Modern Optimization Techniques

Lucas Rego Drumond

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

Subgradient Methods

Outline

1. Subgradients
2. Subgradient Method
3. Subgradient Method Examples

Outline

Methods seen so far

- If a function is differentiable we can optimize it using Gradient Descent and Stochastic Gradient Descent (1st order information)
- If a function is smooth and twice differentiable we can optimize it using Newton's method (2nd order information)

- What if the the function is not differentiable?

1st-order condition for Convexity

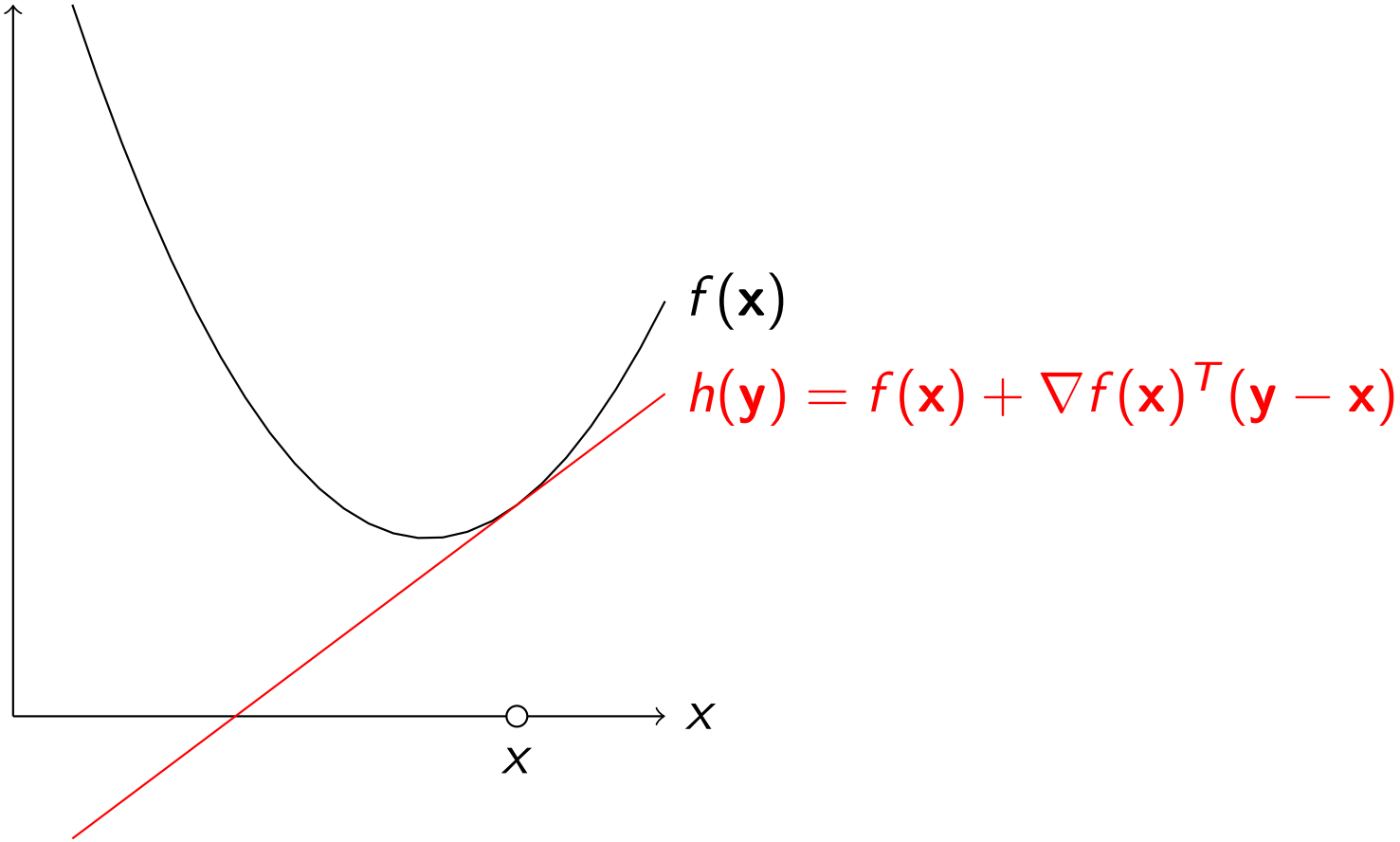
1st-order condition: a differentiable function f is convex iff

