Lecture 4: Convex Functions

Gonzalo De La Torre Parra, Ph.D.

Fall 2021

Linear Methods in Regression (Review)

- **Data**: We are given n training point $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Now, $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$.
- **Linear regression or ordinary least square**: The training process for the search for the best linear predictor is given by the optimization problem

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2.$$

  Here, $\beta = (\beta_1, \ldots, \beta_p)$.
- **Ridge regression**:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2 + \lambda \sum_{i=1}^{p} \beta_i^2.$$

  Keywords: Stability, Invertibility, Shrinkage.
- **LASSO (least absolute shrinkage and selection operator)**:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2 + \lambda \sum_{i=1}^{p} |\beta_i|.$$

  Keywords: Variable selection, soft-thresholding.
- **Elastic Net**

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2 + \lambda_1 \sum_{i=1}^{p} \beta_i^2 + \lambda_2 \sum_{i=1}^{p} |\beta_i|.$$

Convex Functions and Optimality

Definition (Convex Set)

A set C in $\mathbb{R}^p$ is called convex if

$$x, y \in C \quad \Longrightarrow \quad \alpha x + (1-\alpha)y \in C, \quad \text{for any } \alpha \in [0,1].$$

Definition (Convex Function)

A function $f(x) : \mathbb{R}^p \to \mathbb{R}$ is convex if it is defined on a convex set C and satisfies for $\alpha \in [0,1]$ and $x, y \in C$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y).$$

- $f(x) = x^2$ is convex on $\mathbb{R}$.
- $f(x) = |x|$ is convex on $\mathbb{R}$.
- $f(x) = e^x$ is convex $\mathbb{R}$.
- $f(x) = \log(x)$ is **not** convex on $(0, \infty)$.
- $f(x) = x^3$ is **not** convex on $(-\infty, \infty)$, but **is** convex on $(0, \infty)$.

Simple Examples Of Convex Functions

- **Proof that x^2 is convex**: We need to show that

$$(\alpha x + (1-\alpha)y)^2 \leq \alpha x^2 + (1-\alpha)y^2.$$

The above equation is true if all the following equations are true:

$$\begin{aligned}
\alpha^2x^2 + (1-\alpha)^2y^2 + 2\alpha(1-\alpha)xy &\leq \alpha x^2 + (1-\alpha)y^2 \\
(1-\alpha)^2y^2 + 2\alpha(1-\alpha)xy &\leq \alpha x^2 - \alpha^2x^2 + (1-\alpha)y^2 \\
2\alpha(1-\alpha)xy &\leq \alpha(1-\alpha)x^2 + \alpha(1-\alpha)y^2 \\
2xy &\leq x^2 + y^2 \\
0 &\leq x^2 + y^2 - 2xy \\
0 &\leq (x-y)^2.
\end{aligned}$$

The last equation is always true. So, the first equation is true and the function $f(x) = x^2$ is convex.

- **Proof that $|x|$ is convex**: We need to show that

$$|\alpha x + (1-\alpha)y| \leq \alpha|x| + (1-\alpha)|y|.$$

  But, this is trivially true because for any two numbers a and b

$$|a+b| \leq |a| + |b|.$$

- **Verification not always straightforward**: It is not always easy to verify if a given function is convex. The following theorem sometimes helps.

- **Positive semidefinite matrix**: A $p \times p$ symmetric matrix A is called positive semidefinite if for any $h \in \mathbb{R}^p$

$$h^T A\, h \geq 0.$$

  For example, the matrices I_p (identity matrix) and $A^T A$ are positive semidefinite:

$$h^T I_p\, h = h^T h \geq 0, \qquad h^T A^T A\, h = (Ah)^T (Ah) \geq 0.$$

Characterization of Convex Functions

Theorem

A function $f(x) : \mathbb{R}^p \to \mathbb{R}$ is convex if it is defined on a convex set C and satisfies any of these three equivalent conditions:

1. $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$*, for any* $\alpha \in [0,1]$ *and any* $x, y \in C$.
2. $f(y) \geq f(x) + \nabla f(x)^T (y - x)$*, for any* $x, y \in C$. *(When* $\nabla f(x)$ *exists)*
3. $\nabla^2 f(x)$ *is positive semidefinite for any* $x \in C$. *(When* $\nabla^2 f(x)$ *exists)*

If $x = (x_1, \ldots, x_p)$ has p components then

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_p} \end{bmatrix} \qquad \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2}, \cdots, & \frac{\partial^2 f(x)}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2}, \cdots, & \frac{\partial^2 f(x)}{\partial x_2 \partial x_p} \\ \vdots & & \\ \frac{\partial^2 f(x)}{\partial x_p \partial x_1} & \frac{\partial^2 f(x)}{\partial x_p \partial x_2}, \cdots, & \frac{\partial^2 f(x)}{\partial x_p^2} \end{bmatrix}$$

Characterization of Convex Functions

Theorem

A function $f(x) : \mathbb{R}^p \to \mathbb{R}$ is convex if it is defined on a convex set C and satisfies any of these three equivalent conditions:

1. $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$, *for any* $\alpha \in [0,1]$ *and any* $x, y \in C$.
2. $f(y) \geq f(x) + \nabla f(x)^T(y - x)$, *for any* $x, y \in C$. *(When* $\nabla f(x)$ *exists)*
3. $\nabla^2 f(x)$ *is positive semidefinite for any* $x \in C$. *(When* $\nabla^2 f(x)$ *exists)*

Applications of the Theorem

- e^x **is convex**: Taking the derivative twice gives us

$$\frac{d^2}{dx^2} e^x = e^x \geq 0, \quad \text{for all } x.$$

- x^3 **is not convex**:

$$\frac{d^2}{dx^2} x^3 = 6x < 0, \quad \text{for } x < 0.$$

- $\log(x)$ **is not convex**:

$$\frac{d^2}{dx^2} \log(x) = -\frac{1}{x^2} < 0, \quad \text{for all } x.$$

- **Note that we cannot check if $|x|$ is convex by differentiating it because it is not differentiable.**

Optimization of a Convex Function

Theorem

A convex differential function $f(x) : \mathbb{R}^p \to \mathbb{R}$ defined on a convex set C achieves its global minimum at a point x^* if the gradient at x^* is equal to the all zero vector

$$\nabla f(x^*) = 0.$$

- **Proof**: The result follows from the previous theorem on characterization on convexity because

$$f(y) \geq f(x^*) + \nabla f(x^*)^T (y - x^*), \text{ for any } y \in C.$$

Thus, $\nabla f(x^*) = 0$ implies

$$f(y) \geq f(x^*), \text{ for any } y \in C.$$

Hessian Calculation

- Definition

$$\nabla^2 f(\beta) = \begin{bmatrix} \frac{\partial^2 f(\beta)}{\partial \beta_1^2} & \frac{\partial^2 f(\beta)}{\partial \beta_1 \partial \beta_2}, \cdots, \frac{\partial^2 f(\beta)}{\partial \beta_1 \partial \beta_p} \\ \frac{\partial^2 f(\beta)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 f(\beta)}{\partial \beta_2^2}, \cdots, \frac{\partial^2 f(\beta)}{\partial \beta_2 \partial \beta_p} \\ \vdots & \\ \frac{\partial^2 f(\beta)}{\partial \beta_p \partial \beta_1} & \frac{\partial^2 f(\beta)}{\partial \beta_p \partial \beta_2}, \cdots, \frac{\partial^2 f(\beta)}{\partial \beta_p^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_1} \nabla f(\beta) & \frac{\partial}{\partial \beta_2} \nabla f(\beta) & \cdots & \frac{\partial}{\partial \beta_p} \nabla f(\beta) \end{bmatrix}$$

- Now let us assume that $a_1 = (a_{11}, \cdots, a_{1p})$, $a_2 = (a_{21}, \cdots, a_{2p})$, etc are the rows of A. Then,

$$A\beta = \begin{bmatrix} a_{11} & \dots & a_{1p} \\ \vdots & & \\ a_{p1} & \dots & a_{pp} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \beta^T a_1 \\ \vdots \\ \beta^T a_p \end{bmatrix}.$$

