + \nabla g(x_k)^T y + \frac{\beta}{2} \|y - x_k\|_2^2 \right\} \\ &= \operatorname*{argmin}_y \left\{ h(y) + \frac{\beta}{2} \|y - (x_k - \frac{1}{\beta} \nabla g(x_k))\|_2^2 \right\}. \end{aligned}$$

# Proximal Operator

Equivalent formulation:

$$x_{k+1} = \mathbf{prox}_{\beta^{-1}h}\left(x_k - \beta^{-1}\nabla g(x_k)\right) \tag{8}$$

where

$$\mathbf{prox}_h(x) = \underset{y}{\operatorname{argmin}} \left\{ h(y) + \frac{1}{2}\|y - x\|_2^2 \right\} \tag{9}$$

is so-called *proximal operator* associated with the function $h$.

# Optimality Conditions

First-order necessary optimality condition for (7) implies that a stationary point $x^\star$ must satisfy

$$-\nabla g(x^\star) \in \partial h(x^\star) \tag{10}$$

where

$$\partial h(x) = \{v \in \mathbf{R}^n \mid h(y) \geq h(x) + v^T(y - x) \text{ for all } y \in \mathbf{R}^n\} \tag{11}$$

is the *subdifferential* of $h$.[1]
The stationarity condition for the majorization $\bar{f}(y \mid x^\star)$, *i.e.*,

$$-\nabla g(x^\star) - \beta(y - x^\star) \in \partial h(y),$$

agree with optimality conditions when $y = x^\star$.

Hence a fixed-point of the iteration (8) is a stationary point $x^\star$ of $f = g + h$.

---

[1] If $h$ is differentiable at $x$, then the subdifferential of $h$ at $x$ is the singleton $\{\nabla h(x)\}$.

# Convergence Rate

When both $g$ and $h$ are convex, the proximal gradient iteration satisfies the bound

$$f(x^{(k)}) - p^\star = O(1/k).$$

An overview of other variants of the proximal gradient method, including a detailed analysis of both the convex and the nonconvex case, can be found in Beck [2017].

# Constrained Optimization Problems

Consider

$$\begin{array}{ll} \text{minimize} & g(x) \\ \text{subject to} & x \in C \end{array} \tag{12}$$

where $C$ is a convex subset of $\mathbf{R}^n$.

Problem equivalent to (7) if we define $h(x) = I_C(x)$ where $I_C$ indicator function of set $C$ defined as

$$I_C(x) = \begin{cases} 0, & x \in C \\ \infty, & x \notin C. \end{cases}$$

In this special case, proximal operator $\mathbf{prox}_h(x)$ becomes Euclidean projection of $x$ onto $C$, and resulting method is referred to as the *projected gradient* or *gradient projection* method.

# Accelerated Proximal Gradient (APG) Method

Proximal gradient method can be combined with a momentum technique, Nesterov (1983), inspired by so-called *heavy ball* method by Polyak (1964).

Improved worst-case bound

$$f(x_k) - p^\star = O(1/k^2)$$

One such method is *accelerated proximal gradient* method:

$$x_{k+1} = \mathbf{prox}_{\beta^{-1}h}(y_k - \beta^{-1}\nabla g(y_k))$$

$$\gamma_{k+1} = \frac{1 + \sqrt{1 + 4\gamma_k^2}}{2}$$

$$y_{k+1} = x_k + \frac{\gamma_k - 1}{\gamma_{k+1}}(x_k - x_{k-1})$$

with initial values $\gamma_0 = 1$, $y_{-1} = 0$, and $y_0 = x_0$.

The APG method shown here is not a descent method.

# Second-Order Method

Consider minimize$f(x)$with additional assumption that $f$ is twice differentiable.

Seecond-order Taylor expansion:

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x. \qquad (13)$$

Derivative of right hand side with respect to $\Delta x$ equal to zero:

$$\nabla^2 f(x) \Delta x = -\nabla f(x). \qquad (14)$$

Compare with the first-order Taylor expansion of gradient:

$$\nabla f(x + \Delta x) \approx \nabla f(x) + \nabla^2 f(x) \Delta x$$

# Newton's Method

In the convex case Newton's method uses as search direction $\Delta x = -\nabla^2 f(x)^{-1} \nabla f(x)$ which leads to the iteration

$$x_{k+1} = x_k - t_k \nabla^2 f(x_k)^{-1} \nabla f(x_k), \quad k = 0, 1, 2, \ldots \tag{15}$$

where $x_0$ is a initial guess, and $t_k > 0$ is the step size at iteration $k$.

