Optimisation algorithms in Statistics II, lecture 1

Frank Miller, Department of Statistics; Stockholm University

March 23, 2021

Stockholm University

Course schedule

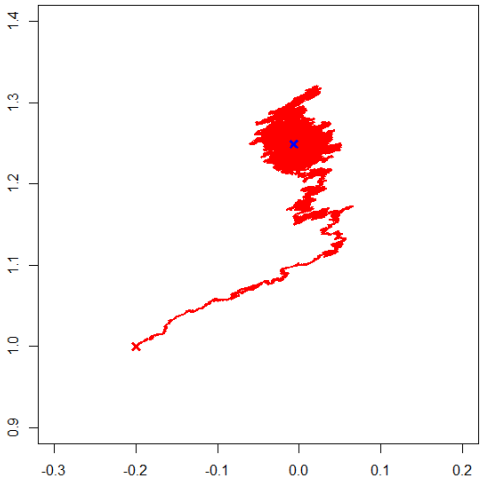
- Topic 1: **Stochastic gradient descent and quasi-Newton algorithms**
  Lectures: March 23; Time 10-12, 13-15 (online, Zoom)
- Topic 2: Lectures: April 13; Time 10-12, 13-15
- Topic 3: Lectures: April 27; Time 10-12, 13-15

Course homepage:
http://gauss.stat.su.se/phd/oasi/optimisation2.html

Includes reading material, lecture notes, assignments

Today's schedule

- Stochastic gradient descent (SGD) algorithms
  - Quick recap
  - Convergence analysis for SGD
- Exercise session
- Quasi-Newton method
  - Idea and definition
  - The BFGS method



Stockholm University

Multivariate optimisation – Gradient descent method

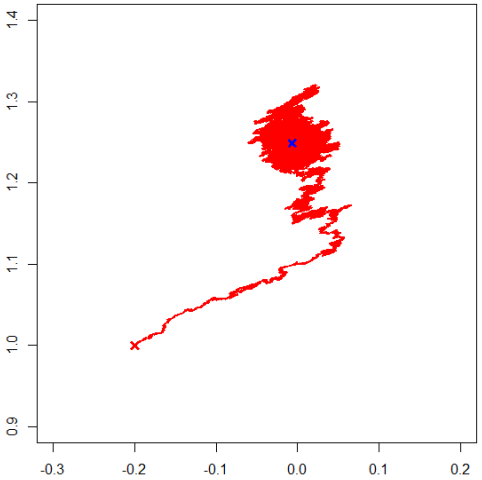
- Iteration: $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \, \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$

- Optimisation problem (finite sum case):
  - $\boldsymbol{x}$ p-dimensional vector, g_i: $\mathbb{R}^p \to \mathbb{R}$ functions
  - We search $\boldsymbol{x}^*$ with $g(\boldsymbol{x}^*) = \min g(\boldsymbol{x})$ where $g = \frac{1}{n}\sum_{i=1}^{n} g_i$

- If n large: Takes time to evaluate gradient $\boldsymbol{g}' = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}'_i$

Stochastic gradient descent method

- Iteration:
  - Choose $i \in \{1, \ldots, n\}$ randomly
  - $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \boldsymbol{g'}_i(\boldsymbol{x}^{(t)})$
- α_t is a predefined sequence, either
  - constant step size $\alpha_t = \alpha$ or
  - decreasing step size e.g. $\alpha_t = \alpha/t$



- Convergence (to a local minimum) can be shown if step size fulfils $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ (example: $\alpha_t = \alpha/t$)

  -> this claim will be made more specify shortly

Stockholm University

Stochastic gradient descent method

- Function to be **minimised**: $g = \frac{1}{n}\sum_{i=1}^{n} g_i$
- Predefined sequence of step sizes: α_t, $t = 1,2, \ldots$
- Starting value: $\boldsymbol{x}^{(\boldsymbol{0})}$
- Sequence of random numbers: $R^{(t)} \in \{1, \ldots, n\}, t = 1,2, \ldots$
- Iteration: $\boldsymbol{x}^{(\boldsymbol{t+1})} = \boldsymbol{x}^{(\boldsymbol{t})} - \alpha_t \boldsymbol{g'}_{R(t)}\left(\boldsymbol{x}^{(\boldsymbol{t})}\right)$

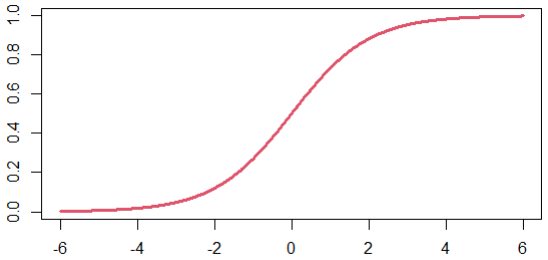
- Now: Convergence analysis
- We assume in the lecture:
  $R^{(t)}$ uniformly distributed on {1, …, n}, all $R^{(t)}$ independent
- We note that $\boldsymbol{E}\boldsymbol{g'}_{R(t)}\left(\boldsymbol{x}^{(\boldsymbol{t})}\right) = \boldsymbol{g'}\left(\boldsymbol{x}^{(\boldsymbol{t})}\right)$