- Thus,

$$\begin{aligned} \nabla^2_\beta \left(\beta^T A\beta\right) &= \left[\tfrac{\partial}{\partial\beta_1}\nabla f(\beta) \quad \tfrac{\partial}{\partial\beta_2}\nabla f(\beta) \quad \dots \quad \tfrac{\partial}{\partial\beta_p}\nabla f(\beta)\right] \\ &= \left[\tfrac{\partial}{\partial\beta_1}2A\beta \quad \tfrac{\partial}{\partial\beta_2}2A\beta \quad \dots \quad \tfrac{\partial}{\partial\beta_p}2A\beta\right] \\ &= 2\begin{bmatrix} a_1^T \\ \vdots \\ a_p^T \end{bmatrix} = 2A. \end{aligned}$$

Linear Regression

- **Training data**: $x \in \mathbb{R}^p$, and $y \in \mathbb{R}$:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n).$$

- **Linear Predictors**: For a given fixed $\beta = (\beta_1, \cdots, \beta_p) \in \mathbb{R}^p$, a linear predictor maps $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ to $\beta^T x_i$, for each i:

$$h(x_i) = \beta^T x_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad i = 1, 2, \ldots, n.$$

- **Training**: Training for linear regression or ordinary least square (OLS) is the search for the best linear predictor or the best β:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2.$$

Verify that you will get the same solution if you solve the problem without the scaling factor $\frac{1}{n}$:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2.$$

- **Matrix $\mathbb{X}$ and the vector $\mathbb{Y}$:**

$$\mathbb{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbb{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}.$$

- **The vector** $\mathbb{X}\beta$:

$$\mathbb{X}\beta = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} x_1^T \beta \\ \vdots \\ x_n^T \beta \end{bmatrix}.$$

- **The difference vector** $\mathbb{Y} - \mathbb{X}\beta$:

$$\mathbb{Y} - \mathbb{X}\beta = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} x_1^T \beta \\ \vdots \\ x_n^T \beta \end{bmatrix} = \begin{bmatrix} y_1 - x_1^T \beta \\ \vdots \\ y_n - x_n^T \beta \end{bmatrix}.$$

- **The training or optimization problem in matrix form**:

$$
\begin{aligned}
\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(\beta^T x_i - y_i)^2 &= \min_{\beta\in\mathbb{R}^p} \frac{1}{n}\left(\mathbb{Y}-\mathbb{X}\beta\right)^T\left(\mathbb{Y}-\mathbb{X}\beta\right)\\
&= \min_{\beta\in\mathbb{R}^p} \frac{1}{n}\left(\mathbb{Y}^T\mathbb{Y}+\beta^T\mathbb{X}^T\mathbb{X}\beta-2\beta^T\mathbb{X}^T\mathbb{Y}\right).
\end{aligned}
$$

- **How do you solve this optimization problem?**
  - First, check if it is a convex function.
  - Second, use the fact that a convex differentiable function can be optimized by setting the gradient to the all-zero vector.
  - Third, identify conditions under which the equation in which we set the gradient equal to zero can be explicitly solved for the optimal solution.

Theorem

A function $f(x) : \mathbb{R}^p \to \mathbb{R}$ is convex if it is defined on a convex set C and satisfies any of these three equivalent conditions:

1. $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$, *for any* $\alpha \in [0,1]$ *and any* $x, y \in C$.
2. $f(y) \geq f(x) + \nabla f(x)^T (y - x)$, *for any* $x, y \in C$. *(When* $\nabla f(x)$ *exists)*
3. $\nabla^2 f(x)$ *is positive semidefinite for any* $x \in C$. *(When* $\nabla^2 f(x)$ *exists)*

- **Calculate Hessian**:

$$\nabla^2_\beta \left(\mathbb{Y}^T \mathbb{Y} + \beta^T \mathbb{X}^T \mathbb{X} \beta - 2\beta^T \mathbb{X}^T \mathbb{Y} \right) = 2\mathbb{X}^T \mathbb{X}.$$

- **Check positive-semidefiniteness**: For any $h \in \mathbb{R}^p$,

$$2h^T \mathbb{X}^T \mathbb{X} h = 2\langle \mathbb{X}h, \mathbb{X}h \rangle \geq 0.$$

This implies that the Hessian is positive semidefinite and the function is convex in β.

