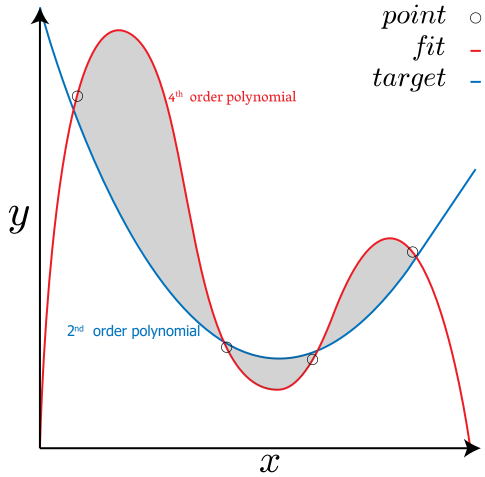
מבוא ללמידה חישובית (20942) | סיכום הרצאה 9

מנחה: ד"ר שי מימון
סמסטר: 2022א'
נכתב על ידי: מתן כהן

1 התאמת יתר - Overfitting

הגדרה: מצב בו מתאימים את המודל לסט הדוגמאות **יותר מהנדרש** - משמע, מצב בו **משפרים** את המודל על ה- training data **ומקבלים ביצועים טובים מאוד** אך בזמן בדיקה על ה- out of sample (דוגמאות שלא נראו עדיין) **המודל מניב ביצועים גרועים**.

דוגמה 1: בעיית רגרסיה ממימד 1 עם 4 נקודות:



נרצה להתאים מודל שיביא למינימום את השגיאה על סט הדוגמאות הנ"ל, ודרך אחת זה להעביר את הסט הפיצ'רים שלנו לממד גדול יותר על ידי הוספת חזקות של x ולאחר מכן לפתור בעיית רגרסיה לינארית עם מספר פיצ'רים גדול יותר.

עם זאת, הנקודות עצמן בדוגמה מעלה נוצרו על ידי פולינום ממעלה שניה בתוספת רעש כלשהו.

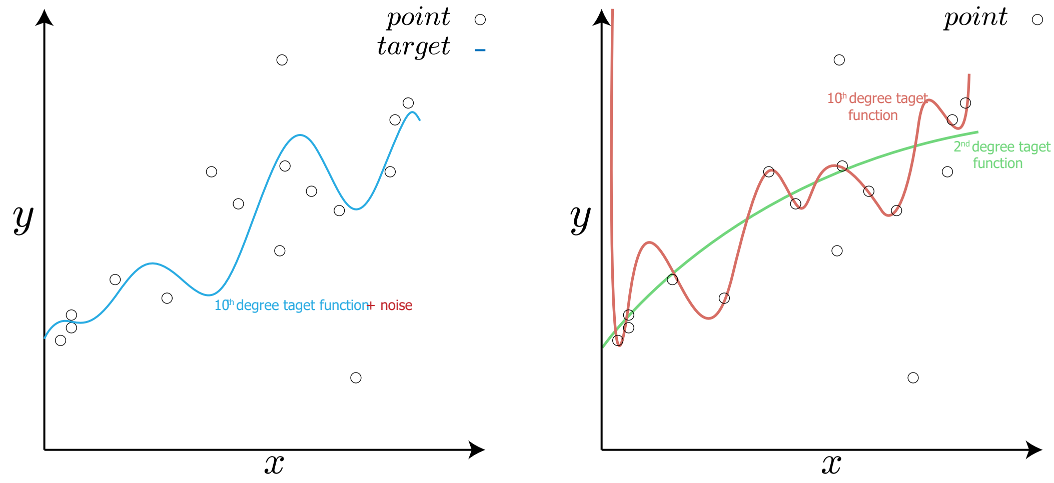
על פי אותו רעש שניסינו לצמצם בפועל מצאנו פולינום ממעלה 4 שבסופו של דבר יגרום ל- overfit, משמע כשנרצה לבדוק עוד נקודות שנוצרו על ידי אותו פולינום נקבל שגיאה עבור נקודות בתחום האפור.

