Large scale and distributed optimization - Part 1

Bigmath[1] Advanced Course 4

Nataša Krejić

[1] This project has received funding from the European Union Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No 812912.

Outline

- Machine learning and Optimization
- Nonlinear optimization problems, optimality conditions
- Line search methods
- First order methods
- Second order methods
- Optimality conditions for constrained problems
- Special classes of constrained problems
- Penalty methods

Machine Learning and Optimization

- A data set for analysis

$$D = \{(a_i, y_i), i = 1, \dots, N\}$$

- $a_i \in \mathbb{R}^n$ - vector of features
- y_i - labels (observations)
- Prediction function Φ such that

$$\Phi(a_i) \approx y_i, i = 1, \dots, N$$

- *approx* in some optimal sense
- Data set is a sample.

- Supervised learning
  - $y_i \in \mathbb{R}$ - regression problem
  - $y_i \in \{1, \ldots, M\}$ - classification problem
  - $y_i \in \{-1, 1\}$ - binary classification
- Unsupervised learning - the labels are not known; clustering, extracting interesting information from the data
- Choice of features

The prediction function Φ properties:

- $$\Phi(a_i) \approx y_i, i = 1, \ldots, N$$
- reliable prediction for new (unseen) data
- features selection (important ones)
- online learning (streaming data)

Prediction function Φ depends on parameters $x \in \mathbb{R}^n$ that we need to learn.

- Data set = training + testing
- $$\Phi(a_i) \approx y_i, i = 1, \ldots, N$$
- Loss function $\ell(a_i, y_i, x)$ - measures discrepancy between $\Phi(a_i)$ and y_i
- $$\min \sum_{i=1}^{N} \ell(a_i, y_i, x)$$
- $$L(a, y, x) = \sum_{i=1}^{N} \ell(a_i, y_i, x)$$

Robustness

- Φ should be a good predictor on unseen data
- Overfitting should be avoided
- Adding a regularizer
- $$\min \sum_{i=1}^{N} \ell(a_i, y_i, x) + \lambda \|x\|_2^2$$
- $$\min \sum_{i=1}^{N} \ell(a_i, y_i, x) + \lambda \|x\|_1$$

Regressions

- Linear regression

$$\Phi(x) = a^T w + b, \; x = (w, b)$$

- Loss function

$$L(x) = \sum_{i=1}^{N} (a_i^T w + b - y_i)^2$$

- corresponds to maximum likelihood solution if $y = a^T w + b + \varepsilon, \; \varepsilon : \mathcal{N}(0, \sigma^2\}$
- a_i i.i.d.

- Ridge regression

$$L(x) = \sum_{i=1}^{N} (a_i^T w - y_i)^2 + \lambda \|w\|^2$$

- Lasso regression (enforces sparsity)

$$L(x) = \sum_{i=1}^{N} (a_i^T w - y_i)^2 + \lambda \|w\|_1$$

- Logistic regression - maximizes likelihood of belonging to one class or another

$$\ell_L(a, y, w, b) = \log(1 + exp(-y(w^T a - b))$$

$$\min_{w,b} \frac{1}{N} \sum_{i=1}^{N} \ell_L(a_I, y_i, w, b) + \frac{\lambda}{2} \|(w, b)\|^2$$

- Neural networks - many, many loss functions ...
- Activation function, number of hidden layers, type of network
- Training - minimization of the loss function

Stochastic optimization

$$\min E_{\xi}[f(x, \xi)]$$

- Sample of i.i.d. $\xi_1, \ldots, \xi_N$
- Sample average approximation (SAA) approximation

$$E_{\xi}[f(x, \xi)] \approx \frac{1}{N} \sum_{i=1}^{N} f(x, \xi_i)$$

$$\min \frac{1}{N} \sum_{i=1}^{N} f(x, \xi_i)$$

Stochastic approximation (SA) methods

$$\min F(x), \text{ subject to noise, not available}$$

$$f(x) = F(x) + \xi(x)$$

$$\min f(x)$$

- $f(x), g(x) \approx \nabla f(x), H(x) \approx \nabla^2 f(x)$ - noisy values that are available

Nonlinear optimization problem

$$\min_{x \in S} f(x), \tag{1}$$

- $f : D \to \mathbb{R}$ and $D, S \subseteq \mathbb{R}^n$.
- f - objective function
- $x \in \mathbb{R}^n$ - decision variable
- S - feasible set
- $S = \mathbb{R}^n$ - unconstrained problem, S - proper subset of $\mathbb{R}^n$ - constrained problem

Constrained versus unconstrained problems

- $$\min x^2$$
- $$\min x^2 \text{ s.t. } x \leq 2$$
- $$\min x^2 \text{ s.t. } x > -1$$

Theorem

(Bolzano-Weierstrass) Every real, continuous function attains its global minimum on any compact subset of $\mathbb{R}^n$.

Definition

A point x^* *is a global solution of the problem (1) if* $f(x^*) \leq f(x)$ *for every* $x \in S$. *If* $f(x^*) < f(x)$ *for every* $x \in S, \ x \neq x^*$, *then* x^* *is a strict global solution.*

Definition

A point x^* *is a local solution of the problem (1) if there exists* $\varepsilon > 0$ *such that* $f(x^*) \leq f(x)$ *for every* $x \in S$ *such that* $\|x - x^*\| \leq \varepsilon$. *If* $f(x^*) < f(x)$ *for every* $x \in S, \ x \neq x^*$ *such that* $\|x - x^*\| \leq \varepsilon$, *then we say that* x^* *is strict local solution.*

Optimality conditions

$$\min_{x \in \mathbb{R}^n} f(x), \tag{2}$$

Theorem

Suppose that $f \in C^1(\mathbb{R}^n)$. If x^* is a local solution of (2), then $\nabla f(x^*) = 0$.

Theorem

Suppose that $f \in C^2(\mathbb{R}^n)$. If x^* is a local solution of (2), then

a) $\nabla f(x^*) = 0$;

b) $\nabla^2 f(x^*) \succeq 0$.

Theorem

Suppose that $f \in C^2(\mathbb{R}^n)$. If

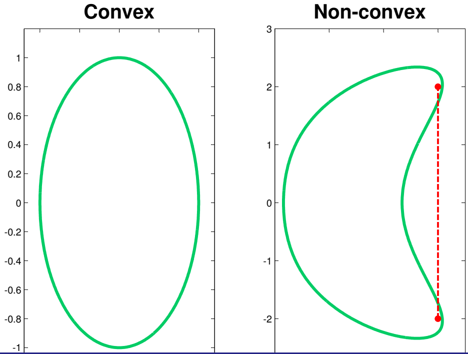
1. $\nabla f(x^*) = 0$ *and*
2. $\nabla^2 f(x^*) \succ 0$,

then x^* is a strict local solution of (2).

