Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 10 - Optimization Foundations Applied to Regression Formulations

Foundations: Necessary Condition 2

- *Is* $\nabla^2 f(\mathbf{w}^*)$ *positive definite ?*
  i.e. $\forall \mathbf{x} \neq 0$, *is* $\mathbf{x}^T \nabla f(\mathbf{w}^*)\mathbf{x} > 0$? (A sufficient condition for local minimum)
  (Any positive definite matrix is also positive semi-definite)
  `(Cite : Section 3.12 & 3.12.1)`[1]

$$
\begin{aligned}
\nabla^2 f(\mathbf{w}^*) &= 2\Phi^T\Phi + 2\lambda I && (1)\\
\implies \mathbf{x}^T \nabla^2 f(\mathbf{w}^*)\mathbf{x} &= 2\mathbf{x}^T(\Phi^T\Phi + \lambda I)\mathbf{x} && (2)\\
&= 2\left((\Phi + \sqrt{\lambda} I)\mathbf{x}\right)^T \Phi\mathbf{x} && (3)\\
&= 2\left\|(\Phi + \sqrt{\lambda} I)\mathbf{x}\right\|^2 \geq 0 && (4)
\end{aligned}
$$

- And with $\lambda = 0$, if Φ **has full column rank** ,

$$\Phi\mathbf{x} = 0 \quad \textit{iff} \quad \mathbf{x} = 0 \qquad (5)$$

$\therefore$ If $\mathbf{x} \neq 0$, $\quad \mathbf{x}^T \nabla^2 f(\mathbf{w}^*)\mathbf{x} > 0$

[1] `CS725/notes/classNotes/LinearAlgebra.pdf`

Conclusion based on discussion of solution to problem 5 of Tuts 3 & 4:

① $\lambda_k(\phi^T\phi + \lambda I) = \lambda_k(\phi^T\phi) + \lambda$ (note: $\lambda \geq 0$)

② $\Rightarrow$ Each eigenvalue of $(\phi^T\phi + \lambda I)$ will be positive & $(\phi^T\phi + \lambda I)$ will be positive definite since: $v^T(\phi^T\phi + \lambda I)v = \|\phi v\|^2 + \lambda\|v\|^2 > 0$ $\Big\}$ if $v \neq 0$

③ $\Rightarrow$ Hessian $\phi^T\phi + \lambda I$ will be "more" positive definite

Smaller condition # $\Rightarrow$ more stable computation

④ $\Rightarrow$ Also ratio $\dfrac{\lambda_1(\phi^T\phi + \lambda I)}{\lambda_n(\phi^T\phi + \lambda I)} \leq \dfrac{\lambda_1(\phi^T\phi)}{\lambda_n(\phi^T\phi)}$

Example of linearly correlated features

- Example where Φ doesn't have a full column rank,

$$\Phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \tag{6}$$

- This is the simplest form of linear correlation of features, and it is not at all desirable.
- Effect of a nonzero λ with such Φ is that

Though $\phi^T\phi$ is positive semidefinite (& NOT positive def)

$(\phi^T\phi + \lambda I)$ WILL be positive definite for $\forall \lambda \geq 0$

Example of linearly correlated features

- Example where Φ doesn't have a full column rank,

$$\Phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \tag{6}$$

- This is the simplest form of linear correlation of features, and it is not at all desirable.
- Effect of a nonzero λ with such Φ is that it tends to make the Hessian more positive definite

Do Closed-form solutions Always Exist?

- Linear regression and Ridge regression both have closed-form solutions
  - For linear regression,

$$w^* = (\Phi^\top \Phi)^{-1} \Phi^\top y$$

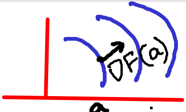
  - For ridge regression,

$$w^* = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top y$$

  (for linear regression, $\lambda = 0$)

- What about optimizing the formulations (constrained/penalized) of Lasso (L_1 norm)? And support-based penalty (L_0 norm)?: Also requires tools of Optimization/duality

Gradient Descent Algorithm



Gradient descent is based on our previous observation that if the multivariate function $F(\mathbf{x})$ is defined and differentiable in a neighborhood of a point $\mathbf{a}$, then $F(\mathbf{x})$ decreases fastest if one proceeds from $\mathbf{a}$ in the direction of the negative of the gradient of F at $\mathbf{a}$,i.e. $-\nabla F(\mathbf{a})$.