- הרעש הקטן גרם לטעות בלמידה
- אם לא היה רעש - היינו מקבלים פולינום fit שהיה מתאים בדיוק לפולינום ה $target$
- מקרה overfitting קלאסי - מודל שמתמש בדרגות החופש שלו על מנת ללמוד את הרעש שקיים בפונקצית המטרה.

נזכור תמיד כי:

Zero in-sample error but Huge out-of-sample error $\Rightarrow$ Bad generalization

דוגמה 2: פונקצית מטרה של פולינום ממעלה 10 עם 15 נקודות ושני מודלים מאומנים



כאשר:

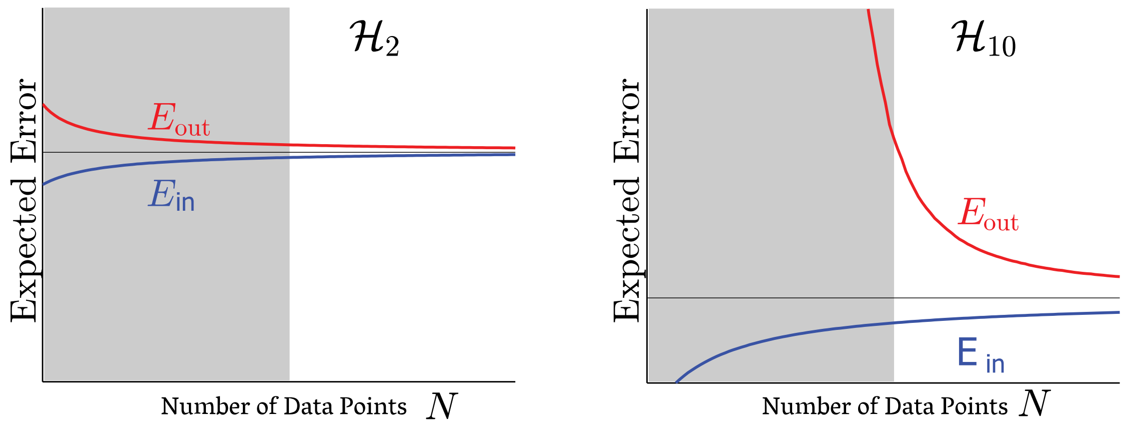
	2nd Order	10th Order
E_{in}	0.050	0.034
E_{out}	0.127	**9.00**

- הפולינום מדרגה 10 בהתאמת יתר גבוהה מאוד
- הפולינום מדרגה 2 תופס את המגמה של הדוגמאות בצורה יותר כללית וטובה

מסקנה: בהינתן שני מודלים O ו- R, אפילו בהנחה שהם **מודעים** לכך שפונקצית המטרה הוא פולינום ממעלה 10 ו-O בוחר היפותזה $\mathcal{H}_{10}$ ו- R בוחר היפותזה $\mathcal{H}_2$ עדיין המודל R **יכליל** בצורה טובה יותר את הדוגמאות.

המסקנה העיקרית היא שיותר חשוב מפונקצית המטרה, הוא:

- כמות המדידות
- איכות המדידות - רעש גדול או קטן
  - משמע - התאמה של פולינום ממעלה 10 ל-15 נקודות רועשות זה "overkill" (כפי שנצפה מעלה)



התחום האפור הוא התחום בו יש לנו overfitting. רואים כי אמנם ה- E_{in} יורד אך ה- E_{out} עולה, השוני הוא עד כמה העליה והירידה קיצוניים.

מצד שני, בתחום הנותר רואים שככל שנגדיל את **כמות המדידות** כך נוכל להקטין את השגיאה גם ב- E_{out}.