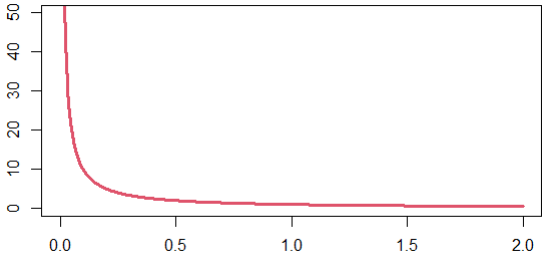
Lipschitz continuous functions

- A function f is called *Lipschitz continuous* with Lipschitz constant L>0, if for all $\mathbf{x}_1$, $\mathbf{x}_2$,

$$\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|_2 \leq L \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2.$$

- If $f: (a, b) \to \mathbb{R}$ is differentiable, the following is true: f Lipschitz continuous with constant L if and only if |f'(x)|≤L for all x
- Example: 1/(1+exp(-x)) is Lipschitz continuous with L=0.25
- Example: 1/x is not Lipschitz continuous on (0,∞)





Stockholm University

Lipschitz continuous functions

- A function f is called *Lipschitz continuous* with Lipschitz constant L>0, if for all $\mathbf{x}_1$, $\mathbf{x}_2$,

$$\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|_2 \leq L \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2.$$

- If f has a derivative (gradient) f′ which is Lipschitz continuous with L>0, then f itself is called *L-smooth*. Further,

$$f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \leq \boldsymbol{f}'(\boldsymbol{x}_2)^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \frac{L}{2} \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2.$$

- If f has a Hessian matrix **f′′** with a bounded spectral norm (by L), the gradient **f′** is Lipschitz continuous with L:

$$\|\boldsymbol{f}''(\boldsymbol{x})\|_{Spectral} \leq L \text{ for all } \boldsymbol{x} \Rightarrow \boldsymbol{f}' \text{Lipschitz continuous with } L$$

Matrix norms

- Most often when writing $\|\cdot\|$, we have a norm for a vector inside the norm-signs (and $\|\boldsymbol{x}\|_2$ can be interpreted as length of vector $\boldsymbol{x}$)
- There are also matrix-norms, and the spectral norm is one example: $\|\boldsymbol{A}\|_{Spectral} = \sqrt{\lambda_{max}(\boldsymbol{A}^T\boldsymbol{A})}$ where $\lambda_{max}(\cdot)$ is the largest eigenvalue of the matrix inside
- Spectral norm and Euclidian norm are *compatible* in the sense that for any $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{n}}$ and $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}$, we have

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{A}\|_{Spectral}\|\boldsymbol{x}\|_2$$

SGD's expected decrease per iteration

- Minimisation of $g = \frac{1}{n}\sum_{i=1}^{n} g_i$ with SGD

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \boldsymbol{g'}_{R(t)}\left(\boldsymbol{x}^{(t)}\right) \qquad \text{(SGD)}$$

- Lemma 1 (Bottou et al): Let g be L-smooth with L>0. Given $\mathbf{x}^{(t)}$, the expected decrease in an SGD iteration is bounded:

$$E\left[g\left(\boldsymbol{x}^{(t+1)}\right)\right] - g\left(\boldsymbol{x}^{(t)}\right) \leq -\alpha_t \left\|\boldsymbol{g'}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 + \alpha_t^2 \frac{L}{2} E\left\|\boldsymbol{g'}_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2$$

SGD's expected decrease per iteration

- Detailed proof of Lemma 1 (Bottou et al):
- We have:

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right) \qquad \text{(SGD)}$$

$$g(\boldsymbol{x}_1) - g(\boldsymbol{x}_2) \leq \boldsymbol{g}'(\boldsymbol{x}_2)^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \tfrac{L}{2} \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 \text{ for all } \boldsymbol{x}_1,\ \boldsymbol{x}_2 \qquad \text{(Lsmooth)}$$

R(t) uniformly distributed on {1, …, n} (R)

- Using (Lsmooth) for $\boldsymbol{x}_1 = \boldsymbol{x}^{(t+1)}$ and $\boldsymbol{x}_2 = \boldsymbol{x}^{(t)}$ (conditional on $R^{(t)}$ and $\boldsymbol{x}^{(t)}$),

$$g\left(\boldsymbol{x}^{(t+1)}\right) - g\left(\boldsymbol{x}^{(t)}\right) \leq \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)^T\left(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right) + \tfrac{L}{2} \cdot \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|_2^2$$

- Using (SGD),

$$= \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)^T\left(-\alpha_t \boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right) + \tfrac{L}{2} \cdot \left\|\alpha_t \boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2$$

$$= -\alpha_t \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)^T \boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right) + \alpha_t^2 \tfrac{L}{2} \cdot \left\|\boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2$$

- Take expectation over $R^{(t)}$ given $\mathbf{x}^{(t)}$

$$E\left[g\left(\boldsymbol{x}^{(t+1)}\right)\right] - g\left(\boldsymbol{x}^{(t)}\right) \leq -\alpha_t \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)^T E\left[\boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right] + \alpha_t^2 \tfrac{L}{2} E\left[\left\|\boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2\right]$$

$$= -\alpha_t \left\|\boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 + \alpha_t^2 \tfrac{L}{2} E\left\|\boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 \text{ since}$$

