Lasso metoda i primjene u visokodimenzionalnoj statistici

Ljubičić, Bruno

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Bruno Ljubičić

LASSO METODA I PRIMJENE U VISOKODIMENZIONALNOJ STATISTICI

Diplomski rad

Voditelj rada:
doc. dr. sc. Hrvoje Planinić

Zagreb, rujan, 2023.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

U suvremenom okruženju obilja podataka, visokodimenzionalna statistika postala je ključna disciplina za razumijevanje složenih veza među varijablama. S obzirom na sve veću prisutnost velikog broja varijabli u analizama podataka, javlja se potreba za tehnikama koje ne samo pravilno modeliraju složene veze, već i omogućavaju selekciju nekolicine najvažnijih. U tom kontekstu, Lasso metoda (skraćeno od "Least Absolute Shrinkage and Selection Operator") izdvaja se kao moćan alat za obradu visokodimenzionalnih podataka.

Uobičajeni procjenitelj najmanjih kvadrata (OLS) često se koristi za prilagodbu linearnog modela podacima. Međutim, ako je broj varijabli strogo veći od broja podataka, OLS nije jedinstven te se rješenja mogu drastično razlikovati, čime je onemogućena interpretabilnost koeficijenata. Ovo dovodi do potrebe za tehnikama selekcije varijabli.

Glavna razlika između OLS-a i Lasso metode je što Lasso dodatno penalizira apsolutnu vrijednost koeficijenata. Ovdje leži i glavnina teškoća analiziranja Lasso metode - apsolutna vrijednost nije diferencijabilna u nuli. Ipak, ispostavlja se da je u analizi dovoljno koristiti alate konveksne analize.

U prvom poglavlju razmatraju se temeljna svojstva procjenitelja najmanjih kvadrata i poopćenje inverza matrice. Drugo poglavlje posvećeno je ključnim rezultatima u polju konveksne analize i odgovara na pitanje egzistencije lasso rješenja. U trećem poglavlju uvodi se pojam subgradijenta konveksne funkcije koji je ključan u traženju minimuma. Konačno, četvrto poglavlje detaljno istražuje glavne karakteristike lasso rješenja, razmatra pitanja vezana uz jedinstvenost rješenja te daje konstruktivan algoritam za lasso putanju.

Poglavlje 1

Procjenitelj najmanjih kvadrata

1.1 Linearni model

Neka je $\left\{(\mathbf{x}_i^T, y_i) : i = 1 \ldots, n\right\}$ uzorak iz linearnog modela

$$y_i = \mathbf{x}_i^T \beta_0 + \varepsilon_i,$$

pri čemu su y_i slučajne varijable, $\mathbf{x}_i$ p-dimenzionalni neslučajni vektori, $\beta_0 \in \mathbb{R}^p$ neslučajni vektor koeficijenata i ε_i međusobno nezavisne i jednakodistribuirane slučajne varijable s očekivanjem 0. Model pišemo kompaktnije kao

$$\mathbf{y} = X\beta_0 + \varepsilon,$$

pri čemu su $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ i

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}.$$

Nadalje, označimo stupce od X sa $X_1, \ldots, X_p \in \mathbb{R}^n$.

Cilj je naći procjenitelj za β_0 na osnovu opaženog uzorka.

Želimo β_0 procijeniti metodom najmanjih kvadrata, tj. točkom minimuma funkcije $g : \mathbb{R}^p \to \mathbb{R}$

$$g(\beta) = \frac{1}{2}\|X\beta - \mathbf{y}\|_2^2.$$

Točku minimuma funkcije g zovemo procjeniteljem najmanjih kvadrata od β_0. Funkcija g je klase $C^\infty(\mathbb{R}^p)$ te je konveksna, pa je svaka stacionarna točka ujedno i točka minimuma. Promotrimo gradijent od g

$$\nabla g(\beta) = X^T(X\beta - \mathbf{y}).$$

Dakle, svaki je procjenitelj najmanjih kvadrata rješenje sustava

$$X^T X\beta = X^T \mathbf{y}.$$

Ako je matrica X punog stupčanog ranga p, tada je $X^T X$ invertibilna te je jedinstveni procjenitelj najmanjih kvadrata dan sa

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}.$$

Situacija kada matrica X nije punog stupčanog ranga je kompliciranija.

1.2 Moore-Penroseov inverz

Definicija 1.2.1. *Neka je $A \in \mathbb{R}^{n\times p}$ matrica. Matricu $A^+ \in \mathbb{R}^{p\times n}$ koja zadovoljava uvjete*

(1) $AA^+A = A$

(2) $A^+AA^+ = A^+$

(3) $(A^+A)^T = A^+A$

(4) $(AA^+)^T = AA^+$

nazivamo Moore-Penroseov inverz matrice A.

Propozicija 1.2.2. *Moore-Penroseov inverz je jedinstven.*

Dokaz. Dokaz je preuzet iz [1]. Neka je $A \in \mathbb{R}^{n\times p}$ proizvoljna matrica i neka su B, C dva Moore-Penrose inverza od A. Promotrimo

$$M = AB - AC = A(B - C) \in \mathbb{R}^{n\times n}.$$

Iz svojstva (4) slijedi da je M simetrična matrica te vrijedi

$$M^2 = (AB - AC)A(B - C) = (ABA - ACA)(B - C) = (A - A)(B - C) = 0.$$

Za proizvoljni $\mathbf{x} \in \mathbb{R}^n$ vrijedi

$$\|M\mathbf{x}\|_2^2 = \mathbf{x}^T M^2 \mathbf{x} = 0 \implies M\mathbf{x} = 0$$

pa je $M = 0$, odnosno $AB = AC$. Koristeći sličan postupak za $M = BA - CA$, dobivamo $BA = CA$. Sada imamo

$$B = BAB = BAC = CAC = C.$$

□

Egzistencija Moore-Penroseovog inverza slijedit će iz sljedećeg teorema:

Teorem 1.2.3. *(SVD dekompozicija) Neka je $A \in \mathbb{R}^{n\times p}$ matrica. Tada postoje ortogonalna matrica $U \in \mathbb{R}^{n\times n}$, ortogonalna matrica $V \in \mathbb{R}^{p\times p}$ i dijagonalna matrica $S \in \mathbb{R}^{n\times p}$ sa nenegativnim elementima na dijagonali, takvi da je*

$$A = USV^T.$$

Za invertibilne matrice, lako se vidi da inverz zadovoljava sva svojstva definicije Moore-Penroseovog inverza. Nadalje, ako je $S \in \mathbb{R}^{n\times p}$ simetrična matrica, tada se lako vidi da matricu S^+ možemo dobiti tako da u matrici S^T svaki element različiti od 0 zamjenimo njegovim inverzom. Također, provjerom uvjeta definicije, lako se vidi da ako za matrice $A \in \mathbb{R}^{n\times p}$ i $B \in \mathbb{R}^{p\times m}$ postoje A^+ i B^+, tada postoji i $(AB)^+$ i jednak je B^+A^+. Korištenjem gornjih razmatranja, lako se vidi da vrijedi sljedeći teorem:

Teorem 1.2.4. *Neka je $A \in \mathbb{R}^{n\times p}$ matrica i neka je*

$$A = USV^T$$

njena SVD dekompozicija. Tada je

$$A^+ = VS^+U^T.$$

Napomena 1.2.5. *Provjerom uvjeta iz definicije 1.2.1, mogu se pokazati i sljedeći identiteti:*

(5) $(A^+)^T = (A^T)^+$

(6) $(A^+)^+ = A$

(7) ako je $\alpha \neq 0$, *tada je* $(\alpha A)^+ = \frac{1}{\alpha}A^+$

(8) $A = AA^T(A^T)^+ = (A^T)^+A^TA$

(9) $A^+ = A^+(A^+)^TA^T = A^T(A^+)^TA^+$

(10) $A^T = A^TAA^+ = A^+AA^T$.

Dokaz se može naći u [1].

Vratimo se na sustav

$$X^T X\beta = X^T \mathbf{y}.$$

Koristeći identitet (10) iz napomene 1.2.5, vidimo da je $\beta = X^+\mathbf{y}$ jedno partikularno rješenje. Kako je $\mathrm{Ker}(X^T X) = \mathrm{Ker}(X)$, vrijedi da je skup rješenja sustava dan sa

$$X^+\mathbf{y} + \mathrm{Ker}(X).$$

Dakle, procjenitelj najmanjih kvadrata uvijek postoji, ali je jedinstven ako i samo ako je X punog stupčanog ranga. Ako X nije punog stupčanog ranga (primjerice kada je $p > n$), tada za proizvoljan $\eta \in \mathrm{Ker}(X)$, $\eta \neq \mathbf{0}$, vrijedi da je i

$$X^+\mathbf{y} + t\eta$$

procjenitelj najmanjih kvadrata za sve $t \in \mathbb{R}$, što pokazuje da sigurno postoje dva procjenitelja najmanjih kvadrata čije se komponente razlikuju u predznacima. Ovo svojstvo je često nepoželjno jer otežava interpretaciju odnosa između zavisnih i nezavisnih varijabli.

1.3 Lasso funkcija cilja

Neka je $\lambda > 0$. Promotrimo funkcicju $g_1 : \mathbb{R}^p \to \mathbb{R}$ danu sa

$$g_1(\beta) = \frac{1}{2}\|X\beta - \mathbf{y}\|_2^2 + \lambda\|\beta\|_1.$$

Točke minimuma ove funkcije zovemo lasso procjeniteljima od β_0. Ključna razlika između funkcije g_1 i funkcije

$$g(\beta) = \frac{1}{2}\|X\beta - \mathbf{y}\|_2^2$$

je diferencijabilnost. Naime, funkcija

$$\mathbf{z} \mapsto \|\mathbf{z}\|_1 = \sum_{i=1}^{p} |z_i|$$

nije diferencijabilna niti za jedan $\mathbf{z} \in \mathbb{R}^p$ koji ima neku komponentu jednaku 0, pa ne možemo koristiti metodu traženja stacionarnih točaka za pronalaženje točki minimuma. Ipak, funkcije g_1 i g dijele bitno svojstvo konveksnosti. Cilj je sljedeća dva poglavlja proučiti svojstva konveksnih skupova i fukncija te naposljetku pokazati da g_1 postiže minimum na $\mathbb{R}^p$.

Poglavlje 2

Konveksna analiza

Sada dajemo kratki pregled rezultata konveksne analize potrebnih za analiziranje lasso metode. Poglavlje prati prvi odjeljak iz [2] te četvrti i deveti odjeljak iz [4].

2.1 Afini potprostori

Definicija 2.1.1. *Kažemo da je $C \subseteq \mathbb{R}^n$ afin skup ako za $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ je*

$$(1-t)\boldsymbol{x}_1 + t\boldsymbol{x}_2 \in C, \quad t \in \mathbb{R}.$$

Definicija 2.1.2. *Ako su $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in C \subseteq \mathbb{R}^n$ vektori, tada svaki vektor oblika*

$$\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i, \quad \sum_{i=1}^{k} \lambda_i = 1, \quad \lambda_i \in \mathbb{R}, i = 1, \ldots, k$$

zovemo afinom kombinacijom vektora iz C. Skup $\mathrm{aff}(C)$ *svih afinih kombinacija vektora iz C zovemo afinom ljuskom od C.*

Teorem 2.1.3. *Svaki je afin skup $C \subseteq \mathbb{R}^n$ zatvoren na afine kombinacije.*

Dokaz. Tvrdnju dokazujemo indukcijom. Baza indukcije, za $k = 2$ vrijedi po definiciji. Neka tvrdnja vrijedi za sve afine kombinacije najviše $k-1$ vektora iz C. Neka su $\mathbf{x}_1, \ldots, \mathbf{x}_k \in C$ te

$$\mathbf{x} = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i, \quad \sum_{i=1}^{k} \lambda_i = 1, \quad \lambda_i \in \mathbb{R}, \;\; i = 1, \ldots, k.$$

Kako je $k \geq 2$, barem je jedan $\lambda_i \neq 1$, pa bez smanjenja općenitosti pretpostavimo da $\lambda_k \neq 1$. Tada je

$$\sum_{i=1}^{k-1} \frac{\lambda_i}{1-\lambda_k} = 1$$

te vrijedi

$$\mathbf{x} = (1-\lambda_k) \sum_{i=1}^{k-1} \frac{\lambda_i}{1-\lambda_k} \mathbf{x}_i + \lambda_k \mathbf{x}_k$$

pa je po pretpostavci indukcije $\mathbf{x} \in C$. □

Teorem 2.1.4. *Ako je $C \subseteq \mathbb{R}^n$ afin skup, te $\boldsymbol{c}_0 \in C$, tada je $C - \boldsymbol{c}_0$ vektorski potprostor od $\mathbb{R}^n$. Nadalje, za $\boldsymbol{c}_1, \boldsymbol{c}_2 \in C$ je*

$$C - \boldsymbol{c}_1 = C - \boldsymbol{c}_2.$$

Dokaz. Neka su $\mathbf{v}_1, \mathbf{v}_2 \in C - \mathbf{c}_0$ i $\alpha \in \mathbb{R}$. Tada postoje $\mathbf{c}_1, \mathbf{c}_2 \in C$ takvi da je

$$\mathbf{v}_i = \mathbf{c}_i - \mathbf{c}_0, \quad i = 1, 2.$$

tada je

$$\alpha \mathbf{v}_1 = \alpha(\mathbf{c}_1 - \mathbf{c}_0) = \underbrace{\alpha \mathbf{c}_1 + (1-\alpha)\mathbf{c}_0}_{\in C} - \mathbf{c}_0 \in C - \mathbf{c}_0$$

te

$$\mathbf{v}_1 + \mathbf{v}_2 = \underbrace{\mathbf{c}_1 - \mathbf{c}_0 + \mathbf{c}_2}_{\in C} - \mathbf{c}_0 \in C - \mathbf{c}_0$$

pa je $C - \mathbf{c}_0$ vektorski prostor.

Neka je $\mathbf{v} = \mathbf{c} - \mathbf{c}_1 \in C - \mathbf{c}_1$. Tada je

$$v = (\mathbf{c} - \mathbf{c}_1 + \mathbf{c}_2) - \mathbf{c}_2 \in C - \mathbf{c}_2$$

pa jedna inkluzija vrijedi, druga se pokazuje analogno. □

Definicija 2.1.5. *Neprazan afin skup zovemo afinim potprostorom. Afin potprostor zovemo pravim ako nije jednak $\mathbb{R}^n$. Afinu dimenziju afinog potprostora $C \subseteq \mathbb{R}^n$, u oznaci* $\dim(C)$*, definiramo kao dimenziju pripadnog vektorskog prostora.*

Iz činjenice da je afina kombinacija afinih kombinacija opet afina kombinacija, slijedi da je za neprazan $S \subseteq \mathbb{R}^n$, $\mathrm{aff}(S)$ najmanji afin potprostor koji sadrži S.

Teorem 2.1.6. *Neka je* $S = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subseteq \mathbb{R}^n$. *Tada je vektorski prostor pridružen* $C = \mathrm{aff}(S)$ *dan sa*

$$L = \mathrm{span}(\{\boldsymbol{x}_1 - \boldsymbol{x}_0, \ldots, \boldsymbol{x}_k - \boldsymbol{x}_0\}).$$

Dokaz. Očito je $S - \mathbf{x}_0 \subseteq C - \mathbf{x}_0$ pa je i $\mathrm{span}(S - \mathbf{x}_0) = L \subseteq C - \mathbf{x}_0$. Neka je sada $v \in C - \mathbf{x}_0$, tada vrijedi

$$\mathbf{v} = \sum_{i=0}^{k} \lambda_i \mathbf{x}_i - \mathbf{x}_0, \quad \sum_{i=0}^{k} \lambda_i = 1$$

za neke skalare $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$. Tada vrijedi

$$\mathbf{v} = \left(1 - \sum_{i=1}^{k} \lambda_i\right)\mathbf{x}_0 + \sum_{i=1}^{k} \lambda_i \mathbf{x}_i - \mathbf{x}_0 = \sum_{i=1}^{k} \lambda_i (\mathbf{x}_i - \mathbf{x}_0) \in L.$$

□

Definicija 2.1.7. *Kažemo da su* $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^n$ *afino nezavisni vektori ako su* $\boldsymbol{v}_1 - \boldsymbol{v}_0, \ldots, \boldsymbol{v}_k - \boldsymbol{v}_0$ *linearno nezavisni, a u suprotnom kažemo da su afino zavisni.*

Napomena 2.1.8. *Koristeći analogone tvrdnje za vektorske prostore, mogu se pokazati sljedeće tvrdnje*

(i) Vektori $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^n$ *su afino nezavisni ako i samo ako je* $\mathrm{aff}(\{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\})$ *dimenzije k.*

(ii) Ako je $A \subseteq \mathbb{R}^n$ *afin potprostor dimenzije k, tada je svaki k + 2-člani skup vektora afino zavisan.*

(iii) Ako je $A \subseteq \mathbb{R}^n$ *afin potprostor dimenzije k, koji sadrži afino nezavisne vektore* $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. *Tada je* $A = \mathrm{aff}(\{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\})$.

(iv) Svaki m + 1 član, afino nezavisan skup vektora iz $\mathbb{R}^n$ *(pri čemu je* $m < n$*) se može nadopuniti do n + 1-članog, afino nezavisnog skupa vektora.*

(v) Ako su $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^n$ *afino nezavisni vektori, tada se svaki* $\boldsymbol{v} \in \mathrm{aff}(\{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\})$ *može na jedinstven način zapisati kao afina kombinacija vektora* $\boldsymbol{v}_i, i = 0, \ldots, k$.

(vi) Presjek proizvoljno mnogo afinih skupova je afin.

Definicija 2.1.9. *Neka su* $\boldsymbol{a} \in \mathbb{R}^n$ $\boldsymbol{a} \neq 0$ *i* $\alpha \in \mathbb{R}$. *Hiperavnina* $H \subseteq \mathbb{R}^n$ *određena sa* $\boldsymbol{a}$ *i* α *je skup*

$$H = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}^T \boldsymbol{x} = \alpha\right\}.$$

Lako se vidi da je svaka hiperravnina zatvorena na afine kombinacije, stoga je i afin potprostor. Pridruženi vektorski prostor je

$$\left\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = 0\right\} = \{\mathbf{a}\}^\perp$$

koji je očito dimenzije $n-1$. Dakle, svaka hiperravnina je dimenzije $n-1$.

Teorem 2.1.10. *Svaki se pravi afini potprostor $C \subseteq \mathbb{R}^n$ može zapisati kao konačni presjek hiperravnina.*

Dokaz. Neka je $\dim(C) = m < n$ i $\mathbf{a} \in C$, te neka je $V = C - \mathbf{a}$ njegov pripadni vektorski prostor. Neka je $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ ortonormirana baza od V. Proširimo tu bazu do ortonormirane baze $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ od $\mathbb{R}^n$. Tada se svaki $\mathbf{x} = \mathbf{v} + \mathbf{a} \in C$ može zapisati kao

$$\mathbf{x} = \mathbf{v} + \mathbf{a} = \sum_{i=1}^{m}\Big((\mathbf{v}^T\mathbf{v}_i)\mathbf{v}_i + (\mathbf{a}^T\mathbf{v}_i)\mathbf{v}_i\Big) + \mathbf{a} - \sum_{i=1}^{m}(\mathbf{a}^T\mathbf{v}_i)\mathbf{v}_i = \sum_{i=1}^{m}(\mathbf{x}^T\mathbf{v}_i)\mathbf{v}_i + \sum_{i=m+1}^{n}(\mathbf{a}^T\mathbf{v}_i)\mathbf{v}_i.$$

Promatramo hiperravnine

$$H_i = \left\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}_i^T\mathbf{x} = \mathbf{v}_i^T\mathbf{a}\right\}, \quad i = m+1, \ldots, n$$

i njihov presjek

$$H = \bigcap_{i=m+1}^{n} H_i = \left\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{v}_i^T\mathbf{x} = \mathbf{v}_i^T\mathbf{a},\ i = m+1, \ldots, n\right\}.$$

Tvrdimo da je $C = H$. Neka je $\mathbf{x} \in C$, iz gornjeg izraza za $\mathbf{x}$ slijedi da je $\mathbf{x}^T\mathbf{v}_i = \mathbf{a}^T\mathbf{v}_i$ za $i = m+1, \ldots, n$ pa vrijedi $C \subseteq H$.

Neka je $\mathbf{v} = \mathbf{x} - \mathbf{a} \in H - \mathbf{a}$ pri čemu je $\mathbf{x} \in H$. Zapišimo $\mathbf{v}$ u bazi

$$\mathbf{v} = \sum_{i=1}^{n}(\mathbf{v}^T\mathbf{v}_i)\mathbf{v}_i = \sum_{i=1}^{n}(\mathbf{x}^T\mathbf{v}_i)\mathbf{v}_i - \sum_{i=1}^{n}(\mathbf{a}^T\mathbf{v}_i)\mathbf{v}_i = \sum_{i=1}^{m}(\mathbf{x}^T\mathbf{v}_i)\mathbf{v}_i - \sum_{i=1}^{m}(\mathbf{a}^T\mathbf{v}_i)\mathbf{v}_i = \sum_{i=1}^{m}(\mathbf{v}^T\mathbf{v}_i)\mathbf{v}_i \in V$$

pa je $H \subseteq V + \mathbf{a} = C$. □

Propozicija 2.1.11. *Svaka hiperravnina je zatvoren skup, praznog interiora.*

Dokaz. Neka je

$$H = \left\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = \alpha\right\}, \quad \mathbf{a} \neq 0, \alpha \in \mathbb{R}.$$

Ako označimo $f : \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ linearni funkcional, tada je $H = f^{-1}(\{\alpha\})$, pa kako je f neprekidna, slijedi da je H zatvoren skup. Nadalje, pretpostavimo da je $\mathbf{x} \in \text{Int}(H)$, tada postoji $r > 0$, takav da je $K(\mathbf{x}, r) \subseteq H$. Tada je

$$\mathbf{y} = \mathbf{x} + \frac{r}{2\|\mathbf{a}\|_2}\mathbf{a} \in K(\mathbf{x}, r) \subseteq H.$$

Međutim,

$$\alpha = \mathbf{y}^T\mathbf{a} = \mathbf{x}^T\mathbf{a} + \frac{r}{2}\|\mathbf{a}\|_2 = \alpha + \frac{r}{2}\|\mathbf{a}\|_2 > \alpha,$$

što je kontradikcija. □

Kombinirajući teorem 2.1.10 i propoziciju 2.1.11 dobivamo

Korolar 2.1.12.

(i) Svaki afin potprostor u $\mathbb{R}^n$ je zatvoren.

(ii) Svaki pravi afin potprostor u $\mathbb{R}^n$ ima prazan interior.

2.2 Konveksni skupovi

Definicija 2.2.1. *Kažemo da je $C \subseteq \mathbb{R}^n$ konveksan, ako za svake dvije točke $\boldsymbol{x}, \boldsymbol{y} \in C$ vrijedi da je $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in C, \quad \forall t \in [0,1]$. Ako je $S \subseteq \mathbb{R}^n$ skup, tada sa* $\text{conv}(S)$ *označavamo konveksnu ljusku - najmanji konveksni skup koji sadrži S.*

Može se pokazati da je $\text{conv}(S)$ ujedno i skup svih konveksnih kombinacija vektora iz S, tj.

$$\text{conv}(S) = \left\{ \sum_{i=1}^{p} \lambda_i \mathbf{x}_i \ : \ p \in \mathbb{N}, \ \sum_{i=1}^{p} \lambda_i = 1, \quad \mathbf{x}_i \in S, \lambda_i \geq 0, \ i = 1, \ldots, p \right\}.$$

Definicija 2.2.2. *Neka je $\boldsymbol{x} \in \mathbb{R}^n$ i $C \subseteq \mathbb{R}^n$. Kažemo da je $\boldsymbol{y}$ projekcija od $\boldsymbol{x}$ na C ako vrijedi*

$$\|\boldsymbol{x} - \boldsymbol{y}\|_2 = \inf_{z \in C} \|\boldsymbol{x} - \boldsymbol{z}\|_2.$$

U kontekstu konveksnih i zatvorenih skupova, projekcija postoji i jedinstvena je.

