Modern Optimization Techniques

2. Unconstrained Optimization / 2.5. Subgradient Methods

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute of Computer Science
University of Hildesheim, Germany

Syllabus

Mon. 30.10.	(0)	0. Overview
		1. Theory
Mon. 6.11.	(1)	1. Convex Sets and Functions
		2. Unconstrained Optimization
Mon. 13.11.	(2)	2.1 Gradient Descent
Mon. 20.11.	(3)	2.2 Stochastic Gradient Descent
Mon. 27.11.	(4)	2.3 Newton's Method
Mon. 4.12.	(5)	2.4 Quasi-Newton Methods
Mon. 11.12.	(6)	2.5 Subgradient Methods
Mon. 18.12.	(7)	2.6 Coordinate Descent
	—	— *Christmas Break* —
		3. Equality Constrained Optimization
Mon. 8.1.	(8)	3.1 Duality
Mon. 15.1.	(9)	3.2 Methods
		4. Inequality Constrained Optimization
Mon. 22.1.	(10)	4.1 Primal Methods
Mon. 29.1.	(11)	4.2 Barrier and Penalty Methods
Mon. 5.2.	(12)	4.3 Cutting Plane Methods

Outline

Outline

Motivation

- If a function is once differentiable
  we can optimize it using
  - Gradient Descent,
  - Stochastic Gradient Descent,
  - Quasi-Newton Methods

  (1st order information)

- If a function is twice differentiable
  we can optimize it using
  - Newton's method

  (2nd order information)

- What if the objective function is not differentiable?

1st-Order Condition for Convexity

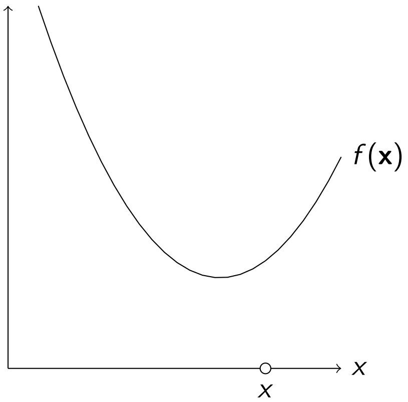
1st-order condition: a differentiable function f is convex iff

- dom f is a convex set and
- for all $\mathbf{x}, \mathbf{y} \in \text{dom } f$

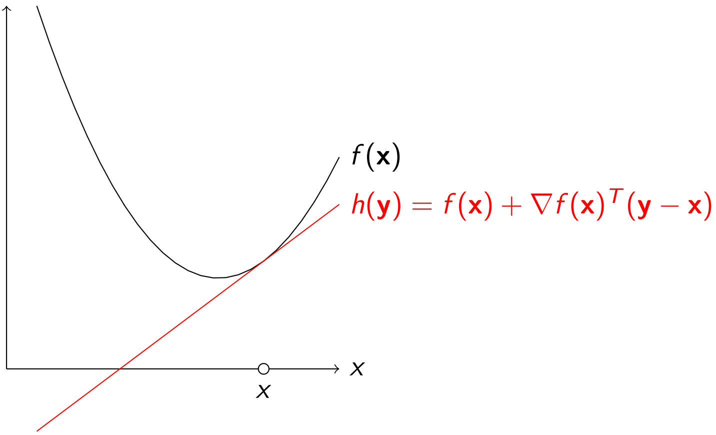
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$$

  - i.e., the tangent (= first order Taylor approximation) of f at $\mathbf{x}$ is a global underestimator