- $E\left[\boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right] = \frac{1}{n}\sum_{i=1}^n \boldsymbol{g}'_i\left(\boldsymbol{x}^{(t)}\right) = \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$ due to (R). □

SGD's expected decrease per iteration

- Minimisation of $g = \frac{1}{n}\sum_{i=1}^{n} g_i$ with SGD

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \boldsymbol{g'}_{R(t)}\left(\boldsymbol{x}^{(t)}\right) \qquad \text{(SGD)}$$

- Lemma 1 (Bottou et al): Let g be L-smooth with L>0. Given $\mathbf{x}^{(t)}$, the expected decrease in an SGD iteration is bounded:

$$E\left[g\left(\boldsymbol{x}^{(t+1)}\right)\right] - g\left(\boldsymbol{x}^{(t)}\right) \leq -\alpha_t \left\|\boldsymbol{g'}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 + \alpha_t^2 \frac{L}{2} E\left\|\boldsymbol{g'}_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2$$

- Proof idea: We apply the **consequence of L-smoothness** for $\boldsymbol{x}_1 = \boldsymbol{x}^{(t+1)}$ and $\boldsymbol{x}_2 = \boldsymbol{x}^{(t)}$ (conditional on $R^{(t)}$ and $\boldsymbol{x}^{(t)}$),

$$g\left(\boldsymbol{x}^{(t+1)}\right) - g\left(\boldsymbol{x}^{(t)}\right) \leq \boldsymbol{g'}\left(\boldsymbol{x}^{(t)}\right)^T \left(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right) + \frac{L}{2} \cdot \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right\|_2^2$$

We **use Equation (SGD)** above, **take expectation over $\mathbf{R^{(t)}}$** (given $\mathbf{x}^{(t)}$ or the history $R^{(t-1)}$, $R^{(t-2)}$, …) and **replace $\boldsymbol{E g'}_{\boldsymbol{R(t)}}\left(\boldsymbol{x}^{(t)}\right)$ by $\boldsymbol{g'}\left(\boldsymbol{x}^{(t)}\right)$**. This shows the claim. □

SGD's expected decrease per iteration

- Minimisation of $g = \frac{1}{n}\sum_{i=1}^{n} g_i$ with SGD

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)$$

- Lemma 2 (Bottou et al): Let g be L-smooth with L>0 and we have following second moment condition:

$$E\left\|\boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 \leq s + w\left\|\boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 \text{ for all } t.$$

Given $\mathbf{x}^{(t)}$, the expected decrease in an SGD iteration is bounded:

$$E\left[g\left(\boldsymbol{x}^{(t+1)}\right)\right] - g\left(\boldsymbol{x}^{(t)}\right) \leq -\alpha_t(1 - \alpha_t Lw/2)\left\|\boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 + \alpha_t^2 \frac{Ls}{2}$$

- Proof: Follows directly from Lemma 1:
- $E\left[g\left(\boldsymbol{x}^{(t+1)}\right)\right] - g\left(\boldsymbol{x}^{(t)}\right) \leq -\alpha_t\left\|\boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 + \alpha_t^2 \frac{L}{2} E\left\|\boldsymbol{g}'_{R(t)}\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2$

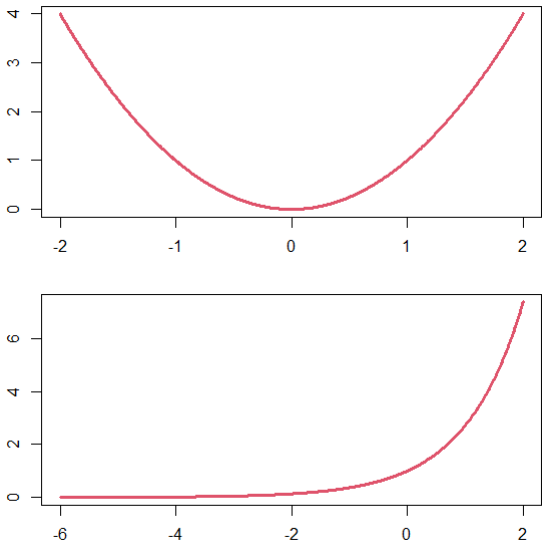
$$\leq -\alpha_t\left\|\boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2 + \alpha_t^2 \frac{L}{2}\left(s + w\left\|\boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)\right\|_2^2\right)$$

□

Strongly convex functions

$$\|y\|_2 = \sqrt{y^2} = |y|$$
$$\|y\|_2^2 = y^2$$
Set $y = x_1 - x_2$:
$$\|x_1 - x_2\|_2^2 = (x_1 - x_2)^2$$

- A differentiable function f is called *m-strongly convex* with m>0, if for all $\mathbf{x}_1$, $\mathbf{x}_2$,
$$(\boldsymbol{f}'(\boldsymbol{x}_1) - \boldsymbol{f}'(\boldsymbol{x}_2))^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) \geq m \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2.$$
- For one-dimensional functions:
$(f'(x_1) - f'(x_2))/(x_1 - x_2) \geq m$ for all x_1, x_2.
- The function $f(x)=x^2$ is m-strongly convex with m=2
- The function f(x)=exp(x) is convex but not m-strongly convex since for $x \to -\infty$, smaller and smaller m would be necessary; no m>0 can be found to fulfil condition above