Teorem 2.2.3. *Neka je $C \subseteq \mathbb{R}^n$ neprazan, konveksan i zatvoren skup te $\boldsymbol{x} \in \mathbb{R}^n$. Tada postoji jedinstvena projekcija od $\boldsymbol{x}$ na C.*

Dokaz. Neka je

$$d(\mathbf{x}, C) = \inf_{\mathbf{y}\in C} \|\mathbf{x} - \mathbf{y}\|_2 =: d.$$

Za $k \in \mathbb{N}$, neka je $\mathbf{z}_k \in C$ takav da

$$\|\mathbf{x} - \mathbf{z}_k\|_2 \leq d + \frac{1}{k}.$$

Za $k, m \in \mathbb{N}$, iz relacije paralelograma slijedi

$$\begin{aligned}
\|\mathbf{z}_k - \mathbf{z}_m\|_2^2 &= 2\|\mathbf{z}_k - \mathbf{x}\|_2^2 + 2\|\mathbf{z}_m - \mathbf{x}\|_2^2 - \|\mathbf{z}_k + \mathbf{z}_m - 2\mathbf{x}\|_2^2 \\
&= 2\|\mathbf{z}_k - \mathbf{x}\|_2^2 + 2\|\mathbf{z}_m - \mathbf{x}\|_2^2 - 4\left\|\frac{\mathbf{z}_k + \mathbf{z}_m}{2} - \mathbf{x}\right\|_2^2 \\
&\leq 2\|\mathbf{z}_k - \mathbf{x}\|_2^2 + 2\|\mathbf{z}_m - \mathbf{x}\|_2^2 - 4d^2
\end{aligned}$$

Budući da desna strana teži prema 0, $(\mathbf{z}_k)_{k\in\mathbb{N}}$ je Cauchyjev niz, a budući da je $\mathbb{R}^n$ potpun, on je i konvergentan. Označimo mu limes sa $\hat{\mathbf{x}}$. Zbog zatvorenosti je $\hat{\mathbf{x}} \in C$, a zbog neprekidnosti norme je

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 = d.$$

Dakle $\hat{\mathbf{x}}$ je projekcija od $\mathbf{x}$ na C.

Neka su $\mathbf{y}$ i $\mathbf{z}$ dvije projekcije od $\mathbf{x}$ na C, tada koristeći jednakost paralelograma kao i prije, slijedi

$$\|\mathbf{y} - \mathbf{z}\|^2 \leq 2\|\mathbf{y} - \mathbf{x}\|_2^2 + 2\|\mathbf{z} - \mathbf{x}\|_2^2 - 4d^2 = 0$$

pa vrijedi $\mathbf{y} = \mathbf{z}$. □

Teorem 2.2.4. *Neka je $C \subseteq \mathbb{R}^n$ neprazan, konveksan i zatvoren skup te $\boldsymbol{x} \in \mathbb{R}^n$, tada je $\hat{\boldsymbol{x}}$ projekcija od $\boldsymbol{x}$ na C ako i samo ako vrijedi*

$$(\forall \boldsymbol{z} \in C) \quad (\boldsymbol{z} - \hat{\boldsymbol{x}})^T(\boldsymbol{x} - \hat{\boldsymbol{x}}) \leq 0.$$

Dokaz. Neka je $\mathbf{z} \in C$, tada za $t \in [0, 1]$ definiramo $\mathbf{z}_t = (1 - t)\hat{\mathbf{x}} + t\mathbf{z} \in C$. Tada vrijedi

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \leq \|\mathbf{x} - \mathbf{z}_t\|_2^2 = \|(\mathbf{x} - \hat{\mathbf{x}}) - t(\mathbf{z} - \hat{\mathbf{x}})\|_2^2 = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + t^2\|\mathbf{z} - \hat{\mathbf{x}}\|_2^2 - 2(\mathbf{x} - \hat{\mathbf{x}})^T(\mathbf{z} - \hat{\mathbf{x}})$$

što povlači

$$(\mathbf{x} - \hat{\mathbf{x}})^T(\mathbf{z} - \hat{\mathbf{x}}) \leq \frac{t^2}{2}\|\mathbf{z} - \hat{\mathbf{x}}\|_2^2.$$

Puštajući $t \to 0$, slijedi tvrdnja.

Obratno, ako je $\mathbf{x} \in C$, uvrštavajući $\mathbf{z} = \mathbf{x}$ u nejednakost iz iskaza, slijedi

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \leq 0$$

pa je $\mathbf{x} = \hat{\mathbf{x}}$ projekcija od $\mathbf{x}$ na C. Inače $\mathbf{x} \notin C$ pa je

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 > 0.$$

Neka je $\mathbf{z} \in C$, tada vrijedi

$$\begin{aligned} 0 &\geq (\mathbf{z} - \hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}}) \\ &= [(\mathbf{z} - \mathbf{x}) + (\mathbf{x} - \hat{\mathbf{x}})]^T(\mathbf{x} - \hat{\mathbf{x}}) \\ &= (\mathbf{z} - \mathbf{x})^T(\mathbf{x} - \hat{\mathbf{x}}) + \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \\ &\geq -\|\mathbf{z} - \mathbf{x}\|_2\|\mathbf{x} - \hat{\mathbf{x}}\|_2 + \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2, \end{aligned}$$

što povlači

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \|\mathbf{x} - \mathbf{z}\|_2.$$

Dakle $\hat{\mathbf{x}}$ je projekcija od $\mathbf{x}$ na C. □

Definicija 2.2.5. *Za $\boldsymbol{a} \neq 0$ i $\alpha \in \mathbb{R}$, hiperravnina*

$$H = \left\{ z \in \mathbb{R}^n \ : \ \boldsymbol{a}^T z = \alpha \right\}$$

razdvaja $\mathbb{R}^n$ na poluprostore

$$H^+ = \left\{ z \in \mathbb{R}^n \ : \ \boldsymbol{a}^T z \geq \alpha \right\}, \quad H^- = \left\{ z \in \mathbb{R}^n \ : \ \boldsymbol{a}^T z \leq \alpha \right\}.$$

Kažemo da je hiperravnina H potporna za $C \subseteq \mathbb{R}^n$ ako $H \cap C \neq \emptyset$ i C leži sa jedne strane hiperravnine (tj. $C \subseteq H^+$ ili $C \subseteq H^-$).

Primjerice, ako je $C \subseteq \mathbb{R}^n$ neprazan, konveksan i zatvoren i $\mathbf{y} \notin C$ te $\hat{\mathbf{y}}$ njena projekcija na C. Ako stavimo $\mathbf{a} = (\mathbf{y} - \hat{\mathbf{y}})$, tada je

$$H_{\mathbf{y}} = \left\{ \mathbf{x} \in R^n \ : \ \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\hat{\mathbf{y}} \right\}$$

potporna hiperravina od C jer je po prethodnom teoremu

$$C \subseteq H_{\mathbf{y}}^- = \left\{ \mathbf{x} \in R^n \ : \ \mathbf{a}^T\mathbf{x} \leq \mathbf{a}^T\hat{\mathbf{y}} \right\}$$

Teorem 2.2.6. *Neka je $C \subseteq \mathbb{R}^n$ neprazan, konveksan i zatvoren skup te $\boldsymbol{y} \in \partial C$. Tada postoji potporna hiperravnina od C koja sadrži $\boldsymbol{y}$.*

Dokaz. Budući da je $\mathbf{y} \in \partial C$, postoji niz $(\mathbf{y}_k)_{k\in\mathbb{N}}$ u $\mathbb{R}^n \setminus C$ koji konvergira prema $\mathbf{y}$. Neka je $\hat{\mathbf{y}}_k$ projekcija od $\mathbf{y}_k$ na C te

$$\mathbf{a}_k = \frac{\mathbf{y}_k - \hat{\mathbf{y}}_k}{\left\|\mathbf{y}_k - \hat{\mathbf{y}}_k\right\|_2}.$$

Iz definicije od $\mathbf{a}_k$ i prethodnog teorema slijedi

$$\mathbf{a}_k^T \mathbf{y}_k > \mathbf{a}_k^T \hat{\mathbf{y}}_k \geq \mathbf{a}_k^T \mathbf{z}, \quad \mathbf{z} \in C.$$

Budući da je $(\mathbf{a}_k)_{k\in\mathbb{N}}$ niz na jediničnoj sferi, koja je kompaktan skup, postoji njegov konvergentan podniz $(\mathbf{a}_{p_k})_{k_\in\mathbb{N}}$ s limesom $\mathbf{a}$. Budući da je skalarni produkt neprekidna funkcija, iz gornje nejednakosti slijedi

$$\mathbf{a}^T \mathbf{y} \geq \mathbf{a}^T \mathbf{z}, \quad \mathbf{z} \in C.$$

Dakle

$$H := \left\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T \mathbf{y} = \mathbf{a}^T \mathbf{x}\right\}$$

je potporna hiperravnina od C koja sadrži $\mathbf{y}$. □

Lema 2.2.7. *Ako je $C \subseteq \mathbb{R}^n$ konveksan skup takav da je* $\mathrm{Int}(C) = \emptyset$*, tada je C sadržan u nekom pravom afinom potprostoru od $\mathbb{R}^n$.*

Dokaz. Promotrimo $\mathrm{aff}(C)$. Ako je $\dim(\mathrm{aff}(C)) < n$ onda smo gotovi, pa pretpostavimo da je $\dim(\mathrm{aff}(C)) = n$. Neka je

$$K := \{\mathbf{x}_0, \ldots, \mathbf{x}_n\} \subseteq C$$

neka afina baza od $\mathrm{aff}(C)$. Tada je zbog konveksnosti $\mathrm{conv}(K) \subseteq C$. No može se provjeriti da je interior od $\mathrm{conv}(K)$ skup

$$\left\{\sum_{i=0}^{n} \lambda_i \mathbf{x}_i \ : \ \sum_{i=0}^{n} \lambda_i = 1, \quad \lambda_i > 0\right\},$$

što je kontradikcija jer je interior od C prazan. □

Lema 2.2.8. *Neka je $C \subseteq \mathbb{R}^n$ konveksan skup te $\boldsymbol{x} \in \mathrm{Int}(C)$ i $\boldsymbol{y} \in \mathrm{Cl}(C)$. Tada je $[\boldsymbol{x}, \boldsymbol{y}\rangle \subseteq \mathrm{Int}(C)$.*

Dokaz. Neka je $t \in [0, 1\rangle$, $\mathbf{z} = (1 - t)\mathbf{x} + t\mathbf{y}$ i $(\mathbf{y}_k)_{k\in\mathbb{N}}$ niz u C takav da je

$$\left\|\mathbf{y} - \mathbf{y}_k\right\|_2 < \frac{1}{k}.$$

Ako označimo $\mathbf{z}_k = (1 - t)\mathbf{x} + t\mathbf{y}_k$, tada je

$$\|\mathbf{z} - \mathbf{z}_k\| < \frac{t}{k} < \frac{1}{k}.$$

Kako je $\mathbf{x} \in \text{Int}(C)$, postoji $r > 0$ takva da je kugla $K(\mathbf{x}, r) \subseteq C$. Tvrdimo da je $K(\mathbf{z}_k, (1-t)r) \subseteq C$. Neka je $\mathbf{w} \in K(\mathbf{z}_k, (1-t)r)$, tada je za $\tilde{\mathbf{w}} = \frac{\mathbf{w}-t\mathbf{y}_k}{(1-t)}$

$$\|\mathbf{x} - \tilde{\mathbf{w}}\|_2 = \frac{1}{1-t}\left\|(1-t)\mathbf{x} + t\mathbf{y}_k - \mathbf{w}\right\|_2 = \frac{1}{1-t}\|\mathbf{z}_k - \mathbf{w}\|_2 < r.$$

Dakle, $\tilde{\mathbf{w}} \in C$, a kako je $\mathbf{w} \in [\tilde{\mathbf{w}}, \mathbf{y}_k]$, slijedi da je $\mathbf{w} \in C$.

Dakle, dokazali smo da za svaki $k \in \mathbb{N}$ je $K(\mathbf{z}_k, (1-t)r) \subseteq C$. Ako uzmemo $k \in \mathbb{N}$ takav da je $(1-t)r < \frac{1}{k}$, tada je $\mathbf{z} \in K(\mathbf{z}_k, (1-t)r) \subseteq C$, odnosno, $\mathbf{z} \in \text{Int}(C)$. □

Lema 2.2.9. *Ako je $C \subseteq \mathbb{R}^n$ konveksan skup, tada je $\partial\text{Cl}(C) = \partial C$.*

Dokaz. Kako je $\partial\text{Cl}(C) = \text{Cl}(C) \setminus \text{Int}(\text{Cl}(C))$, dovoljno je dokazati da je

$$\text{Int}(\text{Cl}(C)) \subseteq \text{Int}(C).$$

Ako je $\text{Int}(C) = \emptyset$, po lemi 2.2.7, C je sadržan u pravom afinom potprostoru od $\mathbb{R}^n$. Budući da je svaki afin potprostor od $\mathbb{R}^n$ zatvoren, slijedi da je i $\text{Cl}(C)$ isto u tom pravom afinom potprostoru od $\mathbb{R}^n$, no kako svaki pravi afini potprostor od $\mathbb{R}^n$ ima prazan interior, vrijedi $\text{Int}(\text{Cl}(C)) = \emptyset$.

Neka je sad $\text{Int}(C) \neq \emptyset$. Neka je $\mathbf{y} \in \text{Int}(\text{Cl}(C))$, tada postoji $r > 0$ takav da je $K(\mathbf{y}, r) \subseteq \text{Cl}(C)$. Uzmimo proizvoljan $\mathbf{x} \in \text{Int}(C)$. Ako je $\mathbf{y} = \mathbf{x} \in \text{Int}(C)$, onda smo gotovi, dakle pretpostavimo $\mathbf{y} \neq \mathbf{x}$. Označimo

$$\varepsilon = \frac{r}{2\|\mathbf{y} - \mathbf{x}\|_2} > 0,$$

pa promotrimo

$$\mathbf{y}' = \mathbf{x} + (1+\varepsilon)(\mathbf{y} - \mathbf{x}) = \mathbf{y} + \varepsilon(\mathbf{y} - \mathbf{x}).$$

Uočimo da vrijedi

$$\|\mathbf{y} - \mathbf{y}'\|_2 = \frac{r}{2} < r$$

pa je $\mathbf{y}' \in \text{Cl}(C)$ te vrijedi

$$\mathbf{y} = \frac{1}{1+\varepsilon}\mathbf{y}' + \frac{\varepsilon}{1+\varepsilon}\mathbf{x},$$

odnosno $\mathbf{y} \in \langle \mathbf{x}, \mathbf{y}' \rangle$. Budući da je $\mathbf{x} \in \text{Int}(C)$, a $\mathbf{y}' \in \text{Cl}(C)$, po prethodnoj lemi je $\mathbf{y} \in \text{Int}(C)$. □

Definicija 2.2.10. *Kažemo da hiperravnina*

$$H = \left\{ \boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{a}^T \boldsymbol{x} = \alpha \right\}$$

separira skupove $A, B \subseteq \mathbb{R}^n$ ako je $A \subseteq H^+$ i $B \subseteq H^-$ (ili obratno), odnosno da ih jako separira ako postoje $\beta_1, \beta_2 \in \mathbb{R}$ takvi da

$$\boldsymbol{a}^T \boldsymbol{x} \leq \beta_1 < \beta_2 \leq \boldsymbol{a}^T \boldsymbol{y}, \quad \boldsymbol{x} \in A, \ \boldsymbol{y} \in B.$$

Teorem 2.2.11. *Neka su $A, B \subseteq \mathbb{R}^n$ konveksni i disjunktni. Tada su separabilni hiperravninom.*

Dokaz. Ako je bilo koji od skupova prazan, tvrdnja je trivijalna, pa pretpostavimo da nisu prazni. Lako se vidi da je $A - B$ isto konveksan skup, pa je to onda i $K := \mathrm{Cl}(A - B)$. Imamo dvije mogućnosti:

- $\mathbf{0} \notin K$. Kako je K zatvoren, konveksan i neprazan, označimo sa $\hat{\mathbf{a}}$ projekciju nule na K. Pa po teoremu 2.2.4,

  $$H_{\mathbf{0}} = \left\{ \mathbf{x} \in \mathbb{R}^n \ \mid \ \mathbf{x}^T \hat{\mathbf{a}} = \|\hat{\mathbf{a}}\|_2^2 \right\}$$

  separira $\{\mathbf{0}\}$ i K.

- $\mathbf{0} \in K$. Zbog disjunktnosti $\mathbf{0} \notin A - B$, pa mora biti $\mathbf{0} \in \partial(A - B)$. Po prethodnoj lemi je $\partial(A - B) = \partial K$, a po teoremu 2.2.6, postoji potporna hiperravnina od K koja sadrži $\mathbf{0}$, pa ona separira $\mathbf{0}$ i $A - B$.

Dakle, u oba slučaja, postoji hiperravnina koja separira $\mathbf{0}$ i $A - B$, pa postoji $\mathbf{z} \neq 0$ takav da za $\mathbf{a} \in A$ i $\mathbf{b} \in B$ je

$$\mathbf{z}^T(\mathbf{a} - \mathbf{b}) \leq \mathbf{z}^T \mathbf{0} = 0 \iff \mathbf{z}^T \mathbf{a} \leq \mathbf{z}^T \mathbf{b}.$$

□

Teorem 2.2.12. *Ako su $A, B \subseteq \mathbb{R}^n$ disjunktni konveksni skupovi, takvi da je A zatvoren, a B kompaktan. Tada postoji hiperravnina koja ih jako separira.*

Dokaz. Metrika $d : \mathbb{R}^n \times \mathbb{R}^n$ dana kao

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$$

je neprekidna funkcija jer je norma neprekidna funkcija. Neka su $\mathbf{a} \in A$ i $\mathbf{b} \in B$ proizvoljne točke i $r_1 = d(\mathbf{a}, \mathbf{b}) > 0$. Također, budući da je B kompaktan, sadržan je u nekoj kugli $K(\mathbf{b}, r_2)$ za neki $r_2 > 0$. Promotrimo sada skup $S = A \cap \overline{K}(\mathbf{a}, r_1 + r_2)$ koji je presjek skupa A i zatvorene kugle. S je kompaktan jer presjek zatvorenog skupa sa kompaktnim skupom te je neprazan jer $\mathbf{a} \in S$. Budući da je d neprekidna funkcija, ona postiže svoj minimum na kompaktnome skupu $S \times B$ u nekim točkama $\mathbf{a}_0 \in S, \mathbf{b}_0 \in B$.

Kada bi postojao par točaka $\mathbf{a}' \in A$ i $\mathbf{b}' \in B$ takvih da je

$$d(\mathbf{a}', \mathbf{b}') < d(\mathbf{a}_0, \mathbf{b}_0) \leq r_1,$$

bilo bi

$$d(\mathbf{a}', \mathbf{b}) \leq d(\mathbf{a}', \mathbf{b}') + d(\mathbf{b}, \mathbf{b}') < r_1 + r_2.$$

što implicira $\mathbf{a}' \in S$. To je kontradikcija sa činjenicom da $(\mathbf{a}_0, \mathbf{b}_0)^T$ minimizira d na $S \times B$. Po definiciji imamo da je $\mathbf{a}_0$ projekcija od $\mathbf{b}_0$ na A i $\mathbf{b}_0$ je projekcija od $\mathbf{a}_0$ na B.

Stavimo

$$\mathbf{z} = \mathbf{b}_0 - \mathbf{a}_0, \quad \beta_1 = \mathbf{z}^T \mathbf{a}_0, \quad \beta_2 = \mathbf{z}^T \mathbf{b}_0$$

pa po teoremu 2.2.4 imamo

$$\begin{aligned} \beta_2 - \beta_1 &= \|\mathbf{z}\|_2^2 > 0 \\ \mathbf{z}^T \mathbf{a} &\leq \beta_1, \qquad \forall \mathbf{a} \in A \\ -\mathbf{z}^T \mathbf{b} &\leq -\beta_2, \quad \forall \mathbf{b} \in B \end{aligned}$$

□

2.3 Konveksne funkcije

Definicija 2.3.1. *Neka je $C \subseteq \mathbb{R}^n$ konveksan skup i $f : C \to \mathbb{R}$. Kažemo da je f konveksna ako vrijedi*

$$(\forall \boldsymbol{x}, \boldsymbol{y} \in C)(\forall t \in [0,1])\ f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq t f(\boldsymbol{x}) + (1-t) f(\boldsymbol{y}).$$

Kažemo da je f strogo konveksna ako vrijedi

$$(\forall \boldsymbol{x}, \boldsymbol{y} \in C)(\forall t \in [0,1])\ \boldsymbol{x} \neq \boldsymbol{y} \implies f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) < t f(\boldsymbol{x}) + (1-t) f(\boldsymbol{y}).$$

Propozicija 2.3.2. *Neka je $C \subseteq \mathbb{R}^n$ konveksan skup i $f : C \to \mathbb{R}$ strogo konveksna funkcija. Ako f poprima minimum na C, onda je taj minimum jedinstven.*

Dokaz. Pretpostavimo suprotno, tj. da postoje točke $\mathbf{x}, \mathbf{y} \in C, \mathbf{x} \neq \mathbf{y}$ takve da je

$$f(\mathbf{x}) = f(\mathbf{y}) = \min_{\mathbf{z} \in C} f(\mathbf{z}).$$

Budući da je C konveksan, $\frac{\mathbf{x}+\mathbf{y}}{2} \in C$, pa stoga vrijedi

$$f\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) < \frac{f(\mathbf{x}) + f(\mathbf{y})}{2} = \min_{\mathbf{z} \in C} f(\mathbf{z}),$$

što je kontradikcija. □

Propozicija 2.3.3. *Neka je $C \subseteq \mathbb{R}^n$ konveksan skup i $f : C \to \mathbb{R}$ diferencijabilna funkcija. Tada je f konveksna ako i samo ako je*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in C.$$

Dokaz. Neka je f konveksna, te $\mathbf{x}, \mathbf{y} \in C$ te $t \in \langle 0, 1\rangle$. Dakle, vrijedi

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{x}),$$

odnosno

$$\frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \leq f(\mathbf{y}) - f(\mathbf{x}).$$

Budući da je f diferencijabilna, limes kada pustimo $t \to 0$ postoji i jednak je $\nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$.

Obratno, neka su $\mathbf{x}, \mathbf{y} \in C$ te $t \in \langle 0, 1\rangle$. Ako definiramo

$$\mathbf{z} = t\mathbf{x} + (1-t)\mathbf{y}$$

tada vrijedi

$$\mathbf{x} - \mathbf{z} = (1-t)(\mathbf{x} - \mathbf{y}), \quad \mathbf{y} - \mathbf{z} = t(\mathbf{y} - \mathbf{x}).$$

Po pretpostavci vrijedi

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^T(\mathbf{x} - \mathbf{z}) = f(\mathbf{z}) + (1-t)\nabla f(\mathbf{z})^T(\mathbf{x} - \mathbf{y})$$
$$f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^T(\mathbf{y} - \mathbf{z}) = f(\mathbf{z}) + t\nabla f(\mathbf{z})^T(\mathbf{y} - \mathbf{x}).$$

Ako pomnožimo prvu nejednakost sa t, a drugu sa $(1-t)$ i zbrojimo ih, dobivamo

$$tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \geq f(t\mathbf{x} + (1-t)\mathbf{y}).$$

□

Korolar 2.3.4. . *Neka je $C \subseteq \mathbb{R}^n$ konveksan skup i $f : C \to \mathbb{R}$ diferencijabilna i konveksna funkcija. Tada je svaka stacionarna točka globalni minimum od f na C.*

Teorem 2.3.5. *Neka je $C \subseteq \mathbb{R}^n$ konveksan skup. Tada je $f : C \to \mathbb{R}$ konveksna funkcija ako i samo ako je epigraf funkcije f*

$$\operatorname{epi}(f) = \left\{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} \mid \boldsymbol{x} \in C,\ y \geq f(\boldsymbol{x})\right\}$$

konveksan.

Dokaz. Pretpostavimo da je f konveksna i neka su $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \in \operatorname{epi}(f)$ te $t \in [0, 1]$. Tada je, zbog konveksnosti od C, $(1-t)\mathbf{x}_1 + t\mathbf{x}_2 \in C$ te vrijedi

$$(1-t)y_1 + ty_2 \geq (1-t)f(\mathbf{x}_1) + tf(\mathbf{x}_2).$$

Dakle, $(1-t)(\mathbf{x}_1, y_1) + t(\mathbf{x}_2, y_2) \in \operatorname{epi}(f)$, tj. $\operatorname{epi}(f)$ je konveksan.