Theorem

A convex differential function $f(x) : \mathbb{R}^p \to \mathbb{R}$ defined on a convex set C achieves its global minimum at a point x^* if the gradient at x^* is equal to the all zero vector

$$\nabla f(x^*) = 0.$$

- **The gradient**:

$$\nabla_\beta \left(\mathbb{Y}^T \mathbb{Y} + \beta^T \mathbb{X}^T \mathbb{X} \beta - 2\beta^T \mathbb{X}^T \mathbb{Y} \right) = 2\mathbb{X}^T \mathbb{X} \beta - 2\mathbb{X}^T \mathbb{Y}.$$

- **Linear regression solution**: Setting gradient equal to zero we get

$$\mathbb{X}^T \mathbb{X} \beta = \mathbb{X}^T \mathbb{Y}.$$

Can we explicitly solve the above equation? We can if the matrix $\mathbb{X}^T \mathbb{X}$ is invertible.

When is the matrix $\mathbb{X}^T \mathbb{X}$ invertible? We will come back to this topic soon.

Ridge Regression

- **Linear regression**:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(\beta^T x_i - y_i)^2 = \min_{\beta\in\mathbb{R}^p} \frac{1}{n}\left(\mathbb{Y}-\mathbb{X}\beta\right)^T\left(\mathbb{Y}-\mathbb{X}\beta\right)$$

$$= \min_{\beta\in\mathbb{R}^p} \frac{1}{n}\left(\mathbb{Y}^T\mathbb{Y} + \beta^T\mathbb{X}^T\mathbb{X}\beta - 2\beta^T\mathbb{X}^T\mathbb{Y}\right).$$

- **Ridge regression**: Fix $\lambda > 0$, remove scaling by $\frac{1}{n}$ and solve

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n}(\beta^T x_i - y_i)^2 + \lambda\sum_{i=1}^{p}\beta_i^2$$

$$= \min_{\beta\in\mathbb{R}^p} \left(\mathbb{Y}-\mathbb{X}\beta\right)^T\left(\mathbb{Y}-\mathbb{X}\beta\right) + \lambda\beta^T\beta$$

$$= \min_{\beta\in\mathbb{R}^p} \left(\mathbb{Y}^T\mathbb{Y} + \beta^T\mathbb{X}^T\mathbb{X}\beta - 2\beta^T\mathbb{X}^T\mathbb{Y} + \lambda\beta^T\beta\right).$$

Thus, in ridge regression there is an extra term. If we set $\lambda = 0$, then ridge regression reduces to linear regression.

- **Ridge objective function**:

$$\begin{aligned} &\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n}(\beta^T x_i - y_i)^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \\ &\quad = \min_{\beta\in\mathbb{R}^p} \left(\mathbb{Y}^T\mathbb{Y} + \beta^T\mathbb{X}^T\mathbb{X}\beta - 2\beta^T\mathbb{X}^T\mathbb{Y} + \lambda\beta^T\beta\right). \end{aligned}$$

- **Convexity**: To check convexity, we compute the Hessian to get

$$\nabla_\beta^2 \left(\mathbb{Y}^T\mathbb{Y} + \beta^T\mathbb{X}^T\mathbb{X}\beta - 2\beta^T\mathbb{X}^T\mathbb{Y} + \lambda\beta^T\beta\right) = 2\left(\mathbb{X}^T\mathbb{X} + \lambda I_p\right).$$

The Hessian is positive semidefinite because for any $h \in \mathbb{R}^p$,

$$2h^T\left(\mathbb{X}^T\mathbb{X} + \lambda I_p\right)h = 2\langle\mathbb{X}h, \mathbb{X}h\rangle + 2\lambda h^T h \geq 0.$$

Thus, the objective function is convex.