Strongly convex functions

- A differentiable function f is called *m-strongly convex* with m>0, if for all $\mathbf{x}_1$, $\mathbf{x}_2$,

$$(\boldsymbol{f}'(\boldsymbol{x}_1) - \boldsymbol{f}'(\boldsymbol{x}_2))^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) \geq m \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2.$$

- An equivalent condition is

$$(f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)) \geq \boldsymbol{f}'(\boldsymbol{x}_2)^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) + \frac{m}{2} \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2.$$

- An m-strongly convex function f has a unique minimum $\mathbf{x}^*$ and the following holds true:

$$2m\big(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\big) \leq \|\boldsymbol{f}'(\boldsymbol{x})\|_2^2.$$

Assumptions

- Optimisation problem (finite sum case):
  - $\boldsymbol{x}$ p-dimensional vector, g_i: $\mathbb{R}^p \to \mathbb{R}$ functions
  - We search $\boldsymbol{x}^*$ with $g(\boldsymbol{x}^*) = \min g(\boldsymbol{x})$ where $g = \frac{1}{n}\sum_{i=1}^{n} g_i$

- Assumptions (A):
  - g is differentiable and L-smooth with L>0,
  - g is m-strongly convex with m>0
  - For all **x**: $\mathrm{E}\left\|\boldsymbol{g}'_{R(t)}(\boldsymbol{x})\right\|_2^2 \leq s + w\|\boldsymbol{g}'(\boldsymbol{x})\|_2^2$
- Assumptions (B):
  - g_i are differentiable and L-smooth with $L_i > 0$,
  - g is m-strongly convex with m>0
  - $\mathrm{E}\left\|\boldsymbol{g}'_{R(t)}(\boldsymbol{x}^*)\right\|_2^2 = s$

Convergence analysis for fixed step size

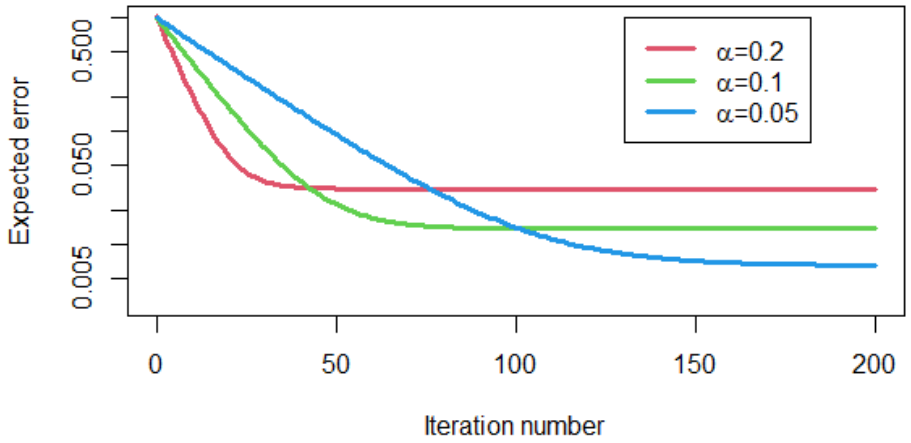
- Theorem 1 (Bottou et al): Consider the finite sum case of the optimization problem, assume Assumptions (A) and that the step size is constant, $\alpha_t = \alpha \leq 1/\{L \max(w, 1)\}$. Then, we have the following convergence result:

$$\mathrm{E}\left[g\left(\boldsymbol{x}^{(t)}\right)\right] - g(\boldsymbol{x}^*) \leq \frac{\alpha L s}{2m} + (1 - \alpha m)^t \{g\left(\boldsymbol{x}^{(0)}\right) - g(\boldsymbol{x}^*) - \frac{\alpha L s}{2m}\}$$

- Proof: Based on Lemma 2, see Bottou et al (2018), https://arxiv.org/pdf/1606.04838.pdf

Convergence analysis for fixed step size

- Theorem 2 (Needell et al): Consider the finite sum case of the optimization problem, assume Assumptions (B) and that the step size is constant, $\alpha_t = \alpha \leq \frac{1}{\max(L_i)}$. Then, we have the following convergence result:
  $\mathrm{E}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|_2^2 \leq \frac{\alpha s}{m\{1-\alpha \max(L_i)\}} + (1 - \alpha m\{1 - \alpha \max(L_i)\})^t \left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|_2^2$
- Proof: See Needell et al (2016), here an arXiv version: https://arxiv.org/abs/1310.5715