Neka je epi(f) konveksan i neka su $\mathbf{x}_1, \mathbf{x}_2 \in C$ te $t \in [0,1]$. Zbog konveksnosti epigrafa je

$$(1-t)(\mathbf{x}_1, f(\mathbf{x}_1)) + t(\mathbf{x}_2, f(\mathbf{x}_2)) \in \text{epi}(f),$$

odnosno

$$f((1-t)\mathbf{x}_1 + t\mathbf{x}_2) \leq (1-t)f(\mathbf{x}_1) + tf(\mathbf{x}_2),$$

tj. f je konveksna. □

Budući da je analiza konveksnih funkcija mnogo zahtjevnija kada im domena nije cijeli $\mathbb{R}^n$, nećemo promatrati takve funkcije jer je lasso funkcija cilja ionako definirana na cijelom $\mathbb{R}^n$. Za pregled općenitije teorije vidi [2]. Jedna pogodnost kod ovakvog pristupa je sljedeći teorem.

Teorem 2.3.6. *Svaka konveksna funkcija $f : \mathbb{R}^n \to \mathbb{R}$ je neprekidna.*

Dokaz. Neka je $S \subseteq \mathbb{R}^n$ omeđen skup. Tada postoji $M > 0$ takav da je

$$S \subseteq [-M, M]^n = \text{conv}(\{-M, M\}^n).$$

Svaki $\mathbf{z} \in \text{conv}(\{-M, M\}^n)$ može se napisati kao konveksna kombinacija vektora iz $\{-M, M\}^n$, tj.

$$\mathbf{z} = \sum_{\mathbf{s} \in \{-M,M\}^n} \lambda_{\mathbf{s}} \mathbf{s}, \qquad \sum_{\mathbf{s} \in \{-M,M\}^n} \lambda_{\mathbf{s}} = 1, \quad (\forall \mathbf{s} \in \{-M, M\}^n)\lambda_{\mathbf{s}} \geq 0.$$

Zbog konveksnosti od f vrijedi

$$f(\mathbf{z}) \leq \sum_{\mathbf{s} \in \{-M,M\}^n} \lambda_{\mathbf{s}} f(\mathbf{s}) \leq \max_{\mathbf{s} \in \{-M,M\}^n} f(\mathbf{s}).$$

Dakle, f je omeđena odozgo na svakom ograničenom podskupu od $\mathbb{R}^n$.

Neka je $\mathbf{x}^* \in \mathbb{R}^n$ proizvoljan i neka je $M > 0$ takav da

$$\sup_{\mathbf{z} \in \overline{K}(\mathbf{x}^*, 1)} f(\mathbf{z}) < M.$$

pri čemu je $\overline{K}(\mathbf{x}^*, 1)$ zatvorena kugla radijusa 1 oko $\mathbf{x}^*$. Neka je sada $\mathbf{x} \in K(\mathbf{x}^*, 1)$, $\mathbf{x} \neq \mathbf{x}^*$. Tada je

$$0 < r := \|\mathbf{x}^* - \mathbf{x}\|_2 < 1$$

te promotrimo

$$\mathbf{x}^- = \mathbf{x}^* - \frac{\mathbf{x} - \mathbf{x}^*}{r}, \quad \mathbf{x}^+ = \mathbf{x}^* + \frac{\mathbf{x} - \mathbf{x}^*}{r}.$$

Lako se vidi da su $\mathbf{x}^-, \mathbf{x}^+ \in \overline{K}(\mathbf{x}^*, 1)$. Tada je

$$\mathbf{x}^* = \frac{r}{r+1}\mathbf{x}^- + \frac{1}{r+1}\mathbf{x},$$

pa zbog konveksnosti od f vrijedi

$$f(\mathbf{x}^*) \leq \frac{r}{r+1}f(\mathbf{x}^-) + \frac{1}{r+1}f(\mathbf{x}) \implies f(\mathbf{x}) - f(\mathbf{x}^*) \geq r\big(f(\mathbf{x}^*) - f(\mathbf{x}^-)\big) \geq r(f(\mathbf{x}^*) - M).$$

Slično,

$$\mathbf{x} = r\mathbf{x}^+ + (1-r)\mathbf{x}^*$$

pa zbog konveksnosti od f vrijedi

$$f(\mathbf{x}) \leq rf(\mathbf{x}^+) + (1-r)f(\mathbf{x}^*) \implies f(\mathbf{x}) - f(\mathbf{x}^*) \leq r\big(f(\mathbf{x}^+) - f(\mathbf{x}^*)\big) \leq r(M - f(\mathbf{x}^*)).$$

Kombinirajući gornje nejednakosti dobivamo

$$|f(\mathbf{x}) - f(\mathbf{x}^*)| \leq r|f(\mathbf{x}^*) - M| = \|\mathbf{x} - \mathbf{x}^*\|_2 |f(\mathbf{x}^*) - M|.$$

Za $\varepsilon > 0$, uzmimanjem

$$\delta = \frac{\varepsilon}{2|f(\mathbf{x}^*) - M|} > 0,$$

po definiciji slijedi da je f neprekidna. □

2.4 Recesivni konusi i nivo skupovi

Definicija 2.4.1. *Neka je $C \subseteq \mathbb{R}^n$ neprazan konveksan skup. Skup*

$$\operatorname{recc}(C) = \{\boldsymbol{y} \in \mathbb{R}^n \mid (\forall \boldsymbol{x} \in C)(\forall \lambda \geq 0)\, \boldsymbol{x} + \lambda \boldsymbol{y} \in C\}$$

nazivamo recesivni konus od C, a njegove elemente recesivnim smjerovima od C.

Lako se vidi da svaki recesivni konus sadrži $\mathbf{0}$, zatvoren je na množenje nenegativnim skalarima te je konveksan (tj. svaki recesivan konus je konveksan konus).

Teorem 2.4.2. *(Karakterizacija recesivnih smjerova) Neka je $C \subseteq \mathbb{R}^n$ neprazan, konveksan i zatvoren skup. Tada je $\boldsymbol{y} \in \mathbb{R}^n$ recesivan smjer ako i samo ako postoji $\boldsymbol{x} \in C$ takav da je*

$$\boldsymbol{x} + \lambda \boldsymbol{y} \in C, \quad \forall \lambda \geq 0.$$

Dokaz. Neka je $\mathbf{y} \in \mathbb{R}^n$ takav da postoji $\mathbf{x} \in C$ takav da

$$\mathbf{x} + \lambda \mathbf{y} \in C, \quad \forall \lambda \geq 0.$$

Ako je $\mathbf{y} = \mathbf{0}$, onda je sigurno u recesivnom konusu, pa pretpostavimo da $\mathbf{y} \neq \mathbf{0}$. Neka je $\mathbf{x}_1 \in C$ proizvoljan. Pokazat ćemo da je $\mathbf{x}_1 + \mathbf{y} \in C$. Definirajmo niz

$$\mathbf{z}_k := \mathbf{x} + k\mathbf{y} \in C, \quad \forall k \in \mathbb{N}.$$

Ako je $\mathbf{x}_1 = \mathbf{z}_k$ za neki $k \in \mathbb{N}$, tada smo gotovi jer je onda

$$\mathbf{x}_1 + \mathbf{y} = \mathbf{x} + (k+1)\mathbf{y} \in C$$

pa pretpostavimo da to nije slučaj. Promotrimo niz

$$\mathbf{y}_k = \frac{\mathbf{z}_k - \mathbf{x}_1}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \|\mathbf{y}\|_2, \quad \forall k \in \mathbb{N}.$$

Vrijedi

$$\mathbf{x}_1 + \mathbf{y}_k = \frac{\|\mathbf{y}\|_2}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \mathbf{z}_k + \frac{\|\mathbf{z}_k - \mathbf{x}_1\|_2 - \|\mathbf{y}\|_2}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \mathbf{x}_1,$$

odakle se vidi da je $\mathbf{x}_1 + \mathbf{y}_k$ afina kombinacija $\mathbf{z}_k$ i $\mathbf{x}_1$. Također, za sve $k \in \mathbb{N}$ za koje je

$$\|\mathbf{y}\|_2 \leq \|\mathbf{z}_k - \mathbf{x}_1\|_2$$

vrijedi da $\mathbf{x}_1 + \mathbf{y}_k$ leži na segmentu $[\mathbf{z}_k, \mathbf{x}_1] \subseteq C$ jer je konveksna kombinacija rubova segmenta. Kako je $\mathbf{z}_k$ neograničen niz, isto vrijedi i za $\|\mathbf{z}_k - \mathbf{x}_1\|_2$ pa po prethodnom razmatranju, postojat će $k_0 \in \mathbb{N}$ takav da je

$$\mathbf{x}_1 + \mathbf{y}_k \in C, \quad \forall k \in \mathbb{N},\ k \geq k_0.$$

Nadalje vrijedi

$$\begin{aligned}
\frac{\mathbf{y}_k}{\|\mathbf{y}\|_2} &= \frac{\mathbf{z}_k - \mathbf{x}_1}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \\
&= \frac{\mathbf{z}_k - \mathbf{x}}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} + \frac{\mathbf{x} - \mathbf{x}_1}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \\
&= \frac{k\mathbf{y}}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} + \frac{\mathbf{x} - \mathbf{x}_1}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \\
&= \frac{\|\mathbf{z}_k - \mathbf{x}\|_2}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \frac{\mathbf{y}}{\|\mathbf{y}\|_2} + \frac{\mathbf{x} - \mathbf{x}_1}{\|\mathbf{z}_k - \mathbf{x}_1\|_2},
\end{aligned}$$

gdje smo u zadnjem redu iskoristili da je $k = \frac{\|\mathbf{z}_k - \mathbf{x}\|_2}{\|\mathbf{y}\|_2}$. Kako je $\mathbf{z}_k$ neograničen niz, vrijedi $\|\mathbf{z}_k - \mathbf{x}_1\|_2 \to \infty$, pa korištenjem obrnute nejednakosti trokuta dobivamo

$$\frac{\|\mathbf{z}_k - \mathbf{x}\|_2}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \to 1, \quad \frac{\mathbf{x} - \mathbf{x}_1}{\|\mathbf{z}_k - \mathbf{x}_1\|_2} \to \mathbf{0},$$

odnosno vrijedi

$$\mathbf{y}_k \to \mathbf{y}.$$

Dakle, našli smo konvergentan niz

$$(\mathbf{y}_k + \mathbf{x}_1)_{k \geq k_0}$$

u C koji konvergira prema $\mathbf{y} + \mathbf{x}_1$, pa kako je C zatvoren, slijedi da je $\mathbf{y} + \mathbf{x}_1 \in C$.

Dakle, dokazali smo da je $\mathbf{y} + \mathbf{x}_1 \in C$ za svaki $\mathbf{x}_1 \in C$. Neka je sada $\lambda \geq 0$ proizvoljan. Ako je $\lambda = 0$, tada je sigurno

$$\mathbf{x}_1 + \lambda\mathbf{y} \in C, \quad \forall \mathbf{x}_1 \in C$$

pa pretpostavimo da je $\lambda > 0$. No tada ponovimo gornji dokaz za $\lambda\mathbf{y}$ umjesto $\mathbf{y}$ i dobijemo da je

$$\mathbf{x}_1 + \lambda\mathbf{y} \in C, \quad \forall \mathbf{x}_1 \in C.$$

□

Teorem 2.4.3. *Neka je $C \subseteq \mathbb{R}^n$ neprazan konveksan i zatvoren skup. Tada je C ograničen ako i samo ako je* $\mathrm{recc}(C) = \{\boldsymbol{0}\}$.

Dokaz. Ako je C ograničen i $\mathbf{y} \in \mathrm{recc}(C)$, $\mathbf{y} \neq 0$, tada je za proizvoljan $\mathbf{x} \in C$ niz

$$\mathbf{z}_k = \mathbf{x} + k\mathbf{y}$$

neograničen, što je kontradikcija.

Obratno, pretpostavimo da je $\mathrm{recc}(C) = \{\mathbf{0}\}$, ali da je C neograničen. Neka je $\mathbf{x} \in C$ proizvoljan i $\mathbf{z}_k$ neki neograničen niz iz C, rastućih normi, takav da je $\mathbf{x} \neq \mathbf{z}_k, \forall k \in \mathbb{N}$. Promotrimo niz

$$\mathbf{y}_k = \frac{\mathbf{z}_k - \mathbf{x}}{\|\mathbf{z}_k - \mathbf{x}\|_2}, \quad \forall k \in \mathbb{N}.$$

Budući da je $\mathbf{y}_k$ niz na jediničnoj sferi koja je kompaktan skup, slijedi da sadrži konvergentan podniz $(\mathbf{y}_{p_k})_{k \in \mathbb{N}}$ sa limesom $\mathbf{y}$ na jediničnoj sferi. Neka je $\alpha \geq 0$ proizvoljan pa promatramo

$$\mathbf{x} + \alpha\mathbf{y}_k = \mathbf{x} + \alpha\frac{\mathbf{z}_k - \mathbf{x}}{\|\mathbf{z}_k - \mathbf{x}\|_2} = \frac{\alpha}{\|\mathbf{z}_k - \mathbf{x}\|_2}\mathbf{z}_k + \frac{\|\mathbf{z}_k - \mathbf{x}\|_2 - \alpha}{\|\mathbf{z}_k - \mathbf{x}\|_2}\mathbf{x}.$$

Dakle, $\mathbf{x} + \alpha\mathbf{y}_k$ je afina kombinacija vektora $\mathbf{x}$ i $\mathbf{z}_k$ te je za one $k \in \mathbb{N}$ takve da je $\alpha \leq \|\mathbf{z}_k - \mathbf{x}\|_2$,

$$\mathbf{x} + \alpha\mathbf{y}_k \subseteq [\mathbf{z}_k, \mathbf{x}] \subseteq C.$$

Budući da je $\mathbf{z}_k$ neograničen niz, postoji $k_0 \in \mathbb{N}$ takav da je

$$\alpha \leq \|\mathbf{z}_k - \mathbf{x}\|_2, \quad k \geq k_0,$$

odnosno zbog prethodnih razmatranja

$$\mathbf{x} + \alpha\mathbf{y}_k \in C, \quad k \geq k_0.$$

Tada je $(\mathbf{x} + \alpha\mathbf{y}_{p_k})_{k\geq k_0}$ konvergentan niz u C, pa kako je C zatvoren, slijedi $\mathbf{x} + \alpha\mathbf{y} \in C$. Budući da je jediničnoj sferi $\mathbf{y} \neq \mathbf{0}$, a kako zadovoljava uvjete karakterizacije recesivnih smjerova, slijedi da je $\mathbf{y}$ recesivan smjer. Dakle, došli smo do kontradikcije. □

Definicija 2.4.4. *Neka $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija. Za $\alpha \in \mathbb{R}$ definiramo α-nivo skup od f kao $V_\alpha(f) = f^{-1}(\langle -\infty, \alpha])$.*

Kako je svaka konveksna funkcija na $\mathbb{R}^n$ neprekidna, slijedi da je svaki nivo skup zatvoren. Koristeći definiciju konveksnosti, lako se pokaže da je svaki nivo skup konveksne funkcije konveksan.

Teorem 2.4.5. *Neka je $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija. Tada svaki neprazan nivo skup ima isti recesivni konus jednak*

$$\operatorname{recc}(V_\alpha(f)) = \{\mathbf{y} \in \mathbb{R}^n \ : \ (\mathbf{y}, 0) \in \operatorname{recc}(\operatorname{epi}(f))\}.$$

Dokaz. Neka je $\alpha \in \mathbb{R}$ takav da je $V_\alpha(f)$ neprazan, neka je $\mathbf{y}$ njegov recesivan smjer te $\mathbf{x} \in V_\alpha(f)$. Kako je $f(\mathbf{x}) \leq \alpha$, slijedi da je $(\mathbf{x}, \alpha) \in \operatorname{epi}(f)$. Nadalje, budući da je $\mathbf{y}$ recesivan smjer, slijedi da je

$$f(\mathbf{x} + t\mathbf{y}) \leq \alpha, \quad \forall \mathbf{t} \geq 0,$$

odnosno

$$(\mathbf{x} + t\mathbf{y}, \alpha) \in \operatorname{epi}(f), \quad \forall \mathbf{t} \geq 0.$$

Dakle, pokazali smo

$$(x, \alpha) + t(\mathbf{y}, 0) \in \operatorname{epi}(f), \quad \forall t \geq 0$$

pa kako je $\operatorname{epi}(f)$ konveksan, zatvoren i neprazan skup, po karakterizaciji recesivnih smjerova, slijedi $(\mathbf{y}, 0) \in \operatorname{recc}(\operatorname{epi}(f))$.

Obratno, neka je $(\mathbf{y}, 0) \in \text{recc}(\text{epi}(f))$ te $(\mathbf{x}, \alpha) \in \text{epi}(f)$. Tada je

$$(\mathbf{x} + t\mathbf{y}, \alpha) \in \text{epi}(f), \quad \forall t \geq 0,$$

tj.

$$f(\mathbf{x} + t\mathbf{y}) \leq \alpha, \quad \forall t \geq 0,$$

odnosno

$$x + t\mathbf{y} \in V_\alpha(f), \quad \forall t \geq 0.$$

Kako je $V_\alpha(f)$ zatvoren, konveksan i neprazan, po karakterizaciji recesivnih smjerova, slijedi da je $\mathbf{y} \in \text{recc}(V_\alpha(f))$. □

Korolar 2.4.6. *Neka je $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija. Ako je bilo koji nivo skup ograničen, svaki nivo skup je ograničen.*

Dokaz. Neka je $V_\alpha(f)$ ograničen za neki $\alpha \in \mathbb{R}$. Tada je po teoremu 2.4.3 njegov recesivan konus jednak $\{\mathbf{0}\}$. No tada, po prethodnom teoremu, svaki nivo skup ima recesivan konus jednak $\{\mathbf{0}\}$. Tada je po teoremu 2.4.3 svaki nivo skup ograničen. □

Definicija 2.4.7. *Neka je $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija. Recesivni konus od f definiramo kao*

$$\text{recc}(f) = \{\boldsymbol{y} \in \mathbb{R}^n \mid (\boldsymbol{y}, 0) \in \text{recc}(\text{epi}(f))\},$$

a njegove elemente nazivamo recesivni smjerovi od f.

Propozicija 2.4.8. *Neka je $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija i $\boldsymbol{x} \in \mathbb{R}^n$. Ako $\boldsymbol{y} \notin \text{recc}(f)$, tada postoji $t \geq 0$ takav da vrijedi*

$$t \leq s_1 < s_2 \implies f(\boldsymbol{x} + s_1\boldsymbol{y}) < f(\boldsymbol{x} + s_2\boldsymbol{y})$$

te

$$\lim_{s \to \infty} f(\boldsymbol{x} + s\boldsymbol{y}) = \infty.$$

Dokaz. Budući da $\mathbf{y}$ nije recesivan smjer od f, po teoremu 2.4.5 on nije recesivan smjer niti jednog nepraznog nivo skupa od f pa posebno nije niti recesivan smjer skupa $V_{f(\mathbf{x})}(f) = f^{-1}(\langle -\infty, f(\mathbf{x})])$. Stoga, po karakterizaciji recesivnih smjerova, postoji $t > 0$ takav da $\mathbf{x} + t\mathbf{y} \notin V_{f(\mathbf{x})}(f)$, tj. vrijedi

$$f(\mathbf{x} + t\mathbf{y}) > f(\mathbf{x}).$$

Neka je $s > t$ proizvoljan te $\mathbf{u} = \mathbf{x} + t\mathbf{y}$ i $\mathbf{v} = \mathbf{x} + s\mathbf{y}$. Ako stavimo $r = \frac{t}{s} \in \langle 0, 1\rangle$, tada je

$$\mathbf{u} = (1 - r)\mathbf{x} + r\mathbf{v},$$

tj. zbog konveksnosti funkcije f vrijedi

$$f(\mathbf{u}) \leq (1-r)f(\mathbf{x}) + rf(\mathbf{v}) < (1-r)f(\mathbf{u}) + rf(\mathbf{v}) \implies f(\mathbf{u}) < f(\mathbf{v}).$$

Ako je $t \leq s_1 < s_2$, tada

$$f(\mathbf{x} + s_1\mathbf{y}) \geq f(\mathbf{x} + t\mathbf{y})$$

te $\mathbf{x} + s_1\mathbf{y}$ leži na segmentu između $\mathbf{x}$ i $\mathbf{x} + s_2\mathbf{y}$ pa koristeći sličnu argumentaciju kao gore, slijedi

$$f(\mathbf{x} + s_2\mathbf{y}) > f(\mathbf{x} + s_1\mathbf{y}).$$

Ako postoji $M \in \mathbb{R}$ takav da

$$f(\mathbf{x} + s\mathbf{y}) \leq M, \quad \forall s \geq 0,$$

tada po karakterizaciji recesivnih smjerova vrijedi $\mathbf{y} \in \operatorname{recc}(V_M(f))$. No $\mathbf{y}$ nije recesivan smjer niti jednog nepraznog nivo skupa od f. □

Teorem 2.4.9. *Neka je $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija. Tada je skup*

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \left\{ \boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = \min_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{y}) \right\}$$

neprazan i kompaktan ako i samo ako f nema recesivnih smjerova različitih od $\boldsymbol{0}$.

Dokaz. Neka je $\operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$ neprazan i kompaktan te označimo $p^* = \min_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{y})$. Tada je

$$\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = V_{p^*}(f).$$

Kako je f neprekidna i konveksna, $V_{p^*}(f)$ je neprazan, konveksan i kompaktan, pa je po teoremu 2.4.3 njegov recesivan konus jednak $\{\mathbf{0}\}$. Budući da je recesivan konus od f jednak recesivnom konusu bilo kojeg nepraznog nivo skupa of f, slijedi da f nema recesivnih smjerova različitih od $\mathbf{0}$.

Obratno, neka je $\alpha \in \mathbb{R}$ takav da je nivo skup $V_\alpha(f)$ neprazan. Budući da je f neprekidna i $\operatorname{recc}(f) = \{\mathbf{0}\}$, slijedi da je $V_\alpha(f)$ zatvoren i ograničen, pa kompaktan. Po Bolzano-Weierstrass teoremu, budući da je f neprekidna, f poprima minimum $\mathbf{x}^*$ na kompkatnom skupu $V_\alpha(f)$. No, $\mathbb{R}^n \setminus V_\alpha(f) = f^{-1}(\langle \alpha, \infty \rangle)$ pa je $\mathbf{x}^*$ točka minimuma of f na cijelom $\mathbb{R}^n$. Ako označimo $p^* = f(\mathbf{x}^*)$, tada je

$$\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = V_{p^*}(f)$$

te je svaki nivo skup od f kompaktan. □

Vratimo se lasso funkciji cilja

$$g(\beta) = \frac{1}{2}\|X\beta - \mathbf{y}\|_2^2 + \lambda\|\beta\|_1.$$

Ako uzmemo proizvoljan recesivni smjer $\mathbf{p}$ od g, tada je to recesivni smjer svakog nepraznog nivo skupa od g. Ako je $V_\alpha(g)$ neprazan, to znači da je

$$g(t\mathbf{p}) = \frac{1}{2}\|tX\mathbf{p} - \mathbf{y}\|_2^2 + \lambda t\|\mathbf{p}\|_1 \leq \alpha, \quad \forall t \geq 0.$$

Međutim, za $\mathbf{p} \neq \mathbf{0}$ je $\lim_{t\to\infty} g(t\mathbf{p}) = \infty$, pa g nema netrivijalnih recesivnih smjerova. Po prethodnom teoremu, postoji minimum funkcije g.

Poglavlje 3

Subgradijenti

3.1 Osnovna svojstva subgradijenata

U ovom poglavlju želimo poopćiti pojam gradijenta na konveksne funkcije. Kao motivaciju koristimo propoziciju 2.3.3 koju ponovno navodimo.
Ako je $f : \mathbb{R}^n \to \mathbb{R}$ diferencijabilna i konveksna funkcija tada vrijedi

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Poglavlje slijedi točku 9.16. iz [4], točku 5.4. iz [2] te [6] prilagođeno za funkcije čija je domena $\mathbb{R}^n$.