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

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

- The first order Taylor approximation of f at $\mathbf{x}$ is a global underestimator

1st-order approximation as a global underestimator



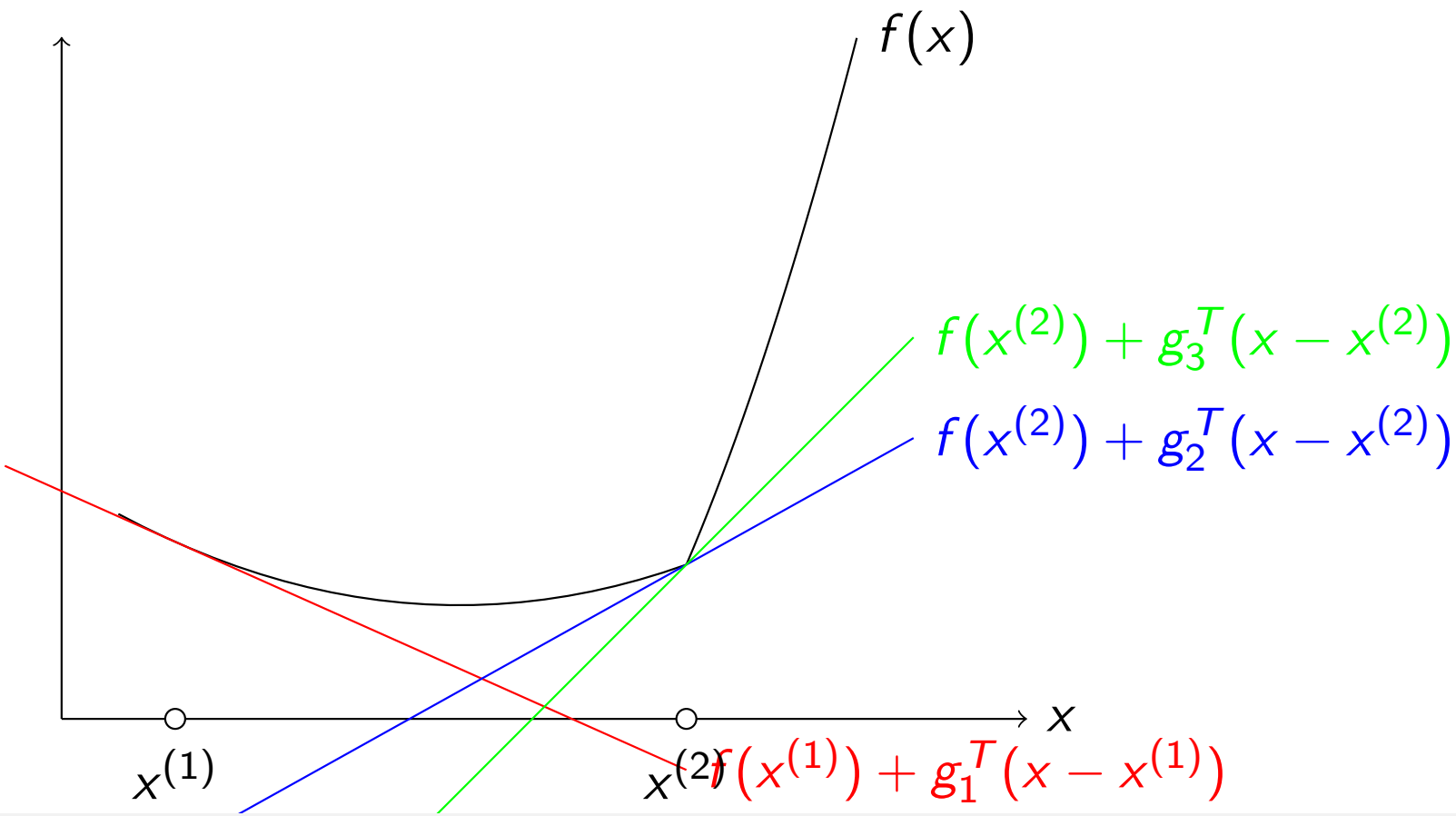
What happens if f is not differentiable?