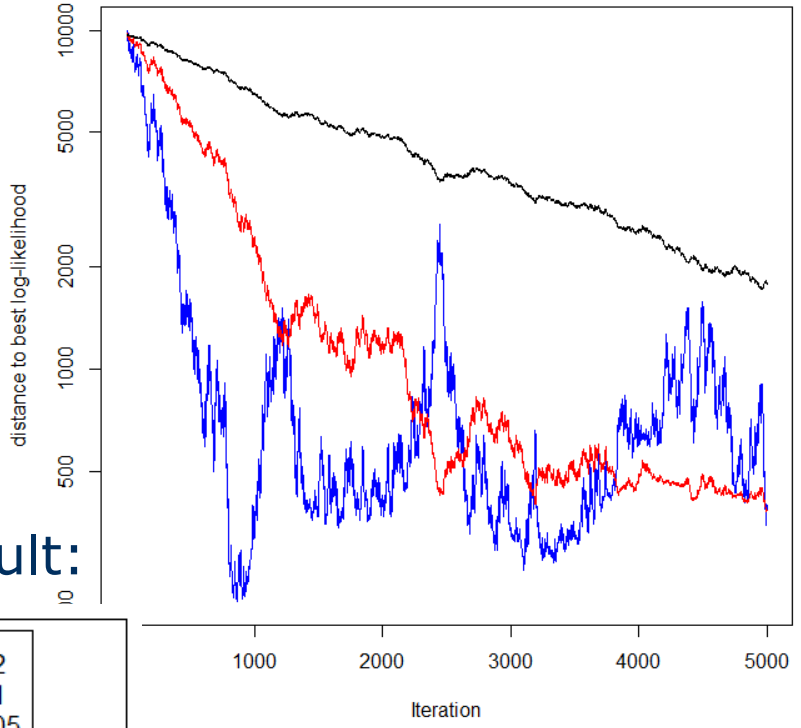


Theoretical behaviour of above bound for expected distance to optimum for s=0.5, m=2, $\max(L_i)=2$, $\varepsilon_0 = \left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|_2^2 = 1$

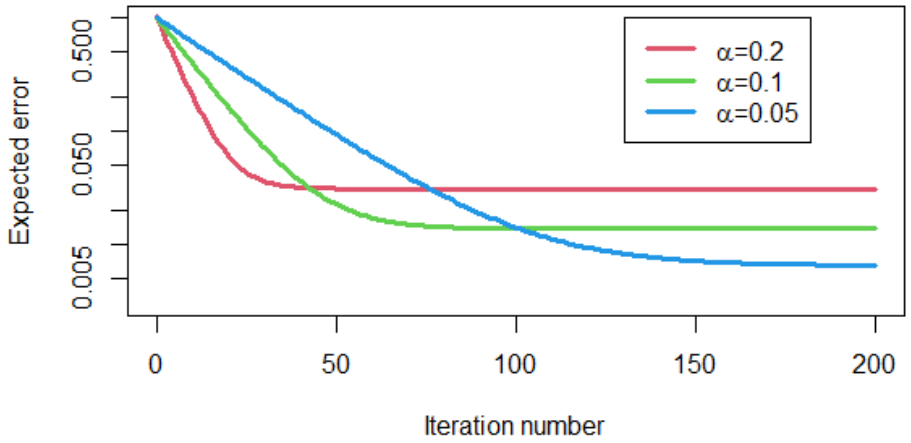
Stockholm University

Stochastic gradient descent method – empirical examples from OASI-L2

- Constant step size $\alpha^{(t)} = \alpha$
- Step size
  - $\alpha = 0.0006$ (black)
  - $\alpha = 0.002$ (red)
  - $\alpha = 0.006$ (blue)



- Compare with theoretical result:

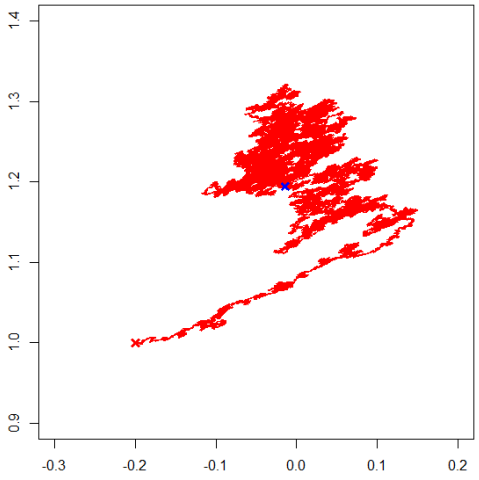


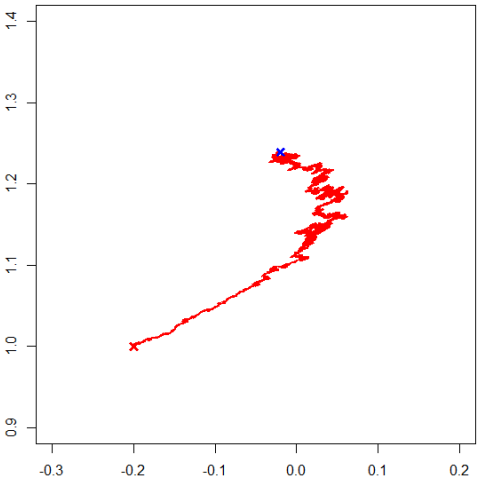
Theoretical behaviour of bound in Theorem 2 for expected distance to optimum for s=0.5, m=2, max(L_i)=2, ε_0=1