Therefore,

$$\underbrace{\Delta \mathbf{w}^{(\mathbf{k})}}_{\text{step direction}} = -\nabla \mathbf{E}(\mathbf{w}^{(\mathbf{k})}) \tag{7}$$

Hence,

$$\mathbf{w}^{(\mathbf{k+1})} = \mathbf{w}^{(\mathbf{k})} + \mathbf{2t}^{(\mathbf{k})}(\mathbf{\Phi}^{\mathbf{T}}\mathbf{y} - \mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi}\mathbf{w}^{(\mathbf{k})} - \lambda \mathbf{w}^{(\mathbf{k})}) \tag{8}$$

$$w^{(k+1)} = w^{(k)} + t^{(k)} \Delta w^{(k)}$$

step size

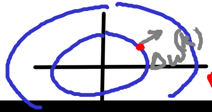
Gradient Descent Algorithm

Find $w = \arg\min_{w} E(w)$

Find starting point $\mathbf{w}^{(0)} \epsilon \mathcal{D}$

- $\Delta \mathbf{w}^{\mathbf{k}} = -\nabla \varepsilon(\mathbf{w}^{(\mathbf{k})})$
- Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search.
- Obtain $\mathbf{w}^{(\mathbf{k+1})} = \mathbf{w}^{(\mathbf{k})} + \mathbf{t}^{(\mathbf{k})} \boldsymbol{\Delta} \mathbf{w}^{(\mathbf{k})}$.
- Set $k = k + 1$. **until** stopping criterion (such as $\| \nabla \varepsilon(w^{(k+1)}) \| \leq \epsilon$) is satisfied → ideal stopping criterion: $\nabla E(w^{(w)}) = 0$ $\|\nabla E\|$ is a proxy

Exact: $t^{(k)} = \arg\min_{t} \underbrace{E(w^{(k)} + t\,\Delta w^{(k)})}_{\text{Reduced } n \text{ dim problem to 1-dimension!}}$



fn(t) →

Tut problem 3: $E(w) = \|\phi w - y\|^2$

$w^{(0)} = 0$

$\nabla E(w^{(k)}) = 2\phi^T\phi w^{(k)} - 2\phi^T y, \quad \nabla E(w^{(0)}) = -2\phi^T y$

$$t^{(k)} = \underset{t}{\operatorname{argmin}}\, E\left[w^{(k)} - t\,\nabla E(w^{(k)})\right]$$

$$t^{(0)} = \underset{t}{\operatorname{argmin}}\, E\left[w^{(0)} + 2t\,\phi^T y\right]$$

$$= \underset{t}{\operatorname{argmin}}\, E\left[2t\,\phi^T y\right]$$

$$= \underset{t}{\operatorname{argmin}}\, \|\phi(2t\phi^T y) - y\|^2 = \underset{t}{\operatorname{argmin}}\, \|2t\,\phi\phi^T y - y\|^2$$

$=$

Gradient Descent Algorithm

Exact line search algorithm to find $t^{(k)}$

- The line search approach first finds a descent direction along which the objective function f will be reduced and then computes a step size that determines how far **x** should move along that direction.
- In general,

$$t^{(k)} = \arg\min_t f\left(\mathbf{w}^{(\mathbf{k+1})}\right) \tag{9}$$

- Thus,

$$t^{(k)} = \arg\min_t f(w^{(k)} + t\Delta w^{(k)})$$

Exact line search algorithm to find $t^{(k)}$

- The line search approach first finds a descent direction along which the objective function f will be reduced and then computes a step size that determines how far **x** should move along that direction.
- In general,

$$t^{(k)} = \arg\min_t f\left(\mathbf{w^{(k+1)}}\right) \tag{9}$$

- Thus,

$$t^{(k)} = \arg\min_t \left(\mathbf{w^{(k)}} + \mathbf{2t}\left(\mathbf{\Phi^T y} - \mathbf{\Phi^T}\phi\mathbf{w^{(k)}} - \lambda\mathbf{w^{(k)}}\right)\right) \tag{10}$$

