Unconstrained Optimization

Pravesh Biyani

IIIT Delhi

September 15, 2014

Outline

- Definitions
- Unconstrained minimization
- First and second order optimality conditions
- First algorithm: gradient descent
- Least square regression.

Definitions

- Local minimum x^*: $\exists \epsilon > 0$ s.t $f(x) \geq f(x^*)$, for all $||x - x^*|| < \epsilon$.
- Strict local minimum x^*: $\exists \epsilon > 0$ s.t $f(x) > f(x^*)$, for all $||x - x^*|| < \epsilon$.
- Global minimum: $f(x) \geq f(x^*)$, for all $x \in \Re^n$.
- Geometrically ?

Relaxation and Approximation

- The simplest goal is to find local minimum of a differential function
- Majority of methods based on the idea of relaxation.
- Generate a sequence $f(x_k)_{k=0}^{k=\infty}$ such that
- Advantages
  - If $f(x)$ is lower bounded, convergence is guaranteed.
  - We decrease the objective function in every step.
- Optimality Conditions useful in reducing the search for the optimal.

Approximations

A first order *local* approximation:

$$f(y) = f(x_0) + \langle f'(x_o), y - x_0 \rangle + O(||y - x_0||)$$

where, $O(r)$ is a vector valued function such that

$$\lim_{r \to 0} \frac{1}{r} O(r) = 0$$

The direction $-f'(x)$ is the direction of the *fastest local decrease* of the function at x.
Proof ?

First Order Optimality Condition

Theorem
Let x^* be a local minimum of differentiable function $f(x)$. Then,

$$f'(x^*) = 0.$$

- Only a necessary condition.
- Points satisfying this condition called *stationary* points.
- Examples: least square, x^3, $x^2 - x^4$

Second Order Optimality Condition

If the function $f(x)$ be twice differentiable.

$$f(y) = f(x_0) + \langle f'(x_o), y - x_0 \rangle + \left\langle f''(x_0)(y - x_o), y - x_0 \right\rangle + O(||y - x_0||^2)$$

The quadratic function above is the second order approximation. $f''(x_0)$ is also called Hessian. It is a symmetric matrix.

Theorem

Let x^* be a local minimum of differentiable function $f(x)$. Then,

$$f'(x^*) = 0, \quad f''(x^*) \succcurlyeq 0.$$

Necessary and Sufficient Conditions

Theorem

Let function $f(x)$ be twice differentiable and let x^* satisfy the following conditions:

$$f'(x^*) = 0 \quad f''(x^*) \succ 0.$$

Then x^* is a strict local minimum of $f(x)$.

Proposition: There exist scalars $\gamma > 0$ and $\epsilon > 0$ such that

$$f(x) > f(x^*) + \gamma/2||x - x^*||^2, \quad ||x - x^*|| < \epsilon$$

Algorithm: Gradient Method

$$\begin{aligned} Choose: &\quad x_0 \in R^n \\ Iterate: &\quad x_{k+1} = x_k - h_k f'(x_k) \quad k = 0, 1, \ldots \end{aligned}$$

h_k is called the step size.
Can be chosen in advanced or can adapt.
Does the gradient method converge to the local minima always?

Generalizing the Gradient Methods

- One can consider a half line of vectors

$$x_\alpha = x + \alpha d \quad \forall \alpha > 0$$

where the direction $d \in R^n$ makes an angle with $\nabla f(x)$ that is greater than 90 degrees that is,

$$\nabla f(x)^T d < 0$$

- Leading to the following algo:

$$\boxed{\begin{array}{rl} \textit{Choose}: & x_0 \in R^n \\ \textit{Iterate}: & x_{k+1} = x_k - \alpha_k H_k d_k \quad k = 0, 1, \ldots \end{array}}$$

- D_k is a PSD matrix

Line Search: How to choose the stepsize

Two simple line search mechanisms.

Exact Line Search: $t = arg\min_{s>0} f(x + sd)$

Find the line which has maximum decrease in the objective function for a given descent direction v.