- **Gradient**: The gradient is

$$\begin{aligned}
&\nabla_\beta \left(\mathbb{Y}^T \mathbb{Y} + \beta^T \mathbb{X}^T \mathbb{X} \beta - 2\beta^T \mathbb{X}^T \mathbb{Y} + \lambda \beta^T \beta \right) \\
&= 2\mathbb{X}^T \mathbb{X} \beta - 2\mathbb{X}^T \mathbb{Y} + 2\lambda\beta \\
&= 2\left(\mathbb{X}^T \mathbb{X} + \lambda I_p \right) \beta - 2\mathbb{X}^T \mathbb{Y}.
\end{aligned}$$

- **Ridge regression solution**: Setting gradient equal to zero we get

$$\left(\mathbb{X}^T \mathbb{X} + \lambda I_p \right) \beta = \mathbb{X}^T \mathbb{Y}.$$

Can we explicitly solve the above equation? We can if the matrix $\left(\mathbb{X}^T \mathbb{X} + \lambda I_p \right)$ is invertible.

When is the matrix $\mathbb{X}^T \mathbb{X} + \lambda I_p$ invertible? It turns out, it is always invertible!

Definition

A $p \times p$ square symmetric matrix A is called positive definite if for any non-zero $h \in \mathbb{R}^p$,

$$h^T A\, h > 0.$$

Lemma

A square matrix is invertible if it is positive definite.

- **Check positive definiteness**: Let $h \neq 0$. Then

$$h^T \left(\mathbb{X}^T\mathbb{X} + \lambda I_p\right) h = \langle \mathbb{X}h, \mathbb{X}h \rangle + \lambda h^T h > 0.$$

- **Ridge solution**: By the above lemma, the matrix $\left(\mathbb{X}^T\mathbb{X} + \lambda I_p\right)$ is invertible because it is positive definite and the ridge solution can be obtained easily:

$$\beta_{ridge} = \left(\mathbb{X}^T\mathbb{X} + \lambda I_p\right)^{-1} \mathbb{X}^T\mathbb{Y}.$$

The ridge solution always exists!

Convexity of LASSO

- **Objective function**:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2 + \lambda \sum_{i=1}^{p} |\beta_i|.$$

  - We already know $\sum_{i=1}^{n} (\beta^T x_i - y_i)^2$ is convex in β.
  - Let us show that $\sum_{i=1}^{p} |\beta_i|$ is convex in β.

- **Proof that $\sum_{i=1}^{p} |\beta_i|$ is convex**: Let $f(\beta) = \sum_{i=1}^{p} |\beta_i|$ and $a = [a_1, \ldots, a_p]$ and $b = [b_1, \ldots, b_p]$ be two vectors. Then,

$$\begin{aligned} f(\alpha a + (1-\alpha) b) &= \sum_{i=1}^{p} |\alpha a_i + (1-\alpha) b_i| \\ &\leq \alpha \sum_{i=1}^{p} |a_i| + (1-\alpha) \sum_{i=1}^{p} |b_i| = \alpha f(a) + (1-\alpha) f(b). \end{aligned}$$

- **Proof that $\lambda \sum_{i=1}^{p} |\beta_i|$ is convex**: We have already shown above that $f(\beta) = \sum_{i=1}^{p} |\beta_i|$ is convex. Let us show that if a function $f(\beta)$ is convex, and $\lambda \geq 0$, then $\lambda f(\beta)$ is also convex. This follows because

$$\begin{aligned} \lambda f(\alpha x + (1-\alpha) y) & \lambda \left(\alpha f(x) + (1-\alpha) f(y) \right) \\ & \leq \alpha \lambda f(x) + (1-\alpha) \lambda f(y). \end{aligned}$$

- **Objective function**:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2 + \lambda \sum_{i=1}^{p} |\beta_i|.$$