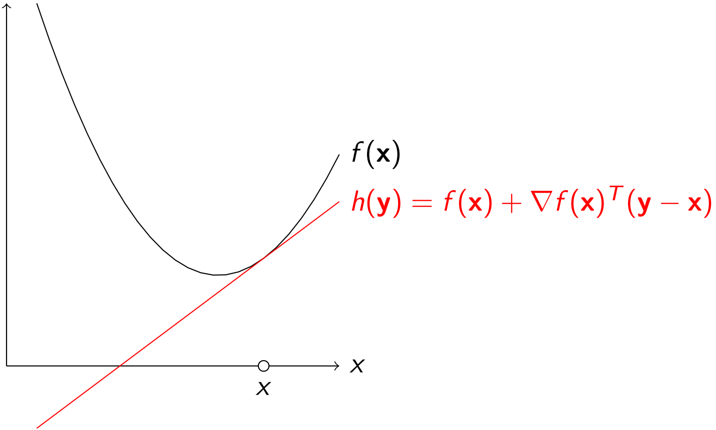
Tangent as a global underestimator



Tangent as a global underestimator



Tangent as a global underestimator

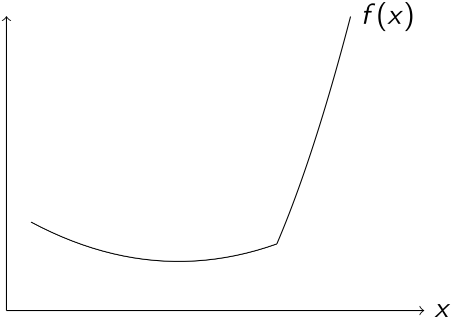


What happens if f is not differentiable?

Subgradient

Given a function f and a point $\mathbf{x} \in \operatorname{dom} f$,
$\mathbf{g} \in \mathbb{R}^n$ is called a **subgradient** of f at $\mathbf{x}$ if:
the hypersurface with slopes $\mathbf{g}$ through $(\mathbf{x}, f(\mathbf{x}))$ is a global underestimator of f, i.e.

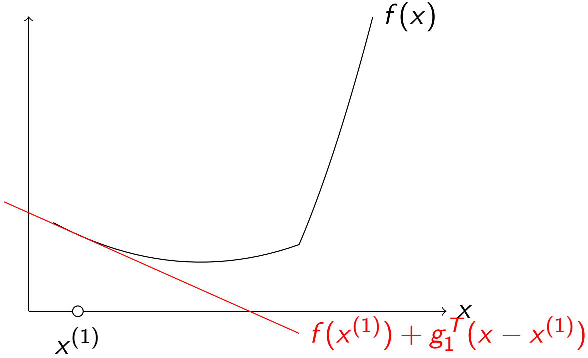
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}), \quad \text{for all } \mathbf{y} \in \operatorname{dom} f$$



Subgradient

Given a function f and a point $\mathbf{x} \in \text{dom}\, f$,
$\mathbf{g} \in \mathbb{R}^n$ is called a **subgradient** of f at $\mathbf{x}$ if:
the hypersurface with slopes $\mathbf{g}$ through $(\mathbf{x}, f(\mathbf{x}))$ is a global underestimator of f, i.e.

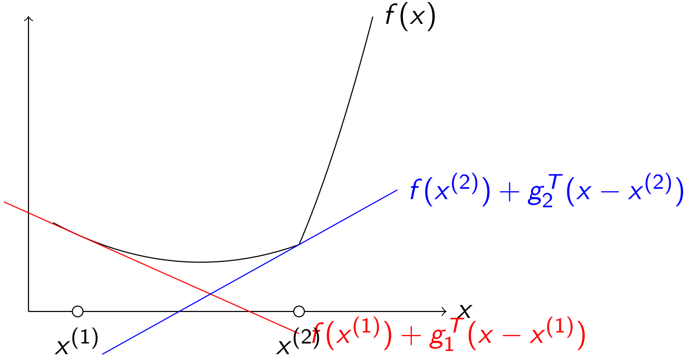
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}), \quad \text{for all } \mathbf{y} \in \text{dom}\, f$$



Subgradient

Given a function f and a point $\mathbf{x} \in \operatorname{dom} f$,
$\mathbf{g} \in \mathbb{R}^n$ is called a **subgradient** of f at $\mathbf{x}$ if:
the hypersurface with slopes $\mathbf{g}$ through $(\mathbf{x}, f(\mathbf{x}))$ is a global underestimator of f, i.e.

