סיכום באופטימיזציה

תזכורת ממדעי הנתונים:

הגדרה: מטריצה תקרא PD (positive semi-definite) אם :

$\forall x \neq 0 \quad x^T A x > 0$.

הגדרה: מטריצה A תקרא סימטרית אם $A = A^T$.

הגדרה: מטריצה A תקרא SPD (סימטרית חיובית) אם
- $A = A^T$.
- $\forall x \neq 0 \quad x^T A x > 0$.

הגדרה: הtranspose של מטריצה A יסומן A^T ויוגדר להיות: $(A^T)_{ij} = (A)_{ji}$.

הגדרה: סקלר λ ווקטור $v \neq 0$ ייקראו ערכים עצמיים ווקטורים עצמיים בהתאמה אם:

$$Av = \lambda v$$

הגדרה: מטריצה A דומיננטית באלכסון אם $|a_{ii}| > \Sigma_{i \neq j} |a_{ij}|$

משפט: לכל מטריצה A , אם λ הוא ערך עצמי של A אז $1 - \lambda$ הוא ערך עצמי של מטריצה $I - A$.

משפט: A היא מטריצה *full Rank* אם ורק אם $A^T A$ הפיך.

משפט: A היא מטריצה *full Rank* אם ורק אם $A^T A$ היא *SPD*.

משפט: אם $\lambda > 0$ אז המטריצה $A^T A + \lambda I$ היא חיובית ממש.

<u>קיום פתרון למשוואה $Ax = b$:</u>

1. פתרון יחיד- A הפיכה והמטריצה לא סינגולרית והיא *full-Rank* .
2. אין פתרון- בהכרח קיימת תלות ליניארית בעמודות A (A לא הפיכה).
3. אינסוף פתרונות- המטריצה A סינגולרית כי קיים וקטור e כלשהו שמקיים $Ae = 0$.

נורמה ווקטורית

נורמה מוגדרת על ווקטור $v \in R^n$ אם היא מקיימת את האקסיומות הבאות:

1. אי שליליות- $\forall v \in R^n , ||v|| \geq 0 \; and \; ||v|| = 0 \leftrightarrow v = 0$.
2. הומוגניות: $\forall v \in R^n , \alpha \in R , ||\alpha v|| = |\alpha| ||v||$.
3. אי שוויון המשולש- $\forall u, v \in R^n \quad ||u + v|| \leq ||v|| + ||u||$.

הנורמות הנפוצות:

$$\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|, \quad \|\mathbf{v}\|_2 = \left(\sum_{i=1}^{n} v_i^2 \right)^{\frac{1}{2}}, \quad \|\mathbf{v}\|_\infty = \max_i |v_i|$$

מכפלה פנימית

מכפלה פנימית בין שני ווקטורים מוגדרת ע"י שלושת האקסיומות הבאות:

1. סימטריות $< u, v > \; = \; \overline{< v, u >}$.
2. ליניאריות $< u, v_1 + v_2 >=< u, v_1 > + < u, v_2 > , < \alpha u, \beta v >= \bar{\alpha} \beta < u, v >$.
3. אי שליליות $< u, u > \geq 0 \; and < u, u >= 0 \; iff \; u = 0$.

המ"פ הנפוצה:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_i \bar{u}_i v_i = \mathbf{u}^* \mathbf{v}.$$

נורמה של מטריצה:

$$\|A\|_F = \left(\sum_i \sum_j |a_{ij}|^2 \right)^{\frac{1}{2}}.$$

או

$$\|A\|_a \triangleq \max_{\mathbf{x} \in \mathbb{C}^n} \frac{\|A\mathbf{x}\|_a}{\|\mathbf{x}\|_a}$$

משפט:

לכל נורמה מושרית, $||Ax||_a \leq ||A||_a ||x||_a$.

משפט:
לכל נורמה מושרית, $||AB||_a \leq ||A||_a ||B||_a$.

הגדרה: רדיוס הספקטרלי מוגדר להיות הערך העצמי המקסימלי של מטריצה A בערך מוחלט.
$p(A) = \max|\lambda_i|$.

משפט: בהינתן נורמה מושרית, היא חוסמת מלמעלה את הרדיוס הספקטרלי.
$||A|| \geq p(A)$

נושא: *Condition number*

$$\boldsymbol{cond(A) = \left|\left|A^{-1}\right|\right| \, ||A||}$$

$$cond(A) \geq \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \geq 1$$

הגדרה: ווקטור השגיאה הוא $e = x - x^* = A^{-1}r$.

הגדרה: ווקטור השארית הוא $r = Ax - b$. באופן דומה נקבל כי : $r = Ae$.

תזכורת על *Least squares*

עלינו לפתור את המשוואה $x^* = \ argmin_x \left|\left|Ax - b\right|\right|_2^2$

•הפונקציה הזו אי שלילית ולכן חסומה מלמטה ע"י 0 ולכן בהכרח יש פתרון למשוואה.

•על מנת לפתור את המשוואה נגזור, נשווה ל-0 ונקבל:

$$A^T Ax = A^T b$$

•נוכל להפוך של $A^T A$ כיוון שהיא לא סינגולרית.