  - We already know $\sum_{i=1}^{n} (\beta^T x_i - y_i)^2$ is convex in β.
  - We just showed that $\sum_{i=1}^{p} |\beta_i|$ is convex in β.
  - We also showed that $\lambda \sum_{i=1}^{p} |\beta_i|$ is convex in β.
  - We will now show that the sum of two convex functions is convex.
- **Proof that the sum of two convex functions is convex**: Let $f_1(\beta)$ and $f_2(\beta)$ be two convex functions. Then,

$$\begin{aligned}(f_1 + f_2)(\alpha x + (1-\alpha)y) &= f_1(\alpha x + (1-\alpha)y) + f_2(\alpha x + (1-\alpha)y) \\ &\leq (\alpha f_1(x) + (1-\alpha) f_1(y)) + (\alpha f_2(x) + (1-\alpha) f_2(y)) \\ &\leq \alpha (f_1 + f_2)(x) + (1-\alpha)(f_1 + f_2)(y).\end{aligned}$$

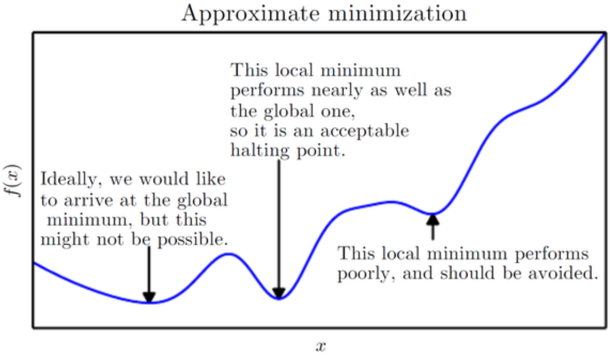
- **This completes the proof of the fact that the LASSO objective function is convex.**
- In your HW, you will prove that the Elastic Net objective is convex as well!

- **Local Minima**: A point β^* is called a local minima of a function $f(\beta)$ if

$$f(\beta^*) \leq f(\beta), \quad \text{for all } \beta \text{ close enough to } \beta^*$$

- **Global Minima**: A point β^* is called a global minima of a function $f(\beta)$ if

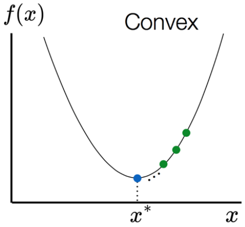
$$f(\beta^*) \leq f(\beta), \quad \text{for all } \beta.$$



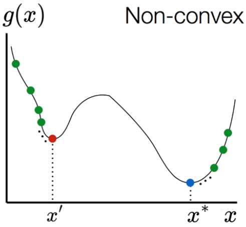
- We are, of course, always looking for the global minimum, if it exists. In most practical problems, we have to contend with a local minima and the effort is always to avoid a bad local minima (one for which the function value is too high to be acceptable).
- **For convex cost functions, as in the case of LASSO, things are better because of the following (ALREADY PROVED) lemma**:

Lemma

Any local minima for a convex function is also a global minima.



Any local minimum is a global minimum



Multiple local minima may exist

- This makes it easier to compute the LASSO solution numerically.

What does the matrix $\mathbb{X}^T\mathbb{X}$ looks like?

- Let us go back to the solution of linear regression and understand how the solution looks like.
- **Recall the $\mathbb{X}^T\mathbb{X}$ from the two variable $(p = 2)$ case (see Week 3 slides 11-15)**: The equation

$$\mathbb{X}^T\mathbb{X}\beta = \mathbb{X}^T\mathbb{Y}$$

for the case $p = 2$ is

$$\begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \end{bmatrix}.$$

Thus, for $p = 2$, we have

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix}.$$

- **Important: Note that the diagonals are NOT $x_i^T x_i$, the inner product of the x vectors with themselves. These are a bit different: the summation runs from $i = 1$ to $i = n$. So, what are these numbers? We need to understand the columns of $\mathbb{X}$.**