Backtracking Line Search $0 < \beta < 1, 0 < \alpha < 0.5$

- starting with $t = 1$, $t = \beta t$
- until $f(x + td) < f(x) + t\alpha f'(x)^T d$
- Also called Armijo rule.

Diminishing Step Size: $h_k \to 0$ but $\sum h_k = \infty$

The Newton Step

The Newton step at x_k is

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

- Minimizes the second order expansion at x at every step.
- Convergence is faster than a simple gradient descent.
- Can be combined with backtracking or exact line search.
- Variants: damped newton and Quasi-Newton.

Convergence Results – key insights

- Generally gradient method is slow, but converges with exact or back tracking line search.
- Newton method converges rapidly (quadratic) when $\nabla^2 f(x) \geq m\mathbf{I}$
- A constant step size method requires stricter conditions for convergence also called Lipschitz conditions

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$$

- Lipschitz condition also helps in diminishing step size case.

Convergence for a constant step size

Let $\{x_k\}$ be a sequence generated by a gradient method $x_{k+1} = x_k + \alpha_k d$. Assume that for some constant $L > 0$, we have

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y|| \quad \forall x, y \in R^n$$

and that for all k we have $d_k \neq 0$ and

$$\epsilon \leq \alpha_k \leq (2 - \epsilon)\hat{\alpha_k}$$

where

$$\hat{\alpha_k} = \frac{|\nabla f(x_k)^T d_k|}{L||d_k||^2}$$

Then every limit point of $\{x_k\}$ is a stationary point of f.

- For steepest descent, the condition on α_k is

$$\epsilon \leq \alpha_k \leq \frac{2 - \epsilon}{L}$$

Some Examples

- Unconstrained Quadratic Minimization

$$\min \quad x^T P x + 2q^T x + r$$

- Unconstrained Geometric Programming

$$\min \quad \log \sum_{i=1}^{m} e^{a_i^T x + b_i}$$

Conjugate Direction Methods

- Generally used for quadratic minimization problems.
- Faster than steepest descent, avoiding overhead of Newton methods.

$$\min \quad \frac{1}{2}x^T Qx - b^T x$$

- Equivalently solve $Qx = b$. (Q is PSD).
- Conjugate direction method solves in atmost N iterations.

Conjugate Direction..

- Vectors $d_1, d_2 \ldots d_k$ are Q conjugate if,

$$d_i^T Q d_j = 0 \quad \forall i \neq j$$

- And they are also linearly independent !!

Conjugate direction method:

$$\begin{aligned} Choose: &\quad x_0 \in R^n \\ Iterate: &\quad x_{k+1} = x_k + \alpha_k d_k \quad k = 0, 1, \ldots \\ \alpha_k(LineSearch): &\quad \min_{\alpha} f(x_k + \alpha d_k) \end{aligned}$$

Conjugate Direction ..

- Successive iterates minimizes f over a progresively expanding space that eventually includes the global minimum of a quadratic f.
- Eventually easy to show that

$$x_{k+1} = arg \min_{x \in M^k} f(x)$$

where,

$$M_k = \{x | x_0 + v, \quad v \in \{subspace\ spanned\ by\ d_0, d_1 \ldots d_k\}\}$$

From Conjugate Direction to CG

- How to generate Conjugate directions?
- Gram Schmidt Orthonalization:
- Take any set of linearly independent vectors $u_0, u_1 \dots u_{n-1}$ and generate Q conjugate directions using them!
- CG method is obtained by applying the GS procedure to the gradient vectors $u_k = -g_k = -\nabla f(x_k)$..

$$g_k = Qx_k - b$$

$$d_k = -g_k + \sum_{i=0}^{k-1} \frac{g_k^T Q d_i}{d_i^T Q d_i} d_i$$

CG Method

CG method:

$$\begin{aligned}
\textit{Choose}: &\quad x_0 \in R^n \\
\textit{Iterate}: &\quad x_{k+1} = x_k + \alpha_k d_k \quad k = 0, 1, \ldots \\
d_k: &\quad d_k = -g_k + \beta_k d_{k-1} \\
\beta_k: &\quad \beta_k = \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}}
\end{aligned}$$