If $x_k$ is sufficiently close to stationary point of $\nabla f$, then full Newton step can be shown to yield descentm provided that $\nabla^2 f$ is Lipschitz continuous. However, this is not always the case if $x_k$ is far away from stationary point.

# Newton Decrement

Directional derivative of $f$ in Newton direction $\Delta x_{\mathrm{nt}}$:

$$\frac{d}{dt} f(x + t\Delta x_{\mathrm{nt}})\Big|_{t=0} = \nabla f(x)^T \Delta x_{\mathrm{nt}} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

is negative if $\nabla f(x) \succ 0$ and $\nabla f(x) \neq 0$.

Thus, descent is possible with a sufficiently small step size.

When Hessian is positive definite directional derivative equal to $-\lambda(x)^2$ where $\lambda(x)$ is so-called *Newton decrement* of $f$ at $x$:

$$\lambda(x) = \|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2} \quad (16)$$

# Stopping Criterion

Difference between $f(x)$ and the minimum of quadratic approximation $\tilde{f}$ may be expressed as

$$f(x) - \inf_y \tilde{f}(y; x) = \frac{1}{2}\lambda(x)^2, \tag{17}$$

which allows us to view $\lambda(x)^2/2$ as estimate of $f(x) - p^\star$.

Motivates stopping criterion $\lambda(x) \leq \epsilon$, where $\epsilon > 0$ is given tolerance.

# Backtracking Line Search

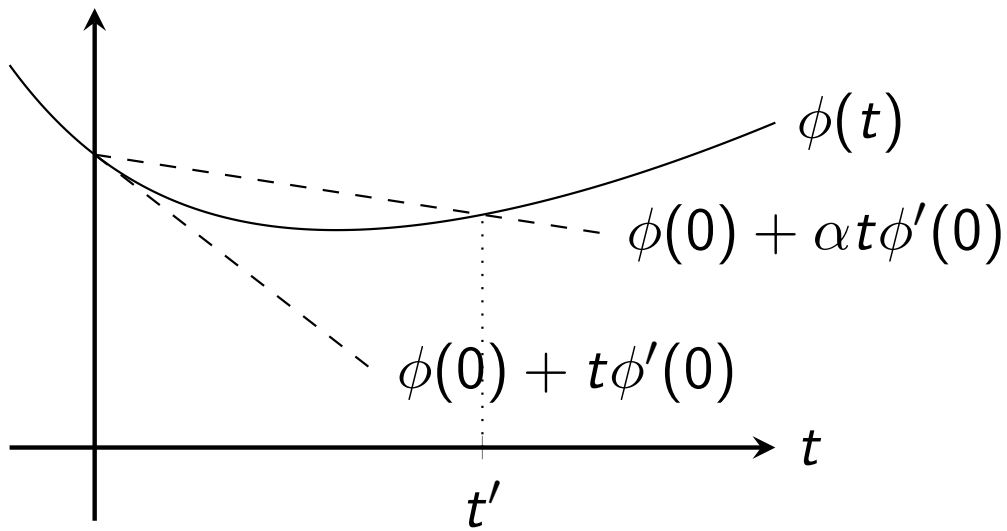
Given descent direction $\Delta x$ at $x_k$ consider $\phi : \mathbf{R} \to \mathbf{R}$ defined as

$$\phi(t) = f(x_k + t\Delta x).$$

Since $\Delta x$ descent direction, we have $\phi'(0) = \nabla f(\Delta x)^T \Delta x < 0$.

*Armijo* descent condition:

$$\phi(t) \leq \phi(0) + \alpha t \phi'(0). \tag{18}$$



# Backtracking Line Search Algorithm

**Require:** Initial step size $t > 0$ and parameters $\alpha \in (0, 1/2]$ and $\beta \in (0, 1)$

**Ensure:** $t > 0$ such that $\phi(t) \leq \phi(0) + \alpha t \phi'(0)$

  **while** $\phi(t) > \phi(0) + \alpha t \phi'(0)$ **do**

    $t \leftarrow \beta t$

  **end while**

# Upper Bound on $\alpha$

Upper bound $\alpha \leq 1/2$ can be motivated by considering case where

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(0).$$

Exact minimizer of $\phi(t)$ is then $t^\star = -\phi'(0)/\phi''(0)$. The Armijo condition (18) imposes an upper bound on $t$ which may be expressed in terms of $t^\star$ as $t \leq 2(1-\alpha)t^\star$.

It follows that $t = t^\star$ does not satisfy the descent condition if $\alpha > 1/2$.