Subgradient

Given a function f and a point $\mathbf{x} \in \operatorname{dim} f$, $\mathbf{g}$ is a **subgradient** of f at $\mathbf{x}$ if:

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

for all $\mathbf{y} \in \operatorname{dom} f$



Subgradient

Given a function f and a point $\mathbf{x} \in \operatorname{dim} f$, $\mathbf{g}$ is a **subgradient** of f at $\mathbf{x}$ if:

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

for all $\mathbf{y} \in \operatorname{dom} f$

In the last example, $\mathbf{g_1}$ is a subgradient of f in $x^{(1)}$ while g_2 and g_3 are subgradients in $x^{(2)}$

For a **convex** function f:

- The subgradient always exist
- If f is differentiable at $\mathbf{x}$, then $\mathbf{g} = \nabla f(\mathbf{x})$

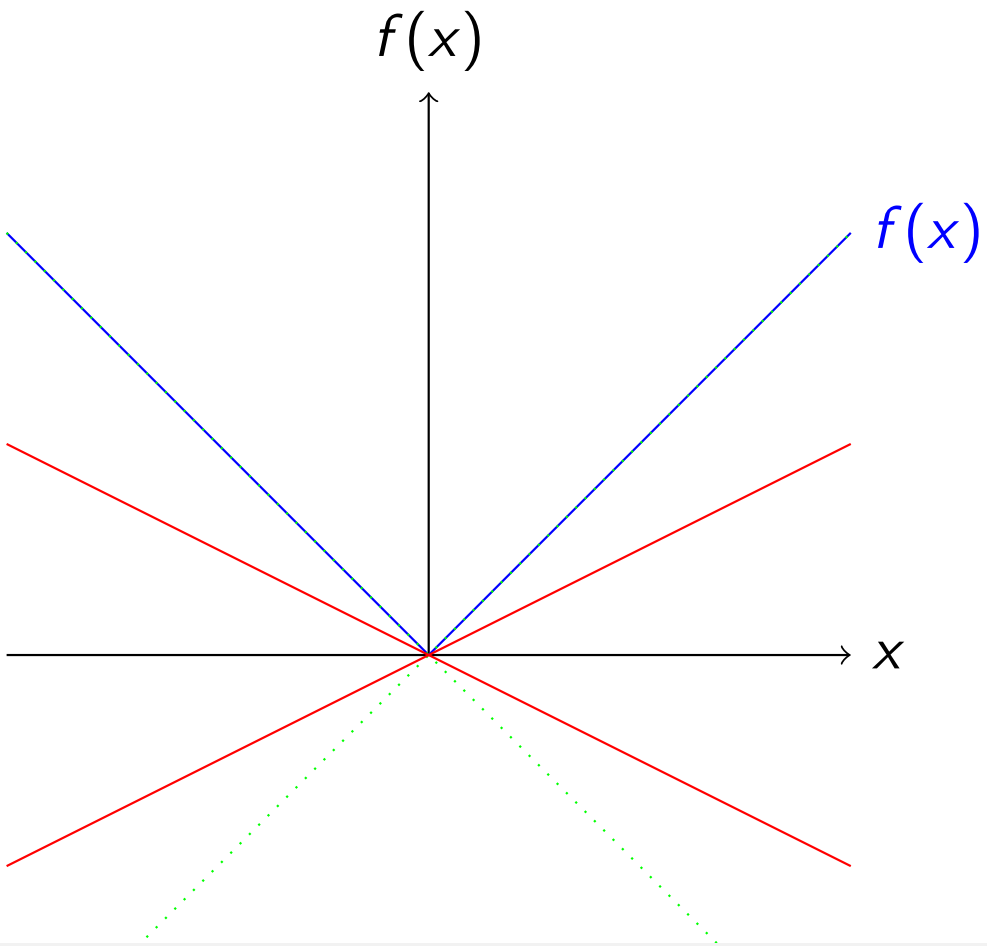
For a **non-convex** function f:

- The same applies, except that ...
- ... the subgradient does **not** always exist

Example

For $f : \mathbb{R} \rightarrow \mathbb{R}$ 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