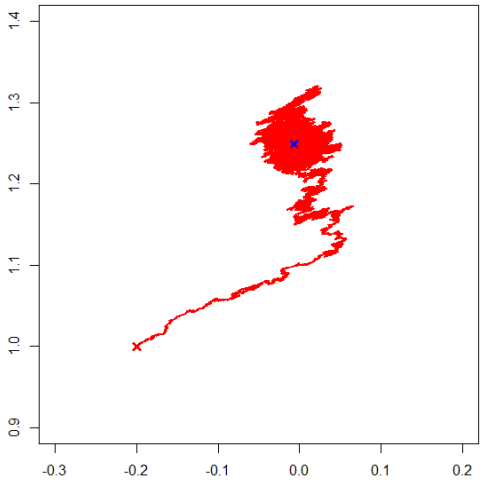
Stockholm University

Stochastic gradient descent method – empirical examples from OASI-L2

- Problem 1.3 with n= 1 000 000 (data duplicated 100 000 x); start at $(\beta_0, \beta_1) = (-0.2, 1)$; best $(\hat{\beta}_0, \hat{\beta}_1) = (-0.009, 1.263)$
- 100 000 iterations,
  $\alpha = 0.002$,
  $\alpha = 0.0006$
  1 000 000 iterations,
  $\alpha = 0.0006$







- (-0.014, 1.195) (-0.020, 1.239) (-0.007, 1.248)

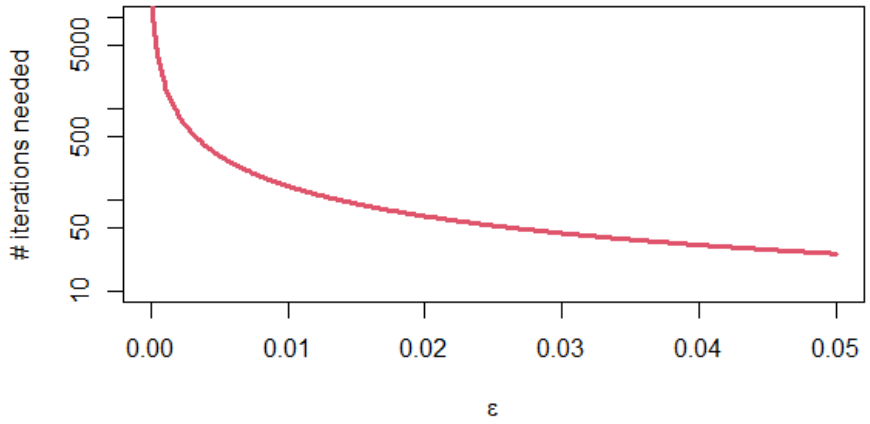
Stockholm University

Convergence analysis for fixed step size

- Corollary: Consider the finite sum case of the optimization problem, assume Assumptions (B) and set $\varepsilon_0 = \left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|_2^2$. For any desired ε>0, using constant step size, $\alpha_t = \alpha = \frac{m\varepsilon}{2m\epsilon \max(L_i)+2s}$, we have after

$$t = 2\log(2\varepsilon_0/\varepsilon)\left(\frac{\max(L_i)}{m} + \frac{s}{m^2\varepsilon}\right)$$

iterations, $\mathrm{E}\left[\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|_2^2\right] \leq \varepsilon$.



Number of iterations needed to bring expected error below ε for s=0.5, m=2, max(L_i)=2, ε_0=1

eps	# iter	alpha
0.00001	305177	0.00002
0.00002	143935	0.00004
0.00004	67646	0.00008
0.00008	31666	0.00016
0.00016	14759	0.00032
0.00032	6846	0.00064
0.00064	3160	0.00127
0.00128	1452	0.00253
0.00256	664	0.00502

Convergence analysis for decreasing step size

- Theorem 3 (Bottou et al): Consider the finite sum case of the optimization problem, assume Assumptions (A) and that the step size is decreasing as $\alpha_t = \frac{\beta}{t+\gamma}$ with $\beta > \frac{1}{m}$, $\gamma > 0$, $\alpha_0 \leq 1/\{L \max(w, 1)\}$. Then, we have the following convergence result:

$$\mathrm{E}\left[g\left(\boldsymbol{x}^{(t)}\right)\right] - g(\boldsymbol{x}^*) \leq \nu/(\gamma + t),$$

where

$$\nu = \max\left\{\frac{\beta^2 L s}{2(\beta m - 1)}, (\gamma + 1)(g\left(\boldsymbol{x}^{(0)}\right) - g(\boldsymbol{x}^*))\right\}.$$

- Proof: See Bottou et al (2018), https://arxiv.org/pdf/1606.04838.pdf

Convergence analysis for decreasing step size

- Note that

$$\mathrm{E}\left[g\left(\boldsymbol{x}^{(t)}\right)\right] - g(\boldsymbol{x}^*) \approx \nu/(\gamma + t),$$

means sublinear convergence since

$$\frac{\left\{\mathrm{E}\left[g\left(\boldsymbol{x}^{(t+1)}\right)\right] - g(\boldsymbol{x}^*)\right\}}{\left\{\mathrm{E}\left[g\left(\boldsymbol{x}^{(t)}\right)\right] - g(\boldsymbol{x}^*)\right\}} \approx \frac{\gamma+t}{\gamma+t+1} \to 1 \ \ (\text{for } t \to \infty)$$

- A bound like

$$\mathrm{E}\left[g\left(\boldsymbol{x}^{(t)}\right)\right] - g(\boldsymbol{x}^*) \leq \nu^t \ \text{ with } 0 < \nu < 1$$

would lead to linear convergence

- So, SGD with $\alpha_t = \frac{\beta}{t+\gamma}$ gives only sublinear convergence
- But in reality, we never go to infinity with t…