נגזרות נפוצות:

1. $f(x) = x^T x, \ \ \nabla f = 2x$.
2. $f(x) = v^T x \ , \nabla f = v$.
3. $f(x) = v^T A^T Ax, \nabla f = A^T Av$.
4. $f(x) = x^T A^T Ax, \nabla f = 2A^T Ax$.

:Weighted least squares

מוטיבציה: לפעמים נרצה לתת דגש לחלק מהמשוואות.

פונקציית המשקל : $||r||_W^2 = r^T W r$

נפתור את המשוואה הבאה: $x^* = (A^T W A)^{-1} A^T W b$.

:Regularized least squares

מוטיבציה: לפעמים נצטרך מידע נוסף כדי "לסדר את הפתרון", עבור מטריצות סינגולריות ייתכן כי המחשב לא יצליח לפתור בגלל חלוקה ב-0.

עבור $\lambda > 0$ נפתור את המשוואה הבאה: $x^* = (A^T A + \lambda I)^{-1} A^T b$.

•המטריצה $A^T A + \lambda I$ תמיד הפיכה.

נושא: פתרון משוואות ליניאריות בצורה ישירה

שיטה ראשונה היא השיטה שלמדנו עד כה באלגברה של דירוג המטריצה עד הגעה למדורגת קנונית.

פירוק LU:

נפרק את A למכפלה של מטריצות נוחות יותר לשימוש $A = LU$.

משפט: לכל מטריצה ריבועית A קיים LU כך ש$PA = LU$ כאשר P זו מטריצת הפרמוטציות, L משולשית תחתונה וU משולשית עליונה.

פירוק צ'ולסקי:

נפרק את A למכפלה של מטריצות נוחות יותר לשימוש $A = LL^T$

משפט: אם A היא SPD אז אלגוריתם צ'ולסקי לא ייכשל (קיים פירוק $A = LL^T$).

אלגוריתם לפירוק צ'ולסקי :

Algorithm: Cholesky
$\# A \in \mathbb{C}^{n\times n}$
Initialize: $L = 0^{n\times n}$.
for $i = 1, \ldots, n$ **do**
  $l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}\bar{l}_{ki}\right)^{\frac{1}{2}}$.
  for $j = i+1, \ldots, n$ **do**
    $l_{ij} = \frac{1}{l_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}\bar{l}_{kj}\right)$,
  end
end

הגדרה: u, v אורתוגונליים אם $< u, v >= 0$ מעל מרחב מ"פ כלשהו.

הגדרה: מטריצה אורתוגונלית אם $Q^T Q = I$.

טענה: $set\{v_i\}$ אורתוגונליים בגודל n מהווה בסיס למרחב R^n.

תהליך גרהם שמידט:

מטרה: לקחת קבוצה של ווקטורים ולהפוך אותם לבסיס אורתוגונלי למרחב.

- Step 1: $\mathbf{q}_1 = \mathbf{a}_1$.
- Step 2: $\mathbf{q}_2 = \mathbf{a}_2 - \frac{\langle \mathbf{a}_2, \mathbf{q}_1 \rangle}{\|\mathbf{q}_1\|^2}\mathbf{q}_1$.
- Step 3: $\mathbf{q}_3 = \mathbf{a}_3 - \frac{\langle \mathbf{a}_3, \mathbf{q}_1 \rangle}{\|\mathbf{q}_1\|^2}\mathbf{q}_1 - \frac{\langle \mathbf{a}_3, \mathbf{q}_2 \rangle}{\|\mathbf{q}_2\|^2}\mathbf{q}_2$.

- Step i: $\mathbf{q}_i = \mathbf{a}_i - \sum_{j=1,..,i-1} \frac{\langle \mathbf{a}_i, \mathbf{q}_j \rangle}{\langle \mathbf{q}_j, \mathbf{q}_j \rangle}\mathbf{q}_j$.

פירוק QR

$A = QR$ עבור Q מטריצה אורתוגונלית ו-R מטריצה משולשית עליונה.

R משולשית ולכן בהכרח הפיכה.

פתרון LS ע"י פירוק זה יהיה על ידי המשוואה : $x^* = R^{-1}Q^T b$.

פירוק SVD

משפט: לכל מטריצה A קיימות מטריצות U, V, Σ כך ש- Σ אלכסונית, V,U מטריצות אורתוגונליות ומתקיים $A = U\Sigma V^T$.

•ערכי האלכסון של Σ הם שורשי הע"ע של $A^T A$.

•Σ^2 היא מטריצה SPD .

פתרון LS ע"י פירוק זה יהיה על ידי המשוואה : $\Sigma V^T x^* = U^T b$.

אם A היא *full-Rank* נקבל $x^* = A^t b \, , A^t = V\Sigma^{-1}U^T$.
אם A אינה *full-Rank* נקבל $x^* = A^t b, A^t = V(\Sigma)^t U^T$ כך :

$$(\Sigma^\dagger)_{ii} = \begin{cases} (\Sigma_{ii})^{-1} & \Sigma_{ii} \neq 0 \\ 0 & \Sigma_{ii} = 0 \end{cases}$$

נושא: פתרון משוואות ליניאריות בצורה איטרטיבית

מוטיבציה: לא תמיד רוצים פתרון מדויק אלא מקורב שיגיע מהר.

