Optimization Methods		Date:	*6th April 2018*
Instructor: *Sujit Prakash Gujar*		Scribes:	Roopal Vaid Rohit Gajawada Saurav Malani

Lecture 21: Constrained optimization

1 Recap

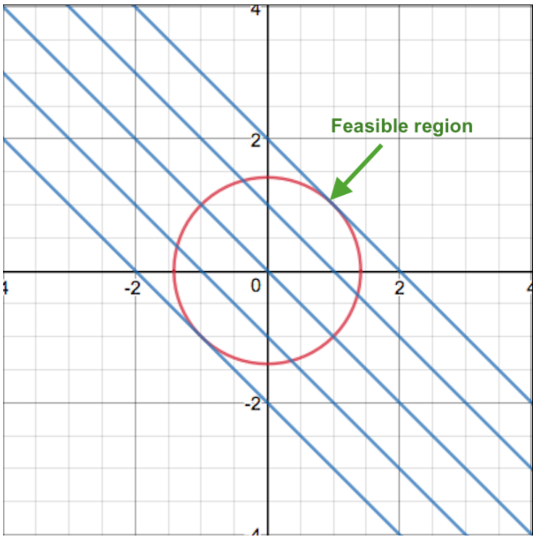
- Conjugate Gradient Method - It is a method useful for optimization of both linear and non-linear systems.
- Conjugate Gradient Descent - The Conjugate Gradient can be used if and only if the function is a Quadratic and Q is Symmetric Real Positive Definite.
- Quasi Newton Method - The quasi-Newton methods use an approximation to $H1$ in place of the true inverse. So, the update equation becomes $x_{k+1} = x_k - \alpha_k B_k^1 \nabla f(x_k)$ where B_k is an $n * n$ positive definite matrix and α_k is a positive search parameter. α_k should satisfy Armijo-Wolfe conditions. Different update rules for B_k are -
  - Rank 1 (B_{k+1} may not be positive definite)
  - Broyden, Fletcher, Goldfarb, and Shanno (BFGS - this can be computed efficiently)
- Stochastic Gradient Descent - Stochastic Gradient descent runs faster, but may lead to convergence at a suboptimal minima. In this method the gradient is updated at each sample.

2 Equality constraint optimization

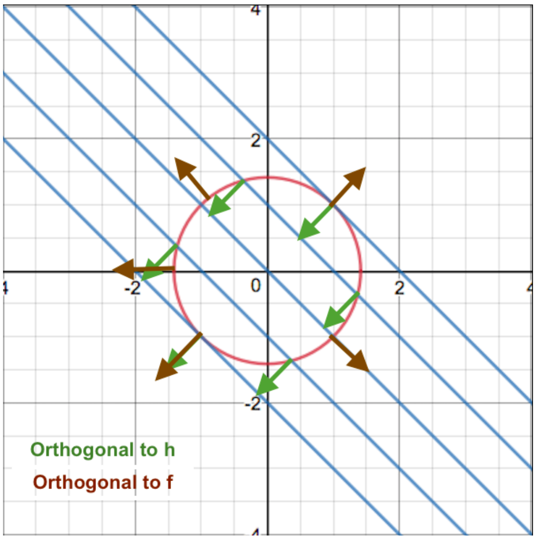
- Here we consider the optimization problem of the form
  $min \quad f(x)$
  $subjectto \quad h_i(x) = 0, \quad i = 1, ..., m$
- Let us assume functions f:$\mathbb{R}^n \rightarrow \mathbb{R}$, h_i:$\mathbb{R}^n \rightarrow \mathbb{R}$ are continuous and differentiable.
- For f to be local minimum at $x*$ without any constraints the necessary and sufficient conditions are -
  - f has zero gradient at $x*$, i.e. $\nabla f(x*) = 0$
  - Hessian of f is positive semi definite at x* i.e $d^T \nabla^2 f(x^*) d \geq 0$
- For example:

  $f(x) = x_1 + x_2; \quad h(x) = x_1^2 + x_2^2 - 2 = 0$

  $\nabla f(x) = [1 \quad 1]^T; \nabla h(x) = [2x_1 \quad 2x_2]^T$



- Suppose we start at $(1,1)$, we move locally tangentially in direction d such that $f(x + d^T\alpha) < f(x); \quad d^T\nabla f(x) < 0$.
- At local optima x^*, $d^T\nabla f(x^*) = 0$. So d is orthogonal to $\nabla f(x)$
- d is orthogonal to h at x^*. $\nabla f(x^*) = \lambda \nabla h(x^*)$, for a scalar λ. So we cannot move from $x*$ by satisfying the constraint and decreasing the f simultaneously.