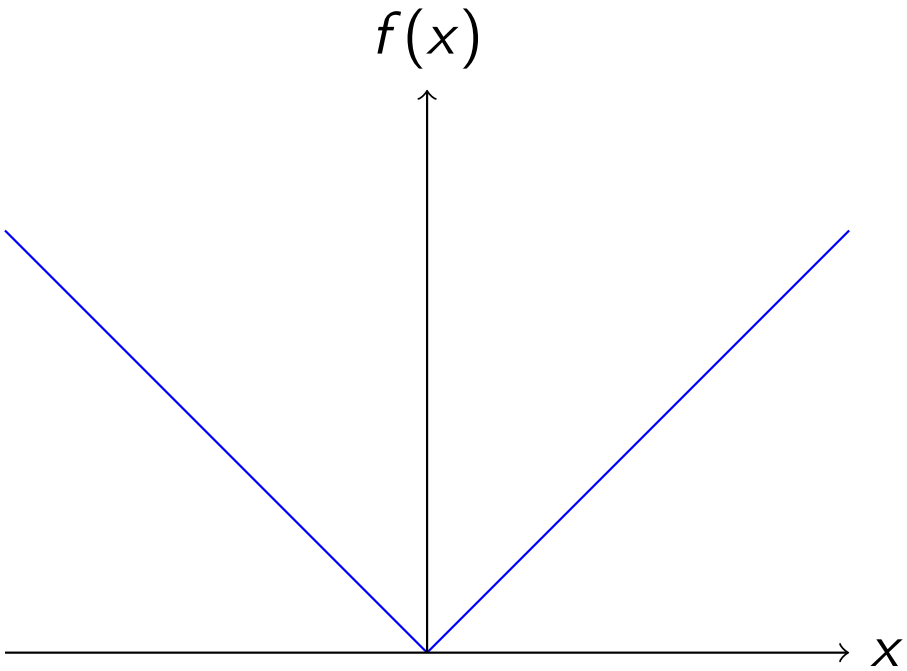
Subdifferential $\partial f(\mathbf{x})$: set of all subgradients of f at $\mathbf{x}$

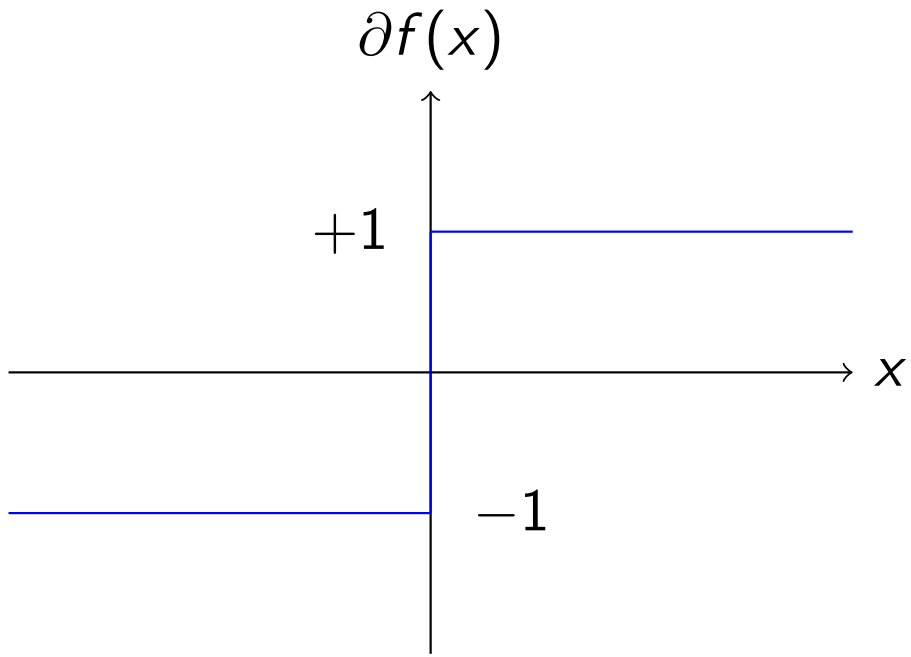
If f is convex:

- $\partial f(\mathbf{x})$ is nonempty
- $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$ if f is differentiable at x
- If $\partial f(\mathbf{x}) = \{\mathbf{g}\}$, then f is differentiable at x and $\mathbf{g} = \nabla f(\mathbf{x})$