שיטה איטרטיבית מוגדרת כך (חד נקודתית או רב נקודתית בהתאם לכמה פתרונות היא "מתחשבת בהם")

$$x^{(k+1)} = \phi\left(x^{(k)}\right) \; or \; x^{(k+1)} = \phi\left(x^{(k)}, \dots, x^{(0)}\right)$$

$$\lim_{k\to\infty}\left\{x^{(k)}\right\} = x^*$$

הגדרה: קצב התכנסות מוגדר להיות

$$\lim_{k\to\infty}\frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|^p} = C$$

•תנאי הכרחי להתכנסות הוא $|C| < 1$.

הגדרה: ווקטור השגיאה מוגדר $e^{(k)} = x^* - x^{(k)}$.

הגדרה: ווקטור השארית מוגדר להיות $r^{(k)} = b - Ax^{(k)}$.

$$\lim_{k\to\infty}\{\mathbf{e}^{(k)}\} = \lim_{k\to\infty}\{\mathbf{r}^{(k)}\} = 0.$$

•תנאי עצירה אפשריים:

$$\frac{\|A\mathbf{x}^{(k)} - \mathbf{b}\|}{\|\mathbf{b}\|} < \epsilon \quad \text{or} \quad \frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|} < \epsilon.$$

שיטות "פשוטות"

מבצעים פירוק של מטריצה A $A = M + N$.
מוטיבציה: ככל שמטריצה M מקורבת יותר ל-A נקבל את הפתרון בפחות איטרציות.
מטרה: לבחור מטריצה M נוחה להיפוך כדי להקל על החישובים.
בחירת M מתבצעת כך ש-M הפיכה.
*הרדיוס הספקטרלי הוא קבוע ההתכנסות עבור שיטות אלו.

תנאי הכרחי להתכנסות: $p(I - M^{-1}A) < 1$

$$\mathbf{x}^{(k+1)} = M^{-1}(\mathbf{b} - N\mathbf{x}^{(k)}) = \mathbf{x}^{(k)} + M^{-1}(\mathbf{b} - A\mathbf{x}^{(k)}),$$

שיטת יעקובי:

M נבחרת להיות האלכסון של A .

$$A = D + L + U\ , M = D, N = L + U$$

$$\mathbf{x}^{(k+1)} = D^{-1}(\mathbf{b} - (L+U)\mathbf{x}^{(k)}) = \mathbf{x}^{(k)} + D^{-1}(\mathbf{b} - A\mathbf{x}^{(k)}).$$

שיטת גאוס-זיידל:

בשיטה זו M היא משולשית תחתונה .

$$A = D + L + U, M = L + D, N = U$$

$$(L+D)\mathbf{x} = \mathbf{b} - U\mathbf{x} \Rightarrow (L+D)\mathbf{x}^{(k+1)} = \mathbf{b} - U\mathbf{x}^{(k)},$$

משפט: אם A מטריצה דומיננטית באלכסון אזי שיטת יעקובי וגאוס-זיידל מתכנסות.

משפט: אם A היא SPD אז גאוס-זיידל מתכנס.

שיטות יעקובי וגאוס-זיידל ממשוקלות:

ניתן להוסיף משקול לצעד, עבור $0 < w \le 1$ הצעד יהיה :

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega D^{-1}(\mathbf{b} - A\mathbf{x}^{(k)}),$$

קצב התכנסות:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha(\mathbf{b} - A\mathbf{x}^{(k)}),$$

במקרה זה מתקיים כי $C = p(I - \alpha A)$

טענה: $p(I - \alpha A) = \max\{|1 - \alpha\lambda_{max}|, |1\alpha\ \lambda_{min}|\}$.

נקבל כי $\alpha_{opt} = \frac{2}{\lambda_{max}+\lambda_{min}}$.

α_{opt} לא ריאלי לחישוב ולכן נחשב לכל איטרציה α בצורה חמדנית.

$$\alpha^{(k)} = \frac{< r^{(k)}, r^{(k)} >}{< r^{(k)}, Ar^{(k)} >}$$

הפונקציה שאותה ממזערים:

$$f(x) = \frac{1}{2}||x - x^*||_A^2 = \frac{1}{2}x^T AX - x^T b + \frac{1}{2}(x^*)^T b$$

טענה: אם A מטריצה חיובית מוגדרת אז גאוס זיידל מתכנס והפונקציה $f(x^{(k)})$ מונוטונית עולה.

שיטת Steepest Descent

• רלוונטית עבור A שהיא SPD .

הצעד באלגוריתם מוגדר להיות:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)} + \alpha(\mathbf{b} - A\mathbf{x}^{(k)}),$$

עבור $\alpha^{(k)}$ שחושבה למעלה .

במקרה זה השגיאה המתקבלת היא :

$$\mathbf{e}^{(k+1)} = \mathbf{e}^{(k)} - \frac{\langle \mathbf{r}^{(k)}, A\mathbf{e}^{k} \rangle}{\langle \mathbf{r}^{(k)}, A\mathbf{r}^{(k)} \rangle} \mathbf{r}^{(k)}.$$

ומתקיים :

$$\langle \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \rangle = 0,$$

שיטת Conjugate Gradients (CG)

מטרה: השגיאה בכל צעד היא A-אורתוגונלית לווקטור השארית מהאיטרציה שלפני.