Convexity

Definition

A set $S \subseteq \mathbb{R}^n$ *is convex if for any* $x, y \in S$ *and any* $\lambda \in [0, 1]$ *there holds* $\lambda x + (1 - \lambda) y \in S$.



Definition

Let S be a convex set. A function $f : S \to \mathbb{R}$ is convex on S if for any $x, y \in S$ and any $\lambda \in [0, 1]$ there holds

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

Moreover, we say that the function is strictly convex if the previous inequality is strict for all $x \neq y$ and $\lambda \in (0, 1)$.

Theorem

Suppose that $f \in C^1(S)$ where $S \subseteq \mathbb{R}^n$ is a convex set. Then, the function f is convex on S if and only if the following inequality holds for all $x, y \in S$

$$f(y) \geq f(x) + \nabla^T f(x)(y - x). \tag{3}$$

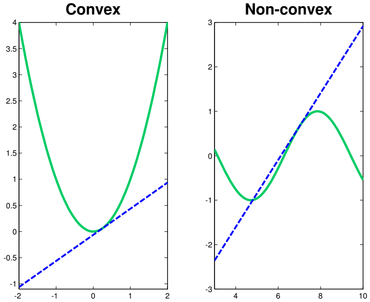


Figure: Convex and non-convex functions.

Theorem

Suppose that $f \in C^2(S)$ where $S \subseteq \mathbb{R}^n$ is a convex set. Then, the following statements hold.

a) *If $\nabla^2 f(x) \succeq 0$ for every $x \in S$, then f is convex on S.*
b) *If $\nabla^2 f(x) \succ 0$ for every $x \in S$, then f is strictly convex on S.*
c) *If S is open and f is convex on S, then $\nabla^2 f(x) \succeq 0$ for every $x \in S$.*

Theorem

Suppose that f is convex on a convex set S. Then, every local minimizer of the function f is also the global minimizer.

Definition

A function f is strongly convex with parameter $m > 0$ on a convex set S if for any $x, y \in S$ and any $\lambda \in [0, 1]$ there holds

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{m}{2}\lambda(1-\lambda)\|x-y\|^2.$$

$$f(y) \geq f(x) + \nabla^T f(x)(y-x) + \frac{m}{2}\|x-y\|^2,$$

Line search methods

Definition

Consider a point x such $\nabla f(x) \neq 0$. A direction d is called descent direction for f at the point x if there exists $\alpha > 0$ such that

$$f(x + \alpha d) < f(x).$$

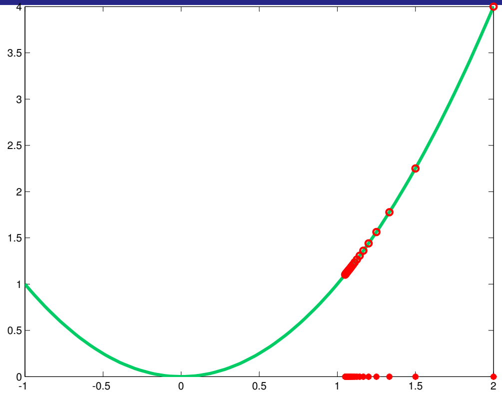


Figure: Insufficient decrease - small steps

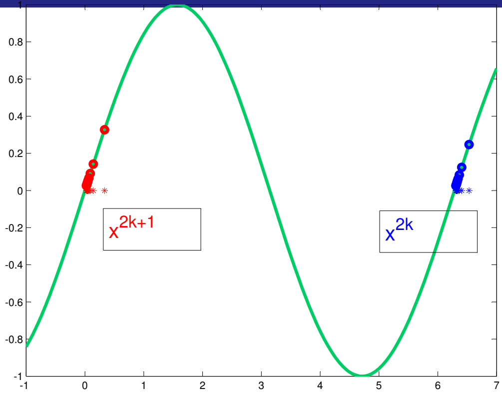


Figure: Insufficient decrease - large steps

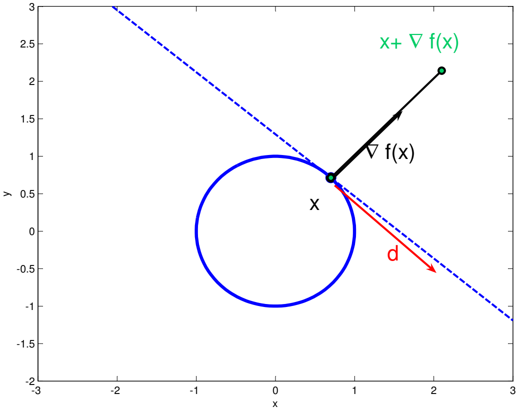


Figure: Insufficient decrease - insufficiently descent direction.

$$\|d^k\| \geq \sigma \|\nabla f(x^k)\|, \tag{4}$$

$$\nabla^T f(x^k) d^k \leq -\theta \|\nabla f(x^k)\| \|d^k\| \tag{5}$$

$$f(x^k + \alpha_k d^k) \leq f(x^k) + \eta \nabla f(x^k) d^k \text{ Armijo condition} \tag{6}$$

$$\nabla f(x^k + \alpha_k d^k) \geq c \nabla f(x^k), c \in (, \eta) \text{ Wolfe condition} \tag{7}$$

Algorithm LS with backtracking

Step 0 Input parameters: $x^0 \in \mathbb{R}^n$, $\beta, \eta \in (0,1)$, $\theta \in (0,1]$, $\sigma > 0$, $k = 0$

Step 1 Stopping criterion: If $\nabla f(x^k) = 0$ STOP.

Step 2 Search direction: Choose d^k such that $\|d^k\| \geq \sigma \|\nabla f(x^k)\|$ and $\nabla^T f(x^k) d^k \leq -\theta \|\nabla f(x^k)\| \|d^k\|$.

Step 3 Given $\beta \in (0,1)$, find the smallest nonnegative integer j such that $\alpha_k = \beta^j$ satisfies

$$f(x^k + \alpha_k d^k) \leq f(x^k) + \eta \alpha_k \nabla^T f(x^k) d^k.$$

Step 4 Update: Set $x^{k+1} = x^k + \alpha_k d^k$, $k = k + 1$.

Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1(\mathbb{R}^n)$ *and* $\nabla^T f(x^k) d^k < 0$. *Moreover, assume that the function* f *is bounded from bellow on the line* $\{x^k + \alpha d^k \mid \alpha > 0\}$. *Then, there exists* $\bar{\alpha} > 0$ *such that the Armijo condition holds for all* $\alpha \in (0, \bar{\alpha}]$.

.

Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1(\mathbb{R}^n)$ and f is bounded from bellow. Moreover, assume that the sequence of search directions $\{d^k\}_{k\in\mathbb{N}}$ is bounded. Then, either the Algorithm LS with backtracking terminates after a finite number of iterations $\bar{k}$ at the stationary point $x^{\bar{k}}$ or every accumulation point of the sequence $\{x^k\}_{k\in\mathbb{N}}$ is a stationary point of the function f.

Gradient method

$$d^k = -\nabla f(x^k). \tag{8}$$

$$\alpha_k = \arg\min_{\alpha>0} f(x^k + \alpha d^k) \text{ - exact line search}$$

Theorem

Suppose that $f \in C^2(\mathbb{R})$ and that the gradient method with the exact line search converges to a point x^* such that $\nabla^2 f(x^*)$ is positive definite, with m and M being the smallest and largest eigenvalues. Then Then

$$f(x^{k+1}) - f(x^*) \leq \left(\frac{M-m}{M+m}\right)^2 (f(x^k) - f(x^*)).$$

Gradient method with fixed step size

$$x^{k+1} = x^k - \alpha \nabla f(x^k). \tag{9}$$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{10}$$

Theorem

Suppose that $f \in C^2(\mathbb{R}^n)$ is convex and that (10) holds. Then, if $\alpha < 1/L$, the fixed step size negative gradient method defined with (9) satisfies

$$f(x^k) - f(x^*) \leq \frac{\|x^0 - x^*\|^2}{2\alpha k}.$$

Minimizing finite sums

$$\min_{x \in \mathbb{R}^n} f(x), \tag{11}$$

$$f(x) = f_N(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x), \tag{12}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz smooth function
- $f_i : \mathbb{R}^n \to \mathbb{R}$.
- f is bounded from below in $\mathbb{R}^n$.
- N is very large

Subsampling

$$f_k = \frac{1}{N_k} \sum_{i \in \mathcal{N}_k} f_i(x_k), \tag{13}$$

$$g_k = \frac{1}{N_k} \sum_{i \in \mathcal{N}_k} \nabla f_i(x_k). \tag{14}$$

- Smaller N_k - cheaper method
- Eventually $N_k = N$ for k large enough
- Many, many scheduling strategies

Stochastic gradient method

- Standard gradient is expensive (N is large)
- Training set might be redundant
- Replace the full gradient with an inexpensive stochastic approximation - minibatch gradient g_k

Algorithm SGD

Step 0 Choose an initial point x_0 and a sequence of strictly positive steplengths $\{\alpha_k\}$. Set $k = 0$.

Step 1 Choose randomly and uniformly $i_k \in \{1, \ldots, N\}$. Set $g_k = \nabla f_{i_k}(x_k)$.

Step 2 Set $x_{k+1} = x_k - \alpha_k g_k$, $k = k + 1$.

Variance condition:

$$E[\|g_k\|^2] \leq M_1 + M_2\|\nabla f(x_k)\|^2, \tag{15}$$

Theorem

Suppose that f has Lipschitz continuous gradient and that it is strongly convex. Let x_* be the minimizer of f. Assume that (15) holds at each iteration. Then, if SGD *is run with $\alpha_k = \frac{\beta}{\gamma+k}$, $\beta > \frac{1}{\mu}$ and $\gamma > 0$ such that $\alpha_1 \leq \frac{1}{LM_2}$, there exists a constant $\nu > 0$ such that*

$$E[f(x_k)] - f(x_*) \leq \frac{\nu}{\gamma + k}. \tag{16}$$

Stochastic variance reduction gradient (SVRG) method

- SGD converges sublinearly (very slow)
- The variance of random sampling implies (very) small step size
- Nonconvex problems: $\sum_k \alpha_k = \infty, \ \sum_k \alpha_k^2 = 0$
- Larger N_k in subsampled gradient might reduce the variance but it is more expensive

Algorithm SVRG

Step 0 Choose an initial point $x_0 \in \mathbb{R}^n$, an inner loop size $m > 0$, a steplength $\alpha > 0$, the option for the iterate update. Set $k = 1$.

Step 1 Outer iteration, full gradient evaluation.
Set $\tilde{x}_0 = x_{k-1}$. Compute $\nabla f_N(\tilde{x}_0)$.

Step 2 Inner iterations
For $t = 0, \ldots, m-1$
  Uniformly and randomly choose $i_t \in \{1, \ldots, N\}$.
  Set $\tilde{x}_{t+1} = \tilde{x}_t - \alpha(\nabla f_{i_t}(\tilde{x}_t) - \nabla f_{i_t}(\tilde{x}_0) + \nabla f_N(\tilde{x}_0))$.

Step 3 Outer iteration, iterate update.
Set $x_k = \tilde{x}_m$ (Option I), $k = k + 1$.
Set $x_k = \tilde{x}_t$ for randomly chosen $t \in \{0, \ldots, m-1\}$ (Option II), $k = k + 1$.

- Outer iterations (epochs) - full gradient is computed
- Inner iterations (m steps) - an unbiased approximation of the gradient is updated randomly

$$\nabla f_{i_t}(\tilde{x}_t) - \nabla f_{i_t}(\tilde{x}_0) + \nabla f_N(\tilde{x}_0)$$

- Inner iterations $m = 2n$ (convex), $m = 5n$ (non-convex)
- Full gradient can be replaced by mini-batch gradient
- Two option for the final approximation

Theorem

Suppose that f has Lipschitz continuous gradient and that it is strongly convex. Let x_* be the minimizer of f. If m and α satisfy

$$\theta = \frac{1}{\mu\alpha(1 - 2L\alpha)m} + \frac{2L\alpha}{1 - 2L\alpha} < 1, \tag{17}$$

then Algorithm **SVRG** *with Option II generates a sequence which converges linearly in expectation*

$$E[f(x_k) - f(x_*)] \leq \theta^k (f(x_0) - f(x_*)).$$

Theorem

Suppose that f has Lipschitz continuous gradient and that it is strongly convex. Let x_* be the minimizer of f. If m and α satisfy

$$\theta = (1 - 2\alpha\mu(1 - \alpha L)^m) + \frac{4\alpha L^2}{\mu(1 - \alpha L)} < 1,$$

then Algorithm **SVRG** *with Option I generates a sequence which converges linearly in expectation*

$$E[x_k - x_*] \leq \theta^k (x_0 - x_*).$$

SAG method

- Stochastic Average Gradient tracking method
- Cost of SGD, convergence of FGD

Algorithm SAG

Step 0 Initialization. Choose an initial point $x_0 \in \mathbb{R}^n$, positive steplengths $\{\alpha_k\}$, $y_i = 0$, for $i = 1, \ldots, N$. Set $k = 0$.