1.1 האפקט של דרגת הרעש ומספר הנקודות

בהינתן:

- דרגת רעש (שונות הרעש) σ^2
- פונקצית מטרה מסדר Q_f
- סט דוגמאות בגודל N

נוכל להתבונן בפונקצית המטרה בצורה הבאה:

$$y = f(x) + \underbrace{\varepsilon(x)}_{\sigma^2} = \underbrace{\sum_{q=0}^{Q_f} \alpha_q x^q}_{normalized} + \varepsilon(x)$$

נתבונן בסט נקודות $(x_1, y_1), \ldots, (x_N, y_N)$ וניזכר בסטים של ההיפתוזות ממקודם:

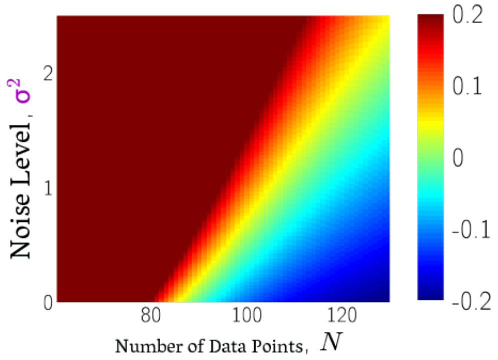
$\mathcal{H}_2$ – 2nd-order polynomials $\mathcal{H}_{10}$ – 10th-order polynomials

נשווה בין ה- E_{out} של שני החזאים:

$g_2 \in \mathcal{H}_2$ and $g_{10} \in \mathcal{H}_{10}$

בעזרת **Overfit-Measure** עובר ערכי N ו- σ^2 שונים:

$$\text{Overfit-Measure} = E_{out}(g_{10}) - E_{out}(g_2)$$



כאשר הצבע יותר אדום כך ה-overfit יותר גדול.

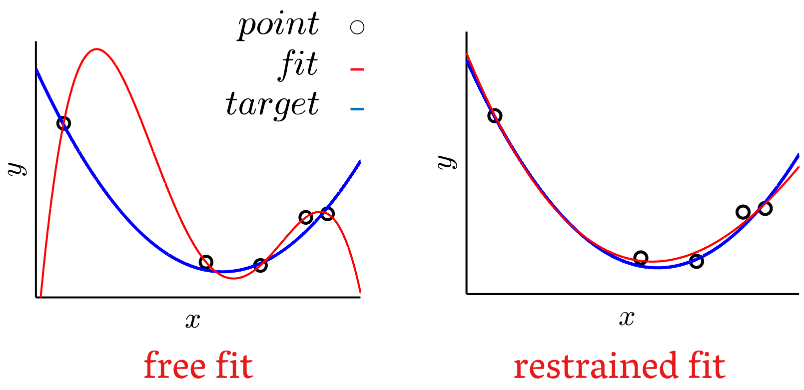
ככל שמספר הנקודות (N) גדל ה-overfit יקטן והמצב משתפר.

כמו-כן, ככל שרמת הרעש עולה כך קשה לנו יותר ויותר ללמוד את פונקצית המטרה והמשימה קשה יותר.

מסקנה: נעדיף להשתמש במודלים שהם כמה שיותר **פשוטים** כדי להתמודד עם בעיות שכאלו.

2 הכללה - Regularization

- אחד מה"נשקים" שלנו למלחמה ב- overfitting
- מאלץ את תהליך הלמידה שלנו (האלגוריתם) על מנת לשפר את ה- E_{in}



כאשר יש לנו פי'צר x ומרחיבים אותו לפיצ'ר ממימד אחר של פולינומים $\mathbf{z}$, נקבל סט היפותזות $\mathcal{H}_Q$ של פולינומים מדרגה Q.

$$\mathbf{z} = \begin{bmatrix} 1 \\ L_1(x) \\ \vdots \\ L_Q(x) \end{bmatrix} \qquad \mathcal{H}_Q = \left\{ \sum_{q=0}^{Q} w_q L_q(x) \right\}$$

$$(x_1, y_1), \ldots, (x_N, y_N) \longrightarrow (\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_N, y_N)$$

2.1 אילוץ קשיח - Hard Constraint

כזכור לנו היו 2 אפשרויות:

(1) לעבוד עם פולינומים מדרגה 10 (מתוך סט ההיפותזות $\mathcal{H}_{10}$)

(2) לעבוד עם פולינומים מדרגה 2 (מתוך סט ההיפותזות $\mathcal{H}_2$) - טובה יותר

עם זאת, נבחין כי פולינומים מסדר שני מוכלים בפולינומים מסדר עשירי, למצב כזה קוראים **Hard Constraint**
על מנת להשתמש באילוץ קשיח בדוגמה שלנו נוכל לאפס את המקדמים w_q עבור $q > 2$ ובכך להשאר בפולינומים בדרגה 2.