נמצא כיוונים שיהיו בסיס אורתוגונלי מווקטורי השארית בכל איטרציה.

$$\mathbf{p}^{(k)} = \mathbf{r}^{(k)} - \sum_{i=0}^{k-1} \frac{\langle \mathbf{r}^{(k)}, A\mathbf{p}^{(i)} \rangle}{\langle \mathbf{p}^{(i)}, A\mathbf{p}^{(i)} \rangle} \mathbf{p}^{(i)}.$$

הצעד באלגוריתם יהיה :

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{p}^{(k)}, \; \mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha^{(k)} A\mathbf{p}^{(k)}.$$

טענה: אם A היא SPD ו-*full Rank* אזי שיטת CG תתכנס לכל היותר ב-n איטרציות (מימד המטריצה) לכל ניחוש התחלתי שנבחר.

תתי המרחבים של קרילוב

עד כה בכל השיטות התבססנו רק על הפתרון מהאיטרציה הקודמת, הפעם נתבסס על כל הפתרונות כך:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(0)} + \sum_{i=0}^{k} \alpha^{(i)} \mathbf{r}^{(i)}.$$

במקרה זה ווקטורי השגיאה יהיו כך :

$$e^{(k+1)} \in e^{(0)} + span\{r^{(0)}, Ar^{(0)}, \dots, A^k r^{(0)}\}$$

•בהנחה ש-A היא SPD מתקיים שמרחב זה הוא ממימד $k+1$.

נושא: אופטימיזציה ללא אילוצים

נושא : $x^* = argmin_{x \in R^n} f(x)$ עבור פונקציות לא לינאריות, כך ש- $f(x)$ היא פונקציית המחיר אותה אנחנו מעוניינים למזער.

קצת חדוו"א:

כיף מספר 1 -טורי טיילור:

במשתנה 1:

$$f(x+\epsilon) = f(x) + f'(x)\epsilon + \frac{1}{2}f''(x)\epsilon^2 + \frac{1}{3!}f'''(c)\epsilon^3 \quad , c \in [x, x+\epsilon]$$

בשני משתנים:

$$f(x_1+\epsilon_1, x_2+\epsilon_2) = f(x_1, x_2) + \frac{\partial f}{\partial x_1}\epsilon_1 + \frac{\partial f}{\partial x_2}\epsilon_2 + \frac{1}{2}\frac{\partial^2 f}{\partial x_1^2}\epsilon_1^2 + \frac{1}{2}\frac{\partial^2 f}{\partial x_2^2}\epsilon_2 + \frac{\partial f}{\partial x_1 x_2}\epsilon_1\epsilon_2$$

עבור פונקציה ווקטורית:

$$f_i(\mathbf{x}+\varepsilon) \approx f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \varepsilon \rangle.$$

ולכל הווקטור נשתמש בנוסחה

$$\delta \mathbf{f} = \mathbf{f}(\mathbf{x}+\varepsilon) - \mathbf{f}(\mathbf{x}) \approx \mathbf{J}\varepsilon,$$

כיף מספר 2-מטריצת הסיין:

$$\nabla^2 f = H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

מוגדר גם כ- $\frac{1}{2} < \epsilon, H\epsilon >$ ולכן פונקציית טיילור יכולה להיראות כך :

$$f(\mathbf{x}+\varepsilon) = f(\mathbf{x}) + \langle \nabla f, \varepsilon \rangle + \frac{1}{2}\langle \varepsilon, H\varepsilon \rangle + O(\|\varepsilon\|^3),$$

כיף מספר 3- מטריצת יעקוביאן:

$$\mathbf{J} = \begin{bmatrix} - & \nabla f_1(\mathbf{x}) & - \\ - & \nabla f_2(\mathbf{x}) & - \\ & \vdots & \\ - & \nabla f_m(\mathbf{x}) & - \end{bmatrix} \qquad \mathbf{J}_{i,j} = \frac{\partial f_i}{\partial x_j}.$$

קצת תיכון:

הגדרה: נקודה x^* תקרא מינימום לוקאלי (מקומי) של $f(x)$ אם קיימת סביבה $R > 0$ כך ש-$f(x^*) \le f(x)$ $\forall x\ s.t: ||x - x^*|| < R$.

הגדרה: נקודה x^* תקרא מינימום גלובאלי של $f(x)$ אם לכל $x \in R^n$ מתקיים $f(x^*) \le f(x)$.

משפט: נניח ש-$\nabla^2 f$ רציפה בסביבה של x^* ומתקיים $\nabla f(x^*) = 0$ ו-$\nabla^2 f(x^*)$ היא חיובית מוגדרת אזי x^* הוא מינימום לוקאלי ממש של f .

•לפי ערן- בקורס שלנו ניתן להניח שהכל גזיר ורציף ונחמד.