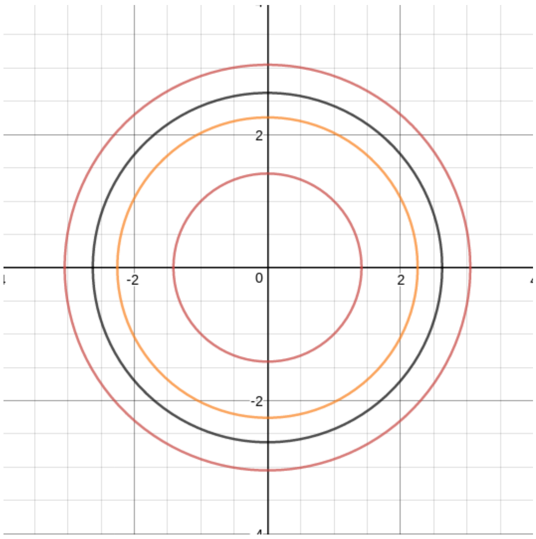
2.1 Lagrange Multiplier

- Single constraint
  - $\min\ f(x)$; subject to $h(x) = 0$
  - Lagrange multiplier for $\mathcal{L}(x,\mu) = f(x) + \mu h(x)$ is μ.
  - At local minima $\mathcal{L}(x^*,\mu^*)$
    * $\nabla_x \mathcal{L}(x^*,\mu^*) = 0 \implies \nabla_{x^*} f(x^*) = \mu^* \nabla_{x^*} h(x^*)$
    * $\nabla_\mu \mathcal{L}(x^*,\mu^*) = 0 \implies h(x^*) = 0$

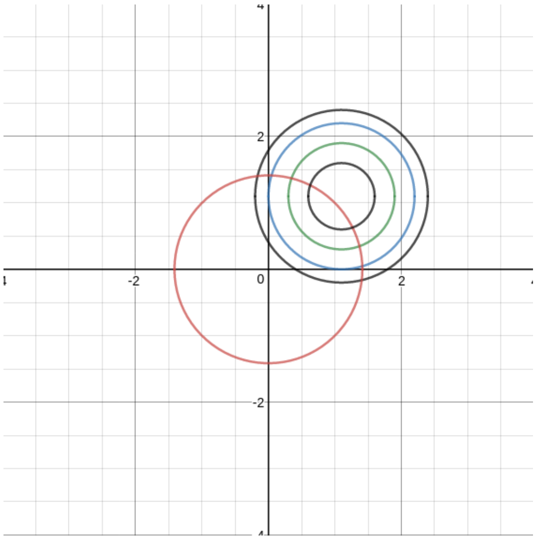
- Multiple constraints
  - $\min\ f(x)$; subject to $h_i(x) = 0 \quad \forall \quad i = 1, ...m$
  - The lagrange multiplier for the optimization is
    $\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x)$
    Here λ_i is lagrange multiplier for each equality constraint $h_i(x) = 0$
  - If (x^*, λ^*) be the local minimum for $\mathcal{L}$
    * $\nabla_x \mathcal{L}(x^*, \lambda_i^*) = 0 \implies \nabla_{x^*} f(x^*) = \lambda_i^* \nabla_{x^*} h_i(x^*)$
    * $\nabla_\lambda \mathcal{L}(x^*, \lambda_i^*) = 0 \implies h_i(x^*) = 0$

3 Inequality Constraint Optimization

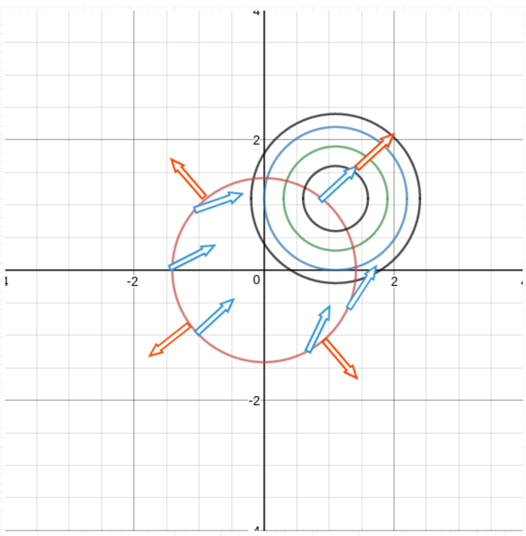
- We now want to consider a function that we would like to minimize which has both equality and inequality constraints:
  $minf(x)$ subject to $h(x) = 0$ and $g(x) \leq 0$
- There are two possible scenarios for this situation.