Step 1 Stochastic gradient update. Uniformly and randomly choose $i_k \in \{1, \ldots, N\}$. Set $y_{i_k} = \nabla f_{i_k}(x_k)$.

Step 2 Iteration update. Set

$$x_{k+1} = x_k - \frac{\alpha_k}{N} \sum_{i=1}^{N} y_i.$$

Set $k = k + 1$.

Theorem

Suppose that f has Lipschitz continuous gradient and that it is strongly convex. Let x_* be the minimizer of f. If $\alpha_k = \alpha = 1/(16L)$ then

$$E[f(x_k)] - f(x_*) \leq \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8N}\right\}\right)^k C_0,$$

where $C_0 > 0$ depends on x_*, x_0, f_N, L, N.

SARAH method

- accumulation of stochastic gradient information
- variance reduction
- biased gradient approximation

Algorithm SARAH

Step 0 Initialization. Choose an initial point $x_0 \in \mathbb{R}^n$, an inner loop size $m > 0$, a steplength $\alpha > 0$. Set $k = 1$.

Step 1 Outer iteration, full gradient evaluation. Set $\tilde{x}_0 = x_{k-1}$. Compute $y_0 = \nabla f_N(\tilde{x}_0)$. Set $\tilde{x}_1 = \tilde{x}_0 - \alpha y_0$.

Step 2 Inner iterations.
For $t = 1, \dots, m-1$
  Uniformly and randomly choose $i_t \in \{1, \dots, N\}$.
  Compute $y_t = \nabla f_{i_t}(\tilde{x}_t) - \nabla f_{i_t}(\tilde{x}_{t-1}) + y_{t-1}$.
  Set $\tilde{x}_{t+1} = \tilde{x}_t - \alpha y_t$.

Step 3 Outer iteration, iterate update.
Take $x_k = \tilde{x}_t$ for randomly chosen $t \in \{0, \dots, m\}$ and set $k = k + 1$.

Theorem

Suppose that f has Lipschitz continuous gradient and that it is strongly convex and that each function f_i, $1 \leq i \leq N$ is convex. Let x_* be the minimizer of f.If α and m are such that

$$\sigma = \frac{1}{\mu\alpha(m+1)} + \frac{\alpha L}{2-\alpha L} < 1, \tag{18}$$

then the sequence $\{\|\nabla f(x_k)\|\}$ generated by Algorithm **SARAH** *satisfy*

$$E[\|\nabla f(x_k)\|^2] \leq \sigma^k \|\nabla f(x_0)\|^2.$$

Theorem

Suppose that f has Lipschitz continuous gradient and each function f_i, $1 \leq i \leq N$ is μ-strongly convex with $\mu > 0$. If $\alpha \leq 2/(\mu + L)$ then for any $t \geq 1$

$$E[\|y_t\|^2] \leq \left(1 - \frac{2\mu L \alpha}{\mu + L}\right)^t E[\|\nabla f(x_0)\|^2].$$

The Newton method

$$\min f(x)$$
$$\nabla f(x^{k+1}) = 0$$
$$\nabla f(x^k + d^k) \approx \nabla f(x^k) + \nabla^2 f(x^k) d^k.$$

The Newton equation

$$\nabla f(x^k) + \nabla^2 f(x^k) d^k = 0. \tag{19}$$
$$x^{k+1} = x^k + d^k \text{ or } \; x^{k+1} = x^k + \alpha_k d^k \tag{20}$$

- Local quadratic convergence
- Expensive (compute $\nabla^2 f(x^k)$, solve (19))
- Suppose that the function f is quadratic and strongly convex. Then, the Newton method provides a global minimizer of function f in one iteration with arbitrary x^0.

Local convergence

Theorem

Suppose that the function $f \in C^2(\mathbb{R}^n)$ *and there exists* $\delta > 0$ *such that* $\nabla^2 f(x) \succ 0$ *and* $\nabla^2 f(x)$ *is Lipschitz continuous with the constant L for all* $x \in B(x^*, \delta)$. *Then there exists* $\epsilon > 0$ *such that the Newton method converges quadratically to the solution* x^* *for all* $x^0 \in B(x^*, \epsilon)$. *Moreover, the sequence of the gradient norms converges quadratically to zero.*

Line search Newton method

- $f \in C^2$ - strongly convex function
- d^k - descent direction
- Line search can be applied
- Global convergence
- Local (quadratic) rate of convergence $\alpha_k = 1, k \geq k_0$

Quasi Newton methods

- The main idea: approximate the Hessian matrix with $B_k \in \mathbb{R}^{n\times n}$ using the first order information

$$s^k = x^{k+1} - x^k \text{ and } y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$$

Mean-value theorem

$$y^k = \int_0^1 \nabla^2 f(x^k + ts^k)s^k dt$$

$$B_{k+1}s^k \approx \int_0^1 \nabla^2 f(x^k + ts^k)s^k dt$$

Secant equation

$$B_{k+1}s^k = y^k \tag{21}$$

Least change secant update

$$B_{k+1} = \arg\min \|B - B_k\| \text{ s.t. } B_{k+1}s^k = y^k, B = B^T, \text{sparsity...}$$

BFGS formula

$$B_{k+1} = B_k + \frac{y^k(y^k)^T}{(y^k)^T s^k} - \frac{B_k s^k (s^k)^T B_k}{(s^k)^T B_k s^k} \tag{22}$$

DFP formula

$$B_{k+1} = \left(I - \frac{y^k(s^k)^T}{(y^k)^T s^k}\right) B_k \left(I - \frac{y^k(s^k)^T}{(y^k)^T s^k}\right) + \frac{y^k(y^k)^T}{(y^k)^T s^k} \tag{23}$$