SGD convergence analysis – exercise

Optimisation in a least squares situation,

- $g(\boldsymbol{b}) = \frac{1}{n}\|\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{b})$ with $g_i(\boldsymbol{b}) = \left(\boldsymbol{x}_i^T\boldsymbol{b} - y_i\right)^2$
- $\boldsymbol{g}'(\boldsymbol{b}) = \frac{2}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i{}'(\boldsymbol{b})$ with $\boldsymbol{g}_i'(\boldsymbol{b}) = 2(\boldsymbol{x}_i^T\boldsymbol{b} - y_i)\boldsymbol{x}_i$
- $\boldsymbol{g}''(\boldsymbol{b}) = \frac{2}{n}\boldsymbol{X}^T\boldsymbol{X}$

R uniformly distributed on {1, …, n}
Compute for (i) general $\boldsymbol{x}_i$, (ii) $\boldsymbol{x}_i = \begin{pmatrix} 1 \\ w_i \end{pmatrix}$ (straight line regr.):

a) $\|\boldsymbol{g}_i{}'(\boldsymbol{b})\|_2^2 = 4\left(\boldsymbol{x}_i^T\boldsymbol{b} - y_i\right)^2\boldsymbol{x}_i^T\boldsymbol{x}_i = 4\,(b_1 + b_2 w_i - y_i)^2(1 + w_i^2)$

b) $\mathrm{E}\|\boldsymbol{g}_R{}'(\boldsymbol{b})\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i{}'(\boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n} 4\left(\boldsymbol{x}_i^T\boldsymbol{b} - y_i\right)^2\boldsymbol{x}_i^T\boldsymbol{x}_i = \cdots$

Compute for general $\boldsymbol{x}_i, \boldsymbol{X}$:

c) $\mathrm{E}[\boldsymbol{g}_R'(\boldsymbol{b})] = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i{}'(\boldsymbol{b}) = \boldsymbol{g}'(\boldsymbol{b})$

d) $\|\boldsymbol{g}''(\boldsymbol{b})\|_{Spectral} = \frac{2}{n}\sqrt{\lambda_{\max}\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^T(\boldsymbol{X}^T\boldsymbol{X})\right)} = \frac{2}{n}\,\lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})$

e) $\|\boldsymbol{g}'(\boldsymbol{b})\|_2^2 = \left(\frac{2}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y})\right)^T \frac{2}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y})$

$= \frac{4}{n^2}(\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y}) = \cdots$

Quasi-Newton method

Quasi-Newton method

- Steepest descent and Newton method have iteration
$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \left(\boldsymbol{M}^{(t)}\right)^{-1} \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$$
with $\boldsymbol{M}^{(t)} = \boldsymbol{g}''\left(\boldsymbol{x}^{(t)}\right)$ for the Newton method and with $\left(\boldsymbol{M}^{(t)}\right)^{-1} = \alpha_t \boldsymbol{I}$ for the steepest descent method
- A disadvantage of Newton is the need to calculate the Hessian $\boldsymbol{g}''\left(\boldsymbol{x}^{(t)}\right)$ in each iteration
- A disadvantage of steepest descent is that no information about the curvature is used
- We can monitor the computed gradients $\boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$ and their change gives information about the curvature of g

Quasi-Newton method

- Steepest descent and Newton method have iteration
$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \left(\boldsymbol{M}^{(t)}\right)^{-1} \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$$
- Newton ($\boldsymbol{M}^{(t)} = \boldsymbol{g}''\left(\boldsymbol{x}^{(t)}\right)$) was motivated with the multidimensional Taylor expansion
$$\boldsymbol{g}'(\boldsymbol{x}^*) \approx \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right) + \boldsymbol{g}''\left(\boldsymbol{x}^{(t)}\right)\left(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right)$$
or
$$\boldsymbol{g}'(\boldsymbol{x}^*) - \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right) \approx \boldsymbol{g}''\left(\boldsymbol{x}^{(t)}\right)\left(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right)$$
- We want to use approximations $\boldsymbol{M}^{(t+1)}$ to $\boldsymbol{g}''\left(\boldsymbol{x}^{(t)}\right)$ which fulfil this relation when $\boldsymbol{x}^*$ is replaced by $\boldsymbol{x}^{(t+1)}$:
$$\boldsymbol{g}'\left(\boldsymbol{x}^{(t+1)}\right) - \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right) = \boldsymbol{M}^{(t+1)}\left(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right)$$
- This condition is called secant condition
- There are multiple solutions to the secant condition

Quasi-Newton method

- Steepest descent and Newton method have iteration

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \left(\boldsymbol{M}^{(t)}\right)^{-1} \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$$

- Secant condition:

$$\boldsymbol{g}'\left(\boldsymbol{x}^{(t+1)}\right) - \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right) = \boldsymbol{M}^{(t+1)}\left(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right)$$

- Or, with $\boldsymbol{y}^{(t)} = \boldsymbol{g}'\left(\boldsymbol{x}^{(t+1)}\right) - \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$ and $\boldsymbol{z}^{(t)} = \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}$:

$$\boldsymbol{y}^{(t)} = \boldsymbol{M}^{(t+1)}\boldsymbol{z}^{(t)}$$

- Suggestion from Broyden, Fletcher, Goldfarb, and Shanno (BFGS; 4 publications in 1970) fulfilling secant condition:

$$\boldsymbol{M}^{(t+1)} = \boldsymbol{M}^{(t)} - \frac{\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}\left(\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}\right)^T}{\boldsymbol{z}^{(t)^T}\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}} + \frac{\boldsymbol{y}^{(t)}\boldsymbol{y}^{(t)^T}}{\boldsymbol{y}^{(t)^T}\boldsymbol{z}^{(t)}}$$