2.2 אילוץ רך - Soft Constraint

ראינו שאילוץ קשיח יכול בהחלט לעזור במצב בו יש המון רעש וסט דוגמאות יחסית קטן.
עם זאת, במקום לדרוש שכמות לא מבוטלת של מקדמי הפולינומים תתאפס, נוכל פשוט להגדיר אילוצים על אותם מקדמים על מנת להקטין אותם ובכך ליצור אילוצים פחות חזקים על הפולינומים ולהשאר עם פולינומים מדרגה 10:

$$\text{Soft version:} \quad \sum_{q=0}^{Q} w_q^2 \leq C$$

בכך נוכל לצמצם **מאוד** את מספר ההיפותזות בסט $\mathcal{H}_{10}$ ובכך לא נוכל להגיע לאותו E_{in} שיגרום ל-overfitting
אינטואיציה: ככל שנקטין את מקדמי הפולינומים, ככה נקבל פונקציות יותר "חלקות" ובכך המודל שלנו יהיה **פחות מותאם לרעש.**

הערה: אילוץ מוכר שמגביל מקדמים של דרגות גבוהות הוא $\sum_{q=0}^{Q} q \cdot w_q^2 \leq C$

2.3 Ridge Regression

פונקצית המטרה:

$$\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} x_{nj}w_j\right)^2$$
$$\textbf{subject to} \sum_{j=1}^{d} w_j^2 \le C$$

בד"כ נהוג להכתב בצורה ללא אילוצים עם $\lambda > 0$:

$$\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} x_{nj}w_j\right)^2 + \lambda\cdot\sum_{j=1}^{d} w_j^2$$

כאשר $\lambda \to \infty$ נקבל משקלים w_j **מאוד** קטנים, וכאשר λ יהיה מאוד קטן, נעדיף להקטין את המרחק בין המדידות לבין הפונקציה שלנו על פני הקטנת המשקלים.

הערה: בד"כ, נהוג מראש לקחת את המידע שלנו ולעשות לו סטנדרטיזציה (נניח לגרום לכך שכל הפיצ'רים יהיו עם שונות 1 על ידי החסרת הממוצע וחלוקת התוצאה בסטיית התקן) לפני שמתמודדים עם האילוצים של הרגולריזציה.

יתרה מזאת:

יש לזכור שכאשר מבצעים סטנדרטיזציה על קבוצת האימון אנחנו צריכים לבצע גם סטנדרטיזציה על המידע שעוד לא נראה, אך עם זאת יש לזכור כי את הסטנדרטיזציה שנבצע על המידע שעוד לא נראה **מתבססת על המידע שעליו התאמנו** - אסור לבצע סטנדרטיזציה שמבוססת על המידע החדש!!

נפתח את פונקצית המטרה, נתמקד בפונקציה עצמה ללא המינימום, נגדיר $\bar{x}_j = \frac{1}{N}\sum_{n=1}^{N} x_{nj}$:

$$\sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} x_{nj}w_j\right)^2 + \lambda\cdot\sum_{j=1}^{d} w_j^2 = \sum_{n=1}^{N}\left(y_n - \overbrace{w_0 - \sum_{j=1}^{d} w_j\bar{x}_j}^{w_0^c} - \sum_{j=1}^{d} w_j\left(x_{nj} - \bar{x}_j\right)\right)^2 + \lambda\cdot\sum_{j=1}^{d} w_j^2$$