קמירות:

הגדרה: קבוצה קמורה היא $S \in R^n$ אם כל קו שמחבר בין 2 נק' בקבוצה נמצא ב-S .

$$\alpha x + (1-\alpha)y \in S\ , \forall 0 \le \alpha \le 1\ , x, y \in S$$

הגדרה: f תקרא פונקציה קמורה אם התחום שעליה היא מוגדרת קמור ולכל 2 נק' x,y בתחום הגרף של f נמצא מתחת למיתר בין x,y .

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y)\ , \forall 0 \le \alpha \le 1\ , x, y \in S$$

•בקורס שלנו נתעסק רק בR^n שהוא תחום קמור.

הגדרות שקולות אלטרנטיביות:

1.הגדרה לא שימושית: נניח כי f גזירה, f תהיה קמורה מעל תחום קמור S אם
$\forall x_1, x_2 \in S:\ \ f(x_1) \ge f(x_2) + < \nabla f(x_2), x_1 - x_2 >$

2.הגדרה סופר שימושית: נניח כי f גזירה פעמיים. F תהיה קמורה מעל תחום קמור S אם
$\forall x \in S:\ \ H = \nabla^2 f(x)\ \ is\ positive\ semi-definite$

משפט: עבור f פונקציה קמורה, כל x^* מינימום לוקאלי הוא גם מינימום גלובאלי של f .
בנוסף, אם f היא גזירה, אז כל נקודה x^* כך ש-$\nabla f(x^*) = 0$ היא נק' מינימום גלובאלי של f .

•אם f קמורה ממש אז יש נקודה יחידה , אחרת יכולות להיות אינסוף.

נושא: שיטות איטרטיביות לאופטימיזציה

•אנחנו חושבים על בעיות קמורות למרות שהשיטות עובדות (בערך) גם עבור בעיות לא קמורות.

•בכל השיטות נרצה להשתמש בכיוון ירידה d שיקיים: $< \nabla f, d > < 0$.

שיטת *Steepest Descent*

הצעד באלגוריתם הוא :

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \nabla f(\mathbf{x}^{(k)})$$

d מוגדר להיות ווקטור יחידה שמייצג את כיוון הירידה .
α מוגדר להיות גודל הצעד.

$$d_{SD} = -\nabla f(x^{(k)})$$

מדובר למעשה ב"משפחת dים" שכל אחד מהם סבבה מבחינתנו ככיוון ירידה.

עבור $\alpha > 0$ מספיק קטן נקבל $< \nabla f(x^{(k)}), d > < 0 \leftrightarrow f(x^{(k)} + \alpha d) < f(x^{(k)}$

מסקנה: לכל כיוון $d = -M\nabla f$ עבור M שהיא SPD נקבל כיוון ירידה.

פירוש של SD כ-LS :

$$d_{SD}^{(k)} = -\alpha \nabla f(x^{(k)})$$

והצעד יהיה : $x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)})$.

שיטת ניוטון:

מוטיבציה: שיטה זו הרבה יותר חזקה מSD אבל יקרה כי אנחנו מבצעים יותר חישובים.
מתי מומלץ? הפיכת H (אפילו בצורה מקורבת) הוא לא יקר מדיי, ואם בשיטה יש מעט פרמטרים.

$$d_N^{(k)} = -(\nabla^2 f(x^{(k)}))^{-1} \nabla f(x^{(k)})$$

•ברור כי מדובר בכיוון ירידה כי הוא מהצורה $d = -M\nabla f$ עבור $M = H = \nabla^2 f(x^{(k)})$.

שיטת *Quasi Newton*:

מוטיבציה: שיטת ניוטון דרשה המון חישובים, שיטת SD דרשה מעט מדיי חישובים, השיטה הזו נמצאת באמצע שלהן.

בוחרים מטריצה $M^{(k)}$ מקורבת להסיין $(\nabla^2 f(x^{(k)}))$ למשל כמו יעקובי שנבחר את האלכסון

חשוב לבחור מטריצה M שהיא SPD על מנת לשמר את כיוון הירידה.

נושא: חיפוש על ישר

מטרה: למצוא את גודל הצעד הטוב ביותר. נגדיר פונקציה שתלויה בα , ננחש $\alpha^{(0)}$ ונוריד אותו עד שנגיע לירידה "מספיקה".

$$\phi(\alpha) = f(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}), \quad \alpha^{(k)} = \arg\min_{\alpha} \phi(\alpha)$$

$$f(\mathbf{x}^{(k)} + \alpha_j \mathbf{d}^{(k)}) \leq f(\mathbf{x}^{(k)}) + c\alpha_j \langle \nabla f, \mathbf{d}^{(k)} \rangle.$$

עבור $0 < c < 1$.

שיטת Coordinate descent

מטרה: לעבור על כל המשתנים x_i ועבור כל אחד למצוא מינימום סקלארי בהינתן שמקבעים את האחרים (דומה לגזירה ב2 משתנים)

הצעד:

$$x_i^{(k+1)} \leftarrow \arg\min_{x_i} f(x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i, x_{i+1}^{(k)}, \dots, x_n^{(k)})$$

• שיטה זו היא בדיוק גאוס-זיידל לא ליניארי.