- Let us take an example for the first scenario:
  $f(x) = x_1^2 + x_2^2$ and $g(x) = x_1^2 + x_1^2 - 2$. In this case, the unconstrained local minimum lies in the feasible region.
  - If x^* is local minimum for this example and $g(x^*) < 0$, this implies that the constraint $g(x)$ is inactive.
  - for this x^*, the necessary and sufficient conditions for f remain the same:
    * The function f has zero gradient at x^* i.e $\nabla f(x^*) = 0$.
    * The hessian of f is positive semi definite at x^* i.e $d^T \nabla^2 f(x^*) d \geq 0$.
    * Basically, the constraint here is inactive since at any x: $g(x) < 0$, f has zero gradient at x^*. Even at local minimum, the constraints still remains inactive i.e the local minimum can be deduced by the same conditions similar to the unconstrained case.



- For the second scenario:
  $f(x) = (x_1 - 1.1)^2 + (x_2 - 1.1)^2$ and $g(x) = x_1^2 + x_1^2 - 1$
  - We can clearly observe that the unconstrained local minimum lies outside the feasible region.
  - Basically, the constrained local minimum lies on the surface of the constraint surface and effectively we obtain an optimization problem with an equality constraint $g(x) = 0$.
  - We are able to obtain a local minima when $-\nabla f(x)$ and $\nabla g(x)$ are parallel. This means $\nabla f(x) + \mu \nabla g(x) = 0$.
  - Also for any z that is orthogonal to $\nabla g(x)$, $z^T \nabla^2 f(x^*) z \geq 0$
  - For x^*, $mu^* g(x^*) = 0$. This is the complimentary slackness condition.
  - W.k.t $\mathcal{L}(x, \mu^) = f(x) + \sum_{i=1}^{m} \mu_i g_i(x)$ where $\mu_i \geq 0$.
  - $\nabla_x \mathcal{L}(x^*, \mu_i^*) = 0 \implies \nabla_{x^*} f(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla_{x^*} g_i(x^*) = 0$
  - So according to complimentary slackness condition, $mu_i^* g_i(x^*) = 0$. Also, $\nabla_\mu \mathcal{L}(x^*, \mu_i^*) = 0$. This implies if $\mu_i > 0$, then $g_i(x^*) = 0$. Otherwise, $\mu_i^* = 0$



4 Karush Kuhn Tucker (KKT) conditions

- We can generalize an unconstrained optimization problem with inequalities using certain conditions called KKT conditions:
- Given a constrained optimization problem: $minf(x)$ subject to $h_i(x) = 0$ and $g_j(x) \leq 0$ for $i = 1, ..l$ and $j = 1, ...m$
- The Lagrangian is $\mathcal{L}(x, \mu, \lambda) = f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{i=1}^{m} \lambda_i h_i(x)$
- The following conditions are the KKT conditions:
  - First Order Condition:
    $\nabla_x^* f(x^*) + \sum_{i=1}^{m} \mu_i^* g_i(x^*) + \sum_{i=1}^{m} \lambda_i^* h_i(x^*) = 0$
  - Feasiblility conditions:
    $h_j(x^*) = 0 \forall j = 1, 2, ...l$
    $g_i(x^*) \leq 0 \forall i = 1, 2, ...l$
  - Complimentary slackness:
    $\mu_i^* g_i(x^*) forall i = 1, 2....m$
  - $\mu_i^* \geq 0$
  - The Lagrangian should be positive definite.