Definicija 3.1.1. *Subgradijent konveksne funkcije $f : \mathbb{R}^n \to \mathbb{R}$ u točki $\boldsymbol{x} \in \mathbb{R}^n$ je svaki vektor $\boldsymbol{g} \in \mathbb{R}^n$ takav da vrijedi*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{y} - \boldsymbol{x}), \forall \boldsymbol{y} \in \mathbb{R}^n.$$

Skup svih subgradijenata u točki $\boldsymbol{x} \in \mathbb{R}^n$ označavamo sa $\partial f(\boldsymbol{x})$.

Propozicija 3.1.2. *Neka je $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija i $\boldsymbol{x} \in \mathbb{R}^n$. Tada je $\partial f(\boldsymbol{x})$ zatvoren, konveksan i neprazan.*

Dokaz. Promatramo

$$\partial f(\mathbf{x}) = \left\{\mathbf{g} \in \mathbb{R}^n \mid (\forall \mathbf{y} \in \mathbb{R}^n) f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x})\right\}.$$

Neka je $(\mathbf{g}_k)_{k\in\mathbb{N}}$ konvergentan niz u $\partial f(\mathbf{x})$ sa limesom $\mathbf{g}$. Za proizvoljan $\mathbf{y} \in \mathbb{R}^n$ vrijedi

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}_k^T(\mathbf{y} - \mathbf{x})$$

pa iz neprekidnosti skalarnog produkta, puštanjem limesa $k \to \infty$ imamo

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}).$$

Dakle, $\mathbf{g} \in \partial f(\mathbf{x})$ pa je $\partial f(\mathbf{x})$ je zatvoren.

Neka su $\mathbf{g}_1, \mathbf{g}_2 \in \partial f(\mathbf{x})$, $t \in [0,1]$ te $\mathbf{y} \in \mathbb{R}^n$. Tada vrijedi

$$tf(\mathbf{y}) \geq tf(\mathbf{x}) + t\mathbf{g}_1^T(\mathbf{y} - \mathbf{x})$$
$$(1-t)f(\mathbf{y}) \geq (1-t)f(\mathbf{x}) + (1-t)\mathbf{g}_2^T(\mathbf{y} - \mathbf{x}).$$

Zbrajanjem nejednakosti slijedi da je $t\mathbf{g}_1 + (1-t)\mathbf{g}_2 \in \partial f(\mathbf{x})$. Dakle, $\partial f(\mathbf{x})$ je konveksan.

Promatramo epigraf od f

$$\text{epi}(f) = \Big\{(\mathbf{x}, y) \in \mathbb{R}^{n+1} \mid y \geq f(\mathbf{x})\Big\}.$$

Zbog konveksnosti i neprekidnosti od f, epi(f) je zatvoren, konveksan i neprazan podskup. Lako se pokaže da je

$$\partial\text{epi}(f) = \Big\{(\mathbf{x}, y) \in \mathbb{R}^{n+1} \mid y = f(\mathbf{x})\Big\}.$$

Neka je sada $\mathbf{x} \in \mathbb{R}^n$ proizvoljna. Tada je $(\mathbf{x}, f(\mathbf{x})) \in \partial\text{epi}(f)$, pa po teoremu 2.2.6, postoji potporna hiperravnina

$$H = \Big\{(\mathbf{z}, y) \in \mathbb{R}^n \;\mid\; \mathbf{a}^T\mathbf{z} + by = \alpha\Big\}, \quad (\mathbf{a}, b) \neq \mathbf{0}$$

od epi(f) takva da je

$$(\mathbf{x}, f(\mathbf{x})) \in H, \quad \text{epi}(f) \subseteq H^+,$$

što je ekvivalentno sa

$$\mathbf{y} \in \mathbb{R}^n \implies \mathbf{a}^T\mathbf{y} + bf(\mathbf{y}) \geq \alpha = \mathbf{a}^T\mathbf{x} + bf(\mathbf{x}). \tag{3.1}$$

Ako je $b \neq 0$, dijeljenjem s b dobivamo da je

$$f(\mathbf{y}) \geq f(\mathbf{x}) - \frac{1}{b}\mathbf{a}^T(\mathbf{y} - \mathbf{x}), \quad \mathbf{y} \in \mathbb{R}^n$$

pa slijedi

$$-\frac{1}{b}\mathbf{a} \in \partial f(\mathbf{x}).$$

S druge strane, ako je $b = 0$, tada je $\mathbf{a} \neq 0$ pa iz (3.1) slijedi

$$\mathbf{a}^T\mathbf{y} \geq \mathbf{a}^T\mathbf{x}, \quad \mathbf{y} \in \mathbb{R}^n.$$

Ako je $\mathbf{a}^T\mathbf{x} = 0$, dobivamo kontradikciju za $\mathbf{y} = -\mathbf{a}$, a ako je $\mathbf{a}^T\mathbf{x} \neq 0$, dobivamo kontradikciju za

$$\mathbf{y} = 2\mathbf{x} \quad \text{ili} \quad \mathbf{y} = -2\mathbf{x}.$$

□

Propozicija 3.1.3. *Neka je $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija . Ako je f diferencijabilna u $\boldsymbol{x} \in \mathbb{R}^n$, tada je $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$.*

Dokaz. Neka je f diferencijabilna u $\mathbf{x}$. Tada po propoziciji 2.3.3 slijedi da je $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$. Neka je sada $\mathbf{g} \in \partial f(\mathbf{x})$ proizvoljan te $\mathbf{h} \in \mathbb{R}^n$ i $\lambda > 0$. Iz definicije subgradijenata slijedi

$$f(\mathbf{x} + \lambda\mathbf{h}) - f(\mathbf{x}) \geq \lambda \mathbf{g}^T \mathbf{h},$$

što nakon dijeljenja s $\lambda > 0$ povlači

$$\frac{f(\mathbf{x} + \lambda\mathbf{h}) - f(\mathbf{x})}{\lambda} \geq \mathbf{g}^T \mathbf{h}.$$

Budući da je f diferencijabilna u $\mathbf{x}$, puštanjem limesa kada $\lambda \to 0$, dobivamo

$$\nabla f(\mathbf{x})^T \mathbf{h} \geq \mathbf{g}^T \mathbf{h},$$

odnosno

$$(\nabla f(\mathbf{x}) - \mathbf{g})^T \mathbf{h} \geq 0.$$

Uzimanjem $\mathbf{h} = -(\nabla f(\mathbf{x}) - \mathbf{g})$, dobivamo $\nabla f(\mathbf{x}) = \mathbf{g}$. □

Teorem 3.1.4. *Neka su $f, f_1, f_2 : \mathbb{R}^m \to \mathbb{R}$ konveksne funkcije, $A \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$ te $\alpha \geq 0$. Tada vrijedi*

(i) $\partial(\alpha f)(\boldsymbol{x}) = \alpha \partial f(\boldsymbol{x})$

(ii) Ako je $h = \boldsymbol{x} \mapsto f(A\boldsymbol{x} + \boldsymbol{b})$, *tada je* $\partial h(\boldsymbol{x}) = A^T \partial f(A\boldsymbol{x} + \boldsymbol{b})$.

(iii) $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$.

Dokaz.
(i) Neka je $\alpha \geq 0$. Vrijedi niz ekvivalencija

$$\begin{aligned}
\mathbf{g} \in (\partial(\alpha f))(x) &\iff (\forall \mathbf{y} \in \mathbb{R}^m)\ \alpha f(\mathbf{y}) \geq \alpha f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}) \\
&\iff (\forall \mathbf{y} \in \mathbb{R}^m)\ f(\mathbf{y}) \geq f(\mathbf{x}) + \left(\frac{1}{\alpha}\mathbf{g}\right)^T (\mathbf{y} - \mathbf{x}) \\
&\iff \frac{1}{\alpha}\mathbf{g} \in \partial f(\mathbf{x}) \\
&\iff \mathbf{g} \in \alpha \partial f(\mathbf{x}).
\end{aligned}$$

(ii) Neka je $\mathbf{g} \in \partial f(A\mathbf{x} + \mathbf{b})$ i $\mathbf{y} \in \mathbb{R}^n$. Tada je

$$h(\mathbf{y}) = f(A\mathbf{y} + \mathbf{b}) \geq f(A\mathbf{x} + \mathbf{b}) + \mathbf{g}^T(A\mathbf{y} - A\mathbf{x}) = h(\mathbf{x}) + (A^T\mathbf{g})^T(\mathbf{y} - \mathbf{x})$$

pa je $A^T\mathbf{g} \in \partial h(\mathbf{x})$. Dakle, dokazali smo $A^T \partial f(A\mathbf{x} + \mathbf{b}) \subseteq \partial h(\mathbf{x})$.

Dokaz obrata je mnogo složeniji. Neka je $\mathbf{g} \in \partial h(\mathbf{x})$. Pretpostavimo bez smanjenja općenitosti da je $\mathbf{b} = \mathbf{0}$. Promatramo dva skupa

$$C = \left\{(\mathbf{w}, s)^T \in \mathbb{R}^m \times \mathbb{R} \mid s < f(A\mathbf{x}) - f(\mathbf{w})\right\}$$
$$D = \left\{(\mathbf{w}, s)^T \in \mathbb{R}^m \times \mathbb{R} \mid (\exists \mathbf{y} \in \mathbb{R}^n)\, A\mathbf{y} = \mathbf{w},\ \ s > -\mathbf{g}^T(\mathbf{y} - \mathbf{x})\right\}.$$

Ako je $(\mathbf{w}, s)^T \in C \cap D$ tada vrijedi

$$-\mathbf{g}^T(\mathbf{y} - \mathbf{x}) < s < f(A\mathbf{x}) - f(A\mathbf{y}) = h(\mathbf{x}) - h(\mathbf{y})$$

što je u kontradikciji sa izborom od $\mathbf{g}$, pa su C i D disjunktni, a koristeći definiciju, lako se pokaže da su C i D konveksni. Tada po teoremu separacije 2.2.11, postoji vektor $(\mathbf{t}, d) \in \mathbb{R}^m \times \mathbb{R}, (\mathbf{t}, d) \neq \mathbf{0}$ te skalar $\alpha \in \mathbb{R}$ takvi da

$$\mathbf{t}^T\mathbf{w} + sd \leq \alpha \leq \mathbf{t}^T\tilde{\mathbf{w}} + \tilde{s}d, \quad (\mathbf{w}, s)^T \in C,\ (\tilde{\mathbf{w}}, \tilde{s})^T \in D. \tag{3.2}$$

Tvrdimo da mora biti $d > 0$. Pretpostavimo da je $d < 0$. Tada je $\lim_{\tilde{s}\to\infty} \tilde{s}d = -\infty$. Primijetimo da je

$$\{A\mathbf{x}\} \times \langle 0, \infty\rangle \subseteq D,$$

pa postavljanjem $\tilde{\mathbf{w}} = A\mathbf{x}$ i puštanjem $\tilde{s} \to \infty$ u (3.2), dobivamo kontradikciju sa

$$\alpha - \mathbf{t}^T A\mathbf{x} \leq \tilde{s}d.$$

Pretpostavimo da je $d = 0$. Tada $\mathbf{t} \neq \mathbf{0}$ pa uzimanjem $\mathbf{w}_n = n\mathbf{t}$ i $s_n \in \mathbb{R}$ dovoljno malog, tako da $(\mathbf{w}_n, s_n) \in C$, $\forall n \in \mathbb{N}$, iz (3.2) dobivamo

$$n\|\mathbf{t}\|_2^2 \leq \alpha, \forall n \in \mathbb{N},$$

što je kontradikcija. Dakle, mora biti $d > 0$ pa bez smanjenja općenitosti pretpostavljamo da je $d = 1$, tj. uzimamo da vrijedi

$$\mathbf{t}^T\mathbf{w} + s \leq \alpha \leq \mathbf{t}^T\tilde{\mathbf{w}} + \tilde{s}, \quad (\mathbf{w}, s)^T \in C,\ (\tilde{\mathbf{w}}, \tilde{s})^T \in D. \tag{3.3}$$

Za $\mathbf{w} = \tilde{\mathbf{w}} = A\mathbf{x}$ i puštanjem $s \nearrow 0$ i $\tilde{s} \searrow 0$ dobivamo $\alpha = \mathbf{t}^T A\mathbf{x}$, odnosno vrijedi

$$\mathbf{t}^T\mathbf{w} + s \leq \mathbf{t}^T A\mathbf{x} \leq \mathbf{t}^T\tilde{\mathbf{w}} + \tilde{s}, \quad (\mathbf{w}, s)^T \in C,\ (\tilde{\mathbf{w}}, \tilde{s})^T \in D. \tag{3.4}$$

Sada za proizvoljan $\mathbf{w} \in \mathbb{R}^m$ i $\mathbf{y} \in \mathbb{R}^n$ u (3.4) stavimo $\tilde{\mathbf{w}} = A\mathbf{y}$ te pustimo $s \nearrow f(A\mathbf{x}) - f(\mathbf{w})$ i $\tilde{s} \searrow -\mathbf{g}^T(\mathbf{y} - \mathbf{x})$ pa dobivamo

$$\mathbf{t}^T\mathbf{w} + f(A\mathbf{x}) - f(\mathbf{w}) \leq \mathbf{t}^T A\mathbf{x} \leq \mathbf{t}^T A\mathbf{y} - \mathbf{g}^T(\mathbf{y} - \mathbf{x}). \tag{3.5}$$

Budući da je $\mathbf{w} \in \mathbb{R}^m$ proizvoljan, prva nejednakost je ekvivalentna $\mathbf{t} \in \partial f(A\mathbf{x})$. Druga nejednakost je ekvivalentna

$$(A^T\mathbf{t} - \mathbf{g})^T(\mathbf{y} - \mathbf{x}) \geq 0$$

pa uvrštavanjem $\mathbf{y} = \mathbf{x} - A^T\mathbf{t} + \mathbf{g}$ dobivamo da je $\mathbf{g} = A^T\mathbf{t}$, odnosno $\mathbf{g} \in A^T\partial f(A\mathbf{x})$.

(iii) Ako su $\mathbf{g}_1 \in \partial f_1(\mathbf{x})$ i $\mathbf{g}_2 \in \partial f_2(\mathbf{x})$. Tada se lako vidi da je za proizvoljan $\mathbf{y} \in \mathbb{R}^m$ vrijedi

$$(f_1 + f_2)(\mathbf{y}) \geq (f_1 + f_2)(\mathbf{x}) + (\mathbf{g}_1 + \mathbf{g}_2)^T(\mathbf{y} - \mathbf{x}),$$

odnosno $\mathbf{g}_1 + \mathbf{g}_2 \in \partial(f_1 + f_2)(\mathbf{x})$.

Obrat je ponovno mnogo zahtjevniji. Neka je $\mathbf{g} \in \partial(f_1 + f_2)(\mathbf{x})$ i promotrimo skupove

$$C = \{(\mathbf{y}, s) \in \mathbb{R}^m \times \mathbb{R} \mid s < f_1(\mathbf{x}) - f(\mathbf{y})\}$$
$$D = \left\{(\mathbf{y}, s) \in \mathbb{R}^m \times \mathbb{R} \mid s > f_2(\mathbf{y}) - f_2(\mathbf{x}) - \mathbf{g}^T(\mathbf{y} - \mathbf{x})\right\}$$

Kao u dijelu (ii), pokaže se da su C i D disjunktni i konveksni, pa po teoremu separacije 2.2.11 postoje $(\mathbf{t}, d)^T \in \mathbb{R}^m \times \mathbb{R}, (\mathbf{t}, d)^T \neq \mathbf{0}$ i $\alpha \in \mathbb{R}$ takvi da

$$\mathbf{t}^T\mathbf{y} + ds \leq \alpha \leq \mathbf{t}^T\tilde{\mathbf{y}} + d\tilde{s}, \quad (\mathbf{y}, s)^T \in C,\ (\tilde{\mathbf{y}}, \tilde{s}) \in D.$$

Slično kao u koraku (ii), pokaže se da je $d > 0$ pa bez smanjenja općenitosti možemo pretpostaviti $d = 1$, odnosno vrijedi

$$\mathbf{t}^T\mathbf{y} + s \leq \alpha \leq \mathbf{t}^T\tilde{\mathbf{y}} + \tilde{s}, \quad (\mathbf{y}, s)^T \in C,\ (\tilde{\mathbf{y}}, \tilde{s}) \in D$$

pa za $\mathbf{y} = \tilde{\mathbf{y}} = \mathbf{x}$ i $s \nearrow 0$, $\tilde{s} \searrow 0$ dobivamo $\alpha = \mathbf{t}^T\mathbf{x}$, odnosno

$$\mathbf{t}^T\mathbf{y} + s \leq \mathbf{t}^T\mathbf{x} \leq \mathbf{t}^T\tilde{\mathbf{y}} + \tilde{s}, \quad (\mathbf{y}, s)^T \in C,\ (\tilde{\mathbf{y}}, \tilde{s}) \in D.$$

Za proizvoljne $\mathbf{y}, \tilde{\mathbf{y}} \in \mathbb{R}^m$ uzimanjem

$$s \nearrow (f_1(\mathbf{x}) - f_1(\mathbf{y})), \quad \tilde{s} \searrow (f_2(\tilde{\mathbf{y}}) - f_2(\mathbf{x}) - \mathbf{g}^T(\tilde{\mathbf{y}} - \mathbf{x}))$$

dobivamo

$$\mathbf{t}^T\mathbf{y} + f_1(\mathbf{x}) - f_1(\mathbf{y}) \leq \mathbf{t}^T\mathbf{x} \leq \mathbf{t}^T\tilde{\mathbf{y}} + f_2(\tilde{\mathbf{y}}) - f_2(\mathbf{x}) - \mathbf{g}^T(\tilde{\mathbf{y}} - \mathbf{x})$$

Zbog proizvoljnosti od $\mathbf{y}, \tilde{\mathbf{y}} \in \mathbb{R}^m$, prva nejednakost povlači $\mathbf{t} \in \partial f_1(\mathbf{x})$, dok druga povlači $\mathbf{g} - \mathbf{t} \in \partial f_2(\mathbf{x})$. Dakle, dokazali smo da je $\mathbf{g} \in \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x})$. □

Teorem 3.1.5. *Neka su $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ konveksne funkcije te*

$$f(\boldsymbol{x}) = \max_{i=1,\ldots,m} f_i(\boldsymbol{x}).$$

Pretpostavimo da je $\boldsymbol{x} \in \mathbb{R}^n$ takva da su sve f_i diferencijabilne u $\boldsymbol{x}$. Ako je

$$I(\boldsymbol{x}) = \{i \in \{1, \ldots, m\} \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x})\},$$

tada je

$$\partial f(\boldsymbol{x}) = \operatorname{conv}(\nabla f_i(\boldsymbol{x}) \mid i \in I(\boldsymbol{x})).$$

Dokaz. Za

$$\mathbf{g} \in \operatorname{conv}(\nabla f_i(\mathbf{x}) \mid i \in I(\mathbf{x}))$$

postoje skalari $\lambda_i \geq 0, \forall i \in I(\mathbf{x})$ takvi da je

$$\mathbf{g} = \sum_{i \in I(\mathbf{x})} \lambda_i \nabla f_i(\mathbf{x}), \quad \sum_{i \in I(\mathbf{x})} \lambda_i = 1.$$

Sada za $\mathbf{y} \in \mathbb{R}^n$ imamo

$$f(\mathbf{y}) = f(\mathbf{y}) \sum_{i \in I(\mathbf{x})} \lambda_i \geq \sum_{i \in I(\mathbf{x})} \lambda_i f_i(\mathbf{y}) \geq \sum_{i \in I(\mathbf{x})} \lambda_i \Big(f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \Big) = f(\mathbf{x}) + \mathbf{g}^T (\mathbf{y} - \mathbf{x})$$

pa je $\mathbf{g} \in \partial f(\mathbf{x})$.

Obratno, pretpostavimo da $\mathbf{g} \in \partial f(\mathbf{x})$, ali $\mathbf{g}$ nije u $C = \operatorname{conv}(\nabla f_i(\mathbf{x}) \mid i \in I(\mathbf{x}))$. Budući da je konveksna ljuska svakog konačnog skupa kompaktna (kao slika neprekidne funkcije na kompaktu), C je kompaktan i konveksan skup disjunktan zatvorenom konveksnom skupu $\{\mathbf{g}\}$. Po teoremu o strogoj separaciji 2.2.12, postoji $\mathbf{w} \in \mathbb{R}^n, \mathbf{w} \neq \mathbf{0}$ te $\alpha \in \mathbb{R}$ takav da je

$$\mathbf{h}^T \mathbf{w} < \alpha < \mathbf{g}^T \mathbf{w}, \quad \forall \mathbf{h} \in C. \tag{3.6}$$

Po definiciji subgradijenta, za svaki $\lambda > 0$ vrijedi

$$\mathbf{g}^T \mathbf{w} \leq \frac{f(\mathbf{x} + \lambda \mathbf{w}) - f(\mathbf{x})}{\lambda}. \tag{3.7}$$

Uzmimo proizvoljan pozitivan padajući niz $(\lambda_k)_{k \in \mathbb{N}}$ takav da $\lim_{k \to \infty} \lambda_k = 0$. Budući da je $\{1, \ldots, m\}$ konačan, postoji barem jedan $j \in \{1, \ldots, m\}$ takav da je

$$f(\mathbf{x} + \lambda_k \mathbf{w}) = f_j(\mathbf{x} + \lambda_k \mathbf{w}), \quad \text{za beskonačno mnogo } k \in \mathbb{N}.$$

Uzmimo podniz $(\lambda_{p_k})_{k \in \mathbb{N}}$ takav da je

$$f(\mathbf{x} + \lambda_{p_k} \mathbf{w}) = f_j(\mathbf{x} + \lambda_{p_k} \mathbf{w}), \quad \text{za sve } k \in \mathbb{N}.$$

Puštanjem limesa $k \to \infty$ u gornjoj jednakosti (jer je f neprekidna), dobivamo

$$f(\mathbf{x}) = f_j(\mathbf{x})$$

pa mora biti $j \in I(\mathbf{x})$.

Uvrštavanjem $\mathbf{h} = \nabla f_j(\mathbf{x}) \in C$ u (3.6) i korištenjem (3.7) dobivamo

$$\nabla f_j(\mathbf{x})^T \mathbf{w} < \alpha < \frac{f_j(\mathbf{x} + \lambda \mathbf{w}) - f_j(\mathbf{x})}{\lambda}, \quad \forall \lambda > 0.$$

Budući da je f_j diferencijabilna u $\mathbf{x}$, puštanjem $\lambda \searrow 0$ slijedi

$$\nabla f_j(\mathbf{x})^T \mathbf{w} < \alpha \leq \nabla f_j(\mathbf{x})^T \mathbf{w},$$

što je kontradikcija. □

Sada dolazimo do jednostavnog, ali jako važnog teorema.

Teorem 3.1.6. *(Nužan i dovoljan uvjet optimalnosti) Neka je $f : \mathbb{R}^n \to \mathbb{R}$ konveksna funkcija i $\boldsymbol{x}^* \in \mathbb{R}^n$. Tada je*

$$f(\boldsymbol{x}^*) = \min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) \iff \boldsymbol{0} \in \partial f(\boldsymbol{x}^*).$$

Dokaz.

$$\mathbf{0} \in \partial f(\mathbf{x}^*) \iff (\forall \mathbf{y} \in \mathbb{R}^n)\ f(\mathbf{y}) \geq f(\mathbf{x}^*) \iff f(\mathbf{x}^*) = \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}).$$

□

3.2 Primjeri

Primjer 3.2.1. *Promotrimo funkciju $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$. Ona je diferencijabilna na $\mathbb{R} \setminus \{0\}$ pa je u tim točkama po propoziciji 3.1.3 $\partial f(x) = \{\text{sign}(x)\}$. Dok za $g \in \mathbb{R}$ vrijedi*

$$g \in \partial f(0) \iff (\forall y \in \mathbb{R})\ |y| \geq g \cdot y \iff g \in [-1, 1]$$

pa je $\partial f(0) = [-1, 1]$.