# Trust-Region Method

In nonconvex case a trust-region is added:

$$\begin{array}{ll} \text{minimize} & \nabla f(x_k)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x_k) \Delta x \\ \text{subject to} & \|\Delta x\|_2^2 \leq \delta_k^2 \end{array}$$

where $\delta_k$ is the so-called trust-region radius.

Special case of a quadratically constrained quadratic problem (QCQP), and it can be solved numerically to global optimality by means of an eigenvalue decomposition of $\nabla^2 f(x_k)$.

# Equality Constrained Problems

Consider convex optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array} \tag{19}$$

where $x \in \mathbf{R}^n$, $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$ ($A$ has full rank and $m < n$).

Suppose $F \in \mathbf{R}^{n \times p}$ basis for $\mathcal{N}A$ and $\bar{x}$ is any vector: $A\bar{x} = b$, then

$$\{x \in \mathbf{R}^n \,|\, Ax = b\} = \{Fz + \bar{x} \,|\, z \in \mathbf{R}^p\},$$

and hence (19) is equivalent to

$$\text{minimize}_z \quad f(Fz + \bar{x})$$

with variable $z \in \mathbf{R}^p$.

Newton equation:

$$F^T \nabla^2 f(Fz + \bar{x}) F \Delta z = -F^T \nabla f(Fz + \bar{x}).$$

Given a minimizer $z^\star$ of $f(Fz + \bar{x})$, a solution to the equality constrained problem may be obtained as $x^\star = Fz^\star + \bar{x}$.

# KKT Optimality Conditions

Lagrangian $L : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ given by

$$L(x, \nu) = f(x) + \nu^T (Ax - b),$$

KKT conditions:

$$\begin{aligned} \nabla f(x) + A^T \nu &= 0 \\ Ax - b &= 0, \end{aligned}$$

or equivalently, $r(z) = 0$ where $z = (x, \nu)$ and

$$r(z) = \begin{bmatrix} \nabla f(x) + A^T \nu \\ Ax - b \end{bmatrix}.$$

# Infeasible start Newton method

Newton direction $\Delta z = (\Delta x, \Delta \nu)$ for $r(z) = 0$ solution to linearized equation:

$$r(z) + J_r(z)\Delta z = 0$$

where $J_r$ is Jacobian of $r$. Equivalently

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \nu \end{bmatrix} = - \begin{bmatrix} \nabla f(x) + A^T \nu \\ Ax - b \end{bmatrix}. \tag{20}$$

Step size using backtracking line search where

$$\phi(t) = \|r(z + t\Delta z)\|_2$$

is used with the Armijo condition $\phi(t) \leq \phi(0) - \alpha t \phi'(0)$. If $r(z) \neq 0$ we have $\phi'(0) = -\|r(z)\|_2$ and hence Armijo descent condition is

$$\|r(z + t\Delta z)\|_2 \leq (1 - t\alpha)\|r(z)\|_2. \tag{21}$$

# Variable Metric Methods

Cosider more general convex quadratic approximation $\hat{f}(\cdot|\cdot) : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ given by

$$\hat{f}(y \,|\, x_k) = f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2}(y - x_k)^T B_k (y - x_k) \tag{22}$$

where $B_k \succ 0$. We have

$$\hat{f}(x_k \,|\, x_k) = f(x_k), \qquad \nabla \hat{f}(x_k \,|\, x_k) = \nabla f(x_k).$$

Define search direction as

$$\Delta x_k = \operatorname*{argmin}_{\Delta x} \hat{f}(x_k + \Delta x \,|\, x_k) = -B_k^{-1} \nabla f(x_k),$$

and iterations as

$$x_{k+1} = x_k - t_k B_k^{-1} \nabla f(x_k). \tag{23}$$

Letting $B_k = I$ yields the gradient descent method and $B_k = \nabla^2 f(x_k)$ corresponds to Newton's method.

# Secant Equation

The mechanism for updating $B_k$ or $H_k = B_k^{-1}$ is the *secant* equation

$$B_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k). \tag{24}$$

Definition of $\hat{f}$ implies $\nabla \hat{f}(x_{k+1} \,|\, x_{k+1}) = \nabla f(x_{k+1})$, and the secant equation is simply the additional condition

$$\nabla \hat{f}(x_k \,|\, x_{k+1}) = \nabla f(x_k)$$

Define $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ and $s_k = x_{k+1} - x_k$ so that secant equation read

$$B_{k+1} s_k = y_k$$

Drop iteration index $k$.

# Quasi-Newton Update

The secant equation is an underdetermined for $n > 1$.