- **Columns of $\mathbb{X}$**: Recall that

$$\mathbb{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}.$$

Now, define the p vectors for the columns of matrix $\mathbb{X}$:

$$x^{(1)} = \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix}, \quad \ldots, \quad x^{(p)} = \begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix}.$$

For the diabetes data example, $x^{(1)}$ is the collection of the 'AGE' attribute for all the 442 patients; $x^{(2)}$ is 'SEX' attribute, $x^{(3)}$ is 'BMI', etc. With this new notation, we have

$$\mathbb{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x^{(1)} & x^{(2)} & \ldots & x^{(p)} \end{bmatrix}.$$

We say that $\mathbb{X}$ has n rows and p columns.

- Recall the matrices $\mathbb{X}$ and $\mathbb{X}^T\mathbb{X}$ for $p = 2$ are

$$\mathbb{X} = \begin{bmatrix} x_{11} & x_{12} \\ \vdots & \\ x_{n1} & x_{n2} \end{bmatrix}, \qquad \mathbb{X}^T\mathbb{X} = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix}.$$

  Now, the columns of $\mathbb{X}$ are

$$x^{(1)} = \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix}$$

  .

- Using these new notations, we have

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix} = \begin{bmatrix} \langle x^{(1)}, x^{(1)} \rangle & \langle x^{(1)}, x^{(2)} \rangle \\ \langle x^{(1)}, x^{(2)} \rangle & \langle x^{(2)}, x^{(2)} \rangle \end{bmatrix}.$$

- $\mathbb{X}^T\mathbb{X}$ **for the general case**: For arbitrary p, the matrix $\mathbb{X}^T\mathbb{X}$ has a similar expression (you should try to verify it directly, it is quite easy):

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} \langle x^{(1)}, x^{(1)} \rangle & \langle x^{(1)}, x^{(2)} \rangle & \dots & \langle x^{(1)}, x^{(p)} \rangle \\ \langle x^{(2)}, x^{(1)} \rangle & \langle x^{(2)}, x^{(2)} \rangle & \dots & \langle x^{(2)}, x^{(p)} \rangle \\ \vdots & & & \\ \langle x^{(p)}, x^{(1)} \rangle & \langle x^{(p)}, x^{(2)} \rangle & \dots & \langle x^{(p)}, x^{(p)} \rangle \end{bmatrix}.$$

  Note that it is a $p \times p$ symmetric matrix.

- Now that we know what the matrix $\mathbb{X}^T\mathbb{X}$ looks like, we can investigate its properties.

When is the matrix $\mathbb{X}^T\mathbb{X}$ invertible?

We state this in the form of a theorem. We will see a sketch of the proof soon.

Theorem

The matrix $\mathbb{X}^T\mathbb{X}$ is invertible if and only if the columns of $\mathbb{X}^T\mathbb{X}$ are linearly independent if and only if the columns $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ of $\mathbb{X}$ are linearly independent.

- **Linear Independence**: Let us first understand what linear independence is: in words, *a collection of vectors are called linearly independent if they cannot be written as a linear function of each other*.
- **Examples**:
  - Vectors $a = [1, 1]$ and $b = [2, 2]$ are **not** linearly independent vectors in $\mathbb{R}^2$ because we can write $b = 2a$.
  - Vectors $a = [1, 0, 1]$ and $b = [3, 0, 3]$ are **not** linearly independent vectors in $\mathbb{R}^3$ because we can write $b = 3a$.
  - Vectors $a = [1, 0, 1]$, $b = [3, 0, 3]$, $c = [1, 0, 0]$ are **not** linearly independent vectors in $\mathbb{R}^3$ because a and b are not linearly independent.

Definition

A set of vectors $v_1, v_2, \ldots, v_p$ in $\mathbb{R}^n$ are called **linearly independent** if

$$c_1 v_1 + c_2 v_2 + \cdots + c_p v_p = [0, 0, \ldots, 0] \quad \text{(the vector of } n \text{ zeros).}$$

implies that all the constants are zero: $c_1 = c_2 = \cdots = c_p = 0$.