Primjer 3.2.2. *Promotrimo $\|\cdot\|_1 : \mathbb{R}^n \to \mathbb{R}$, odnosno*

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| = \max_{\boldsymbol{s} \in \{-1,1\}^n} \boldsymbol{x}^T \boldsymbol{s}.$$

Primijetimo da je

$$\|\boldsymbol{x}\|_1 = \boldsymbol{x}^T \boldsymbol{s} \iff x_i \neq 0 \implies s_i = \text{sign}(x_i)$$

pa je po teoremu 3.1.5

$$\partial \|\boldsymbol{x}\|_1 = \text{conv}(\boldsymbol{s} \in \{-1, 1\}^n \mid x_i \neq 0 \implies s_i = \text{sign}(x_i)) = \{\boldsymbol{s} \in [-1, 1]^n \mid x_i \neq 0 \implies s_i = \text{sign}(x_i)\}.$$

Primjerice, za $\boldsymbol{x} = (2, -3, 0, 0)^T \in \mathbb{R}^4$ je

$$\partial \|\boldsymbol{x}\|_1 = \{1\} \times \{-1\} \times [-1, 1] \times [-1, 1].$$

Primjer 3.2.3. *(Lasso)*
Za $X \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$ i $\lambda > 0$, promatramo lasso funkciju cilja

$$g(\beta) = \frac{1}{2}\|X\beta - \boldsymbol{y}\|_2^2 + \lambda\|\beta\|_1, \quad \forall \beta \in \mathbb{R}^p.$$

Želimo minimizirati g po svim $\beta \in \mathbb{R}^p$. Pokazali smo da je to ekvivalentno nalaženju $\beta^* \in \mathbb{R}^p$ takvog da je $\boldsymbol{0} \in \partial g(\beta^*)$. Budući da je $\boldsymbol{z} \mapsto \|\boldsymbol{z}\|_2^2$ diferencijabilna funkcija, koristeći svojstva subgradijenata imamo

$$\partial g(\beta^*) = X^T(X\beta^* - \boldsymbol{y}) + \lambda\partial\|\beta^*\|_1 \ni \boldsymbol{0},$$

odnosno

$$X^T(\boldsymbol{y} - X\beta^*) = \lambda \boldsymbol{c} \tag{3.8}$$

gdje je $\boldsymbol{c} = (c_1, \ldots, c_n)^T \in [-1, 1]^n$ takav da $\beta_i^* \neq 0 \implies c_i = \operatorname{sign}(\beta_i^*)$.

Poglavlje 4

Svojstva lasso rješenja

4.1 Osnovna svojstva

U ovom poglavlju navodimo važna svojstva lasso funkcije cilja

$$g(\beta) = \frac{1}{2}\|X\beta - \mathbf{y}\|_2^2 + \lambda\|\beta\|_1, \quad \forall\beta \in \mathbb{R}^p.$$

Poglavlje prati drugi, treći i četvrti odjeljak iz [7], treći odjeljak iz [8] te treće poglavlje iz [9].

Teorem 4.1.1. *Za sve $X \in \mathbb{R}^{n\times p}$,$\boldsymbol{y} \in \mathbb{R}^n$ i $\lambda > 0$ vrijedi*

(i) Funkcija g ima ili jedinstvenu točku minimuma ili neprebrojivo mnogo točaka minimuma.

(ii) Svaka točka minimuma β funkcije g ima jednaku uklopljenu vrijednost $X\beta$.

(iii) Svaka točka minimuma β funkcije g ima jednaku ℓ_1 normu.

Dokaz.
(i) Na kraju drugog poglavlja smo pokazali da uvijek postoji lasso rješenje. Ako g ima dvije točke minimuma $\beta_1 \neq \beta_2$ tada za $t \in \langle 0, 1\rangle$ imamo

$$g(\beta_1) \leq g(t\beta_1 + (1-t)\beta_2) \leq tg(\beta_1) + (1-t)g(\beta_2) = g(\beta_1)$$

pa je i $t\beta_1 + (1-t)\beta_2$ točka minimuma za svaki $t \in \langle 0, 1\rangle$ - dakle postoji neprebrojivo mnogo točaka minimuma.

(ii) Neka su β_1 i β_2 dvije točke minimuma takve da je $X\beta_1 \neq X\beta_2$ te neka je $\alpha \in \langle 0, 1\rangle$. Označimo vrijednost minimuma od g sa c. Budući da je $\mathbf{z} \mapsto \|\mathbf{z} - \mathbf{y}\|_2^2$ striktno konveksna te $X\beta_1 \neq X\beta_2$, imamo

$$\|X((1-\alpha)\beta_1 + \alpha\beta_2) - \mathbf{y}\|_2^2 = \|(1-\alpha)X\beta_1 + \alpha X\beta_2 - \mathbf{y}\|_2^2 < (1-\alpha)\|X\beta_1 - \mathbf{y}\|_2^2 + \alpha\|X\beta_2 - \mathbf{y}\|_2^2$$

pa vrijedi

$$c \leq g((1-\alpha)\beta_1 + \alpha\beta_2) < (1-\alpha)g(\beta_1) + \alpha g(\beta_2) = c,$$

što je kontradikcija.

(iii) Treća tvrdnja trivijalno proizlazi iz prethodne. □

Na kraju trećeg poglavlja smo pokazali da je $\beta \in \mathbb{R}^p$ lasso rješenje ako i samo ako zadovoljava

$$X^T(\mathbf{y} - X\beta^*) = \lambda\mathbf{c} \tag{4.1}$$

gdje je $\mathbf{c} = (c_1, \ldots, c_p)^T \in [-1, 1]^p$ neki vektor takav da $\beta_i \neq 0 \implies c_i = \operatorname{sign}(\beta_i)$. Kako su uklopljene vrijednosti $X\beta$ jednake za svako rješenje $\beta \in \mathbb{R}^p$, vektor $\mathbf{c}$ je također jedinstven po svim rješenjima $\beta \in \mathbb{R}^p$. To za posljedicu ima da dva lasso rješenja ne mogu imati različite predznake na istoj komponenti. U prvom poglavlju smo pokazali da procjenitelj najmanjih kvadrata nema to svojstvo.

4.2 Oblik lasso rješenja

Fiksirajmo jedno lasso rješenje $\beta \in \mathbb{R}^p$ i promotrimo skup ekvikorelacije

$$\mathcal{E} = \left\{i \in \{1, \ldots, p\} \ : \ |X_i^T(\mathbf{y} - X\beta)| = \lambda\right\} = \{i \in \{1, \ldots, p\} \ : \ |c_i| = 1\}.$$

Uvedimo notaciju. Neka je $I \subseteq \{1, \ldots, p\}$. Za vektor $\mathbf{v} \in \mathbb{R}^p$, sa $\mathbf{v}_I \in \mathbb{R}^{|I|}$ označavamo podvektor od $\mathbf{v}$ koji sadrži samo koordinate koje su u I (u istom poretku kao u $\mathbf{v}$), a sa $\mathbf{v}_{-I} \in \mathbb{R}^{p-|I|}$ podvektor od $\mathbf{v}$ koji sadrži samo one koordinate koje nisu u I. Slično, za matricu $A \in \mathbb{R}^{n\times p}$, sa $A_I \in \mathbb{R}^{n\times|I|}$ označavamo podmatricu od A koja sadrži samo one stupce čiji je redni broj u I (u istom poretku kao u A), a sa $A_{-I} \in \mathbb{R}^{n\times(p-|I|)}$ podmatricu od A koja sadrži stupce čiji redni brojevi stupaca nisu u I.

Označimo $\mathbf{s} = \mathbf{c}_\mathcal{E}$. Iz definicije od $\mathcal{E}$ i vektora $\mathbf{s}$ slijedi da je

$$\beta_{-\mathcal{E}} = \mathbf{0}.$$

Nadalje, iz (4.1) slijedi

$$X_\mathcal{E}^T(\mathbf{y} - X_\mathcal{E}\beta_\mathcal{E}) = \lambda\mathbf{s}. \tag{4.2}$$

Dakle, $\lambda\mathbf{s}$ se nalazi u slici od $X_{\mathcal{E}}^T$. Koristeći (1) iz definicije 1.2.1 dobivamo

$$X_{\mathcal{E}}^T(X_{\mathcal{E}}^T)^+\lambda\mathbf{s} = \lambda\mathbf{s}.$$

Kombinirajući gornje dvije jednakosti dobivamo

$$X_{\mathcal{E}}^T X_{\mathcal{E}}\beta_{\mathcal{E}} = X_{\mathcal{E}}^T(\mathbf{y} - (X_{\mathcal{E}}^T)^+\lambda\mathbf{s}).$$

Množenjem slijeva sa $(X_{\mathcal{E}}^T)^+$ i korištenjem svojstava (1) i (4) iz definicije 1.2.1 te svojstva (10) iz napomene 1.2.5 dobivamo da za jedinstvenu uklopljenu vrijednost vrijedi

$$X\beta = X_{\mathcal{E}}\beta_{\mathcal{E}} = (X_{\mathcal{E}}^T)^+X_{\mathcal{E}}^T X_{\mathcal{E}}\beta_{\mathcal{E}} = (X_{\mathcal{E}}^T)^+X_{\mathcal{E}}^T(\mathbf{y} - (X_{\mathcal{E}}^T)^+\lambda\mathbf{s}) = X_{\mathcal{E}}X_{\mathcal{E}}^+(\mathbf{y} - (X_{\mathcal{E}}^T)^+\lambda\mathbf{s}),$$

odakle slijedi da je svako lasso rješenje oblika

$$\beta_{-\mathcal{E}} = 0, \quad \beta_{\mathcal{E}} = X_{\mathcal{E}}^+(\mathbf{y} - (X_{\mathcal{E}}^T)^+\lambda\mathbf{s}) + \mathbf{b}, \quad \mathbf{b} \in \mathrm{Ker}(X_{\mathcal{E}}).$$

Štoviše, svaki $\mathbf{b} \in \mathrm{Ker}(X_{\mathcal{E}})$ daje jedno lasso rješenje $\tilde{\beta}$, pod uvjetom da je zadovoljen uvjet predznaka

$$\tilde{\beta}_i \neq 0 \implies \mathrm{sign}(\tilde{\beta}_i) = s_i.$$

Uvjete zajedno možemo pisati kao

$$\mathbf{b} \in \mathrm{Ker}(X_{\mathcal{E}}), \quad s_i \cdot \left(\left[X_{\mathcal{E}}^+(\mathbf{y} - (X_{\mathcal{E}}^T)^+\lambda\mathbf{s})\right]_i + b_i\right) \geq 0, \quad i \in \mathcal{E}.$$

Dakle, dokazali smo

Teorem 4.2.1. *Za proizvoljne $\boldsymbol{y}, X$ i $\lambda > 0$. Sva lasso rješenja su oblika*

$$\beta_{-\mathcal{E}} = 0, \quad \beta_{\mathcal{E}} = X_{\mathcal{E}}^+(\boldsymbol{y} - (X_{\mathcal{E}}^T)^+\lambda\boldsymbol{s}) + \boldsymbol{b}, \quad \boldsymbol{b} \in \mathrm{Ker}(X_{\mathcal{E}}).$$

pri čemu je $\mathcal{E}$ skup ekvikorelacije i $\boldsymbol{s} \in \{-1, 1\}^p$ jedinstveni vektor iz (4.2).

Nadalje, svaki vektor $\boldsymbol{b} \in \mathrm{Ker}(X_{\mathcal{E}})$ daje jedno rješenje ako zadovoljava uvjet

$$s_i \cdot \left(\left[X_{\mathcal{E}}^+(\boldsymbol{y} - (X_{\mathcal{E}}^T)^+\lambda\boldsymbol{s})\right]_i + b_i\right) \geq 0, \quad i \in \mathcal{E}. \tag{4.3}$$

Ako je $\mathrm{Ker}(X_{\mathcal{E}}) = \{\mathbf{0}\}$, tada je rješenje po prethodnom teoremu jedinstveno i dano kao

$$\beta_{-\mathcal{E}} = 0, \quad \beta_{\mathcal{E}} = X_{\mathcal{E}}^+(\mathbf{y} - (X_{\mathcal{E}}^T)^+\lambda\mathbf{s}).$$

Naravno, nije odmah jasno kako dobiti skup ekvikorelacije $\mathcal{E}$, a time i vektor $\mathbf{s}$. S druge strane, ako $\mathrm{Ker}(X_{\mathcal{E}}) \neq \{\mathbf{0}\}$, nije jasno da $\mathbf{b} = \mathbf{0}$ zadovoljava uvijet (4.3). Kasnije ćemo pokazati da je to istina.

4.3 Dovoljan uvjet za jedinstvenost lasso rješenja

Jedinstvenost rješenja je prema gornjem teoremu ekvivalentna tome da $X_\mathcal{E}$ bude punog stupčanog ranga. Kako nam skup ekvikorelacije nije poznat, želimo naći dovoljne uvjete na matricu X, koja bi osiguravala da je $X_\mathcal{E}$ punog ranga.

Pretpostavimo da je $\text{Ker}(X_\mathcal{E}) \neq \{\mathbf{0}\}$, tada za neki $i \in \mathcal{E}$ vrijedi

$$X_i = \sum_{j \in \mathcal{E}\setminus\{i\}} c_j X_j, \quad c_j \in \mathbb{R},$$

odnosno

$$s_i X_i = \sum_{j \in \mathcal{E}\setminus\{i\}} (s_i s_j c_j) s_j X_j.$$

Skalarnim množenjem gornje jednakosti sa $(\mathbf{y} - X\beta)$, iz jednadžbe (4.2) dobivamo

$$\lambda s_i^2 = \sum_{j \in \mathcal{E}\setminus\{i\}} (s_i s_j c_j) s_j^2 \lambda$$

no $s_i^2 = s_j^2 = 1$ jer su $i, j \in \mathcal{E}$, pa imamo

$$1 = \sum_{j \in \mathcal{E}\setminus\{i\}} (s_i s_j c_j).$$

Dakle, ako $\text{Ker}(X_\mathcal{E}) \neq \{\mathbf{0}\}$, tada postoji $i \in \mathcal{E}$ takav da

$$s_i X_i = \sum_{j \in \mathcal{E}\setminus\{i\}} a_j X_j, \quad \sum_{j \in \mathcal{E}\setminus\{i\}} a_j = 1$$

što znači da $s_i X_i$ leži u affinoj ljusci od $s_j X_j$, $j \in \mathcal{E} \setminus \{i\}$. Dodatno, možemo bez smanjenja općenitosti pretpostaviti da $\mathcal{E} \setminus \{i\}$ ima najviše n elemenata, jer u suprotnom ponovimo gornje argumente za neki podskup od $\mathcal{E}$ sa $n+1$ elemenata (argumentacija se neće promijeniti). Stoga afina ljuska od $s_j X_j$, $j \in \mathcal{E} \setminus \{i\}$ je najviše $n-1$ dimenzionalna. Uvodimo novi pojam

Definicija 4.3.1. *Kažemo da je matrica $X \in \mathbb{R}^{n\times p}$ u generalnoj poziciji ako bilo koji afini potprostor $L \subseteq \mathbb{R}^n$ dimenzije $k < n$ ne sadrži više od $k+1$ elemenata skupa $\left\{\pm X_1, \ldots, \pm X_p\right\}$ pri čemu antipodalne parove $+X_i$ i $-X_i$ računamo zajedno.*

Drugi način da se definira generalna pozicija: za svaki $k < n$, afina ljuska bilo kojih $k+1$ točaka $\sigma_1 X_{i_1}, \ldots, \sigma_{k+1} X_{i_{k+1}}$, za proizvoljne predznake $\sigma_i \in \{-1, 1\}$, ne sadrži niti jedan element od $\{\pm X_i : i \neq i_1, \ldots, i_{k+1}\}$.

Gornjom analizom smo pokazali da $\mathrm{Ker}(X_{\mathcal{E}}) \neq \{\mathbf{0}\}$ implicira da matrica X nije u generalnoj poziciji. Koristeći obrat po kontrapoziciji, zaključujemo da za svaku matricu koja je u generalnoj poziciji vrijedi $\mathrm{Ker}(X_{\mathcal{E}}) = \{\mathbf{0}\}$, odnosno za takve matrice je lasso rješenje jedinstveno.

Sada ćemo pokazati da ako elementi matrice dolaze iz neke apsolutno neprekidne razdiobe, da je ta matrica nužno u generalnoj poziciji. Prije svega dokazujemo lemu:

Lema 4.3.2. *Svaki pravi afini potprostor od $\mathbb{R}^n$ ima Lebesgueovu mjeru 0.*

Dokaz. Kako je svaki pravi afini potprostor konačan presjek hiperravnina, dovoljno je dokazati da svaki afini potprostor dimenzije $n-1$ ima Lebesgueovu mjeru 0. Kako je Lebesgueova mjera invarijantna na translacije, dovoljno je dokazati da svaki vektorski potprostor dimenzije $n-1$ ima Lebesgueovu mjeru 0.

Prvo pokažimo da $R := \mathbb{R}^{n-1} \times \{0\}$ ima Lebesgueovu mjeru 0. Neka je $\varepsilon > 0$ proizvoljan. Lako se vidi da je

$$R \subseteq \bigcup_{n \in \mathbb{N}} \left([-n, n]^{n-1} \times \left[-\frac{\varepsilon}{4^n n^{n-1}}, \frac{\varepsilon}{4^n n^{n-1}} \right] \right)$$

te je Lebesgueova mjera unije na desnoj strani strogo manja od ε.

Neka je sada $V \leq \mathbb{R}^n$ vektorski potprostor dimenzije $n-1$ i $\varepsilon > 0$. Kako su svi konačnodimenzionalni vektorski prostori iste dimenzije međusobno izomorfni, postoji izomorfizam A između R i V kojeg proširimo na $\mathbb{R}^n$. Kako je svaki linearni operator na konačnodimenzionalnim normiranim prostorima ograničen, postoji $M > 0$ takav da je

$$\|A\mathbf{x}\|_\infty \leq M\|\mathbf{x}\|_\infty, \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Označimo sa

$$K^\infty(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{y}\|_\infty < r\}$$

otvorenu kocku oko $\mathbf{x}$ duljine stranice $2r$. Kako je R Lebesgueove mjere 0, postoji prebrojivo mnogo otvorenih kocaka $K_i = K^\infty(\mathbf{x}_i, r_i)$, $i \in \mathbb{N}$, takvih da je

$$R \subseteq \bigcup_{i \in \mathbb{N}} K_i, \quad \sum_{i \in \mathbb{N}} \lambda(K_i) = \sum_{i \in \mathbb{N}} (2r_i)^n < \frac{\varepsilon}{M^n}.$$

Za svaki $\mathbf{x} \in K_i$ vrijedi

$$\|A\mathbf{x} - A\mathbf{x}_i\|_\infty \leq M\|\mathbf{x} - \mathbf{x}_i\|_\infty < Mr_i,$$

pa je

$$V = A(R) \subseteq A\left(\bigcup_{i\in\mathbb{N}} K_i\right) = \bigcup_{i\in\mathbb{N}} A(K_i) \subseteq \bigcup_{i\in N} K^{\infty}(A\mathbf{x}_i, Mr_i).$$

Također vrijedi

$$\sum_{i\in\mathbb{N}} \lambda(K^{\infty}(A\mathbf{x}_i, Mr_i)) = \sum_{i\in\mathbb{N}} (2Mr_i)^n < \frac{\varepsilon}{M^n} M^n = \varepsilon$$

pa je V Lebesgueove mjere 0. □

Pretpostavimo sada da matrica X dolazi iz neprekidne razdiobe na $\mathbb{R}^{n\times p}$. Za $k < n$ i fiksne vektore $X_1, \ldots, X_{k+1}$, $\mathrm{aff}(X_1, \ldots, X_{k+1})$ je pravi afini potprostor pa je po prethodnoj lemi Lebesgueove mjere 0. Jer je distribucija matrice apsolutno neprekidna u odnosu na Lebesgueovu mjeru, vrijedi

$$\mathbb{P}(X_{k+2} \in \mathrm{aff}(X_1, \ldots, X_{k+1}) | X_1, \ldots, X_{k+1}) = 0.$$

Budući da to vrijedi za svake fiksne $X_1, \ldots, X_{k+1}$, zaključujemo da je

$$\mathbb{P}(X_{k+2} \in \mathrm{aff}(X_1, \ldots, X_{k+1})) = 0.$$

Ista argumentacija vrijedi za svaku kombinaciju $k + 2$ stupaca i $k + 2$ predznaka, pa zaključujemo da vrijedi

Teorem 4.3.3. *Ako matrica $X \in \mathbb{R}^{n\times p}$ dolazi iz neke apsolutno neprekidne distribucije na $\mathbb{R}^{np}$, tada je lasso rješenje gotovo sigurno jedinstveno.*

4.4 LARS algoritam za konstruiranje lasso putanje

U ovom poglavlju predstavljamo LARS algoritam koji konstruira jednu (posebnu) lasso putanju. Točnije, za fiksnu matricu $X \in \mathbb{R}^{n\times p}$ i $\mathbf{y} \in \mathbb{R}^n$ konstruiramo LARS lasso rješenje $\beta^L(\lambda)$ za sve $\lambda \in \langle 0, \infty\rangle$. Ranije smo pokazali da je $\beta \in \mathbb{R}^p$ lasso rješenje ako i samo ako zadovoljava

$$X^T(\mathbf{y} - X\beta) = \lambda\mathbf{c}, \quad \mathbf{c} \in [-1, 1]^p, \quad \beta_i \neq 0 \implies c_i = \mathrm{sign}(\beta_i). \tag{4.4}$$

Konstrukcija kreće sa $\lambda = \infty$ gdje je svako lasso rješenje jednako $\beta = \mathbf{0}$ te kako se λ smanjuje, konstruira $\beta^L(\lambda)$ kao po dijelovima linearnu funkciju. Prvo navodimo algoritam, a zatim pokazujemo da je dobivena funkcija $\lambda \mapsto \beta^L(\lambda)$ zaista lasso putanja. Naravno, ako $X_{\mathcal{E}}^T X_{\mathcal{E}}$ nije regularna, onda lasso rješenje nije jedinstveno. Algoritam će generirati lasso rješenje najmanje ℓ^2 norme.