The inverse B_{k+1}^{-1} is computable by SMW formula

$$B_k d^k = -\nabla f(x^k). \quad (24)$$

- Positive definite property of B_k - if $(y^k)^T s^k \geq \delta > 0$
- d^k - descent direction
- superlinear convergence

Theorem

Suppose that $f \in C^2(\mathbb{R}^n)$*. Let* $\{x^k\}$ *be a sequence generated by a quasi Newton method (24) and assume that* $\{x^k\}_{k\in\mathbb{N}}$ *converges to a point* x^* *such that* $\nabla f(x^*) = 0$ *and* $\nabla^2 f(x^*) \succ 0$*. Then* $\{x^k\}_{k\in\mathbb{N}}$ *converges superlinearly if*

$$\lim_{k\to\infty} \frac{\|(B_k - \nabla^2 f(x^*))d^k\|}{\|d^k\|} = 0. \quad (25)$$

Spectral gradient method

Approximate Hessian is a scalar matrix, $B_k = \gamma_k^{-1} I$.
The secant equation yields

$$\gamma_k = \arg\min_{\gamma>0} \|\gamma y^{k-1} - s^{k-1}\|$$

and

$$\gamma_k = \frac{(s^{k-1})^T y^{k-1}}{\|y^{k-1}\|^2}. \tag{26}$$

Safeguard conditions (curvature condition does not hold)

$$\bar{\gamma}_k = \min\{\gamma_{max}, \max\{\gamma_k, \gamma_{min}\}\}$$

- Very efficient, nonmonotone bahaviour

Spectral gradient method for finite sums

$$\min_{x\in\mathbb{R}^n} f(x),$$

$$f(x) = f_N(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

- Stochastic variance reduction with variable (spectral) step size

Algorithm SVRG - BB

Step 0 Initialization. Choose an initial point $x_0 \in \mathbb{R}^n$, an inner loop size $m > 0$, an initial steplength $\alpha_0 > 0$. Set $k = 1$.

Step 1 Outer iteration, full gradient evaluation.
Set $\tilde{x}_0 = x_{k-1}$. Compute $\nabla f_N(\tilde{x}_0)$.
If $k > 0$, then set $\alpha_k = \frac{1}{m} \frac{\|x_k - x_{k-1}\|^2}{(x_k - x_{k-1})^T(\nabla f_N(x_k) - \nabla f_N(x_{k-1}))}$

Step 2 Inner iterations
For $t = 0, \ldots, m-1$
Uniformly and randomly choose $i_t \in \{1, \ldots, N\}$.
Set $\tilde{x}_{t+1} = \tilde{x}_t - \alpha_k(\nabla f_{i_t}(\tilde{x}_t) - \nabla f_{i_t}(\tilde{x}_0) + \nabla f_N(\tilde{x}_0))$

Step 3 Outer iteration, iterate update. Set $x_k = \tilde{x}_m$ and $k = k + 1$.

Theorem

Suppose that f has Lipschitz continuous gradient and that it is strongly convex. Let x_* be the minimizer of f. Define $\theta = (1 - e^{-2\mu/L})/2$. If m is chosen such that

$$m > \max\left\{\frac{2}{\log(1-2\theta)+2\mu/L}, \frac{4L^2}{\theta\mu^2} + \frac{L}{\mu}\right\},$$

then **SVRG-BB** *converges linearly in expectation*

$$E[\|x_k - x_*\|^2] < (1-\theta)^k \|\tilde{x}_0 - x_*\|^2.$$

Inexact Newton method

- The main idea: solve the Newton equation inexactly

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k) + r^k$$

$$\|r^k\| = \|\nabla^2 f(x^k) d^k + \nabla f(x^k)\| \leq \eta_k \|\nabla f(x^k)\| \tag{27}$$

- The rate of convergence depends on η_k
  - $\eta_k = \eta \in (0, 1)$ - linear convergence
  - $\eta_k \to 0$ - superlinear convergence
  - $\eta_k = \mathcal{O}(\|\nabla f(x^k)\|)$ - quadratic convergence

Subsampled Newton method for finite sum minimization

$$f(x) = f_N(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x), \tag{28}$$

- Subsampled (Inexact) Newton method
- Subsampled function, gradient, Hessian approximation

$$\nabla^2 f_{\mathcal{D}_k}(x^k)s^k = -\nabla f_{\mathcal{N}_k}(x^k) + r^k, \; \|r^k\| \leq \eta_k \|\nabla f_{\mathcal{N}_k}(x^k)\|, \tag{29}$$

- The subsample size N_k, D_k
- The choice of forcing term η_k - adaptive

$$\eta_k = \min\{\bar{\eta}, \frac{|f_{\mathcal{N}_k}(x^k) - m_{k-1}(s^{k-1})|}{\|\nabla f_{\mathcal{N}_{k-1}}(x^{k-1})\|}\}, \ \bar{\eta} < 1 \tag{30}$$

Theorem

Assume that $f \in C^2$ is strongly convex and that $\nabla^2 f(x)$ is Lipschitz continuous. Assume that $\mathcal{D}_k$ is chosen such that

$$\max_{\substack{\mathcal{D}:|\mathcal{D}|=D \\ x \in \mathcal{N}_{\delta*}(x^*)}} \|\nabla^2 f_{\mathcal{N}}(x) - \nabla^2 f_{\mathcal{D}}(x)\| \leq C\eta_k$$

holds for some $C < (1/\bar{\eta} - 1)\lambda_1$ and η_k is given by (30). Then $\{x^k\}$ converges to x^* locally superlinearly assuming that $N_k = N$ for k large enough.

Convergence in mean square

- Relaxing the subsampled Hessian error bound

$$\nabla^2 f_{\mathcal{D}}(x) = \frac{1}{D}\sum_{i=1}^{D} \nabla^2 f_i(x)$$

$$E(\nabla^2 f_{\mathcal{D}}(x)) = \nabla^2 f_{\mathcal{N}}(x). \tag{31}$$

- The Bernstein inequality

$$P(\|\nabla^2 f_{\mathcal{D}}(x) - \nabla^2 f_{\mathcal{N}}(x)\| \leq \gamma) \geq 1 - \alpha, \tag{32}$$

for given $\gamma > 0$ and $\alpha \in (0, 1)$.