נגזור על פי w_0^c (לא לשכוח נגזרת פנימית שעל פיה נקבל את המינוס בהתחלה) ונשווה לאפס:

$$\frac{\partial}{\partial w_0^c} = -\sum_{n=1}^{N} 2\cdot\left(y_n - w_0^c - \sum_{j=1}^{d} w_j\left(x_{nj} - \bar{x}_j\right)\right) = 0$$
$$\Rightarrow \sum_{n=1}^{N} y_n - Nw_0^c - \sum_{j=1}^{d} w_j\left(\underbrace{\sum_{n=1}^{N} x_{nj} - N\bar{x}_j}_{=0}\right) = 0$$
$$\Rightarrow w_0^c = \frac{1}{N}\sum_{n=1}^{N} y_n = \bar{y}$$

ונוכל להחליף את w_0^c ב- $\bar{y}$ ולהתמודד עם מודל ללא הטעיה.

כעת נותרנו עם:

$$\sum_{n=1}^{N}\left(\underbrace{y_n-\bar{y}}_{\tilde{y}_n}-\sum_{j=1}^{d}w_j\left(\underbrace{x_{nj}-\bar{x}_j}_{\tilde{x}_{nj}}\right)\right)^2+\lambda\cdot\sum_{j=1}^{d}w_j^2=\boxed{\sum_{n=1}^{N}\left(\tilde{y}_n-\sum_{j=1}^{d}w_j\tilde{x}_{nj}\right)^2+\lambda\cdot\sum_{j=1}^{d}w_j^2}$$

כאשר:

$$\tilde{\mathbf{y}}=\begin{bmatrix}\tilde{y}_1\\\tilde{y}_2\\\vdots\\\tilde{y}_N\end{bmatrix}\quad\tilde{X}=\begin{bmatrix}| & \dots & |\\\tilde{\mathbf{x}}_1 & \dots & \tilde{\mathbf{x}}_d\\| & \dots & |\end{bmatrix}$$

וכעת נוכל לרשום בצורה שכולנו מכירים ואוהבים:

$$\underset{\mathbf{w}\in\mathbb{R}^{\mathbf{d}}}{min}\ \left\|\tilde{\mathbf{y}}-\tilde{X}\mathbf{w}\right\|^2+\lambda\left\|\mathbf{w}\right\|^2$$

נפתור את הבעיה על ידי גזירת $\left(\tilde{\mathbf{y}}-\tilde{X}\mathbf{w}\right)^T\left(\tilde{\mathbf{y}}-\tilde{X}\mathbf{w}\right)+\lambda\cdot\mathbf{w}^T\mathbf{w}$ והשוואה לאפס:

$$\frac{\partial}{\partial\mathbf{w}}=-2\cdot\underbrace{\tilde{X}^T}_{d\times N}\underbrace{\left(\tilde{\mathbf{y}}-\tilde{X}\mathbf{w}\right)}_{N\times 1}+2\lambda\mathbf{w}=0$$

$$\Rightarrow\left(\tilde{X}^T\tilde{X}+\lambda I\right)\mathbf{w}=\tilde{X}^T\tilde{\mathbf{y}}$$

וכמובן אם $\lambda=0$ אז קיבלנו את המשוואות הנורמליות כצפוי מבעית הרגרסיה הלינארית הרגילה והמוכרת - בעית ה Least Squares

ואילו אם $\lambda>0$ (כפי שמוגדר תחת הרגולריזציה) אז במידה והמטריצה $\left(\tilde{X}^T\tilde{X}+\lambda I\right)$ **הפיכה** - ישנו פתרון (כצפוי). אך במידה והמטריצה **לא בהכרח הפיכה** נוכל להוכיח שהיא כן בעזרת הוכחה שהיא חיובית לחלוטין:

לשם כך נגדיר $\mathbf{c}\neq\mathbf{0}\in\mathbb{R}^d$ ונבדוק:

$$\mathbf{c}^T\left(\tilde{X}^T\tilde{X}+\lambda I\right)\mathbf{c}=\underbrace{\left\|\tilde{X}\mathbf{c}\right\|^2}_{\geq 0}+\underbrace{\lambda}_{>0}\underbrace{\left\|\mathbf{c}\right\|^2}_{>0}>0$$

כעת נוכל להכפיל בהפכית ולקבל:

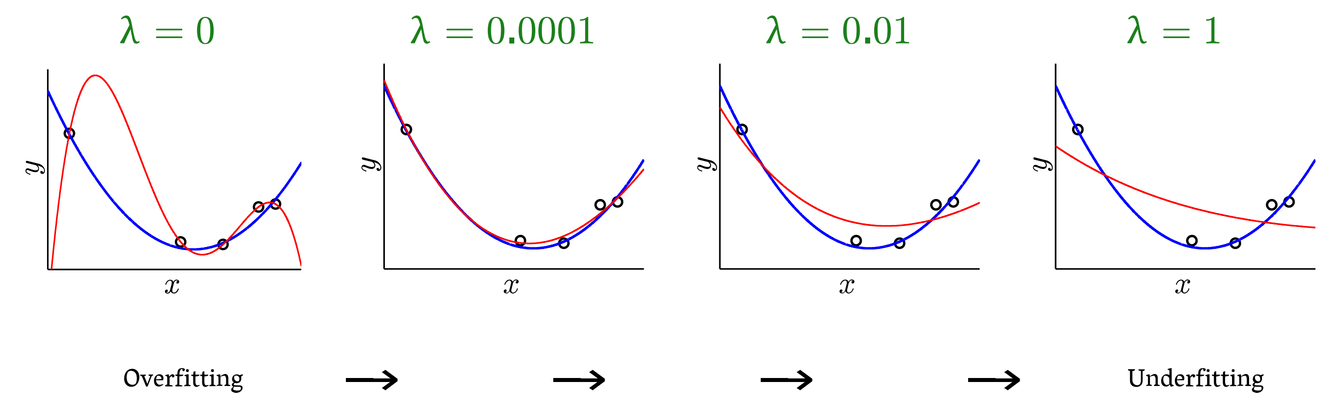
$$\boxed{\mathbf{w}=\left(\tilde{X}^T\tilde{X}+\lambda I\right)^{-1}\cdot\tilde{X}^T\tilde{\mathbf{y}}}$$

כעת כדי לקבל את המשערך שלנו נציב את $\mathbf{w}$ בהגדרה

$$\hat{\tilde{\boldsymbol{y}}}=\tilde{X}\cdot\left(\tilde{X}^T\tilde{X}+\lambda I\right)^{-1}\cdot\tilde{X}^T\tilde{\mathbf{y}}$$

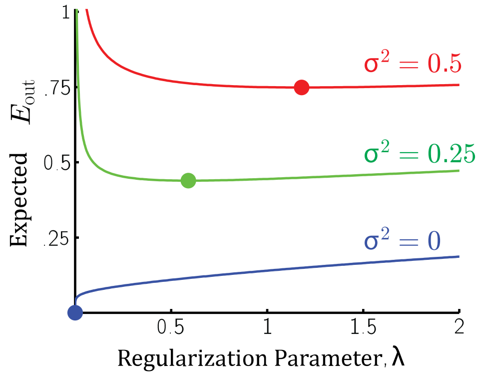
2.3.1 ערכי λ

יש להבחין בעובדה שערכי λ גדולים מידי יכולים לגרום למצב הפוך מ-overfitting שהוא underfitting.
במצב זה הרגולריזציה העודפת גורמת לאלגורים להתקבע על סט היפותזות מאוד מצומצם שבסופו של דבר יגרום לכך שהמודל שלנו לא יצליח להתאים את עצמו כלל.



כמו-כן, במידה וניקח $\lambda \longrightarrow \infty$ הרי שהמינימום שלנו יתקבל עבור $\mathbf{w} = \mathbf{0}$ כיוון שנקבל במטריצה האלכסונית $\left(\tilde{X}^T\tilde{X} + \lambda I\right)$ שאברי האלכסון שואפים ל-∞ ואילו המטריצה ההפכית תניב אברי אלכסון ששואפים ל-$0 = \frac{1}{\infty}$ משמע נקבל $\mathbf{w} = \mathbf{0}$ ונתכנס למצב שבו המודל שלנו מורכב מממוצע כלל הדגימות שלנו.