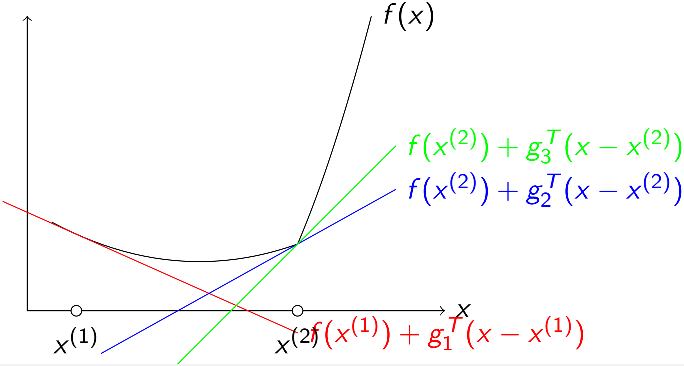
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}), \quad \text{for all } \mathbf{y} \in \operatorname{dom} f$$



Subgradient

Given a function f and a point $\mathbf{x} \in \operatorname{dom} f$,
$\mathbf{g} \in \mathbb{R}^n$ is called a **subgradient** of f at $\mathbf{x}$ if:
the hypersurface with slopes $\mathbf{g}$ through $(\mathbf{x}, f(\mathbf{x}))$ is a global underestimator of f, i.e.

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}), \quad \text{for all } \mathbf{y} \in \operatorname{dom} f$$



Subgradient

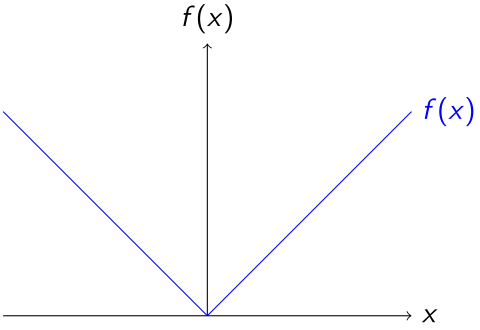
In the last example,

- $\mathbf{g_1}$ is a subgradient of f at $x^{(1)}$
- g_2 and g_3 are subgradients of f at $x^{(2)}$

Example

For $f : \mathbb{R} \to \mathbb{R}$ and $f(x) = |x|$:

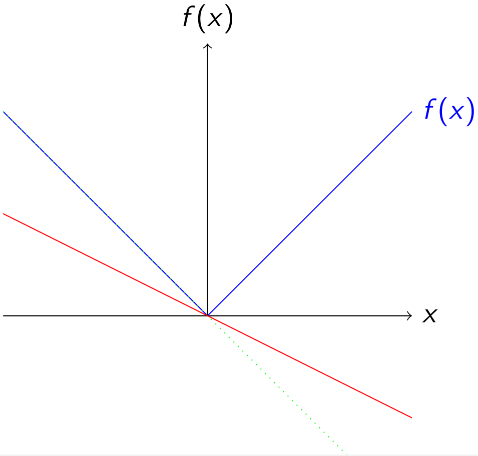
- For $x \neq 0$ there is one subgradient $g = \nabla f(x) = \text{sign}(x)$
- For $x = 0$ the subgradient is $g \in [-1, 1]$



Example

For $f : \mathbb{R} \to \mathbb{R}$ and $f(x) = |x|$:

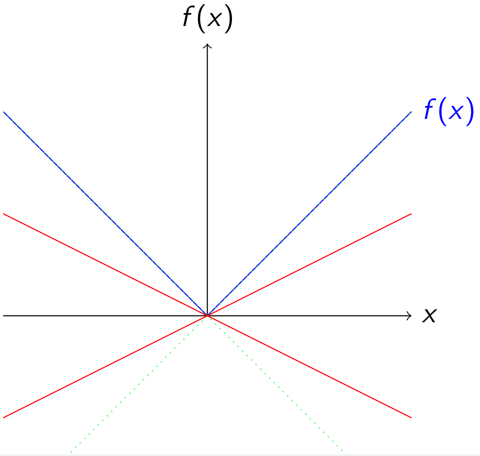
- For $x \neq 0$ there is one subgradient $g = \nabla f(x) = \text{sign}(x)$
- For $x = 0$ the subgradient is $g \in [-1, 1]$