Example

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





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 $a > 0$: $\partial(a \cdot f) = \{a \cdot \mathbf{g} | \mathbf{g} \in \partial(f)\}$
- **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 $f(\mathbf{x}) = \max_{i=1\ldots,m} f_i(\mathbf{x})$ then

$$\partial f(\mathbf{x}) = \text{conv} \bigcup_{i:f_i(\mathbf{x})=f(\mathbf{x})} \partial f_i(\mathbf{x})$$

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

Outline

The Subgradient Method

Be f_0 a nondifferentiable and convex function $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$:

$$\text{minimize} \quad f_0(\mathbf{x})$$

Be $\mathbf{g}^t$ **any** subgradient of f_0 at $\mathbf{x}^t$

1. Start with an initial solution $\mathbf{x}^{(0)}$
2. $t \leftarrow 0$
3. Repeat until convergence
   - 3.1 Find $\mathbf{x}^{t+1} = \mathbf{x}^t - \mu_t \mathbf{g}^t$
   - 3.2 $t \leftarrow t + 1$
4. Return $f_{0\text{best}} = \min_{j=1,\ldots,t} f_0(\mathbf{x}^j)$

The subgradient method is not a descent method!

Optimality Conditions

For a convex $f_0 : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}^*$ is a minimizer iff $\mathbf{0}$ is a subgradient of f_0 at $\mathbf{x}^*$

$$f_0(\mathbf{x}^*) = \min_{\mathbf{x} \in \mathbb{R}^n} f_0(\mathbf{x}) \quad \Leftrightarrow \quad \mathbf{0} \in \partial f_0(\mathbf{x}^*)$$

Proof:

If $\mathbf{0}$ is a subgradient of f_0 at $\mathbf{x}^*$, then for all $\mathbf{y} \in \mathbb{R}^n$:

$$f_0(\mathbf{y}) \geq f_0(\mathbf{x}^*) + \mathbf{0}^T(\mathbf{y} - \mathbf{x}^*)$$
$$f_0(\mathbf{y}) \geq f_0(\mathbf{x}^*)$$

Step size rules

Fixed step-size: keep $\mu_t = \mu$ constant

Fixed length: keep $\mu_t = \frac{\gamma}{||\mathbf{g}^t||_2}$ so that $||\mathbf{x}^{t+1} - \mathbf{x}^t||_2 = \gamma$

Diminishing:

$$\lim_{t \to \infty} \mu_t = 0, \qquad \sum_{t=1}^{\infty} \mu_t = \infty$$

Does the algorithm converge?

The convergence analysis of the subgradient method makes some assumptions:

- $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex
- f_0 is Lipschitz continuous with constant $G > 0$. i.e. for all $\mathbf{x}, \alpha \in \mathbb{R}^n$:

$$|f_0(\mathbf{x}) - f_0(\alpha)| \leq G||\mathbf{x} - \alpha||_2$$

- Equivalently: $||\mathbf{g}||_2 \leq G$ for any subgradient of f_0 at any $\mathbf{x}$
- We know a constant R such that $||\mathbf{x}^t - \mathbf{x}^*||_2 \leq R$

Does the algorithm converge?

Using the definition of the subgradient we have:

$$
\begin{aligned}
||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 &= ||\mathbf{x}^t - \mu_t \mathbf{g}^t - \mathbf{x}^*||_2^2 \\
&= ||\mathbf{x}^t - \mathbf{x}^*||_2^2 - 2\mu_t {\mathbf{g}^t}^T(\mathbf{x}^t - \mathbf{x}^*) + \mu_t^2 ||\mathbf{g}^t||_2^2 \\
&\leq ||\mathbf{x}^t - \mathbf{x}^*||_2^2 - 2\mu_t(f_0(\mathbf{x}^t) - f_0(\mathbf{x}^*)) + \mu_t^2 ||\mathbf{g}^t||_2^2
\end{aligned}
$$

If we iterate that inequality over all the previous steps to $t+1$:

$$
||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 \leq ||\mathbf{x}^1 - \mathbf{x}^*||_2^2 - 2\sum_{i=1}^{t} \mu_i(f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) + \sum_{i=1}^{t} \mu_i^2 ||\mathbf{g}^i||_2^2
$$

Does the algorithm converge?