Theorem

Assume that $f \in C^2$ is strongly convex and that the subsample $\mathcal{D}$ is chosen randomly and uniformly from $\mathcal{N}$. Let $\gamma > 0$ and $\alpha \in (0,1)$ be given. Then

$$P(\|\nabla^2 f_{\mathcal{D}}(x) - \nabla^2 f_{\mathcal{N}}(x)\| \leq \gamma) \geq 1 - \alpha,$$

holds at any point x if the subsample size D satisfies

$$D \geq \frac{2(\ln 2n - \ln \alpha)(\lambda_n^2 + \lambda_n \gamma/3)}{\gamma^2} := \tilde{l}. \tag{33}$$

Take $\mathcal{D}_k$ such that

$$P(\|\nabla^2 f_{\mathcal{D}}(x) - \nabla^2 f_{\mathcal{N}}(x)\| \leq C \max\{\eta_k, \|\nabla f_{\mathcal{N}_k}(x^k)\|\} \geq 1 - \alpha_k \qquad (34)$$

with $\alpha_k \in (0, 1)$.

a) if η_k defined by (30) then

$$E(\|x^{k+1} - x^*\|^2) \leq \left(V_1 \tau^{2k} + V_2 \alpha_k\right) E(\|x^k - x^*\|^2);$$

b) if $\eta_k = \bar{\eta}$ is sufficiently small then

$$E(\|x^{k+1} - x^*\|^2) \leq \left(C_1 \tau^{2k} + C_2 \bar{\eta}^2 + V_2 \alpha_k\right) E(\|x^k - x^*\|^2).$$

Constrained optimization

$$\min_{x \in S} f(x), \; S = \{x \in \mathbb{R}^n \mid h(x) = 0, \; g(x) \le 0\}. \tag{35}$$

$$f^* = \inf_{x \in S} f(x). \tag{36}$$

- Infeasible problems
  - S is empty
  - f is unbounded on S
- Explicit constraints $h(x) = 0, g(x) \le 0$
- Implicit constraints; domain of f

Ex. 1 $f(x) = x^{-2}$. $D = \mathbb{R}\backslash\{0\}$ and $f^* = 0$, but there is no optimal point.

Ex. 2 $f(x) = \ln(x)$. $D = \mathbb{R}_+\backslash\{0\}$ and $f^* = -\infty$.

Ex. 3 $f(x) = x\ln(x)$. $D = \mathbb{R}_+\backslash\{0\}$, $f^* = -e^{-1}$ and the optimal point is $x^* = e^{-1}$.

Ex. 4 $f(x) = x^3 - 3x$. No implicit constraints, the optimal value is $f^* = -\infty$, one local minimum at $\tilde{x} = 1$.

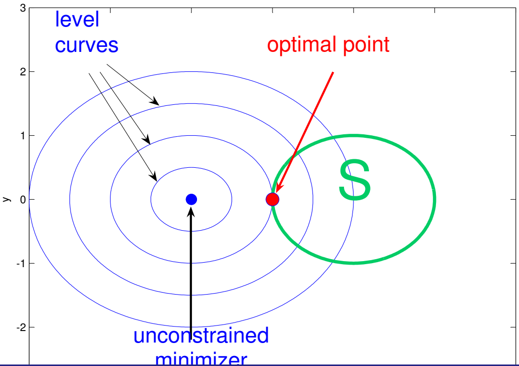
$$\min_{x \in \mathbb{R}^n} - \sum_{i=1}^{k} \ln(b_i - x^T a_i). \tag{37}$$

- No explicit constraints
- Equivalent form

$$\min_{x \in S} - \sum_{i=1}^{k} \ln(b_i - x^T a_i), \; S = \{x \in \mathbb{R}^n \mid x^T a_i < b_i, i = 1, ..., k.\}. \tag{38}$$

Convex problems

The problem (35) is convex if the objective function f and the inequality constraints functions $g_1, \ldots, g_m$ are convex, while the equality constraints functions $h_1, \ldots, h_p$ are affine.



Theorem

Every local solution of a convex constrained problem is a global solution of the same problem.

Theorem

Suppose that $f \in C^1(\mathbb{R}^n)$ *and that the problem is convex. Then,* x^* *is optimal if and only if* $x^* \in S$ *and for every* $y \in S$ *there holds*

$$\nabla^T f(x^*)(y - x^*) \geq 0. \tag{39}$$

Lagrangian function

$$\min_{x \in S} f(x), \;\; S = \{x \in \mathbb{R}^n \mid h(x) = 0, \;\; g(x) \leq 0\}$$

$$L(x, \lambda, \mu) := f(x) + \lambda^T g(x) + \mu^T h(x) = f(x) + \sum_{i=1}^{p} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j h_j(x), \qquad (40)$$

- $\lambda = (\lambda_1, ..., \lambda_p)^T \in \mathbb{R}^p$ - Lagrange multipliers associated to inequality constraints
- $\mu = (\mu_1, ..., \mu_m)^T \in \mathbb{R}^m$ - Lagrange multipliers associated to equality constraints
- λ and μ - dual variables

Daulity

The Lagrange dual function

$$l(\lambda, \mu) := \inf_{x \in D} L(x, \lambda, \mu). \tag{41}$$

The Lagrange dual problem

$$\max_{\lambda \geq 0} l(\lambda, \mu). \tag{42}$$

- LDP is convex
- Unique solution (λ^*, μ^*) - dual optimal, optimal Lagrange multipliers

KKT optimality conditions

Definition

Strong duality holds if the primal and dual optimal values are attained and equal.

Definition

KKT conditions are:

a) $g(x^*) \leq 0$ *(feasibility - inequality constraints).*

b) $h(x^*) = 0$ *(feasibility - equality constraints).*

c) $\lambda^* \geq 0$ *(dual feasibility).*

d) $\lambda_i^* g_i(x^*) = 0, \quad i = 1, ..., p$ *(complementarity).*

e) $\nabla f(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla g_i(x^*) + \sum_{i=j}^{m} \mu_j^* \nabla h_j(x^*) = 0$ *(optimality).*

- Necessary conditions if the strong duality holds

Theorem

Suppose that x^* and (λ^*, μ^*) are such that the KKT conditions are satisfied and the problem (35) is convex. Then x^* is a solution of the problem (35).

- Many, many other optimality conditions...

Linear independence constraint qualification (LICQ)

Definition

LICQ holds at point x^* if the gradients of active constraints at the point x^* are linearly independent.