- **Remark**: If we can find non-zero constants for which the above relation is true then we will be able to write (assuming $c_1 \neq 0$)

  $$c_1 v_1 = -c_2 v_2 - \cdots - c_p v_p,$$

  and v_1 will be a linear function of $v_2, \ldots, v_p$.
- **Examples**:
  - The vectors $e_1 = [1, 0, 0]$, $e_2 = [0, 1, 0]$, $e_3 = [0, 0, 1]$ are linearly independent. Because for constants c_1, c_2, c_3,

    $$c_1 e_1 + c_2 e_2 + c_3 e_3 = [c_1, c_2, c_3] = [0, 0, 0]$$

    if and only if $c_1 = 0$, $c_2 = 0$, and $c_3 = 0$ (violating the non-zero condition in the definition).
  - Check that $v_1 = [1, 1, 0]$, $v_2 = [0, 1, 1]$, $v_3 = [1, 0, 1]$ are also linearly independent!

Lemma

A square matrix is invertible if and only if its columns are linearly independent.

Although the proof of this is not that difficult, we skip this and take this for granted for the rest of the course. Otherwise, this ML course will become a course in linear algebra.

- We saw that the vectors $e_1 = [1, 0, 0]$, $e_2 = [0, 1, 0]$, $e_3 = [0, 0, 1]$ are linearly independent. This means the identity matrix given below is invertible because e_1, e_2, e_3 are the columns of it:

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

- Similarly, because $v_1 = [1, 1, 0]$, $v_2 = [0, 1, 1]$, $v_3 = [1, 0, 1]$ are linearly independent, the following matrix is invertible:

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

- **Can you think of a matrix that is not invertible?**

Linear Combinations of Columns

- Recall that

$$\mathbb{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x^{(1)} & x^{(2)} & \dots & x^{(p)} \end{bmatrix}.$$

  where $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ are the columns of matrix $\mathbb{X}$:

$$x^{(1)} = \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix}, \quad \dots, \quad x^{(p)} = \begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix}.$$

- Then, the matrix multiplication $\mathbb{X}\beta$ is a linear combination of the columns of $\mathbb{X}$:

$$\mathbb{X}\beta = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} \beta_1 + \begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix} \beta_2 + \dots \begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix} \beta_p.$$

Proof of the Theorem

Theorem

The matrix $\mathbb{X}^T\mathbb{X}$ is invertible if and only if the columns $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ of $\mathbb{X}$ are linearly independent.

- **Proof**: We will prove that the columns of $\mathbb{X}$ are linearly dependent if and only if the columns of $\mathbb{X}^T\mathbb{X}$ are linearly dependent.

  If the columns of $\mathbb{X}$ are linearly dependent, then there exists a non-zero vector $c = (c_1, \cdots, c_p)$ such that

  $$\mathbb{X}c = 0.$$

  Multiply by $\mathbb{X}^T$ to get

  $$\mathbb{X}^T\mathbb{X}c = 0.$$

  This proves that the columns of $\mathbb{X}^T\mathbb{X}$ are linearly dependent and that it is not invertible (based on the lemma).

- **Proof Continued**: Now if the columns of $\mathbb{X}^T\mathbb{X}$ are linearly dependent, then there exists a non-zero vector $c = (c_1, \cdots, c_p)$ such that
$$\mathbb{X}^T\mathbb{X}c = 0.$$
Multiple on the left by c^T to get
$$c^T\mathbb{X}^T\mathbb{X}c = 0.$$
But,
$$c^T\mathbb{X}^T\mathbb{X}c = \langle \mathbb{X}c, \mathbb{X}c \rangle.$$
This means that
$$\mathbb{X}c = 0$$
because the inner product of a vector with itself is the sum of square of the components, and if the sum of square of the components is zero, then each of the component must be zero. This means that the columns of $\mathbb{X}$ are linearly dependent. Note that $\mathbb{X}$ is not (or need not be) a square matrix, so the lemma on invertibility does not apply to $\mathbb{X}$.