Remember that $||\mathbf{x}^t - \mathbf{x}^*||_2 \leq R$ and $||\mathbf{g}||_2 \leq G$:

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 \leq ||\mathbf{x}^1 - \mathbf{x}^*||_2^2 - 2\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) + \sum_{i=1}^{t} \mu_i^2 ||\mathbf{g}^i||_2^2$$

$$\leq R^2 - 2\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) + G^2 \sum_{i=1}^{t} \mu_i^2$$

Which we can rearrange as:

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 \leq R^2 - 2\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) + G^2 \sum_{i=1}^{t} \mu_i^2$$

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 + 2\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) \leq R^2 + G^2 \sum_{i=1}^{t} \mu_i^2$$

Does the algorithm converge?

Now we also know that:

$$\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) \geq (f_{0\,\text{best}}^t - f_0(\mathbf{x}^*)) \sum_{i=1}^{t} \mu_i$$

From which the following still holds:

$$2\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) \geq 2(f_{0\,\text{best}}^t - f_0(\mathbf{x}^*)) \sum_{i=1}^{t} \mu_i$$

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 + 2\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) \geq 2(f_{0\,\text{best}}^t - f_0(\mathbf{x}^*)) \sum_{i=1}^{t} \mu_i$$

Does the algorithm converge?

From

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 + 2\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) \leq R^2 + G^2 \sum_{i=1}^{t} \mu_i^2$$

and

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 + 2\sum_{i=1}^{t} \mu_i (f_0(\mathbf{x}^i) - f_0(\mathbf{x}^*)) \geq 2(f_{0\,\text{best}}^t - f_0(\mathbf{x}^*)) \sum_{i=1}^{t} \mu_i$$

We have

$$f_{0\,\text{best}}^t - f_0(\mathbf{x}^*) \leq \frac{R^2 + G^2 \sum_{i=1}^{t} \mu_i^2}{2\sum_{i=1}^{t} \mu_i}$$

Convergence guarantees for different step sizes

For a fixed step size $\mu_t = \mu$

$$f_{0\,\text{best}}^{t} - f_0(\mathbf{x}^*) \leq \frac{R^2 + G^2 \sum_{i=1}^{t} \mu_i^2}{2 \sum_{i=1}^{t} \mu_i} = \frac{R^2 + G^2 t \mu^2}{2t\mu}$$

The error upperbound converges to:

$$\lim_{t \to \infty} \frac{R^2 + G^2 t \mu^2}{2t\mu} = \frac{G^2 \mu}{2}$$

Convergence guarantees for different step sizes

For a diminishing step size such that

$$\sum_{t=1}^{\infty} \mu_t = \infty \qquad \sum_{t=1}^{\infty} \mu_t^2 < \infty$$

$$f_{0\,\text{best}}^{t} - f_0(\mathbf{x}^*) \leq \frac{R^2 + G^2 \sum_{i=1}^{t} \mu_i^2}{2 \sum_{i=1}^{t} \mu_i}$$

The error upperbound converges to:

$$\lim_{t \to \infty} \frac{R^2 + G^2 \sum_{i=1}^{t} \mu_i^2}{2 \sum_{i=1}^{t} \mu_i} = 0$$

This proves that for a diminishing step size with the properties above the algorithm will converge to the optimum

Outline

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 the problem

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

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_0 is convex and non-smooth

Example: L1-Regularized Linear Regression

The subgradients of
$f_0(\mathbf{x}) = -\sum_{i=1}^{m} y_i \log \sigma(\mathbf{x}^T \mathbf{a_i}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}^T \mathbf{a_i})) + \lambda ||\mathbf{x}||_1$ 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. $f_{0\text{best}} \leftarrow f_0(\mathbf{x}^{(0)})$
4. Repeat until convergence
   - 4.1 $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^t - \mu_t(-\mathbf{A}^T(\mathbf{y} - \hat{\mathbf{y}}) + \lambda \mathbf{s})$
   - 4.2 $t \leftarrow t + 1$
   - 4.3 $f_{0\text{best}} \rightarrow \min(f_0(\mathbf{x}^t), f_{0\text{best}})$
5. Return $f_{0\text{best}}$

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 $\mathbf{x}_k = 0$