מציאת λ **אופטימלי** תלויה גם ברמת הרעש שלנו



נבחין כי רמת רעש אפסית נוכל פשוט לוותר על λ ולהשתמש במסווג הרגיל שלנו ואילו עבור רמת רעש הולכת ועולה נרצה להגדיל את λ בהתאם יחד עם התחשבות במספר הדגימות שלנו.

2.4 Least Absolute Shrinkage and Selection Operator Regression

זוהי שיטת רגולריזציה שדי דומה ל- Ridge Regression למעט איבר הרגולריזציה:

$$\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} x_{nj}w_j\right)^2$$

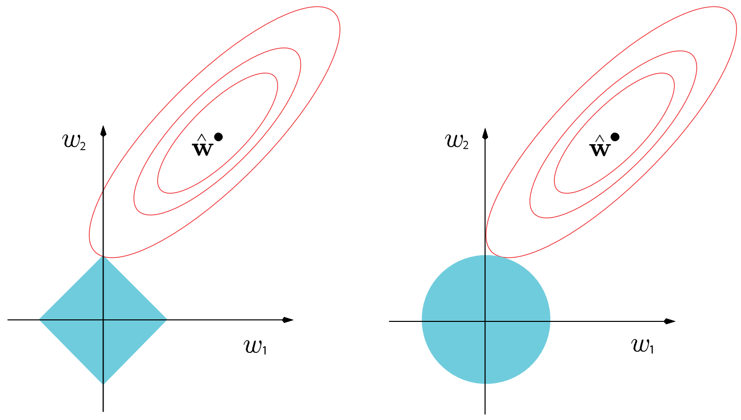
$$\textbf{subject to} \sum_{j=1}^{d}|w_j| \leq C$$

בצורה דומה נוכל להראות שהדבר דומה לבעית האופטימיזציה הלא מאולצת:

$$\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{n=1}^{N}\left(y_n - w_0 - \sum_{j=1}^{d} x_{nj}w_j\right)^2 + \lambda\sum_{j=1}^{d}|w_j|$$

2.5 Ridge vs. Lasso

נוכל להתבונן בהבדלים בין Ridge Regression ל- Lasso Regression



מימין: Ridge Regression משמאל Lasso Regression

מה רואים בתמונה?

- האליפסות האדומות הן השגיאה ביחס לפתרון $\hat{\mathbf{w}}$
- המעוין של ה Lasso הוא בעצם האילוץ

$$\sum_{j=1}^{d} |w_j| \leq C$$

- המעגל של ה- Ridge הוא בעצם האילוץ

$$\sum_{j=1}^{d} w_j^2 \leq C$$

והוא בעל רדיוס $R = \sqrt{C}$

אנחנו נרצה למצוא פתרון שיהיה כמה שיותר קרוב לפונקציתה המטרה (אליפסות קרובות לנקודה $\hat{\mathbf{w}}$) אך ישאר בתוך תחום ההגדרה (מעגל או מעוין)

מסקנה: במידה ונבחר C גדול, הנקודה $\hat{\mathbf{w}}$ תכנס לתוך תחום ההגדרה של C ואז הפתרון לבעיה הוא בעצם הפתרון של ה-Least Squares בפני עצמו.

אז מה ההבדל העיקרי?

כיוון של Lasso-ישנן נקודות השבירה (במקרה הזה קודקודי המעוין) הרי שיש סיכוי לפגוש בנקודה בה אחד המקדמים (בתמונה w_1) **מתאפס!**

לעומת זאת ב- Ridge הכל סימטרי ונקבל פתרונות שבד"כ לא יתאפסו.

לפיכך, ככל שנעלה בממדים שלנו, הצורה שיניב ה-Lasso תהיה בעלת הרבה יותר קודקודים ויהיה קל יותר לפגוש נקודות ששייכות למצב בו סט הפרמטרים דליל.

מסקנה: ניעזר ב-Lasso כשנרצה לעשות feature selection של ממש - מצב בו נרצה למצוא לזרוק חלק מהפיצ'רים שלנו!