Tut 3, prob 2, $w^{(0)} = 0 \Rightarrow t^{(0)} = ?$

Example of Gradient Descent Algorithm

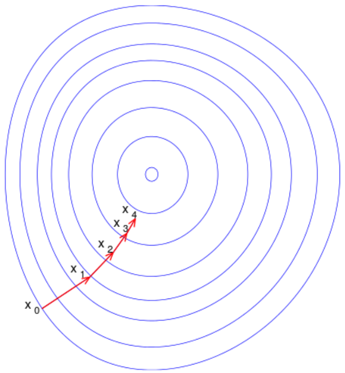


Figure 1: A red arrow originating at a point shows the direction of the negative gradient at that point. Note that the (negative) gradient at a point is orthogonal to the level curve going through that point. We see that gradient descent leads us to the bottom of the bowl, that is, to the point where the value of the function F is minimal. Source: Wikipidea

Constrained Least Squares Linear Regression

Find

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\phi\mathbf{w} - \mathbf{y}\|^2 \;\; s.t. \;\; \|\mathbf{w}\|_p \leq \zeta, \tag{11}$$

where

$$\|\mathbf{w}\|_p = \left(\sum_{i=1}^{n} |w_i|^p\right)^{\frac{1}{p}} \tag{12}$$

Claim: This is an equivalent reformulation of the penalized least squares. Why?

Other motivations → SVR & it dual

→ Lasso $\lambda\|w\|_1$ or $\|w\|_1 \leq \theta$

p-Norm level curves

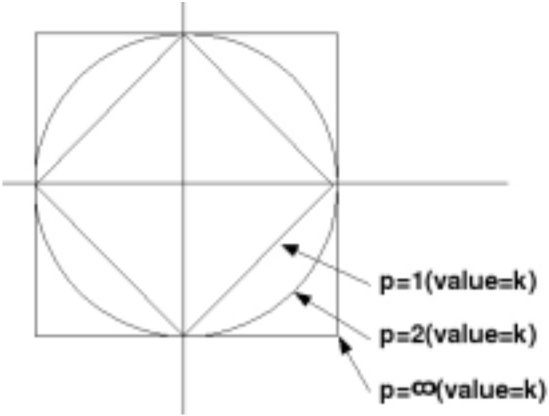


Figure 2: p-Norm curves for constant norm value and different p

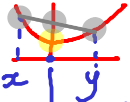
Convex Optimization Problem

- Formally, a convex optimization problem is an optimization problem of the form

$f(\theta x+(1-\theta)y) \leq \theta f(x)+(1-\theta)f(y)$

$$\text{minimize } f(\mathbf{x}) \tag{13}$$

$$\text{subject to } c \in C \tag{14}$$

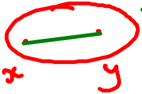


where f is a convex function, C is a convex set, and $\mathbf{x}$ is the optimization variable.

$\theta x+(1-\theta)y$

$\theta\in[0,1]$

- An improved form of the above would be



$\forall \theta\in[0,1]$ $\theta x+(1-\theta)y \in C$

$$\text{minimize } f(\mathbf{x}) \tag{15}$$

$$\text{subject to } g_i(\mathbf{x}) \leq 0,\ i=1,\ldots,m \tag{16}$$

$$h_i(\mathbf{x}) = 0, i=1,\ldots,p \tag{17}$$

convex g_i's & linear h_i $\Rightarrow$ convex C

where f is a convex function, g_i are convex functions, and h_i are affine functions, and $\mathbf{x}$ is the vector of optimization variables.

linear

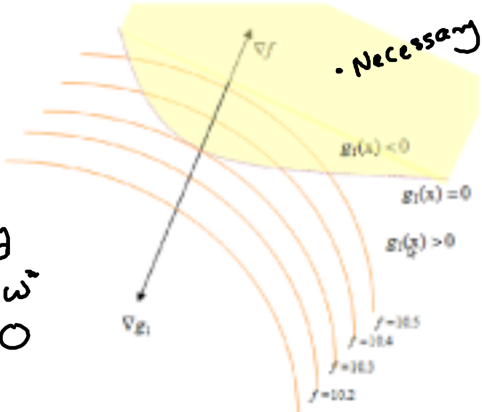
Eg: $g_i(x)=\|w\|_p^2$ $\quad f(x)=\|\phi w-y\|^2$

Q. *How to solve constrained problems of the above-mentioned type?*

A. General problem format :

$$Minimize\ f(\mathbf{w})\ s.t.\ g_1(\mathbf{w}) \leq 0 \tag{18}$$

- Necessary cond $\nabla f(w^*) = 0$ if $g_1(w^*) < 0$



H/W: Think of what should happen at w^* if $g_1(w^*) = 0$