Example

For $f : \mathbb{R} \to \mathbb{R}$ and $f(x) = |x|$:

- For $x \neq 0$ there is one subgradient $g = \nabla f(x) = \text{sign}(x)$
- For $x = 0$ the subgradient is $g \in [-1, 1]$



Subdifferential

Subdifferential $\partial f(\mathbf{x})$: set of all subgradients of f at $\mathbf{x}$

$$\partial f(\mathbf{x}) := \{\mathbf{g} \in \mathbb{R}^n \mid f(\mathbf{y}) \geq f(\mathbf{x}) + g^T(\mathbf{y} - \mathbf{x}) \; \forall \mathbf{y} \in \operatorname{dom} f\}$$

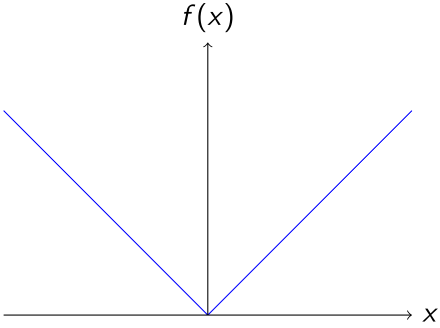
For a **convex** function f:

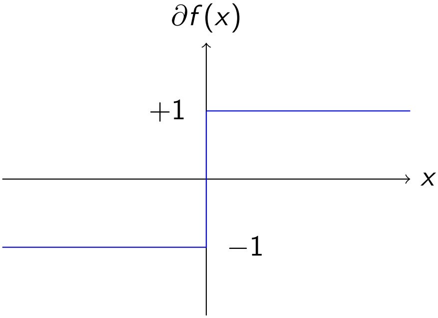
- subgradients always exist: $\partial f(\mathbf{x}) \neq \emptyset$
- f is differentiable at x
  iff the subdifferential contains a single element (the gradient)

$$f \text{ differentiable at } x \Longleftrightarrow \partial f(x) = \{\nabla f(x)\}$$

Example

For $f(x) = |x|$:





Subdifferential

For a **non-convex** function f:

- subgradients make less sense
  - see generalized subgradients, defined on local information

Outline

Subgradient Calculus

Assume f convex and $\mathbf{x} \in \operatorname{dom} f$

Some algorithms require only **one** subgradient for optimizing nondifferentiable functions f

Other algorithms, and optimality conditions require the *whole* subdifferential at $\mathbf{x}$

Tools for finding subgradients:

- **Weak subgradient calculus**: finding *one* subgradient $\mathbf{g} \in \partial f(\mathbf{x})$
- **Strong subgradient calculus**: finding the *whole* subdifferential $\partial f(\mathbf{x})$

Subgradient Calculus

We know that if f is differentiable at $\mathbf{x}$ then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$
There are a couple of additional rules:

- **Scaling**: for $a > 0$: $\partial(a \cdot f) = \{a \cdot \mathbf{g} \mid \mathbf{g} \in \partial(f)\}$
- **Addition**: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- **Affine composition**: for $h(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$ then

$$\partial h(\mathbf{x}) = A^T \partial f(A\mathbf{x} + \mathbf{b})$$

- **Finite pointwise maximum**: if $f(\mathbf{x}) = \max_{m=1\,\ldots,M} f_m(\mathbf{x})$ then

$$\partial f(\mathbf{x}) = \text{conv} \bigcup_{m: f_m(\mathbf{x}) = f(\mathbf{x})} \partial f_m(\mathbf{x})$$

  the subdifferential is the convex hull of the union of subdifferentials of all active functions at $\mathbf{x}$