שיטת Newton לא מדויק

נניח כי H היא לא SPD אזי יש לה ע"ע שליליים.
"נייצב" את ההסיין ע"י ביצוע הזזה β.
אם H לא הפיך אז מומלץ לבצע את ההזזה הזו.

$$\nabla^2 f(\mathbf{x}^{(k)})\mathbf{d} = -\nabla f(\mathbf{x}^{(k)})$$

$$M = (\nabla^2 f + \beta I)$$

נושא: Data fitting problems

מוטיבציה: אנחנו מעוניינים למצוא מינימום לפונקציה שאינה ליניארית ולכן LS בעייתי, ננסה ליצור מצב בו נקבל בעיית מינימום ריבועי ליניארית בכל איטרציה.

$$\min_{\mathrm{x}} f(x) = \min_x \frac{1}{2}\left|\left|f(x) - y^{obs}\right|\right|_2^2$$

כאשר $f(x)$ היא ווקטור של פונקציות שהן לא בהכרח ליניאריות.

עבור מימד 1:

• הצעד בשיטת SD יהיה :

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla_x f(x^{(k)})$$

• הצעד בשיטת ניוטון יהיה:

$$f\left(x^{(k)} + \epsilon\right) = f\left(x^{(k)}\right) + < \nabla f(x^{(k)}, \epsilon > + \frac{1}{2} < \epsilon, \nabla^2 f\left(x^{(k)}\right)\epsilon > + O\left(\left|\left|\epsilon\right|\right|^3\right)$$

• הצעד בשיטת גאוס-ניוטון (לאחר פיתוח טיילור מסדר ראשון והשארת הריבוע) יהיה:

$$f(x+\epsilon) \approx \frac{1}{2}\left(f(x) + f'(x)\epsilon - y^{obs}\right)^2$$

עבור רב מימד :

•הצעד בשיטת גאוס-ניוטון יהיה:

$$f\left(x^{(k)}+d\right) \approx f\left(x^{(k)}\right)+J\left(x^{(k)}\right)d$$

$$d_{GN}^{(k)} = -(J^TJ)^{-1}\nabla(f(x^{(k)})$$

•אבחנה: J^TJ היא מטריצה *positive semi definite* ולכן ייתכן שיהיה לה ערך עצמי 0 והיא לא תהיה הפיכה, אם היינו מבצעים ניוטון ולא גאוס ניוטון יכלו להיווצר בעיות.

•פתרון משודרג עבור המקרה בו J^TJ סינגולרית :

$$\mathbf{d}_{LM}^{(k)} = \arg\min_{\mathbf{d}} \frac{1}{2}\|\mathbf{f}(\theta^{(k)})+\mathbf{J}(\theta^{(k)})\mathbf{d}-\mathbf{y}^{obs}\|_2^2+\frac{\mu}{2}\|\mathbf{d}\|_2^2,$$

עבור $\mu > 0$.

נושא: *IRLS*

מוטיבציה : אנחנו מעוניינים לבצע *LS* בנורמה שאינה l_2 וזה קשה.

נבצע פתרון איטרטיבי כך שבכל איטרציה נבצע *LS* ממושקל עם l_2. ובכל איטרציה נעדכן את מטריצת המשקל בהתאם כדי למצוא את הפתרון הבא.
הצעד באלגוריתם יהיה :

$$x^{(k+1)} = \mathrm{argmin_x}\left||Ax-b|\right|_{W^{(k)}}^2$$

פתרון בl_1:

המשקל בכל צעד יהיה $w_i^{(k)} = \frac{1}{\left(\left|a_i^Tx^{(k)}-b\right|+\epsilon\right)}$. (ונוסיף 1/2 כמקדם של המשוואה בצעד)

פתרון בl_p:

המשקל בכל צעד יהיה $w_i^{(k)} = \frac{1}{\left|a_i^Tx^{(k)}-b\right|^{p-2}+\epsilon}$. (ונוסיף 1/2 כמקדם של המשוואה בצעד)

IRLS באופן כללי:

נרצה לפתור את המשוואה $r^* = argmin_r\Sigma_i\phi(r_i)$.

נבצע *LS* בכל איטרציה ונתקן את המשקל בהתאם. הצעד יהיה :

$$r^{(k+1)} = argmin_r\left||r|\right|_{W^{(k)}}^2$$

במקרה זה $w_i^{(k)} = \frac{\phi'(r_i)}{2r_i}$.

נושא: אופטימיזציה עם אילוצים

מוטיבציה: אנחנו מעוניינים למצוא x מינימאלי ובו בזמן לקיים מספר אילוצים של שוויונים ואי שוויונים.

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{subject to} \quad \begin{cases} c_j^{\text{eq}}(\mathbf{x}) = 0 & j = 1, \dots, m^{\text{eq}} \\ c_l^{\text{ieq}}(\mathbf{x}) \le 0 & l = 1, \dots, m^{\text{ieq}} \end{cases},$$

כאשר $f(x)$ היא פונקציית המטרה ו- $c(x)$ היא פונקציית האילוץ.

•סימן האי שוויון יכול להיות $\leq$ או $\geq$ זה לא משנה כל עוד דואגים להוסיף מינוס ולהשאיר את משמעות המשוואה כמו שהיא.