Theorem

Suppose that x^* is a local solution of the problem (35) and that LICQ holds at the point x^*. Then there are Lagrange multipliers (λ^*, μ^*) such that the KKT conditions are satisfied.

Second order optimality conditions

Let x^* and (λ^*, μ^*) be primal and dual variables that satisfy KKT conditions. Then

$$A_1 = \{d \in \mathbb{R}^n \mid \nabla^T h_i(x^*) d = 0, i = 1, ..., m\}, \tag{43}$$

$$A_2 = \{d \in \mathbb{R}^n \mid \nabla^T g_i(x^*) d = 0 \text{ for all active constraints with } \lambda_i^* > 0\},$$

$$A_3 = \{d \in \mathbb{R}^n \mid \nabla^T g_i(x^*) d \leq 0 \text{ for all active constraints with } \lambda_i^* = 0\},$$

$$A = A_1 \cap A_2 \cap A_3. \tag{44}$$

Theorem

Suppose that x^* is a local solution of the problem (35) and that LICQ holds at the point x^*. Suppose that the Lagrange multipliers (λ^*, μ^*) are such that the KKT conditions hold. Then,

$$d^T \nabla_x^2 L(x^*, \lambda^*, \mu^*) d \geq 0 \text{ for all } d \in A.$$

Theorem

Suppose that x^* and (λ^*, μ^*) are such that the KKT conditions are satisfied and

$$d^T \nabla_x^2 L(x^*, \lambda^*, \mu^*) d > 0 \text{ for all } d \in A \backslash \{0\}.$$

Then x^* is a strict local solution of the problem (35).

Linear constraints

$$\min_{Ax=b} f(x), \tag{45}$$

- $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2(\mathbb{R}^n)$
- f - convex
- $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $rank(A) = m < n$

KKT conditions:

$$\nabla f(x^*) + A^T \mu^* = 0 \quad \text{and} \quad Ax^* = b \tag{46}$$

Box constrained optimization

$$\min_{l \leq x \leq u} f(x), \tag{47}$$

- $l, u \in \mathbb{R}^n_\infty$
- f - continuously differentiable on $S = \{x \in \mathbb{R}^n : l \leq x \leq u\}$

Optimality conditions for box constrained problems

Theorem

Let f be continuously differentiable. If x^* is a local solution of

$$\min f(x) \; s.t. \; l \leq x \leq u$$

then

$$\frac{\partial f}{\partial x} = \begin{cases} \geq 0, & x_i^* = l_i \\ = 0 & l_i < x_i^* < u_l \\ \leq 0 & x_i^* = u_l \end{cases}$$

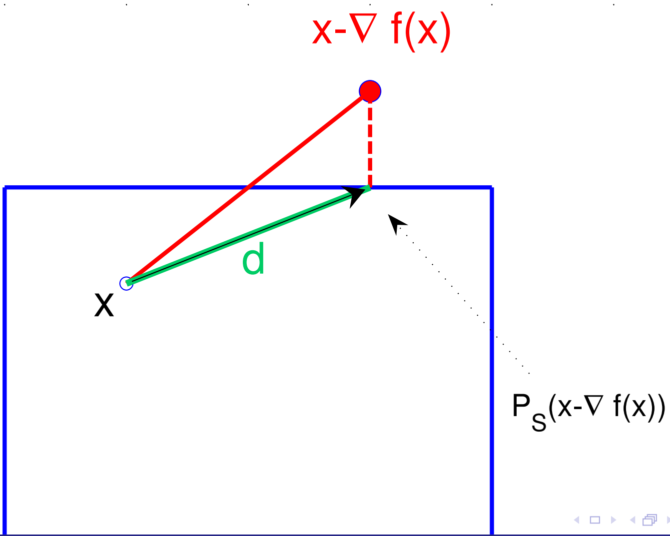
Orthogonal projections

Orthogonal distance

$$dist_S(x) = \inf_{y \in S} \|y - x\|. \quad (48)$$

Orthogonal projection of point x on a set S

$$P_S(x) = \arg\min_{y \in S} \|y - x\|. \quad (49)$$


x-∇ f(x)
d
x
P
S
(x-∇ f(x))

Projected gradient direction

$$d = d(x) = P_S(x - \nabla f(x)) - x. \tag{50}$$

Theorem

Suppose that $f \in C^1(S)$ and $x \in S$. Then the projected gradient direction d defined by (50) satisfies the following:

a) $d^T \nabla f(x) \leq -\|d\|^2$.

b) $d = 0$ *if and only if x is a stationary point for the problem (47).*

Algorithm PG-LS

Step 0 Input parameters: $x^0 \in S$, $\beta, \eta \in (0, 1)$, $k = 0$.

Step 1 Search direction: Compute the projected gradient direction d defined by (50). If $d^k = 0$ STOP.

Step 2 Step size: Find the smallest nonnegative integer j such that $\alpha_k = \beta^j$ satisfies the Armijo condition

$$f(x^k + \alpha_k d^k) \leq f(x^k) + \eta \alpha_k \nabla^T f(x^k) d^k.$$

Step 3 Update: Set $x^{k+1} = x^k + \alpha_k d^k$, $k = k + 1$.

Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$, f is bounded from bellow on the feasible set $S = \{x \in \mathbb{R}^n \mid l \leq x \leq u\}$ and $f \in C^1(S)$. Moreover, assume that the sequence of search directions $\{d^k\}_{k\in\mathbb{N}}$ is bounded. Then, either the Algorithm PG-LS terminates after a finite number of iterations $\bar{k}$ at a stationary point $x^{\bar{k}}$ of the problem (47) or every accumulation point of the sequence $\{x^k\}_{k\in\mathbb{N}}$ is a stationary point of the problem (47).

