7 Continuos Optimization

7.1 Opt using gradient Desc 227-233
7.2 Constrained opt & lagrange 233-236

As ML is implemented by computer - the Math formulations are expresed as numerical optimization methods

Training ML = find good parameters

"good" is defined by the objective func

2 Main branches of opt — unconstrained & constrained

Assumption that obj func is Diferable

Hence we have a gradient @ each location in the space to help us find the optimum value

by convention, most obj funcs in ML are intended to be minimized

global minima

Gradients can be used to help find the minimum - indicating whether we should step left or right

We can solve for all stationary points of a function by calcing derivatives & setting to 0

Stationary points are real ~~points~~ roots of the derivative & points that have zero gradient

Example:

$$l(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$$

$$\frac{dl(x)}{d(x)} = 4x^3 + 21x^2 + 10x - 17$$

this is a cubic equation, it has in general 3 solutions when set to 0

in example 2 = mins, 1 = max

to check whether stationary point is a min or max, we take 2nd Derivative & check whether it eq is pos or neg at stationary point

$$\frac{d^2 l(x)}{dx^2} = 12x^2 + 42x + 10$$

by submit visual est values $x = -4.5$, -1.4, 0.7 we see that middle point < 0
So it is a max
and theres one Minimum

Note - low-order polynomials are solvable analytically high-orders either exist or are too costly

As a ~~fe~~ find the result by starting @ some value $x_0 = -6$ & follow the negative gradient

Gradient tells us the direction to go but not how far (step size)

~~Thar~~ Furthermore if there are > 1 minimum then our start point could bring us down to the wrong stationary minima

7.1 Optimization Using Grad Desc

Consider problem of solving real-valued function: $\min f(x)$

$f = \mathbb{R}^d \to \mathbb{R}$ = ML Problem

f is Differable but no Analytical solution

Gradient Descent is a first order opt Alg
- This means it only uses the first order derivative
- this is cheap, simple & applicable

- Gradient gives direction of Steepest Asc
- the opposite gives steepest Descent

Gradient points in a direction that is orthogonal to the contour lines lines we wish to optimize

$$w_1 = w_0 - \gamma\left((\nabla f)(w_0)\right)^\top$$

Start w/ init w_0 & interate

for a suitable step size the sequence will converge to a minima

Note, near to the minima GD is slow & the rate is inferior to many other methods. Zig zag many times around the min point

7.2 Constrained Opt & Lagrange Mult

Previous 7.1 Solve min of function

$\min f(x)$ where $f: R^p \to R$

Consider additional constraints

$$\min_{x} f(x)$$
$$\text{Subject to } g_i(x) \leq 0 \text{ for all } i = 1, \ldots, m$$

One way of turning constrained prob into unconstrained is to use indic func

$$O(x) = F(x) + \sum_{i=1}^{m} I(g_i(x))$$

$I(z)$ is an infinite Step Func

$$I(z) = \begin{cases} 0 & \text{if } z \leq 0 \\ \infty & \text{otherwise} \end{cases}$$

But this route is also difficult to solve for

the solution is to introduce lagrange multiples

idea of lagrane is to replace step function w/ a linear one

Start Min $f(x)$ Subj to $g_i(x) \leq 0$

lagrangian:
intro λ to each inequal const

$$\hat{f}(x, \lambda) = f(x) + \sum^{m} \lambda_i g_i(x)$$

vectorised $= f(x) + \lambda^T g(x)$

Introduction of lagrangian Duality

in opt, Dual = convert Prob in one set of vars x (Primal) into another problem w/ new set of vars(λ) (Dual)

(Primal)

Start: $\min_{x} f(x)$ Subj to $g_i(x) \leq 0$
for all $i = 1, \ldots, m$

Dual: $\max_{\lambda \in \mathbb{R}^m} D(\lambda)$ Subj to $\lambda \geq 0$

Dual var $= \lambda$ & $D(\lambda) = \min \mathcal{L}(x, \lambda)$