Quasi-Newton update formulae mostly differ in how they construct a matrix $B_{k+1}$ that satisfies the secant equation.

The BFGS update, named after Broyden, Fletcher, Goldfarb, and Shanno, is

$$B_{k+1} = B_k - \frac{B_k s s^T B_k}{s^T B_k s} + \frac{y y^T}{y^T s}. \tag{25}$$

The update $B_{k+1}$ is positive definite, and it can be shown to be the solution of convex optimization problem

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(B B_k^{-1}) - \ln\det(B) \\ \text{subject to} & Bs = y \end{array} \tag{26}$$

where $B \in \mathbf{S}_+^n$.

# Update Formula for Inverse

The BFGS update may also be expressed as

$$H_{k+1} = H_k + \frac{(y^T s + y^T H_k y) s s^T}{(y^T s)^2} - \frac{H_k y s^T + s y^T H_k}{y^T s} \qquad (27)$$

where $H_{k+1} = B_{k+1}^{-1}$.

Hence

$$\Delta x = -H_k \nabla f(x_k)$$

# Davidson, Fletcher, and Powell (DFP) update

DFP update:

$$B_{k+1} = B_k + \frac{ry^T + yr^T}{y^T s} - \frac{(r^T s)yy^T}{(y^T s)^2} \tag{28}$$

where $r = y - B_k s$, or equivalently, as inverse Hessian approximation $H_{k+1} = B_k^{-1}$,

$$H_{k+1} = H_k - \frac{H_k yy^T H_k}{y^T H_k y} + \frac{ss^T}{y^T s}. \tag{29}$$

# Symmetric Rank-1 (SR1) Update

SR1 update:

$$B_{k+1} = B_k + \sigma v v^T$$

where $\sigma \in \{-1, 1\}$. The secant equation $B_{k+1}s = y$ implies

$$r = \sigma(v^T s)v \tag{30}$$

where $r = y - B_k s$. Forming the inner product with $s$ on both sides of this equation yields $\sigma(v^T s)^2 = r^T s$ which, combined with (30), implies

$$vv^T = \frac{rr^T}{\sigma^2(v^T s)^2} = \frac{rr^T}{\sigma r^T s},$$

and hence

$$B_{k+1} = B_k + \frac{rr^T}{r^T s}. \tag{31}$$

Update maintains symmetry, but unlike BFGS and DFP updates, it does not guarantee that $B_{k+1}$ is positive definite.

# Barzilai–Borwein (BB) Step Size Rule

If we require $B_{k+1} = \gamma_{k+1} I$ where $\gamma_{k+1} \in \mathbf{R}_{++}$, the secant condition is generally not satisfiable, but one can choose $\gamma_{k+1}$ such that

$$\gamma_{k+1} = \operatorname*{argmin}_{\gamma > 0} \|\gamma s - y\|_2^2 = \frac{y^T s}{\|s\|_2^2}. \tag{32}$$

Alternatively, we can define an inverse approximation $H_{k+1} = \gamma_{k+1} I$ and get

$$\gamma_{k+1} = \operatorname*{argmin}_{\gamma > 0} \|s - \gamma y\|_2^2 = \frac{y^T s}{\|y\|_2^2}. \tag{33}$$

Combining the gradient descent method with one of the two BB step size rules does not yield a descent method. The method converges if $f$ is quadratic, but it is not guaranteed to converge in general. However, convergence can be established if the BB steps are combined with a line search.

# Proximal Quasi-Newton

Variable metric approach can be extended to $f = g + h$ where $h$ is convex and $g$ is twice differentiable and convex. Let

$$\hat{f}(y \mid x_k) = h(y) + g(x_k) + \nabla g(x_k)^T (y - x_k) + \frac{1}{2}(y - x_k)^T B_k (y - x_k) \tag{34}$$

and let

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_y \left\{ \hat{f}(y \mid x_k) \right\} \\ &= \operatorname*{argmin}_y \left\{ h(y) + \frac{1}{2} \| y - (x_k - B_k^{-1} \nabla g(x_k)) \|_{B_k}^2 \right\} \\ &= \mathbf{prox}_h^{B_k}(x_k - B_k^{-1} \nabla g(x_k)) \end{aligned} \tag{35}$$

where

$$\mathbf{prox}_h^B(x) = \operatorname*{argmin}_y \left\{ h(y) + \frac{1}{2} \| y - x \|_B^2 \right\}. \tag{36}$$

The iteration (35) is known as *proximal quasi-Newton* or *proximal Newton* if $B_k = \nabla^2 f(x_k)$.