Subgradients / More Examples

$$f(x) := ||x||_2$$

$$\partial f(x) =$$

Subgradients / More Examples

$$f(x) := ||x||_2$$

$$\partial f(x) = \begin{cases} \{\frac{x}{||x||_2}\}, & \text{if } x \neq 0_N \\ \{g \in \mathbb{R}^N \mid ||g||_2 \leq 1\}. & \text{if } x = 0_N \end{cases}$$

proof:

$$\begin{aligned} \partial(||x||_2) &= \partial(\max_{z:||z||_2\leq 1} z^T x) \\ &= \text{conv} \bigcup_{z:||z||_2\leq 1, z^T x \text{ max.}} z, \quad \text{for } x = 0 \\ &= \text{conv} \bigcup_{z:||z||_2\leq 1} z \\ &= \{z \in \mathbb{R}^N \mid ||z||_2 \leq 1\} \end{aligned}$$

Outline

Descent Direction

- idea:
  - choose an arbitrary subgradient $g \in \partial f$
  - use its negative $-g$ as next direction
- negative subgradients are in general no descent directions
  - example:

$$f(x) := |x|$$
$$x^{(0)} := 0$$

Optimality Condition

For a convex $f : \mathbb{R}^n \rightarrow \mathbb{R}$:

$$\mathbf{x}^* \text{ is a global minimizer} \quad \Leftrightarrow \quad \mathbf{0} \text{ is a subgradient of } f \text{ at } \mathbf{x}^*$$

$$f(\mathbf{x}^*) = \min_{\mathbf{x} \in \text{dom}\, f} f(\mathbf{x}) \qquad \mathbf{0} \in \partial f(\mathbf{x}^*)$$

Proof:

If $\mathbf{0}$ is a subgradient of f at $\mathbf{x}^*$, then for all $\mathbf{y} \in \mathbb{R}^n$:

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbf{0}^T(\mathbf{y} - \mathbf{x}^*)$$
$$f(\mathbf{y}) \geq f(\mathbf{x}^*)$$

Gradient Descent (Review)

min-gd$(f, \nabla f, x^{(0)}, \mu, \epsilon, K)$:
    for $k := 1, \ldots, K$:
        $\Delta x^{(k-1)} := -\nabla f(x^{(k-1)})$
        if $||\nabla f(x^{(k-1)})||_2 < \epsilon$:
            return $x^{(k-1)}$
        $\mu^{(k-1)} := \mu(f, x^{(k-1)}, \Delta x^{(k-1)})$
        $x^{(k)} := x^{(k-1)} + \mu^{(k-1)} \Delta x^{(k-1)}$
    return "not converged"

where

- f objective function
- ∇f gradient of objective function f
- $x^{(0)}$ starting value
- μ step length controller
- ϵ convergence threshold for gradient norm
- K maximal number of iterations

Subgradient Method

min-subgrad$(f, \partial f, x^{(0)}, \mu, K)$:

$\quad x_{\text{best}}^{(0)} := x^{(0)}$

$\quad$ for $k := 1, \ldots, K$:

$\quad\quad$ if $0 \in \partial f(x^{(k-1)})$:

$\quad\quad\quad$ return $x_{\text{best}}^{(k-1)}$

$\quad\quad$ choose $g \in \partial f(x^{(k-1)})$ arbitrarily

$\quad\quad \Delta x^{(k-1)} := -g$

$\quad\quad \mu^{(k-1)} := \mu_{k-1}$

$\quad\quad x^{(k)} := x^{(k-1)} + \mu^{(k-1)} \Delta x^{(k-1)}$

$$x_{\text{best}}^{(k)} := \begin{cases} x^{(k)}, & \text{if } f(x^{(k)}) < f(x_{\text{best}}^{(k-1)}) \\ x_{\text{best}}^{(k-1)}), & \text{else} \end{cases}$$

$\quad$ return "not converged"

where

- $\mu \in \mathbb{R}^*$ step length schedule

Outline

Slowly Diminishing Stepsizes

Proof of convergence requires **slowly diminishing stepsizes**:

$$\lim_{k\to\infty} \mu^{(k)} = 0, \quad \sum_{j=0}^{\infty} \mu^{(j)} = \infty, \quad \sum_{j=0}^{\infty} (\mu^{(j)})^2 < \infty$$

for example:

$$\mu^{(k)} := \frac{1}{k+1}$$

but not:

- constant stepsizes $\mu^{(k)} := \mu \in \mathbb{R}$
- too fast shrinking stepsizes, e.g., $\mu^{(k)} := \frac{1}{(k+1)^2}$
- adaptive stepsize chosen by a step length controller

Convergence

Theorem (convergence of subgradient method)

Under the assumptions

I. $f : X \to \mathbb{R}$ *is convex,* $X \subseteq \mathbb{R}^n$ *is open*

II. f *is Lipschitz-continuous with constant* $G > 0$*, i.e.*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq G||\mathbf{x} - \mathbf{y}||_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

- *Equivalently:* $||\mathbf{g}||_2 \leq G$ *for any subgradient of* f *at any* $\mathbf{x}$

III. *slowly diminishing stepsizes* $\mu^{(k)}$*, i.e.,*

$$\lim_{k \to \infty} \mu^{(k)} = 0, \quad \sum_{j=0}^{\infty} \mu^{(j)} = \infty, \quad \sum_{j=0}^{\infty} (\mu^{(j)})^2 < \infty$$

the subgradient method converges and

$$f_{best}^{(k)} - f(\mathbf{x}^*) \leq \frac{||\mathbf{x}^{(0)} - \mathbf{x}^*||^2 + G^2 \sum_{j=0}^{k} (\mu^{(j)})^2}{2 \sum_{j=0}^{k} \mu^{(j)}}$$

Convergence / Proof (1/2)

$$
\begin{aligned}
||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_2^2 \\
&= ||\mathbf{x}^{(k)} - \mu^{(k)}\mathbf{g}^{(k)} - \mathbf{x}^*||_2^2 \\
&= ||\mathbf{x}^{(k)} - \mathbf{x}^*||_2^2 - 2\mu^{(k)}(\mathbf{g}^{(k)})^T(\mathbf{x}^{(k)} - \mathbf{x}^*) + (\mu^{(k)})^2||\mathbf{g}^{(k)}||_2^2 \\
&\underset{\text{SG}}{\leq} ||\mathbf{x}^{(k)} - \mathbf{x}^*||_2^2 - 2\mu^{(k)}(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)) + (\mu^{(k)})^2||\mathbf{g}^{(k)}||_2^2 \\
&\underset{\text{rec}}{\leq} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_2^2 - 2\sum_{j=0}^{k}\mu^{(j)}(f(\mathbf{x}^{(j)}) - f(\mathbf{x}^*)) + \sum_{j=0}^{k}(\mu^{(j)})^2||\mathbf{g}^{(j)}||_2^2 \\
&\underset{\text{II}}{\leq} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_2^2 - 2\sum_{j=0}^{k}\mu^{(j)}(f(\mathbf{x}^{(j)}) - f(\mathbf{x}^*)) + G\sum_{j=0}^{k}(\mu^{(j)})^2 \qquad (1)
\end{aligned}
$$

Convergence / Proof (2/2)

$$f_{\text{best}}^{(k)} - f(\mathbf{x}^*) \leq \frac{\sum_{j=0}^{k}(f_{\text{best}}^{(k)} - f(\mathbf{x}^*))\mu^{(j)}}{\sum_{j=0}^{k} \mu^{(j)}}$$

$$\leq \frac{\sum_{j=0}^{k}(f(\mathbf{x}^{(j)}) - f(\mathbf{x}^*))\mu^{(j)}}{\sum_{j=0}^{k} \mu^{(j)}}$$

$$\leq \frac{2\sum_{j=0}^{k}(f(\mathbf{x}^{(j)}) - f(\mathbf{x}^*))\mu^{(j)} + ||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_2^2}{2\sum_{j=0}^{k} \mu^{(j)}}$$