Penalty function

$$\min_{x \in S} f(x), \ S = \{x \in \mathbb{R}^n \mid h(x) = 0, \ g(x) \leq 0\}. \tag{51}$$

$$\min_{x \in \mathbb{R}^n} \Phi(x), \tag{52}$$

$$\Phi(x, \tau) = f(x) + \tau\rho(x), \tag{53}$$

- ρ - measure of constraint violation
- τ - penalty parameter

$$\rho(x) = 0 \quad \iff \quad x \in S. \tag{54}$$

- A sequence of penalty problems of the form

$$\min_{x \in \mathbb{R}^n} \Phi(x, \tau_k), \tag{55}$$

  are solved
- The sequence of penalty parameters tends to infinity, i.e.,

$$\lim_{k \to \infty} \tau_k = \infty. \tag{56}$$

Definition

The penalty function Φ is exact if there exists $\bar{\tau} > 0$ such that for all $\tau \geq \bar{\tau}$ any local solution of the problem (51) is a local minimizer of the penalty function $\Phi(x, \tau)$.

$$Q_1(x, \tau) = f(x) + \tau\left(\sum_{i=1}^{m} |h_i(x)| + \sum_{i=1}^{p} \max\{0, g_i(x)\}\right).$$

Quadratic penalty for equality constrained problems

$$\min_{h(x)=0} f(x). \tag{57}$$

$$Q(x,\tau) = f(x) + \frac{\tau}{2}(\sum_{i=1}^{m}(h_i(x))^2 \tag{58}$$

- Introducing slack variables for inequality constraints
- $$\min_{x\in S} f(x),\ S = \{x \in \mathbb{R}^n \mid h(x) = 0,\ g(x) \leq 0\}$$

$$\min_{y\in \tilde{S}} f(x),\ \tilde{S} = \{(x,s) \in \mathbb{R}^{n+p},\ h(x) = 0,\ g(x) + s = 0,\ s \geq 0\}$$

Algorithm QP

Step 0 Input parameters: Take $x^0 \in \mathbb{R}^n, \varepsilon_0 \geq 0, \tau_0 > 0, k = 0$.

Step 1 Initialization: $x^0_{start} = x^0$.

Step 2 Solve the subproblem $\min Q(x, \tau_k)$ approximately: Start with x^k_{start}, terminate when

$$\|\nabla_x Q(x^k, \tau_k)\| \leq \varepsilon_k. \tag{59}$$

Step 3 Update the penalty parameter: Choose $\tau_{k+1} > \tau_k$.

Step 4 Update the tolerance: Choose $\varepsilon_{k+1} \in [0, \varepsilon_k)$.

Step 5 Update the starting point: Set $x^{k+1}_{start} = x^k$ and $k = k + 1$. Go to Step 2.

Theorem

Suppose that $f, h \in C^1(\mathbb{R}^n)$ and that each x^k is the exact global minimizer of function $Q(x, \tau_k)$. Suppose that (56) holds. Then every accumulation point of the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by Algorithm 12.1 is a solution of the problem (57).

Inexact solution of subproblems

Theorem

Suppose that $f, h \in C^1(\mathbb{R}^n)$ *and that* $\lim_{k\to\infty} \varepsilon_k = 0$*. Suppose that (56) holds. Then every accumulation point* x^* *of the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *generated by Algorithm 12.1 at which LICQ holds is a KKT point of the problem (57). Moreover, Lagrange multipliers associated with* $x^* = \lim_{k\in K} x^k$ *are given by*

$$\lim_{k\in K} \tau_k h(x^k) = \mu^*. \tag{60}$$

- A. Friedlander, N. Krejić, N. Krklec Jerinkić, Fundamentals of Numerical Optimization, University of Novi Sad Faculty of Sciences, 2019, available at https://www.pmf.uns.ac.rs/studije/epublikacije/matinf/friedlander_krejic_krklecjerinkic_lectures_fundamentals_numerical_optimization.pdf
- S.Bellavia, T. Bianconcini, N. Krejić, B. Morini, Subsampled first-order optimization methods with applications in imaging, technical report
- Schmidt M., Le Roux N., Bach F., Minimizing Finite Sums with the Stochastic Average Gradient, Mathematical Programming 162, 1-2, (2017), 83-112.
- Defazio, A., Bach, F., Lacoste-Julien, S., SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives, Advances in Neural Information Processing Systems 27 (NIPS 2014),

- Babanezhad R., Ahmed M. O., Virani A., Schmidt M., Konečný J., Sallinen S., Stop wasting my gradients: Practical SVRG, Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 Pages 2251-2259 2015
- Bellavia, S., Krejić, N., Krklec Jerinkić, N., Subsampled Inexact Newton methods for minimizing large sums of convex function, IMA J. Numer. Anal, to appear, DOI: https://doi.org/10.1093/imanum/drz027
- Bellavia, S., Krejić, N., Morini, B., Inexact restoration with subsampled trust-region methods for finite-sum minimization, arXiv preprint, arXiv: 1902.01710,2019
- Bertsekas D.P., Tsitsiklis J.N., Gradient Convergence in Gradient Methods with Errors, SIAM J. Optimization 10(2) (2000), 627-642.

- Bottou, L., Curtis F.C., Nocedal, J. Optimization Methods for Large-Scale Machine Learning, SIAM Review, 60(2), (2018) 223-311.
- Delyon B., Juditsky A., Accelerated stochastic approximation, SIAM J. Optimization, Vol.3, No.4, 1993, pp. 868-881.
- Lichman M., UCI machine learning repository, `https://archive.ics.uci.edu/ml/index.php`, 2013.
- N. Krejić, Z. Lužanin, Z. Ovcin, I. Stojkovska, Descent direction method with line search for unconstrained optimization in noisy environment. Optimization Methods and Software 30(6): 1164-1184 (2015)
- Nocedal, J., Wright, S. J., Numerical Optimization, Springer Series in Operations Research, Springer, 1999.

- Nguyen, L.M., liu, J., Scheinberg, K., Takač, M., SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient, Proceedings of the 34th International Conference on Machine Learning - Volume 70 Pages 2613-2621
- Robbins H., Monro S., A Stochastic Approximation Method, The Annals of Mathematical Statistics, 22(3) (1951), 400-407.
- Tan, C., Ma, S., Dai, Y., Qian, Y., Barzilai-Bprwein step size for stochastic gradient descent, in Neural Information Processing Systems, pp. 685-693 (2016)
- Yousefian F., Nedic A., Shanbhag U.V., On stochastic gradient and subgradient methods with adaptive steplength sequences, Automatica 48 (1),2012, pp. 56-67.

- Wang, C., Chen, X., Smola, A., Xing, E., Variance Reduction for Stochastic Gradient Optimization, In Advances in Neural Information Processing Systems, 181-189, 2013.
- Johnson, R., Zhang, T., Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1 Pages 315-323, 2013
- L. N. Vicente, S. Gratton, and R. Garmanjani, Concise Lecture Notes on Optimization Methods for Machine Learning and Data Science, ISE Department, Lehigh University, January 2019.