5 Fritz John Conditions

- Fritz John conditions are same as KKT conditions (as mentioned in above) except the $First\ Order\ Condition$:
  $\mu_0 \nabla_x^* f(x^*) + \sum_{i=1}^{m} \mu_i^* g_i(x^*) + \sum_{j=1}^{l} \lambda_j^* h_j(x^*) = 0$
- Here, we have an additional constraint i.e. $\mu_0 >= 0$.This is required when it is impossible to form linearly independent basis. Hence, $\mu_0 > 0$ if $\nabla g_i \ \ and \ \ \nabla h_j$ are linearly independent i.e. when KKT conditions holds.
- Consider the following optimization problem:

  $minimize \ \ f(x)$
  $subject\ to: \ \ g_i(x) \leq 0, \ \ i \in \{1, ..., m\}$
  $subject\ to: \ \ h_j(x) = 0, \ \ j \in \{1, ..., l\}$

  where f is the function to be minimized, g_i is the inequality constraints, and h_j is the equality constraints.
- Unlike KKT, Fritz John conditions are a necessary condition for a solution in nonlinear programming to be optimal. In general, they are used as lemma in the proof of the Karush Kuhn Tucker conditions.
  - Let us take an example: $minimize \ \ x$
    such that: $x^2 = 0$
    so, $f(x) = x$
    $h(x) = x^2$
  - FOC: $\nabla f + \lambda^* \nabla h = 0$
    $=> 1 + \lambda^* 2x = 0$
  - Feasibility conditions: $h(x^*) = 0$
    $=> (x^*)^2 = 0$
    $=> x^* = 0$

- As in this example both the conditions are not satisfied, yet clearly $x^* = 0$ is optimal. Thus, KKT conditions are not necessary, but are optimal conditions.
- Hence, the crucial difference between KKT and FJ of optimality conditions is that when there is at least one nonlinear constraint, a constraint qualification (CQ) condition must be satisfied for the KKT conditions to be necessary for optimality. The Fritz-John conditions hold at any local minimizer regardless of whether a CQ holds or not.

6 Channel Bit Rate Problem: Water Filling Algorithm

- **Problem**:
  - To maximize the transmission rate of a communication system, with 'n' communication channels.
  - And the transmission rate of the communication system is: $\sum_{i=1}^{n} log(\alpha_i + x_i)$.
  - where power of i^{th} channel is x_i and $\alpha_i^{\prime}s$ are constants.
- Therefore, the Optimization Problem is:

$$\begin{gathered} max_x \sum_{i=1}^{n} log(\alpha_i + x_i) \\ \text{s.t. } \sum_{i=1}^{n} x_i = 1 \\ x_i \geq 0 \ \ \forall i \end{gathered}$$

- Lagrangian for above problem is:
  $\mathcal{L}(x, \mu, \lambda) = -\sum_{i=1}^{n} log(\alpha_i + x_i) \quad + \quad \sum_{i=1}^{n} \mu_i(-x_i) \quad + \quad \lambda(\sum_{i=1}^{n} x_i - 1)$
- KKT conditions are:
  - First Order Condition:

$$\frac{-1}{\alpha_i + x_i} - \mu_i^* + \lambda^* = 0$$

  - Feasibility Conditions:

$$\begin{gathered} x_i^* \geq 0 \ \ \forall i \\ \sum_{i=1}^{n} x_i^* = 1 \\ \mu_i^* \geq 0 \ \ \forall i \end{gathered}$$

  - Complementary Slackness Conditions:

$$\mu_i^* x_i^* = 0 \ \ \forall i$$

- Now, as $\lambda^* = \mu^* + \dfrac{-1}{\alpha_i + x_i}$
  but from feasibility conditions: $\mu_i^* \geq 0 \ \ \forall i$
  => $\lambda^* \geq \dfrac{-1}{\alpha_i + x_i}$
  Now, two possible cases i.e. $\lambda^* < \dfrac{1}{\alpha_i}$ or $\lambda^* \geq \dfrac{1}{\alpha_i}$
  Case-1: When $\lambda^* < \dfrac{1}{\alpha_i}$
  => $x_i = \dfrac{1}{\lambda^*} - \alpha_i$ (as from above equations $x_i^* > 0 \mu_i^* = 0$
  Case-2: when $\lambda^* \geq \dfrac{1}{\alpha_i}$
  This case is not possible as, if $x_i^* > 0$
  => $\mu_i^* > 0$ //But, both cannot be greater that zero simultaneously. Hence, this case is not possible.

- Therefore the original problem of maximizing the transmission rate can be simplified to:

$$x_i^* = \frac{1}{\lambda^*} - \alpha_i \text{ when } \lambda^* < \frac{1}{\alpha_i}$$
$$= 0 \text{ other wise and } \sum_{i=1}^{n} x_i^* = 1$$