Ako označimo $\lambda_1 = \left\|X^T\mathbf{y}\right\|_\infty$, primijetimo da je $\beta = \mathbf{0}$ jedno lasso rješenje na $\lambda \in \langle\lambda_1, \infty\rangle$ jer $\mathbf{c} = \frac{X^T\mathbf{y}}{\lambda} \in [-1,1]^p$ zadovoljava jednadžbu (4.4), a kako po teoremu 4.1.1 sva lasso rješenja imaju istu ℓ^1 normu, slijedi da je svako lasso rješenje na tom intervalu jednako $\mathbf{0}$. Dakle, na $\langle\lambda_1, \infty\rangle$ je $\mathcal{E}_0 = \emptyset$, a u $\lambda = \lambda_1$ je

$$\mathcal{E}_1 = \left\{j \in \{1, \ldots, p\} \ : \ |X_j^T\mathbf{y}| = \left\|X^T\mathbf{y}\right\|_\infty\right\}$$

i $\mathbf{s}_1 = \text{sign}(X_{\mathcal{E}}^T\mathbf{y})$. Sada provodimo iteracije. Neka je u λ_k dan skup ekvikorelacije $\mathcal{E}_k$ i $\mathbf{s}_k$. U λ_k postavimo $\beta^L_{-\mathcal{E}_k}(\lambda_k) = \mathbf{0}$ i

$$\beta^L_{\mathcal{E}_k}(\lambda_k) = (X_{\mathcal{E}_k})^+(\mathbf{y} - (X^T_{\mathcal{E}_k})^+\lambda_k\mathbf{s}_k) = c - \lambda_k d$$

pri čemu su $\mathbf{c} = X^+_{\mathcal{E}_k}\mathbf{y}$ i $\mathbf{d} = (X^T_{\mathcal{E}_k}X_{\mathcal{E}_k})^+\lambda_k s_k$. To je ujedno i rješenje najmanje ℓ^2 norme iz skupa

$$\underset{\mathbf{b}\in\mathbb{R}^{|\mathcal{E}_k|}}{\text{argmin}}\left\|\mathbf{y} - (X^T_{\mathcal{E}_k})^+\lambda_k\mathbf{s}_k - X_{\mathcal{E}_k}\mathbf{b}\right\|.$$

Smanjujući λ, držimo $\beta^L_{-\mathcal{E}_k}(\lambda) = \mathbf{0}$ dok ostatak proširujemo afino $\beta^L_{\mathcal{E}_k}(\lambda) = \mathbf{c} - \lambda\mathbf{d}$. Kako λ pada od λ_k, radimo dvije provjere. Prvo, kada bi varijabla izvan ekvikorelacijskog skupa trebala u njega ući tj. kada najprije prestaje vrijediti

$$|X_i^T(\mathbf{y} - X_{\mathcal{E}_k}\mathbf{c} + \lambda X_{\mathcal{E}_k}\mathbf{d})| < \lambda$$

za neki $i \notin \mathcal{E}_k$ i drugo kada bi varijabla trebala izaći iz ekvikorelacijskog skupa, odnosno

$$\beta_i^L(\lambda) = c_i - \lambda d_i = 0$$

za neki $i \in \mathcal{E}_k$. Prvo vrijeme nazivamo vremenom ulaska i lagano se pokaže da je jednako

$$\lambda^{\text{in}}_{k+1} = \max^k_{i\notin\mathcal{E}_k}\frac{X_i^T(X_{\mathcal{E}_k}\mathbf{c} - \mathbf{y})}{X_i^TX_{\mathcal{E}_k}\mathbf{d} \pm 1} = \max^k_{i\notin\mathcal{E}_k}\frac{X_i^T(X_{\mathcal{E}_k}X^+_{\mathcal{E}_k} - I)\mathbf{y}}{X_i^T(X^T_{\mathcal{E}_k})^+\mathbf{s}_k \pm 1}$$

pri čemu $\max^k$ označava maksimum koji je u intervalu $[0, \lambda_k\rangle$ ili ako takvog nema, postavimo ga na 0. Označimo i $i^{\text{in}}_{k+1} \notin \mathcal{E}_k$ indeks na kojemu se postiže maksimum. Drugo vrijeme nazivamo vremenom izlaska i jednako je

$$\lambda^{\text{out}}_{k+1} = \max^k_{i\in\mathcal{E}_k}\frac{c_i}{d_i} = \max^k_{i\in\mathcal{E}_k} = \frac{[X^+_{\mathcal{E}_k}\mathbf{y}]_i}{[(X^T_{\mathcal{E}_k}X_{\mathcal{E}_k})^+\mathbf{s}_k]_i}$$

te označimo pripadni indeks sa i^{out}_{k+1}.

Odredimo λ_{k+1} kao

$$\lambda_{k+1} = \max\left\{\lambda_{k+1}^{\text{in}}, \lambda_{k+1}^{\text{out}}\right\}.$$

Ako je λ_{k+1} vrijeme ulaska, stavimo

$$\mathcal{E}_{k+1} = \mathcal{E}_k \cup \left\{i_{k+1}^{\text{in}}\right\},$$

a ako je vrijeme izlaska

$$\mathcal{E}_{k+1} = \mathcal{E}_k \setminus \left\{i_{k+1}^{\text{out}}\right\}.$$

Konačno stavimo

$$\mathbf{s}_{k+1} = \text{sign}(X_{\mathcal{E}_{k+1}}^T(\mathbf{y} - X_{\mathcal{E}_k}\mathbf{c} + \lambda_{k+1}X_{\mathcal{E}_k}\mathbf{d})).$$

Iteracije prestaju kada $\lambda_{k+1} = 0$.

Lema 4.4.1. *Za proizvoljne $X \in \mathbb{R}^{n\times p}$ i $\mathbf{y} \in \mathbb{R}^n$ algoritam napravi najviše*

$$\sum_{k=0}^{p}\binom{p}{k}2^k = 3^p$$

koraka.

Dokaz. Pokazat ćemo da za $k < k'$ ne može biti $(\mathcal{E}_k, \mathbf{s}_k) = (\mathcal{E}_{k'}, \mathbf{s}_{k'})$ pa će broj koraka algoritma biti određen svim mogućim izborima $\mathcal{E}_k \subseteq \{1, \ldots, p\}$ i $\mathbf{s}_k \in \{-1, 1\}^{|\mathcal{E}_k|}$. Pretpostavimo da je $(\mathcal{E}_k, \mathbf{s}_k) = (\mathcal{E}_{k'}, \mathbf{s}_{k'}) =: (\mathcal{E}, \mathbf{s})$. Tada su za obje iteracije jednaki i vektori $\mathbf{c}$ i $\mathbf{d}$. Po konstrukciji, za $\lambda = \lambda_k$ i $\lambda = \lambda_{k'}$ vrijedi

$$|X_i^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda X_{\mathcal{E}}\mathbf{d})| < \lambda, \quad i \notin \mathcal{E}$$

te

$$\mathbf{s} = \text{sign}(X_{\mathcal{E}}^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda X_{\mathcal{E}}\mathbf{d})).$$

Za $t \in \langle 0, 1\rangle$, $\lambda_t := (1-t)\lambda_k + t\lambda_{k'}$ i proizvoljan $i \notin \mathcal{E}$ je

$$\begin{aligned}|X_i^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda_t X_{\mathcal{E}}\mathbf{d})| &\leq (1-t)|X_i^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda_k X_{\mathcal{E}}\mathbf{d})| + t|X_i^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda_{k'} X_{\mathcal{E}}\mathbf{d})| \\ &< (1-t)\lambda_k + t\lambda_{k'} \\ &= \lambda_t.\end{aligned}$$

Također, za proizvoljan $i \in \mathcal{E}$ je

$$X_i^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda_t\mathbf{d}) = (1-t)X_i^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda_k\mathbf{d}) + tX_i^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda_{k'}\mathbf{d})$$

pa kako su predznaci dva skalarna produkta jendaki, i njihova konveksna kombinacija ima isti predznak, tj. vrijedi

$$\mathbf{s} = \text{sign}(X_{\mathcal{E}}^T(\mathbf{y} - X_{\mathcal{E}}\mathbf{c} + \lambda_t X_{\mathcal{E}}\mathbf{d})).$$

No to je kontradikcija sa činjenicom da su k i k' različite iteracije. □

Napomena 4.4.2. *Može se pokazati da ako je $(\mathcal{E}, \boldsymbol{s})$ u nekoj iteraciji, da u nekoj drugoj iteraciji ne može biti $(\mathcal{E}, -\boldsymbol{s})$ pa se gornja ograda u gornjoj lemi može popraviti do $\frac{3^p+1}{2}$. Također, može se konstruirati primjer matrice X i vektora $\mathbf{y}$ tako da broj iteracija dosegne baš tu gornju ogradu. Za konstrukciju takvog primjera vidi [5].*

Sada dokazujemo ispravnost LARS algoritma, tj. da je LARS putanja ujedno i lasso putanja. Prije svega dokazujemo lemu

Lema 4.4.3. *(Lema ubacivanja i izbacivanja) Neka je $k \geq 1$. Postoje dvije mogućnosti:*

- *(Izbacivanje) Ako u λ_{k+1} varijabla izlazi iz ekvikorelacijskog skupa, tada*

$$\begin{bmatrix} [(X_{\mathcal{E}_k})^+(\mathbf{y} - (X_{\mathcal{E}_k}^T)^+\lambda_{k+1}\boldsymbol{s}_k)]_{\mathcal{E}_k \setminus i_{k+1}^{\text{out}}} \\ [(X_{\mathcal{E}_k})^+(\mathbf{y} - (X_{\mathcal{E}_k}^T)^+\lambda_{k+1}\boldsymbol{s}_k)]_{i_{k+1}^{\text{out}}} \end{bmatrix} = \begin{bmatrix} (X_{\mathcal{E}_{k+1}})^+(\mathbf{y} - (X_{\mathcal{E}_{k+1}}^T)^+\lambda_{k+1}\boldsymbol{s}_{k+1}) \\ 0 \end{bmatrix}, \tag{4.5}$$

- *(Ubacivanje) Ako u λ_{k+1} varijabla ulazi u ekvikorelacijski skup, tada*

$$\begin{bmatrix} (X_{\mathcal{E}_k})^+(\mathbf{y} - (X_{\mathcal{E}_k}^T)^+\lambda_{k+1}\boldsymbol{s}_k) \\ 0 \end{bmatrix} = \begin{bmatrix} [(X_{\mathcal{E}_{k+1}})^+(\mathbf{y} - (X_{\mathcal{E}_{k+1}}^T)^+\lambda_{k+1}\boldsymbol{s}_{k+1})]_{\mathcal{E}_{k+1} \setminus i_{k+1}^{\text{in}}} \\ [(X_{\mathcal{E}_{k+1}})^+(\mathbf{y} - (X_{\mathcal{E}_{k+1}}^T)^+\lambda_{k+1}\boldsymbol{s}_{k+1})]_{i_{k+1}^{\text{in}}} \end{bmatrix}. \tag{4.6}$$

Dokaz. (Izbacivanje) Označimo lijevu stranu od (4.5) sa

$$\begin{bmatrix} \mathbf{b}_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} [(X_{\mathcal{E}_k})^+(\mathbf{y} - (X_{\mathcal{E}_k}^T)^+\lambda_{k+1}\mathbf{s}_k)]_{\mathcal{E}_k \setminus i_{k+1}^{\text{out}}} \\ [(X_{\mathcal{E}_k})^+(\mathbf{y} - (X_{\mathcal{E}_k}^T)^+\lambda_{k+1}\mathbf{s}_k)]_{i_{k+1}^{\text{out}}} \end{bmatrix}.$$

Po konstrukciji vremena izlaska imamo da je $b_2 = 0$. Bez smanjenja općenitosti pretpostavimo da je i_{k+1}^{out} posljednja varijabla u $\mathcal{E}_k$. Tada je

$$\begin{bmatrix} \mathbf{b}_1 \\ b_2 \end{bmatrix} = (X_{\mathcal{E}_k})^+(\mathbf{y} - (X_{\mathcal{E}_k}^T)^+\lambda_{k+1}\mathbf{s}_k).$$

Po konstrukciji, vektor $(\mathbf{b}_1, b_2)^T$ je rješenje najmanje ℓ^2 norme problema najmanjih kvadrata

$$\underset{\mathbf{b} \in \mathbb{R}^{|\mathcal{E}_k|}}{\operatorname{argmin}} \left\| \mathbf{y} - (X_{\mathcal{E}_k}^T)^+\lambda_{k+1}\mathbf{s}_k - X_{\mathcal{E}_k}\mathbf{b} \right\|$$

odnosno rješenje najmanje ℓ^2 norme sustava normalnih jednadžbi

$$X_{\mathcal{E}_k}^T X_{\mathcal{E}_k} \begin{bmatrix} \mathbf{b}_1 \\ b_2 \end{bmatrix} = X_{\mathcal{E}_k}^T \mathbf{y} - \lambda_{k+1}\mathbf{s}_k \tag{4.7}$$

pri čemu smo koristili da je $\mathbf{s}_k$ u slici od $X_{\mathcal{E}_k}^T$ što slijedi iz (4.2). Zapišimo (4.7) u blokovskom obliku

$$\begin{bmatrix} X_{\mathcal{E}_{k+1}}^T X_{\mathcal{E}_{k+1}} & X_{\mathcal{E}_{k+1}}^T X_{i_{k+1}^{\text{out}}} \\ X_{i_{k+1}^{\text{out}}}^T X_{\mathcal{E}_{k+1}} & X_{i_{k+1}^{\text{out}}}^T X_{i_{k+1}^{\text{out}}} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_{\mathcal{E}_{k+1}}^T \\ X_{i_{k+1}^{\text{out}}}^T \end{bmatrix} \mathbf{y} - \lambda_{k+1} \begin{bmatrix} \mathbf{s}_{k+1} \\ z \end{bmatrix} \tag{4.8}$$

pri čemu je $z \in \{-1, 1\}$. Budući da je $b_2 = 0$, vrijedi

$$X_{\mathcal{E}_{k+1}}^T X_{\mathcal{E}_{k+1}} \mathbf{b}_1 = X_{\mathcal{E}_{k+1}}^T \mathbf{y} - \lambda_{k+1} \mathbf{s}_{k+1},$$

odnosno $\mathbf{b}_1$ je oblika

$$\mathbf{b}_1 = \left(X_{\mathcal{E}_{k+1}}^T X_{\mathcal{E}_{k+1}}\right)^+ (X_{\mathcal{E}_{k+1}} \mathbf{y} - \lambda_{k+1} \mathbf{s}_{k+1}) + \eta = X_{\mathcal{E}_{k+1}}^+ \left(\mathbf{y} - (X_{\mathcal{E}_{k+1}}^T)^+ \lambda_{k+1} \mathbf{s}_{k+1}\right) + \eta,$$

pri čemu je $\eta \in \text{Ker}(X_{\mathcal{E}_{k+1}})$. Budući da je $(X_{\mathcal{E}_{k+1}}^T X_{\mathcal{E}_{k+1}})^+$ projektor na stupce od $X_{\mathcal{E}_{k+1}}^T$, a kako su $\text{Im}(X_{\mathcal{E}_{k+1}}^T)$ i $\text{Ker}(X_{\mathcal{E}_{k+1}})$ ortogonalni komplementi, vrijedi da je rješenje minimalne ℓ^2 norme upravo ono za koje je $\eta = \mathbf{0}$ što dokazuje prvu tvrdnju.

(Ubacivanje) Dokaz je sličan, ali nešto teži. Označimo desnu stranu od (4.6) sa

$$\begin{bmatrix} \mathbf{b}_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} [(X_{\mathcal{E}_{k+1}})^+ (\mathbf{y} - (X_{\mathcal{E}_{k+1}}^T)^+ \lambda_{k+1} \mathbf{s}_{k+1})]_{\mathcal{E}_{k+1} \setminus i_{k+1}^{\text{in}}} \\ [(X_{\mathcal{E}_{k+1}})^+ (\mathbf{y} - (X_{\mathcal{E}_{k+1}}^T)^+ \lambda_{k+1} \mathbf{s}_{k+1})]_{i_{k+1}^{\text{in}}} \end{bmatrix}$$

te pretpostavimo da je i_{k+1}^{in} najveći indeks u $\mathcal{E}_{k+1}$. Po konstrukciji $(\mathbf{b}_1, b_2)^T$ je najmanje po ℓ^2 normi rješenje sustava normalnih jednadžbi

$$X_{\mathcal{E}_{k+1}}^T X_{\mathcal{E}_{k+1}} \begin{bmatrix} \mathbf{b}_1 \\ b_2 \end{bmatrix} = X_{\mathcal{E}_{k+1}}^T \mathbf{y} - \lambda_{k+1} \mathbf{s}_{k+1}. \tag{4.9}$$

Ako zapišemo (4.9) po blokovima dobivamo

$$\begin{bmatrix} X_{\mathcal{E}_k}^T X_{\mathcal{E}_k} & X_{\mathcal{E}_k}^T X_{i_{k+1}^{\text{in}}} \\ X_{i_{k+1}^{\text{in}}}^T X_{\mathcal{E}_k} & X_{i_{k+1}^{\text{in}}}^T X_{i_{k+1}^{\text{in}}} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_{\mathcal{E}_k}^T \\ X_{i_{k+1}^{\text{in}}}^T \end{bmatrix} \mathbf{y} - \lambda_{k+1} \begin{bmatrix} \mathbf{s}_k \\ z \end{bmatrix} \tag{4.10}$$

pri čemu je $z \in \{-1, 1\}$.

Rješavajući sustav za $\mathbf{b}_1$ dobivamo

$$\mathbf{b}_1 = (X_{\mathcal{E}_k}^T X_{\mathcal{E}_k})^+ \left(X_{\mathcal{E}_k}^T \mathbf{y} - \lambda_{k+1} \mathbf{s}_k - X_{\mathcal{E}_k}^T X_{i_{k+1}^{\text{in}}} b_2\right) + \eta \tag{4.11}$$

$$= X_{\mathcal{E}_k}^+ \left(\mathbf{y} - \lambda_{k+1} (X_{\mathcal{E}_k}^+)^T \mathbf{s}_k - X_{i_{k+1}^{\text{in}}} b_2\right) + \eta \tag{4.12}$$

pri čemu je $\eta \in \text{Ker}(X_{\mathcal{E}_k})$. Kao u prošlom dijelu, $\eta = \mathbf{0}$ postiže najmanju ℓ^2 normu, pa je još preostalo dokazati da je $b_2 = 0$. Druga jednadžba iz 4.10 je

$$X_{i_{k+1}^{\text{in}}}^T X_{\mathcal{E}_k} \mathbf{b}_1 + X_{i_{k+1}^{\text{in}}}^T X_{i_{k+1}^{\text{in}}} b_2 = X_{i_{k+1}^{\text{in}}}^T \mathbf{y} - \lambda_{k+1} z. \tag{4.13}$$

Uvrštavanjem 4.12 u 4.13 dobivamo sustav

$$X^T_{i^{\text{in}}_{k+1}}(I - X_{\mathcal{E}_k}X^+_{\mathcal{E}_k})X_{i^{\text{in}}_{k+1}}b_2 = X^T_{i^{\text{in}}_{k+1}}\left[(I - X_{\mathcal{E}_k}X^+_{\mathcal{E}_k})\mathbf{y} + \lambda_{k+1}X_{\mathcal{E}_k}(X^+_{\mathcal{E}_k})^T\mathbf{s}_k\right] - \lambda_{k+1}z \tag{4.14}$$

$$= X^T_{i^{\text{in}}_{k+1}}\left[\mathbf{y} - X_{\mathcal{E}_k}X^+_{\mathcal{E}_k}\mathbf{y} + \lambda_{k+1}X_{\mathcal{E}_k}(X^+_{\mathcal{E}_k})^T\mathbf{s}_k\right] - \lambda_{k+1}z \tag{4.15}$$

$$= X^T_{i^{\text{in}}_{k+1}}\left[\mathbf{y} - X_{\mathcal{E}_k}\beta^L_{\mathcal{E}_k}(\lambda_{k+1})\right] - \lambda_{k+1}z \tag{4.16}$$

$$= 0, \tag{4.17}$$

gdje zadnja jednakost dolazi iz definicije vremena ulaska. Neovisno o skalaru $X^T_{i^{\text{in}}_{k+1}}(I - P)X_{i^{\text{in}}_{k+1}}$ (koji može biti 0 ako je matrica $X_{\mathcal{E}_k}$ singularna), $b_2 = 0$ je rješenje gornje jednakosti sa najmanjom ℓ^2 normom. □

Gornja lema pokazuje da su limesi slijeva i zdesna LARS rješenja $\lambda \mapsto \beta^L(\lambda)$ jednaki, tj. LARS putanja je neprekidna.

Teorem 4.4.4. *LARS rješenje je lasso rješenje.*

Dokaz. Dokaz provodimo indukcijom. U konstrukciji LARS putanje smo komentirali da je $\beta = \mathbf{0}$ zaista jedno lasso rješenje na intervalu $[\lambda_1, \infty\rangle$, pri čemu je $\lambda_1 = \left\|X^T\mathbf{y}\right\|_\infty$.

Pretpostavimo da smo dokazali da je LARS rješenje ujedno i lasso rješenje na $\langle\lambda_k, \infty\rangle$, dokažimo da je i lasso rješenje na $\langle\lambda_{k+1}, \infty\rangle$. Znamo da je nužan i dovoljan uvjet da bi $\beta \in \mathbb{R}^p$ bilo lasso rješenje u $\lambda > 0$ dan sa

$$\beta_i = 0 \implies |X_i^T(\mathbf{y} - X\beta)| \leq \lambda \tag{4.18}$$

$$\beta_i \neq 0 \implies X_i^T(\mathbf{y} - X\beta) = \lambda\text{sign}(\beta_i). \tag{4.19}$$

Pokažimo da LARS zadovoljava gornje uvjete u λ_k. Prvo pretpostavimo da je λ_k vrijeme izlaska. Pretpostavimo da $i_1 \in \mathcal{E}_{k-1}$ napušta ekvikorelacijski skup, tj $\mathcal{E}_k = \mathcal{E}_{k-1} \setminus \{i_1\}$. Po definiciji vremena izlaska, $\beta^L_{i_1}(\lambda_k) = 0$, a kako je zbog uvjeta (4.19) vrijedilo

$$|X_{i_1}^T(\mathbf{y} - X\beta^L(\lambda))| = \lambda$$

za $\lambda \in \langle\lambda_k, \infty\rangle$, po neprekidnosti LARS putanje (koju implicira prethodna lema) to vrijedi i za $\lambda = \lambda_k$. Za sve ostale indekse, (4.18) i (4.19) će nastaviti vrijediti po neprekidnosti (jer se predznaci tih indeksa nisu mijenjali).

Ako je λ_k vrijeme ulaska, te $\mathcal{E}_k = \mathcal{E}_{k-1} \cup i_1$. Po neprekidnosti je $(\beta^L(\lambda_k))_{i_1} = 0$ jer je $(\beta^L(\lambda))_{i_1} = 0$ na $\langle\lambda_k, \lambda_{k-1}\rangle$, a kako je uvjet za i_1 (4.18) vrijedio na $\langle\lambda_k, \lambda_{k-1}\rangle$, vrijedi i u λ_k. Iz istog razloga vrijede oba uvjeta za ostale indekse (jer se predznaci tih indeksa nisu mijenjali). Dakle LARS rješenje je lasso rješenje u λ_k.

Neka je sada $\lambda \in \langle\lambda_{k+1}, \lambda_k\rangle$. Po konstrukciji LARS rješenja, $\mathcal{E}_k$ smo birali tako da vrijedi

$$\left\|X^T_{-\mathcal{E}_k}(\mathbf{y} - X\beta^L(\lambda))\right\|_\infty < \lambda, \quad \lambda > \lambda^{\text{in}}_{k+1} \tag{4.20}$$

te

$$X_{\mathcal{E}_k}^T(\mathbf{y} - X\beta^L(\lambda_k)) = \lambda_k \mathbf{s}_k, \quad \mathbf{s}_k = \text{sign}(X_{\mathcal{E}_k}^T(\mathbf{y} - X\beta^L(\lambda_k))). \tag{4.21}$$

Nejednakost (4.20) osigurava da uvjet (4.18) vrijedi na $\langle \lambda_{k+1}, \lambda_k \rangle$. Nadalje vrijedi

$$\begin{aligned} X_{\mathcal{E}_k}^T(\mathbf{y} - X\beta^L(\lambda)) &= X_{\mathcal{E}_k}^T\mathbf{y} - X_{\mathcal{E}_k}^T X_{\mathcal{E}_k} X_{\mathcal{E}_k}^+ \mathbf{y} + X_{\mathcal{E}_k}^T X_{\mathcal{E}_k} X_{\mathcal{E}_k}^+ (X_{\mathcal{E}_k}^T)^+ \lambda \mathbf{s}_k \\ &= X_{\mathcal{E}_k}^T (X_{\mathcal{E}_k}^T)^+ \lambda \mathbf{s}_k \\ &= \lambda \mathbf{s}_k \end{aligned}$$

gdje zadnja jednakost slijedi iz činjenice da je $\mathbf{s}_k \in \text{Im}(X_{\mathcal{E}_k}^T)$. Još treba pokazati da je

$$\mathbf{s}_k = \text{sign}(\beta_{\mathcal{E}_k}^L(\lambda)). \tag{4.22}$$

Budući da smo pokazali da je $\beta^L(\lambda_k)$ lasso rješenje, ako je λ_k vrijeme izlaska, onda su za sve $i \in \mathcal{E}_k$ vrijedi $\beta_i^L(\lambda_k) \neq 0$, pa po (4.19) mora biti

$$\mathbf{s}_k = \text{sign}(\beta_{\mathcal{E}_k}^L(\lambda_k)).$$

Budući da je LARS rješenje neprekidno, predznaci svih indeksa iz $\mathcal{E}_k$ ostaju nepromjenjeni dok varijable ne izađu iz skupa ekvikorelacije. Dakle u ovom slučaju vrijedi (4.22).