•אומגה היא מרחב הפתרונות החוקיים x (שמקיימים את האילוצים) מהם נחפש את הפתרון המינימאלי.

פורמאלי:

$$\Omega = \left\{\mathbf{x} | c_j^{\text{eq}}(\mathbf{x}) = 0, \quad j = 1, \dots, m^{\text{eq}}; \quad c_l^{\text{ieq}}(\mathbf{x}) \le 0, \quad l = 1, \dots, m^{\text{ieq}}\right\}.$$

אילוצים עם שוויונים (כופלי לגראנז'):

קיים $\lambda > 0$ שעבורו $\nabla f(x^*) + \lambda \nabla c_i(x^*) = 0$ אזי x^* הוא מינימום.

קיים $\lambda > 0$ שעבורו $\nabla f(x^*) + \lambda \nabla c_i(x^*) = 0$ אזי x^* הוא מקסימום.

•על מנת ש-d יהיה כיוון ירידה וישמר את האילוצים עליו לקיים:

1. $< \nabla f, d > < 0$.
2. $< \nabla c, d >= 0$.

הגדרה: מטריצת הטלה היא $\left(I - \frac{1}{<\nabla c, \nabla c>} \nabla c \nabla c^T\right) \nabla f$

•אם $\left(I - \frac{1}{<\nabla c, \nabla c>} \nabla c \nabla c^T\right) \nabla f \neq 0$ אזי הנק' x היא לא נקודה חשודה לקיצון.

•מבחינה אלגברית: נרצה ש$\nabla c, \nabla f$ יהיו מקבילים. $\nabla f - \lambda \nabla c = 0$.

הגדרה: פונקציית לגראדיאנט היא $L(x, \lambda) = f(x) + \lambda c(x)$

הגדרה: בהינתן נקודה x^* נאמר ש- $LICQ$ קיים אם אין תלות ליניארית במטריצת הגרדיאנט של האילוצים (יעקוביאן היא *full-Rank*).

משפט: x^* הוא פתרון לוקאלי שמקיים את כל האילוצים ואת *LICQ* אזי קיים וקטור יחיד λ^* שיקרא "כופל לגראנז'" שעבורו מתקיים :

$$\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + (J^{eq})^T \lambda^* = 0$$

משפט: x^* הוא פתרון מינימום ויש λ^* כופל לגראנז' המקיים $\nabla L(x^*, \lambda^*) = 0$ ונניח כי *LICQ* מתקיים אזי H של L היא *SPD*. (בכיוון ש"מותר לנו ללכת בו")
פורמאלית:

$$\mathbf{y}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{y} \ge 0 \quad \forall \quad \mathbf{y} \in \mathbb{R}^n \quad s.t. \quad \mathbf{J}^{\text{eq}} \mathbf{y} = 0.$$

ברב מימד:

$$\min_{\boldsymbol{\lambda}} \|\mathbf{J}^T\boldsymbol{\lambda} + \nabla f\|_2^2 = 0,$$

•אם יהיה את אותו האילוץ מספר פעמיים זה לא ייצור בעיה אבל J לא תהיה הפיכה וזה יכול לפגוע במשפטים שתוארו למעלה.

•אסור שיהיה לנו כיוון ירידה d שמקיים את כל האילוצים.

$$L(x,\lambda) = f(x) + \bar{\lambda}^T \bar{c}(\bar{x})$$

$$\nabla L_x = \nabla_x f + J\bar{\lambda} = 0$$
$$\nabla L_\lambda = \bar{c}(\bar{x}) = 0$$

אילוצים עם אי-שוויונים:

מוטיבציה: מעוניינים למצוא כיוון ירידה d שיקיים את האילוצים.

$$c^{ieq}(x+d) \approx c^{ieq}(x) + < \nabla c^{ieq}(x), d > \le 0$$

הגדרה: סט אקטיבי $A(x)$ בנקודה x שרירותית הוא איחוד אילוצי האי שוויון שמתקיימים בשוויון.

$$\{l \mid c_l^{ieq}(x) = 0\}$$

•אי שוויונים פעילים למעשה מתנהגים כמו שוויונים.

אבחנה: אם מתקיים כי $c^{ieq}(x) < 0$ אז d חופשי לנוע (אפילו אם קצת).

•אם האילוץ $c^{ieq}(x)$ לא פעיל בנקודה x אזי אנחנו יכולים לזוז לכל הכיוונים אך אם הוא כן פעיל אזי מתקיים $c^{ieq}(x) = 0$ ועל מנת לספק את האילוצים עלינו לדרוש $< \nabla c^{ieq}(x), d > \le 0$.
כמו כן, תמיד אנחנו דורשים ש-d יהיה כיוון ירידה ולכן עליו לקיים גם את התנאי $< \nabla f(x), d > < 0$.
•נדרוש גם ש$\nabla c, \nabla f$ יהיו מקבילים. כלומר: $\nabla f + \lambda \nabla c^{ieq}(x^*) = 0$ עבור $\lambda \ge 0$.