$$\underset{(1)}{\leq} \frac{||\mathbf{x}^{(0)} - \mathbf{x}^*||_2^2 + G\sum_{j=0}^{k}(\mu^{(j)})^2}{2\sum_{j=0}^{k} \mu^{(j)}}$$

$$\lim_{k\to\infty} f_{\text{best}}^{(k)} - f(\mathbf{x}^*) \leq \lim_{k\to\infty} \frac{||\mathbf{x}^{(0)} - \mathbf{x}^*||_2^2 + G\sum_{j=0}^{k}(\mu^{(j)})^2}{2\sum_{j=0}^{k} \mu^{(j)}} \underset{\text{III}}{=} 0$$

Further Readings

- Subgradient methods are not covered by Boyd and Vandenberghe [2004]
- Subgradients:
  - [Bertsekas, 1999, ch. B.5 and 6.1]
- Subgradient methods:
  - [Bertsekas, 1999, ch. 6.3.1]

References I

Dimitri P. Bertsekas. *Nonlinear Programming*. Springer, 1999.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge Univ Press, 2004.

Example: Text Classification

Features **A**: normalized word frequecies in text documents

Category **y**: topic of the text documents

$$A_{m,n} = \begin{pmatrix} 1 & a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ 1 & a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & a_{m,1} & a_{m,2} & a_{m,3} & a_{m,4} \end{pmatrix} \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

$$\hat{y}_i = \sigma(\mathbf{x}^T \mathbf{a_i})$$

Text Classification: L1-Regularized Logistic Regression

For $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$ we have the following problem

$$\text{minimize} \quad -\sum_{i=1}^{m} y_i \log \sigma(\mathbf{x}^T \mathbf{a_i}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}^T \mathbf{a_i})) + \lambda ||\mathbf{x}||_1$$

Which can be rewritten as:

$$\text{minimize} \quad -\sum_{i=1}^{m} y_i \log \sigma(\mathbf{x}^T \mathbf{a_i}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}^T \mathbf{a_i})) + \lambda \sum_{k=1}^{n} |x_k|$$

f is convex and non-smooth

Example: L1-Regularized Logistic Regression

The subgradients of
$f(\mathbf{x}) = -\sum_{i=1}^{m} y_i \log \sigma(\mathbf{x}^T \mathbf{a_i}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}^T \mathbf{a_i})) + \lambda ||\mathbf{x}||_1$ are:

$$\mathbf{g} = -\mathbf{A}^T(\mathbf{y} - \hat{\mathbf{y}}) + \lambda \mathbf{s}$$

where $\mathbf{s} \in \partial ||\mathbf{x}||_1$, i.e.:

- $s_k = \text{sign}(\mathbf{x}_k)$ if $\mathbf{x}_k \neq 0$
- $s_k \in [-1, 1]$ if $\mathbf{x}_k = 0$

Example - The algorithm

For $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$ we have the following the problem

$$\text{minimize} \quad -\sum_{i=1}^{m} y_i \log \sigma(\mathbf{x}^T \mathbf{a_i}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}^T \mathbf{a_i})) + \lambda \sum_{k=1}^{n} |x_k|$$

1. Start with an initial solution $\mathbf{x}^{(0)}$
2. $t \leftarrow 0$
3. $f_{\text{best}} \leftarrow f(\mathbf{x}^{(0)})$
4. Repeat until convergence
   - 4.1 $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} - \mu^{(k)}(-\mathbf{A}^T(\mathbf{y} - \hat{\mathbf{y}}) + \lambda \mathbf{s})$
   - 4.2 $t \leftarrow t + 1$
   - 4.3 $f_{\text{best}} \leftarrow \min(f(\mathbf{x}^{(k)}), f_{\text{best}})$
5. Return f_{best}

where $\mathbf{s} \in \partial ||\mathbf{x}||_1$, i.e.:

- $s_k = \text{sign}(\mathbf{x}_k)$ if $\mathbf{x}_k \neq 0$
- $s_k \in [-1, 1]$ if $\mathbf{x}_k = 0$