Pretpostavimo sada da je λ_k vrijeme ulaska, te da je i_k ušao u skup ekvikorelacije. Za sve varijable u $\mathcal{E}_k$ različite od i_k, istom argumentacijom kao u prošlom slučaju možemo zaključiti da zadovoljavaju (4.22) pa je još preostalo dokazati da to vrijedi i za i_k. Zaista,

$$\beta_{\mathcal{E}_k}(\lambda) = \mathbf{c} - \lambda\mathbf{d} = \mathbf{c} - \lambda_k\mathbf{d} + (\lambda_k - \lambda)\mathbf{d} = \beta(\lambda_k) + (\lambda_k - \lambda)(X_{\mathcal{E}_k}^T X_{\mathcal{E}_k})^+ \mathbf{s}_k.$$

Kako je $[\beta(\lambda_k)]_{i_k} = 0$ jer je λ_k vrijeme ulaska od i_k te je $\lambda_k - \lambda > 0$, vrijedi

$$\text{sign}([\beta_{\mathcal{E}_k}(\lambda)]_{i_k}) = \text{sign}([\mathbf{s}_k]_{i_k}),$$

što završava dokaz. □

4.5 Svojstva LARS rješenja

Sada kada smo našli jedno lasso rješenje, znamo i pripadni skup ekvikorelacije za svaki $\lambda > 0$. Koristeći teorem (4.2.1) možemo konstruirati bilo koje lasso rješenje krećući se po $\text{Ker}(X_{\mathcal{E}})$ pazeći jedino da je zadovoljen uvjet (4.3). Primijetimo da smo LARS rješenje dobili tako da smo u tom teoremu birali $\mathbf{b} = \mathbf{0}$. Prisjetimo se da u općem slučaju $\text{Ker} \neq \{\mathbf{0}\}$ nismo znali da $\mathbf{b} = \mathbf{0}$ daje jedno rješenje (nije bilo jasno zadovoljava li (4.3)). LARS rješenje je ipak posebno.

Propozicija 4.5.1. *LARS rješenje je lasso rješenje sa najmanjom ℓ^2 normom.*

Dokaz. Iz teorema (4.2.1) slijedi da je svako lasso rješenje oblika

$$\beta_{-\mathcal{E}} = 0, \quad \beta_{\mathcal{E}} = X_{\mathcal{E}}^{+}(\mathbf{y} - (X_{\mathcal{E}}^{T})^{+}\lambda \mathbf{s}) + \mathbf{b} = \beta_{\mathcal{E}}^{L}(\lambda) + \mathbf{b}, \quad \mathbf{b} \in \mathrm{Ker}(X_{\mathcal{E}}).$$

Iz identiteta (9) u napomeni (1.2.5), slijedi da je

$$\beta_{\mathcal{E}}^{L}(\lambda) \in \mathrm{Im}(X_{\mathcal{E}}^{T}),$$

a kako je $\mathrm{Im}(X_{\mathcal{E}}^{T})$ ortogonalni komplement od $\mathrm{Ker}(X_{\mathcal{E}})$, slijedi

$$\|\beta_{\mathcal{E}}\|_2^2 = \left\|\beta_{\mathcal{E}}^{L}(\lambda)\right\|_2^2 + \|\mathbf{b}\|_2^2 \geq \left\|\beta_{\mathcal{E}}^{L}(\lambda)\right\|_2^2.$$

□

Budući da lema (4.4.1) implicira da algoritam završi u konačno mnogo koraka, algoritam vraća cijelu LARS putanju na $\langle 0, \infty \rangle$. Zbog njene neprekidnosti, ima smisla promatrati $\lim_{\lambda \to 0^+} \beta^L(\lambda)$.

Propozicija 4.5.2. *Za proizvoljne $X \in \mathbb{R}^{n \times p}$ i $\mathbf{y} \in \mathbb{R}^n$*

$$\lim_{\lambda \to 0^+} \beta^L(\lambda)$$

je rješenje problema najmanjih kvadrata

$$\|\mathbf{y} - X\beta\|_2^2 \to \min \tag{4.23}$$

sa najmanjom ℓ^1 normom.

Dokaz. Proširimo LARS rješenje u 0 po neprekidnosti

$$\beta^L(0) := \lim_{\lambda \to 0^+} \beta^L(\lambda).$$

Budući da je LARS rješenje i lasso rješenje, ono zadovoljava uvjete (4.18) i (4.19) na $\lambda \in \langle 0, \infty \rangle$. Posebno vrijedi

$$\left\|X^T(\mathbf{y} - X\beta^L(\lambda))\right\|_\infty \leq \lambda, \quad \lambda \in \langle 0, \infty \rangle.$$

Uzimanjem limesa $\lambda \to 0^+$, vrijedi

$$X^T(\mathbf{y} - X\beta^L(0)) = 0.$$

Odnosno, $\beta^L(0)$ je rješenje sustava normalnih jednadžbi pa je i rješenje od (4.23). Pretpostavimo da postoji rješenje $\beta_0 \in \mathbb{R}^p$ od (4.23) manje ℓ^1 norme

$$\|\beta_0\|_1 < \left\|\beta^L(0)\right\|_1.$$

Po neprekidnosti LARS putanje, postoji $\lambda > 0$ takav da je

$$\|\beta_0\|_1 < \left\|\beta^L(\lambda)\right\|_1.$$

Tada vrijedi

$$\frac{1}{2}\|\mathbf{y} - X\beta_0\|_2^2 + \lambda\|\beta_0\|_1 < \frac{1}{2}\left\|\mathbf{y} - X\beta^L(\lambda)\right\|_2^2 + \lambda\left\|\beta^L(\lambda)\right\|,$$

što je kontradikcija sa činjenicom da je $\beta^L(\lambda)$ lasso rješenje u λ. □

Primijetimo da proizvoljna matrica $A \in R^{n\times p}$ ima nul stupac ako i samo A^+ ima nul redak. Zaista, neka je i-ti stupac od A nul stupac, i neka je e_i i-ti vektor iz kanonske baze. Ideja je gledati SVD dekompoziciju

$$0 = A\mathbf{e}_i = USV^T\mathbf{e}_i \implies 0 = \mathbf{e}_i^+ VS^+U^T = \mathbf{e}_i^T A^+.$$

Kako je $(A^+)^+ = A$, vrijedi i druga implikacija.

Teorem 4.5.3. *Za fiksne $X \in \mathbb{R}^{n\times p}$ i $\lambda > 0$ i gotovo sve $\mathbf{y} \in \mathbb{R}^n$, nosač LARS rješenja $\beta^L(\lambda)$ je skup ekvikorelacije.*

Dokaz. Budući da je $\lambda > 0$, primijetimo da $X_\mathcal{E}$ ne može imati nul stupac jer u suprotnom indeks tog stupca ne bi zadovoljavao definicioni uvjet skupa ekvikorelacije. Prema gornjem razmatranju, to znači da $X_\mathcal{E}^+$ ne može imati nul redak. Promotrimu sada skup

$$\mathcal{N} = \bigcup_{\mathcal{E}} \bigcup_{\mathbf{s}} \bigcup_{i\in\mathcal{E}} \left\{\mathbf{z} \in \mathbb{R}^n \ : \ \mathbf{e}_i^T (X_\mathcal{E}^+)(\mathbf{z} - (X_\mathcal{E}^T)^+\lambda\mathbf{s}) = 0\right\}$$

pri čemu prva unija ide po svim $\mathcal{E} \subseteq \{1, \ldots, p\}$ koji ne sadržavaju indeks nul stupca od X i druga koja ide po svim predznacima $\mathbf{s} \in \{-1, 1\}$. Međutim, $\mathcal{N}$ je onda konačna unija hiperravnina, pa je po 4.3.2 Lebesgueove mjere nula. Primijetimo da je $\mathcal{N}$ skup svih $\mathbf{y} \in \mathbb{R}^n$ koji daju LARS rješenje čiji se nosač razlikuje od skupa ekvikorelacije. □

4.6 Neophodne varijable

Ranije smo pokazali da za dane $X, \mathbf{y}$ i λ ne mogu postojati dva lasso rješenja koja na nekoj komponenti imaju različite predznake. Ipak, može se dogoditi da dva lasso rješenja imaju različite nosače

$$\text{supp}(\beta) = \{i \in \{1, \ldots, p\} \ : \ \beta_i \neq 0\}.$$

Za konstrukciju takvog primjera vidi točku 4.2 u [9].

Po teoremu (4.2.1), svako je lasso rješenje dano sa

$$\beta_{-\mathcal{E}} = \mathbf{0}, \quad \beta_\mathcal{E} = \beta_\mathcal{E}^L(\lambda) + \mathbf{b}, \quad \mathbf{b} \in \text{Ker}(X_\mathcal{E}), \quad s_i(\beta_i^L(\lambda) + b_i) \geq 0, \quad i \in \mathcal{E}. \tag{4.24}$$

Iz teorema (4.2.1) je jasno da je nosač svakog rješenja nužno podskup od $\mathcal{E}$. Također, iz gornjeg razmatranja je jasno da ako je $\beta \in \mathbb{R}^p$ lasso rješenje sa nosačem $\mathcal{A}$ mora biti

$$b_i = -\beta^L(\lambda), \quad i \notin \mathcal{E} \setminus \mathcal{A}$$
$$b_i \neq -\beta^L(\lambda), \quad i \in \mathcal{E} \setminus \mathcal{A}.$$

Postavlja se pitanje, postoji li za svaku varijablu nosač koji ju sadrži i nosač koji ju ne sadrži?

Lema 4.6.1. *Za $X \in \mathbb{R}^{n\times p}, \boldsymbol{y} \in \mathbb{R}^n$ i $\lambda > 0$, neka su $\mathcal{E}$ pripadni ekvikorelacijski skup i pridruženi vektor predznaka $\boldsymbol{s}$. Ako označimo $S = \operatorname{diag}(\boldsymbol{s})$, tada je skup svih rješenja dan sa*

$$\beta_{-\mathcal{E}} = \boldsymbol{0}, \quad \beta_{\mathcal{E}} \in K = \left\{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{E}|} \ : \ (X_{\mathcal{E}}^+ X_{\mathcal{E}})\boldsymbol{x} = \beta_{\mathcal{E}}^L(\lambda), \quad S\boldsymbol{x} \geq 0\right\}.$$

Dokaz. Iz napomene (1.2.5) i teorema (4.2.1) slijedi da je LARS rješenje u $\operatorname{Im}(X_{\mathcal{E}}^T)$. Nadalje, (X^+X) je ortogonalni projektor na $\operatorname{Im}(X_{\mathcal{E}}^T)$, pa je

$$(X_{\mathcal{E}}^+ X_{\mathcal{E}})\mathbf{x} = \beta_{\mathcal{E}}^L(\lambda) \iff \mathbf{x} = \beta_{\mathcal{E}}^L(\lambda) + \mathbf{b}, \quad \mathbf{b} \in (\operatorname{Im}(X_{\mathcal{E}}))^\perp = \operatorname{Ker}(X_{\mathcal{E}})$$

pa je prvi uvjet iz (4.24) zadovoljen dok je drugi uvjet očito ekvivalentan $S\mathbf{x} \geq 0$. □

Skup K iz prethodne leme je poliedarski skup (skup dan sa konačno mnogo linearnih jednadžbi ili nejednažbi). Također, budući da svako lasso rješenje ima istu ℓ^1 normu, on je ograničen pa je politop. Korištenjem linearnog programiranja, za $i \in \mathcal{E}$ mogu se riješiti problemi maksimiziranja i minimiziranja

$$\beta_i^{\max} = \max_{\mathbf{x}\in\mathbb{R}^{|\mathcal{E}|}} x_i, \quad \text{pod uvjetom} \quad (X_{\mathcal{E}}^+ X_{\mathcal{E}})\mathbf{x} = \beta_{\mathcal{E}}^L(\lambda), \quad S\mathbf{x} \geq 0, \tag{4.25}$$

$$\beta_i^{\min} = \min_{\mathbf{x}\in\mathbb{R}^{|\mathcal{E}|}} x_i, \quad \text{pod uvjetom} \quad (X_{\mathcal{E}}^+ X_{\mathcal{E}})\mathbf{x} = \beta_{\mathcal{E}}^L(\lambda), \quad S\mathbf{x} \geq 0. \tag{4.26}$$

Svaka je komponenta $i \in \mathcal{E}$ tada u $\left[\beta_i^{\min}, \beta_i^{\max}\right]$. Pokazali smo da interval $\left[\beta_i^{\min}, \beta_i^{\max}\right]$ ne može sadržavati 0 u interioru jer bi u suprotnom postojala lasso rješenja sa različitim predznacima na varijabli i. Označimo zajedničku ℓ^1 normu svih lasso rješenja sa $L \geq 0$. Tada je

$$\left[\beta_i^{\min}, \beta_i^{\max}\right] \subseteq [0, L]$$

ako je $s_i > 0$, odnosno

$$\left[\beta_i^{\min}, \beta_i^{\max}\right] \subseteq [-L, 0]$$

ako je $s_i < 0$.

Ako je $\beta_i^{\min} > 0$ ili $\beta_i^{\max} < 0$, tada ne postoji lasso rješenje koje ne sadrži i u nosaču. Takvu varijablu nazivamo neophodnom.

4.7 Lasso rješenje kao funkcija vektora odziva

U točkama 4.4 i 4.5, za fiksne $\mathbf{X} \in \mathbb{R}^{n\times p}$ i $\mathbf{y} \in \mathbb{R}^n$ smo promatrali kako se ponaša lasso rješenje $\lambda \mapsto \beta(\lambda)$. U ovom poglavlju držimo $X \in \mathbb{R}^{n\times p}$ i $\lambda > 0$ fiksnima i promatramo kako se ponaša $\mathbf{y} \mapsto \beta(\mathbf{y})$. Za više detalja vidi [9]

Neka su dani $\lambda > 0$, $X \in \mathbb{R}^{n\times p}$ te $\mathbf{y} \in \mathbb{R}^n$. Promotrimo skup

$$C = \left\{\mathbf{u} \in \mathbb{R}^n \ : \ \left\|X^T\mathbf{u}\right\|_\infty \leq \lambda\right\}.$$

Lako se vidi da je C zatvoren i konveksan skup, stoga postoji projekcija od $\mathbf{y}$ na C koju označavamo sa $\hat{\mathbf{y}}$. Neka je $\beta(\lambda, X, \mathbf{y}) = \beta$ jedno lasso rješenje te neka je $\mathbf{c} \in [-1, 1]^p$ takav da je

$$X^T(\mathbf{y} - X\beta) = \lambda\mathbf{c}, \quad \beta_i \neq 0 \implies c_i = \operatorname{sign}(\beta_i).$$

Tvrdimo da je $\hat{\mathbf{y}} = \mathbf{y} - X\beta$. Prema teoremu 2.2.4, to je ekvivalentno

$$(X\beta)^T(\mathbf{u} + X\beta - \mathbf{y}) \leq 0, \quad \forall \mathbf{u} \in C. \tag{4.27}$$

Zaista, vrijedi

$$(X\beta)^T(\mathbf{u} + X\beta - \mathbf{y}) = (X^T\mathbf{u})^T\beta - X^T(\mathbf{y} - X\beta)\beta = (X^T\mathbf{u})^T\beta - \lambda\mathbf{c}^T\beta. \tag{4.28}$$

Budući da je $\mathbf{c}$ vektor predznaka od β na njegovom nosaču, vrijedi

$$\mathbf{c}^T\beta = \max_{\mathbf{s}\in[-1,1]^p} \mathbf{s}^T\beta = \max_{\|\mathbf{s}\|_\infty \leq 1} \mathbf{s}^T\beta$$

pa je

$$\lambda\mathbf{c}^T\beta = \max_{\|s\|_\infty \leq \lambda} \mathbf{s}^T\beta.$$

Sada iz (4.28) slijedi da mora vrijediti (4.27).

Nadalje za $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ vrijedi

$$\begin{aligned}
\|\hat{\mathbf{y}} - \hat{\mathbf{z}}\|_2^2 &= (\hat{\mathbf{y}} - \hat{\mathbf{z}})^T(\hat{\mathbf{y}} - \hat{\mathbf{z}}) \\
&= (\mathbf{z} - \hat{\mathbf{z}} + \hat{\mathbf{y}} - \mathbf{y} + \mathbf{y} - \mathbf{z})^T(\hat{\mathbf{y}} - \hat{\mathbf{z}}) \\
&= (\mathbf{z} - \hat{\mathbf{z}})^T(\hat{\mathbf{y}} - \hat{\mathbf{z}}) + (\mathbf{y} - \hat{\mathbf{y}})^T(\hat{\mathbf{z}} - \hat{\mathbf{y}}) + (\mathbf{y} - \mathbf{z})^T(\hat{\mathbf{y}} - \hat{\mathbf{z}}) \\
&\leq (\mathbf{y} - \mathbf{z})^T(\hat{\mathbf{y}} - \hat{\mathbf{z}}) \\
&\leq \|\mathbf{y} - \mathbf{z}\|_2\|\hat{\mathbf{y}} - \hat{\mathbf{z}}\|_2
\end{aligned}$$

pri čemu smo u prijelazu iz trećeg u četvrti red koristili 2.2.4. Dakle, vrijedi

$$\|\hat{\mathbf{y}} - \hat{\mathbf{z}}\|_2 \leq \|\mathbf{y} - \mathbf{z}\|_2$$

pa je $\mathbf{y} \mapsto \hat{\mathbf{y}}$ Lipschitz neprekidna funkcija, a time i neprekidna funkcija. Kako je

$$X\beta = \mathbf{y} - \hat{\mathbf{y}},$$

vrijedi da je i $\mathbf{y} \mapsto (X\beta)(\mathbf{y})$ neprekidna funkcija.

Fiksirajmo $\lambda > 0$ i $X \in \mathbb{R}^{n\times p}$. Za $\mathbf{y} \in \mathbb{R}^n$ označimo sa $\mathcal{E}(\mathbf{y})$ skup ekvikorelacije lasso rješenja u $(\lambda, X, \mathbf{y})$ te sa $\mathbf{s}(\mathbf{y})$ pripadni skup predznaka.

Teorem 4.7.1. *Za gotovo sve $\boldsymbol{y} \in \mathbb{R}^n$ postoji okolina $U \subseteq \mathbb{R}^n$ od $\boldsymbol{y}$ takva da je*

$$\mathcal{E}(\boldsymbol{y}) = \mathcal{E}(\boldsymbol{y}'), \quad \boldsymbol{s}(\boldsymbol{y}) = \boldsymbol{s}(\boldsymbol{y}'), \quad \forall \boldsymbol{y}' \in U.$$

Dokaz. U teoremu 4.5.3 smo pokazali da za gotovo sve $\mathbf{y} \in \mathbb{R}^n$ je pripadni skup ekvikorelacije ujedno i nosač LARS rješenja u $\mathbf{y}$. Dokazat ćemo tvrdnju za sve takve $\mathbf{y} \in \mathbb{R}^n$. Označimo $\mathcal{E} = \mathcal{E}(\mathbf{y})$, $\mathbf{s}(\mathbf{y}) = \mathbf{s}$ i za $\mathbf{y}' \in \mathbb{R}^n$ stavimo

$$\beta_{-\mathcal{E}}(\mathbf{y}') = \mathbf{0}, \quad \beta_{\mathcal{E}}(\mathbf{y}) = X_{\mathcal{E}}^{+}(\mathbf{y}' - (X_{\mathcal{E}}^{T})^{+}\lambda\mathbf{s}).$$

Želimo pokazati da je $\beta(\mathbf{y}')$ lasso rješenje sa ekvikorelacijskim skupom $\mathcal{E}$ i vektorom predznaka $\mathbf{s}$. Primijetimo da vrijedi

$$X_{\mathcal{E}}^{T}(\mathbf{y}' - X_{\mathcal{E}}\beta_{\mathcal{E}}(\mathbf{y}')) = X_{\mathcal{E}}^{T}(\mathbf{y}' - X_{\mathcal{E}}X_{\mathcal{E}}^{+}(\mathbf{y}' - (X_{\mathcal{E}}^{T})^{+}\lambda\mathbf{s})) = X_{\mathcal{E}}^{+}X_{\mathcal{E}}\lambda\mathbf{s} = \lambda\mathbf{s}, \tag{4.29}$$

pri čemu zadnja jednakost vrijedi jer $\lambda\mathbf{s} \in \mathrm{Im}(X_{\mathcal{E}}^{T})$. Nadalje, kako je

$$\mathbf{z} \mapsto f(\mathbf{z}) := X_{-\mathcal{E}}^{T}(\mathbf{z} - X_{\mathcal{E}}X_{\mathcal{E}}^{+}(\mathbf{z} - (X_{\mathcal{E}}^{T})^{+}\lambda\mathbf{s})) = X_{-\mathcal{E}}^{T}(\mathbf{z} - X_{\mathcal{E}}\beta_{\mathcal{E}}(\mathbf{z}))$$

afina funkcija, posebno je neprekidna. Budući da je $\mathcal{E}$ nosač LARS rješenja u $\mathbf{y}$, mora biti $\|f(\mathbf{y})\|_\infty < \lambda$, pa po neprekidnosti od f, postoji okolina $U_1 \subseteq \mathbb{R}^n$ oko $\mathbf{y}$ takva da je

$$\|f(\mathbf{y}')\|_\infty < \lambda, \quad \forall \mathbf{y}' \in U_1. \tag{4.30}$$

Primijetimo da je $\mathbf{z} \mapsto \beta_{\mathcal{E}}(\mathbf{z}) = X_{\mathcal{E}}^{+}(\mathbf{z} - (X_{\mathcal{E}}^{T})^{+}\lambda\mathbf{s})$ afina funkcija pa je i neprekidna. Dakle, postoji okolina U_2 od $\mathbf{y}$ takva da je

$$\mathrm{sign}(\beta_{\mathcal{E}}(\mathbf{y}')) = \mathbf{s}, \quad \forall \mathbf{y}' \in U_2. \tag{4.31}$$

Ako stavimo $U = U_1 \cap U_2$, vidimo da (4.29), (4.30) i (4.31) osiguravaju da je $\beta(\mathbf{y}')$ lasso rješenje na U sa skupom ekvikorelacije $\mathcal{E}$ i vektorom predznaka $\mathbf{s}$, odnosno $\beta(\mathbf{y}') = \beta^{L}(\mathbf{y}')$. □

Prethodni teorem povlači da za gotovo sve $\mathbf{y} \in \mathbb{R}^n$ postoji okolina oko $\mathbf{y}$ na kojoj je funkcija LARS rješenja $\mathbf{y}' \mapsto \beta^{L}(\mathbf{y}')$ afina.

Može se pokazati i jača tvrdnja za općenito lasso rješenje.

Teorem 4.7.2. *Postoji skup $\mathcal{M} \subseteq \mathbb{R}^n$ Lebesgueove mjere nula sa sljedećim svojstvom: za $\boldsymbol{y} \notin \mathcal{M}$ te nekim lasso rješenjem $\beta(\boldsymbol{y})$ sa nosačem $\mathcal{A}(\boldsymbol{y})$ i skupom predznaka $\boldsymbol{r}(\boldsymbol{y})$, postoji okolina U od $\boldsymbol{y}$ takva da za svaki $\boldsymbol{y}' \in U$ postoji lasso rješenje $\beta(\boldsymbol{y}')$ sa istim nosačem $\mathcal{A}(\boldsymbol{y})$ i skupom predznaka $\boldsymbol{r}(\boldsymbol{y})$.*

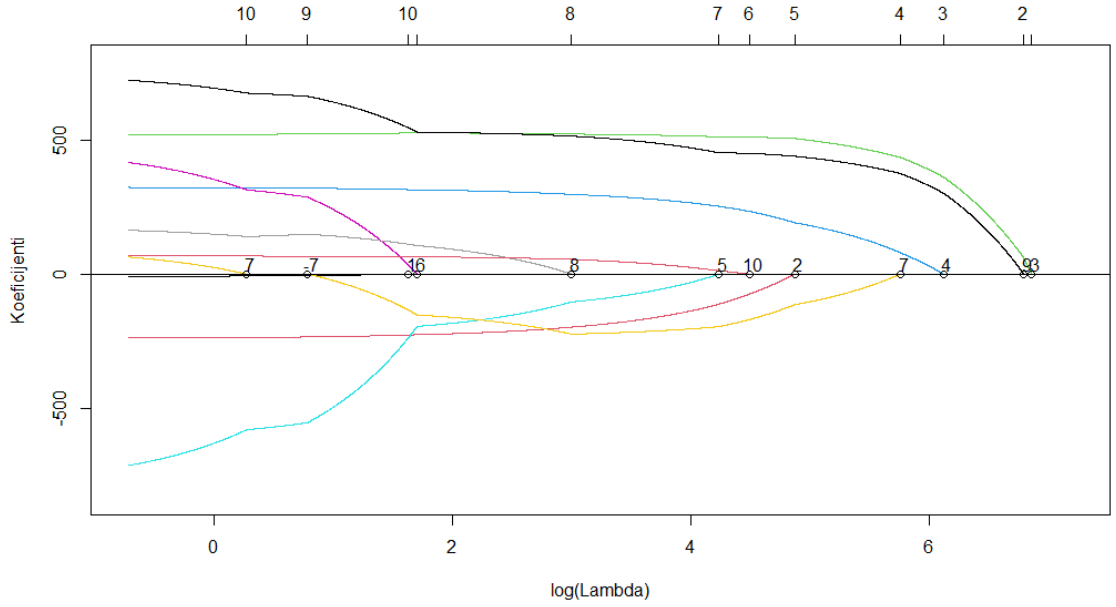
Za dokaz vidi [9].