פונקציית לגראדיאנט:

$$L(x, \lambda^{eq}, \lambda^{ieq}) = f(x) + (\lambda^{eq})^T c^{eq}(x) + \left(\lambda^{ieq}\right)^T c^{ieq}(x)$$

כאשר λ^{eq} הוא וקטור שמימדו הוא כמספר אילוצי השוויון ו-λ^{ieq} הוא וקטור שמימדו הוא כמספר אילוצי האי-שוויון.

משפט (KKT): נניח ש-x^* הוא פתרון של בעיית אופטימיזציה עם אילוצים, ושתנאי הרגולריות $LICQ$ מתקיימים. אזי קיימים כופלי לגראנז' $\lambda^{eq*}, \lambda^{ieq*}$ כך ש-

$$\begin{aligned}
\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^{\text{eq}*}, \boldsymbol{\lambda}^{\text{ieq}*}) = \nabla f(\mathbf{x}^*) + (\mathbf{J}^{\text{eq}})^\top \boldsymbol{\lambda}^{\text{eq}*} + (\mathbf{J}^{\text{ieq}})^\top \boldsymbol{\lambda}^{\text{ieq}*} &= 0 \\
\mathbf{c}^{\text{eq}}(\mathbf{x}^*) &= 0 \\
\mathbf{c}^{\text{ieq}}(\mathbf{x}^*) &\le 0 \\
\boldsymbol{\lambda}^{\text{ieq}*} &\ge 0 \\
(Complementary\ slackness) \quad for\ \ l = 1, \dots, m^{\text{ieq}} \quad \lambda_l^{\text{ieq}*} c_l^{\text{ieq}}(\mathbf{x}) &= 0
\end{aligned}$$

•עבור אילוץ אקטיבי יתקיים כי $\lambda^{ieq^*} > 0$ ועבור אילוץ שאינו אקטיבי יתקיים כי $\lambda^{ieq^*} = 0$.

משפט: נניח ש-x^* הוא פתרון שמקיים את תנאי KKT ונניח כי קיימים כופלי לגראנז' $\lambda^{eq*}, \lambda^{ieq*}$ ושתנאי $LICQ$ מתקיים אזי צריך להתקיים : (הסיין אי שלילי)

$$\mathbf{y}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^{\text{eq}*}, \boldsymbol{\lambda}^{\text{ieq}*})\mathbf{y} \geq 0 \quad \forall \quad \mathbf{y} \in \mathbb{R}^n \quad s.t. \quad \begin{cases} \mathbf{J}^{\text{eq}}\mathbf{y} = 0 \\ \nabla c_l^{\text{ieq}}(\mathbf{x}^*))^\top \mathbf{y} = 0 \quad l \in \mathcal{A}(\mathbf{x}^*) \end{cases}$$

אבחנה: נקודה x^* במקרה זה היא **חשודה לקיצון** אך תחשב מינימום אך ורק אם הביטוי מלמעלה יהיה חיובי.

שלבי הפתרון של בעיית אופטימיזציה עם אילוצי אי-שוויון:

1. נתעלם לגמרי מהאי שוויונים ונחפש פתרון לבעיה.
2. נציב את הפתרון שמצאנו בכל האילוצים . אם הפתרון מקיים את כל האילוצים נוכל **לסיים כאן.**
3. אחרת, קיימת לפחות משוואה אחת שהפתרון לא מקיים, אם קיימות יותר נבחר אחת באקראיות ונציב אותה בפונקציית לגראדיאנט ונחפש פתרון (נצא מנקודת הנחה שבמבחן הוא ייתן מערכת משוואות פתירה)
4. את הפתרון שמצאנו נציב באילוצים ונוודא שהוא מקיים את כל האילוצים, אם כן נוכל **לסיים כאן.**
5. אחרת, ניקח אילוץ אחר או את כל האילוצים שלא התקיימו ונפעיל שוב את השיטה מ-3 .
6. בעזרת הפתרון שמצאנו נמצא את λ (וקטור כופלי לגראנז') ועבור האילוצים הפעילים אנחנו אמורים לקבל $\lambda > 0$ ועבור הלא פעילים $\lambda = 0$.

אבחנה: שיטת פתרון זו לא כיפית בכלל ! ☹

שיטת Penalty:

מוטיבציה: נהפוך בעיית אופטימיזציה עם אילוצים לבעיית אופטימיזציה ללא אילוצים ונזרוק אותה לאלגוריתם ניוטון, SD או גאוס-ניוטון .

נגדיר ρ סקלאר שיהיה המקדם של האילוצים במשוואה ועבור ρ מספיק גדול או שהוא יקיים $c(x) = 0$ או שהפונקציה תקבל ערך מאוד גבוה. כך או כך , ρ תמיד יגרום לכך שנפר את האילוצים כל עוד הם לא מתקיימים.

$$\min f(x) + \rho\big(c(x)\big)^2$$

$$\min f(x) + \rho_1 \big|\big|c(x)\big|\big|_2^2 + \rho_2 || \max\{0, c^{ieq}(x)\}||_2^2$$

נתחיל מ- ρ_0 בגודל סביר ולאט לאט מגדילים. מומלץ לא להתחיל עם ערך גדול מדיי על מנת לא להקשות על האלגוריתם לפתור את הבעיה . $\rho_0 < \rho_1 < \cdots < \rho_k$.