Quasi-Newton method

- The BFGS (quasi-Newton) method has iteration

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \left(\boldsymbol{M}^{(t)}\right)^{-1} \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$$

  and

$$\boldsymbol{M}^{(t+1)} = \boldsymbol{M}^{(t)} - \frac{\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}\left(\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}\right)^T}{\boldsymbol{z}^{(t)^T}\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}} + \frac{\boldsymbol{y}^{(t)}\boldsymbol{y}^{(t)^T}}{\boldsymbol{y}^{(t)^T}\boldsymbol{z}^{(t)}}$$

- Descent is not ensured but stepsize-halving ("backtracking") can be used as for steepest descent to ensure it:

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t\left(\boldsymbol{M}^{(t)}\right)^{-1} \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$$

- The `R` function `optim` includes the quasi-Newton BFGS
- Convergence of quasi-Newton methods is faster than linear but slower than quadratic (some assumptions necessary; see e.g. Nocedal and Wright, 2006, Theorem 3.7)

$$\lim_{t\to\infty} \frac{\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|} = 0, \qquad \frac{\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2} \text{ diverging}$$

Quasi-Newton method

- The BFGS (quasi-Newton) method has iteration

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \left(\boldsymbol{M}^{(t)}\right)^{-1} \boldsymbol{g}'\left(\boldsymbol{x}^{(t)}\right)$$

  and

$$\boldsymbol{M}^{(t+1)} = \boldsymbol{M}^{(t)} - \frac{\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}\left(\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}\right)^T}{\boldsymbol{z}^{(t)^T}\boldsymbol{M}^{(t)}\boldsymbol{z}^{(t)}} + \frac{\boldsymbol{y}^{(t)}\boldsymbol{y}^{(t)^T}}{\boldsymbol{y}^{(t)^T}\boldsymbol{z}^{(t)}}$$

- For higher dimensional problems, computation of inverse could be computationally intensive, but can be avoided by applying the Sherman-Morrison-Woodbury formula
- Defining $\boldsymbol{L}^{(t)} = \left(\boldsymbol{M}^{(t)}\right)^{-1}$, the matrix update can be done by

$$\boldsymbol{L}^{(t+1)} = \left(\boldsymbol{I} - r_k \boldsymbol{z}^{(t)}\left(\boldsymbol{y}^{(t)}\right)^T\right)\boldsymbol{L}^{(t)}\left(\boldsymbol{I} - r_k \boldsymbol{y}^{(t)}\left(\boldsymbol{z}^{(t)}\right)^T\right) + r_k \boldsymbol{z}^{(t)}\boldsymbol{z}^{(t)^T}$$

  with $r_k = 1/(\boldsymbol{y}^{(t)^T}\boldsymbol{z}^{(t)})$
- Starting value could be $\boldsymbol{L}^{(0)} = \boldsymbol{M}^{(0)} = \boldsymbol{I}$

Quasi-Newton method

- Convergence can be extremely fast in praxis
- Example from Nocedal and Wright (2006), chapter 6: Rosenbrock function $g(\boldsymbol{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$, starting point (-1.2, 1), optimum at (1,1).
  #iterations until error < 10^{-5}:
  - Steepest descent 5264
  - BFGS 34
  - Newton 21

Quasi-Newton method

- In high dimensional problems, storage and update of $\boldsymbol{M}^{(t)}$ might be problematic
- L-BFGS modification deals with it (saves not $\boldsymbol{M}^{(t)}$ but sequence of $\boldsymbol{y}^{(t)}$ and $\boldsymbol{z}^{(t)}$)
- Simple box-constraints are handled with BFGS-B

- Together: L-BFGS-B which is option in **`optim`**

Convergence order for deterministic algorithms

- Recall: Convergence order and convergence rate

$$\frac{\{\mathrm{E}[g(\boldsymbol{x}^{(t+1)})] - g(\boldsymbol{x}^*)\}}{\{\mathrm{E}[g(\boldsymbol{x}^{(t)})] - g(\boldsymbol{x}^*)\}^p} \to c \ \ (\text{for } t \to \infty)$$

- p is convergence order (p=1, 0<c<1 linear; p=2, 0<c<1 quadratic)
- c is convergence rate
- Under certain assumption, we have following orders:

unidimensional	**Bisection** order = roughly 1*	**Secant** order=$(1+\sqrt{5})/2$	**Newton** order = 2
multidimensional	**Steepest descent** order = 1	**Quasi-Newton** order > 1**	**Newton** order = 2

*strictly, the above criterion cannot be proven for bisection
**criterion above fulfilled for p=1 and c=0; “superlinear”

Assignments

- Topic 1: March 23 until April 12
- Topic 2: April 13 until April 26
- Topic 3: April 27 until Mai 11
- Second chance for Topic 1-3: until **August 31 (no extension!)**
- I might collect all submissions which have missed the first deadline and look at them in September