4.8 Primjer - Dijabetes

U ovoj točki demonstriramo lasso metodu na skupu podataka o dijabetesu [3]. Uzorak je veličine $n = 442$, a broj varijabli $p = 10$. Matrica dizajna i vektor odziva su centrirani

$$\sum_{i=1}^{n} X_{ij} = 0, \quad \sum_{i=1}^{n} y_i = 0$$

te je dodatno norma svakog stupca matrice X jednaka 1.
Matrica X je punog ranga 10 pa je za svaki $\lambda > 0$ pripadno lasso rješenje jedinstveno i jednako LARS rješenju. Na slici 4.1 prikazan je graf koeficijenata LARS rješenja u ovisnosti o $\log(\lambda)$. Na x-osi u čvorovima su prikazani indeksi koji u danom čvoru ulaze u skup ekvikorelacije, odnosno iz njega izlaze. Iznad grafa prikazan je kardinalitet skupa ekvikorelacije. Ukupno je 12 čvorova od kojih je 11 vrijeme ulaska, i jedno vrijeme izlaska.



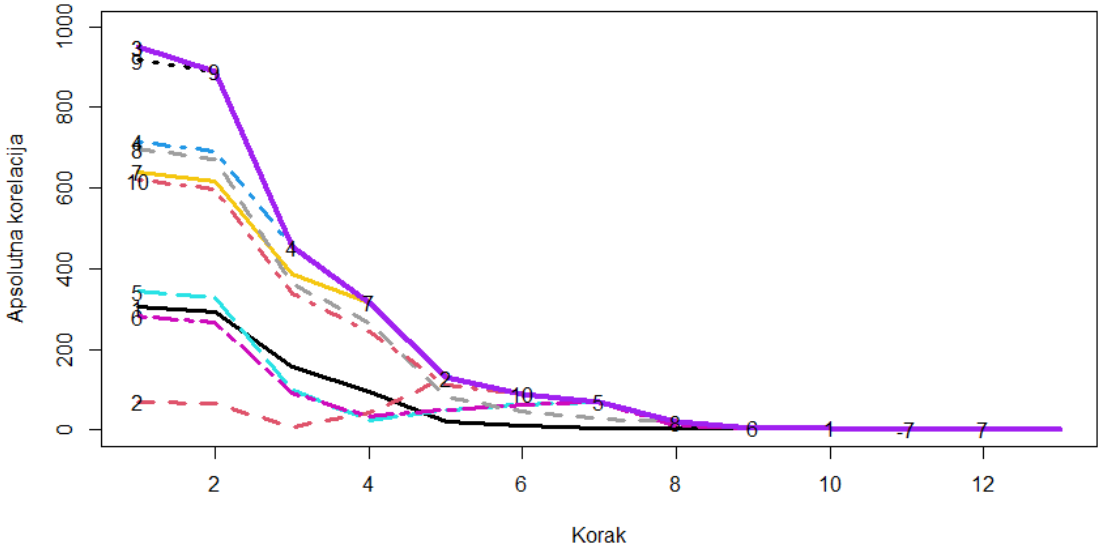
Slika 4.1: lasso putanja za skup podataka o dijabetesu

Na slici 4.2 prikazan je graf apsolutnih korelacija

$$X^T(\mathbf{y} - X\beta^L(\lambda_k))$$

po koracima $k = 1, \ldots, 1$. Debelom ljubičastom bojom prikazana je trenutna maksimalna korelacija za svaki korak. Tu korelaciju dostižu sve varijable koje su u skupu ekvikorelacije. Iz grafa se jasno vidi kako apsolutna korelacija opada. To je jasno iz uvjeta

$$X^T(\mathbf{y} - X\beta^L(\lambda)) = \lambda \mathbf{s}(\lambda)$$

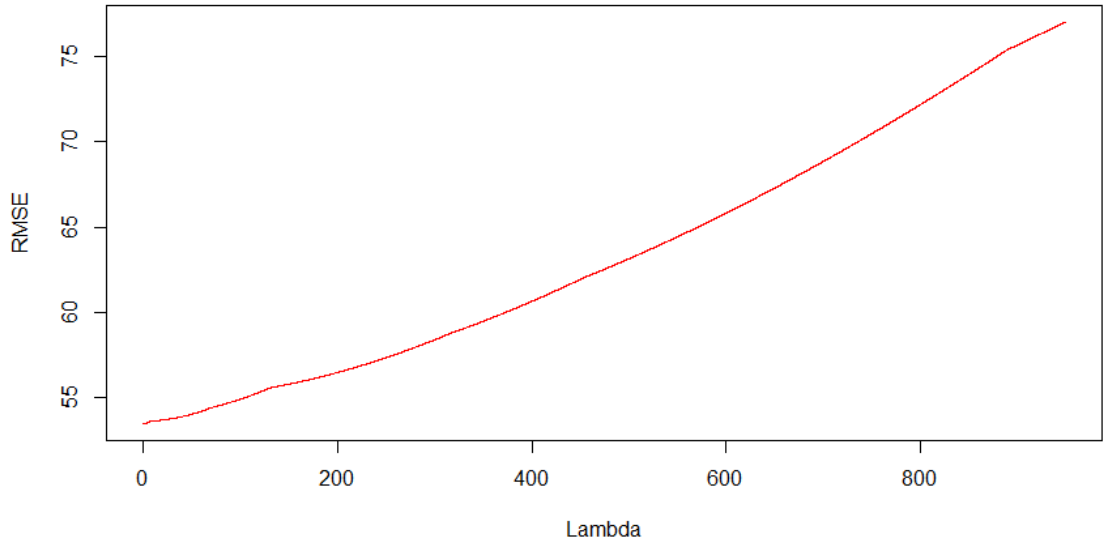


Slika 4.2: Graf apsolutnih korelacija po koracima.

Slika 4.3 prikazuje graf ovisnosti uzoračke korjenovane srednje kvadratne pogreške

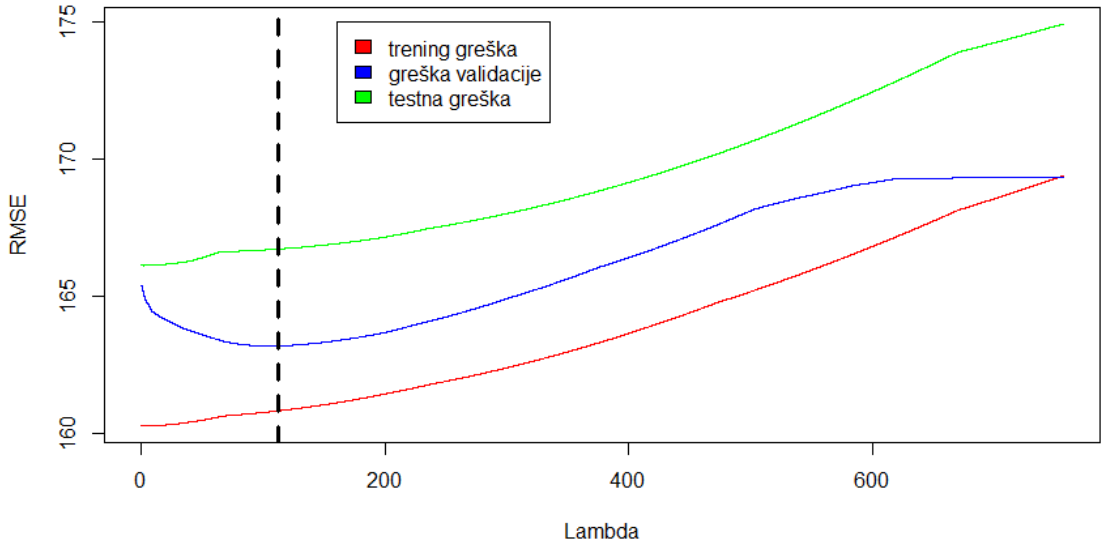
$$\text{RMSE} = \frac{1}{\sqrt{n}}\left\|\mathbf{y} - X\beta^L(\beta)\right\|_2$$

o λ. Iz grafa se jasno vidi kako RMSE monotono pada što više smanjujemo λ. Prisjetimo se, iz propozicije 4.5.2 znamo da je $\lim_{\lambda\to 0^+}\beta^L(\lambda)$ upravo procjenitelj najmanjih kvadrata, pa se RMSE za taj procjenitelj na grafu nalazi na $\lambda = 0$.

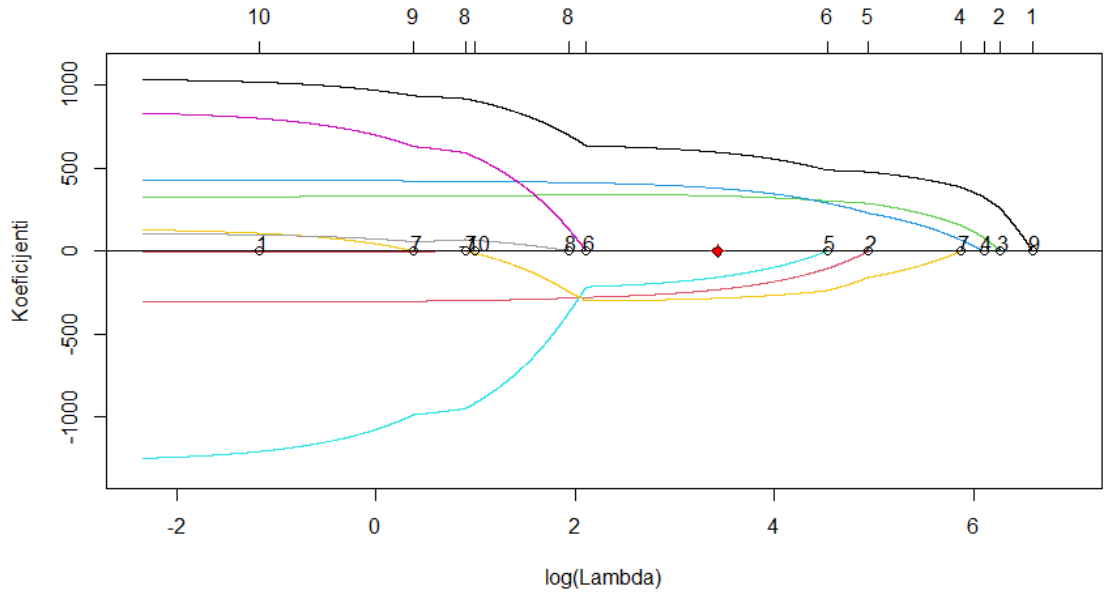


Slika 4.3: RMSE na cijelom skupu podataka.

Sada na istom skupu podataka pokušavamo procijeniti populacijski RMSE i odabrati najbolji $\lambda > 0$. Skup podataka nasumično podijelimo na podatke za trening (80%) i podatke za testiranje (20%). Zatim izračunamo grešku validacije pomoću k-fold unakrsne validacije za $k = 5$. Za najbolji λ biramo onaj koji ima najmanju uzoračku grešku validacije. Na kraju procjenimo testnu grešku validacije pomoću skupa za testiranje. Na slici 4.4 je prikazana ovisnost uzoračkih greški u ovisnosti o λ. Iscrtkana linija je na na mjestu odabranog λ. Na slici 4.5 prikazana je LARS putanja na skupu za trening. Sa crvenim kružićem označen je λ sa najmanjom greškom na skupu za validaciju. Pripadni LARS procjenitelj za taj λ ima 6 varijabli u nosaču i to $\{2, 3, 4, 5, 7, 9\}$.

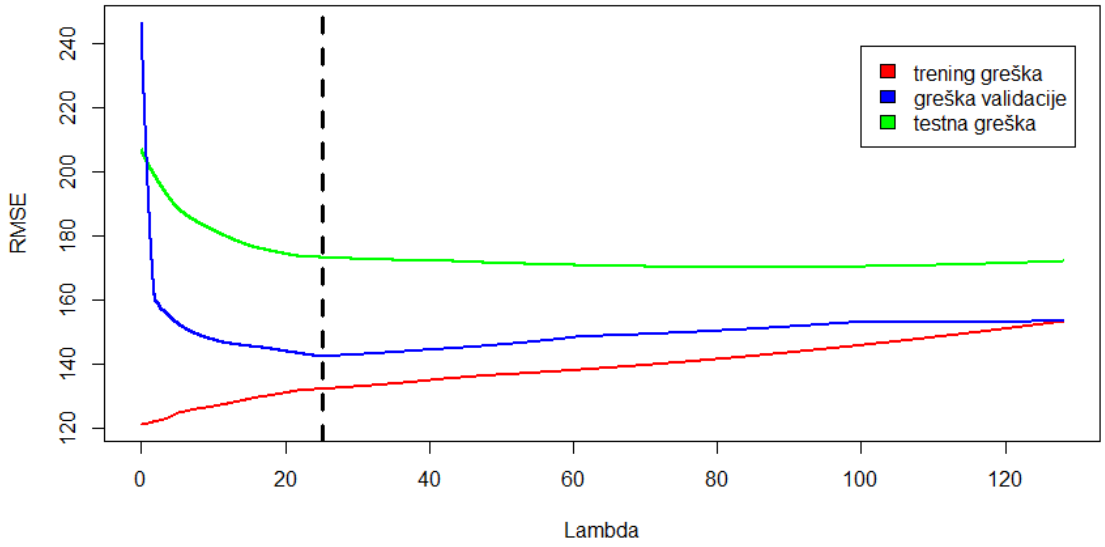


Slika 4.4: Trening greška, greška validacije i testna greška. Iscrtkana linija ima je na mjestu odabranog λ sa najmanjom greškom validacije.

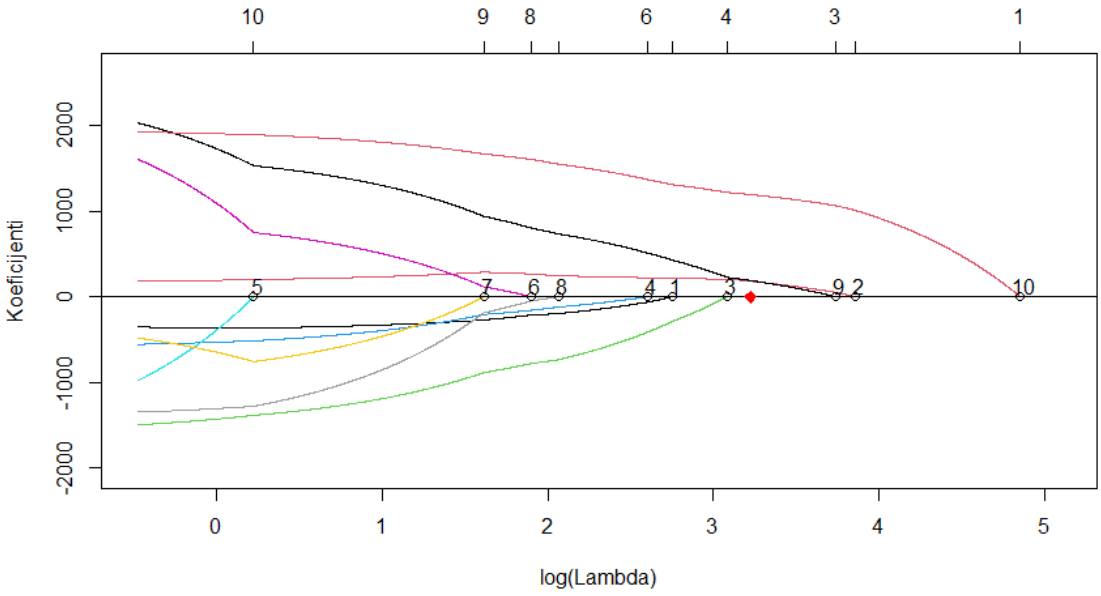


Slika 4.5: LARS putanja na skupu za trening. Na x-osi uz čvorove stoje indeksi koji izlaze ili ulaze u skup ekvikorelacije. Crvenim kružićem je označen λ odabran unakrsnom validacijom.

Lasso postiže još bolje rezultate od procjenitelja najmanjih kvadrata kada je skup podataka manji. Zaista, uzmimo sada da je skup za trening samo veličine 36, a ostatak uzorka koristimo za procjenu testne greške. Ovaj puta koristimo *leave-one-out* unakrsnu validaciju, odnosno k-fold unakrsnu validaciju za $k = 36$. Na slici 4.6 prikazane su pripadne greške. Ovaj puta procjenitelj najmanjih kvadrata ima drastično veću grešku validacije od najboljeg lasso rješenja.
Na slici 4.7 prikazana je LARS putanja na skupu za trening. Sa crvenim kružićem označen je λ sa najmanjom greškom na skupu za validaciju. Pripadni LARS procjenitelj za taj λ ima 3 varijable u nosaču i to $\{2, 9, 10\}$. Primijetimo da 10 nije bio u nosaču modela kojeg je izabrao veći skup za trening sa slike 4.5.



Slika 4.6: Trening greška, greška validacije i testna greška. Iscrtkana linija ima je na mjestu odabranog λ sa najmanjom greškom validacije.



Slika 4.7: LARS putanja na skupu za trening. Na x-osi uz čvorove stoje indeksi koji izlaze ili ulaze u skup ekvikorelacije. Crvenim kružićem je označen λ odabran unakrsnom validacijom.

Bibliografija

[1] João Carlos Alves Barata i Mahir Saleh Hussein, *The Moore–Penrose Pseudoinverse: A Tutorial Review of the Theory*, Brazilian Journal of Physics **42** (2011), br. 1-2, 146–165, https://doi.org/10.1007%2Fs13538-011-0052-z.

[2] D. Bertsekas, *Convex Optimization Theory*, Athena Scientific optimization and computation series, Athena Scientific, 2009, ISBN 9781886529311.

[3] Bradley Efron, Trevor Hastie, Iain Johnstone i Robert Tibshirani, *Least angle regression*, The Annals of Statistics **32** (2004), br. 2, https://doi.org/10.1214%2F009053604000000067.

[4] S. Kumar, *Topics in signal processing*, 2021-2022, https://tisp.indigits.com, posjećena 28.08.2023.

[5] Julien Mairal i Bin Yu, *Complexity Analysis of the Lasso Regularization Path*, 2012.

[6] M. Pilanci S. Boyd, J. Duchi i L. Vandenberghe, *Subgradients, notes for EE364b, Stanford University, Spring 2021-22*, 2022, https://web.stanford.edu/class/ee364b/lectures/subgradients_notes.pdf, posjećena 28.08.2023.

[7] R. J. Tibshirani, *The Lasso Problem and Uniqueness*, 2012, https://arxiv.org/pdf/1206.0313.pdf.

[8] R. J. Tibshirani i L. Wasserman, *Sparsity, the Lasso, and Friends - Statistical Machine Learning, Spring 2017*, 2017, https://www.stat.cmu.edu/~ryantibs/statml/lectures/sparsity.pdf, posjećena 28.08.2023.

[9] Ryan J. Tibshirani i Jonathan Taylor, *Degrees of freedom in lasso problems*, The Annals of Statistics **40** (2012), br. 2, https://doi.org/10.1214%2F12-aos1003.

Sažetak

Ovaj se rad bavi analizom i konstrukcijom lasso metode. Prvi dio rada promatra procjenitelj najmanjih kvadrata i probleme koji se javljaju kada matrica dizajna nije punog ranga, te uvodi koncept Moore-Penrose inverza.

U drugom dijelu rada, daje se pregled pojmova rezultata konveksne analize te dokazi teorema o projekciji te teorema o separaciji. Na kraju poglavlja pokazuje se egzistencija lasso rješenja koristeći rezultate vezane za recesivne smjerove i nivo skupove.

Treći dio rada fokusira se na subgradijente kao analogon gradijentima za nediferencijabilne i konveksne funkcije. Subgradijenti su koristan alat za traženje minimuma takvih funkcija. Ovdje se izvodi nužan i dovoljan uvjet koji lasso rješenja moraju zadovoljavati.

Četvrti dio istražuje svojstva lasso rješenja. Dokazuje se jedinstvenost uklopljenih vrijednosti i jednakost ℓ^1 normi svih lasso rješenja. Uvode se ključni pojmovi poput skupa ekvikorelacije i vektora predznaka te se dokazuje njihova jedinstvenost za sva lasso rješenja. Dalje se pokazuje da, za razliku od procjenitelja najmanjih kvadrata, ne mogu postojati dva lasso rješenja sa različitim predznacima na zajedničkom nosaču. Također se utvrđuje da je lasso rješenje jedinstveno ako matrica dizajna dolazi iz neke apsolutno neprekidne razdiobe.

Nadalje, rad opisuje LARS algoritam koji konstruira po dijelovima linearnu i neprekidnu lasso putanju za fiksnu matricu dizajna i vektor odziva te pokazuje korisna svojstva takve putanje.

Također, u ovom poglavlju se daje algoritam za određivanje gornjih i donjih granica koeficijenata po svim lasso rješenjima te se uvodi pojam neophodnih varijabli koje su u nosaču svakog lasso rjenja.

Konačno, posljednja točka rada promatra lasso rješenje kao funkciju vektora odziva te dokazuje da je uklopljena vrijednost neprekidna u tom kontekstu. Također se pokazuje da je LARS lasso rješenje afina funkcija u okolini gotovo svih vektora odziva.

Na kraju je dan primjer primjene i analize LARS algoritma na skupu podataka o dijabetesu.

Summary

This paper focuses on the analysis and construction of the lasso method. The first part of the paper examines the least squares estimator and the issues arising when the design matrix does not have full rank. Also it introduces the concept of the Moore-Penrose inverse that is used throughout the paper.

The second part of the paper provides an overview of results from convex analysis along with proofs of the projection theorem and the hyperplane separation theorems. At the end of this section, the existence of lasso solutions is proving by using the results related to recession directions and level sets.

The third part of the paper concentrates on subgradients as analogs of gradients for nondifferentiable and convex functions. Subgradients are a valuable tool for finding the minimum of such functions. At the end of this chapter, the necessary and sufficient condition that lasso solutions must satisfy are derived.

The fourth part explores the properties of lasso solutions. It proves uniqueness of fitted values, along with the equivalence of the ℓ^1 norms of all lasso solutions. Key concepts such as the equicorrelation set and the sign vector are introduced, and their uniqueness over all Lasso solutions is established. Furthermore, it is shown that, unlike least squares estimators, two Lasso solutions with different signs on the common support cannot exist. It is also established that the Lasso solution is unique if the design matrix comes from an absolutely continuous distribution.

Moreover, the paper describes the LARS algorithm, which constructs a piecewise linear and continuous Lasso path for a fixed design matrix and response vector and demonstrates useful properties of such solution.

Furthermore, an algorithm for determining upper and lower bounds of coefficients across all Lasso solutions is provided. It also defines the concept of indispensable variables.

Finally, the last part of the paper observes the Lasso solution as a function of the response vector and proves the continuity of the fitted value in this context. It is also shown that the LARS Lasso solution is an affine function in the neighbourhood of almost every response vector.

At the end, there is an example of applying LARS algorithm on diabetes dataset.

Životopis

Rođen sam 20.01.2000. u Osijeku. U istome gradu pohađao sam Osnovnu školu Franje Krežme te nakon toga II. gimnaziju koju sam završio 2018. godine. Iste godine upisujem prediplomski studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Odmah nakon završenog prediplomskog studija 2021. godine upisujem diplomski studij Matematičke statistike na istom fakultetu.