EECS 127/227AT Optimization Models in Engineering UC Berkeley Spring 2023

Homework 12

Self grades are due at 11 PM on April 27, 2023.

1. **Newton's Method, Coordinate Descent and Gradient Descent**

In this question, we will compare three different optimization methods: Newton's method, coordinate descent and gradient descent. We will consider the simple set-up of unconstrained convex quadratic optimization; i.e we will consider the following problem:

$$\min_{\vec{x}\in\mathbb{R}^d} \vec{x}^\top A\vec{x} - 2\vec{b}^\top \vec{x} + c \tag{1}$$

where $A \succ 0$ and $\vec{b} \in \mathbb{R}^d$.

(a) How many steps does Newton's method take to converge to the optimal solution? Recall that the update rule for Newton's method is given by the equation:

$$\vec{x}_{t+1} = \vec{x}_t - (\nabla^2 f(\vec{x}_t))^{-1}\nabla f(\vec{x}_t). \tag{2}$$

when optimizing a function f.

Solution: Newton's method converges in a single step irrespective of the starting point. Let $\vec{x}_0$ be any starting point. We have:

$$\nabla^2 f(\vec{x}_0) = 2A \text{ and } \nabla f(\vec{x}_0) = 2(A\vec{x} - \vec{b}). \tag{3}$$

Therefore, we have:

$$\vec{x}_1 = \vec{x}_0 - A^{-1}(A\vec{x}_0 - \vec{b}) = A^{-1}\vec{b}. \tag{4}$$

Note that since this is an unconstrained convex quadratic optimization problem with A being full rank, we can find the optimum point by setting the derivative of the function to 0. Therefore, we have:

$$\nabla f(\vec{x}^*) = 2(A\vec{x}^* - \vec{b}) = 0 \implies \vec{x}^* = A^{-1}\vec{b}. \tag{5}$$

(b) Now, consider the simple two variable quadratic optimization problem for $\sigma > 0$:

$$\min_{\vec{x}\in\mathbb{R}^2} f(\vec{x}) = \sigma x_1^2 + x_2^2. \tag{6}$$

How many steps does coordinate descent take to converge on this problem? Assume that we start by updating the variable x_1 in the first step, x_2 in step two and so on; therefore, we will update x_1 and x_2 in odd and even iterations respectively:

$$(x_{t+1})_1 = \begin{cases} \text{argmin}_{x_1} f(x_1, (x_t)_2) & \text{for odd t} \\ (x_t)_1 & \text{otherwise} \end{cases} \text{ and } (x_{t+1})_2 = \begin{cases} \text{argmin}_{x_2} f((x_t)_1, x_2) & \text{for even t} \\ (x_t)_2. & \text{otherwise} \end{cases} \tag{7}$$

Here, $(x_t)_2$ represents x_2 at time t and so on.

Solution: On this problem, coordinate descent converges in 2 steps starting from any initialization point. Note that the optimal solution for each of the updates is 0, by setting the gradient to 0. Therefore, coordinate descent converges in two steps, one to update x_1 and the other to update x_2.

(c) We will now analyze the performance of coordinate descent on another quadratic optimization problem:

$$\min_{\vec{x}\in\mathbb{R}^2} f(\vec{x}) = \sigma(x_1 + x_2)^2 + (x_1 - x_2)^2. \tag{8}$$

where we have, as before, $\sigma > 0$. Note that $(0, 0)$ is the optimal solution to this problem. Now, starting from the point $\vec{x}_0 = (1, 1)$, write how each coordinate of $(\vec{x}_{t+1})_i$ relates to $(\vec{x}_t)_i$ for $i = 1, 2$. Use this to show how the algorithm converges from the initial point $(1, 1)$ to $(0, 0)$. What happens when σ grows large? *HINT: First find the update rule for $(\vec{x}_t)_1$, i.e. keep $(\vec{x}_t)_2$ fixed and figure out how $(\vec{x}_t)_1$ changes when t is odd. Then do the same for $(\vec{x}_t)_2$ when $(\vec{x}_t)_1$ is fixed for even t.*

Solution: We first find the update rule for x_1. Note that we only update x_1 when t is odd. Now, by taking the gradient and setting it to 0, we get:

$$\sigma((x_{t+1})_1 + (x_t)_2) + ((x_{t+1})_1 - (x_t)_2) = 0 \implies (x_{(t+1)})_1 = \frac{(1-\sigma)}{(1+\sigma)}(x_t)_2. \tag{9}$$

Note that the function, f, is symmetric in the variables, x_1 and x_2. Therefore, the update rule for x_2 (when t is even) is given by:

$$(x_{(t+1)})_2 = \frac{(1-\sigma)}{(1+\sigma)}(x_t)_1. \tag{10}$$

Therefore, we get for all $t \geq 2$:

$$(x_t)_1 = \left(\frac{1-\sigma}{1+\sigma}\right)^{2\left\lfloor\frac{t+1}{2}\right\rfloor-1} \text{ and } (x_t)_2 = \left(\frac{1-\sigma}{1+\sigma}\right)^{2\left\lfloor\frac{t}{2}\right\rfloor}. \tag{11}$$

When σ grows large, the $\frac{1-\sigma}{1+\sigma}$ goes to -1 and this results in slow convergence as the algorithm converges quickly when $\left|\frac{1-\sigma}{1+\sigma}\right|$ is small.

(d) Now, let $f(\vec{x}) = \frac{1}{2}\vec{x}^\top A\vec{x} + \vec{x}^\top\vec{b} + c$ where A is PD. When we run gradient descent on $f(\vec{x})$, the convergence along each of the unit eigenvectors $\vec{v}_i$ of A is

$$|1 - \eta\left(\lambda_i\{A\}\right)| \tag{12}$$

This can be derived similar to HW 8 Question 1e, which you may reference. Formally, in the current setting, we have

$$(\vec{x}_k - \vec{x}^\star) = (I - \eta A)^k(\vec{x}_0 - \vec{x}^\star)$$

One way we can derive an "optimal" learning rate $\eta^\star$ is to minimize the largest rate of convergence:

$$\eta^\star \in \operatorname*{argmin}_{\eta\in\mathbb{R}} \max_{i\in\{1,\dots,n\}} |1 - \eta\left(\lambda_i\{A\}\right)|. \tag{13}$$

One important property of $\eta^\star$ is that it makes the rates of convergence $|1 - \eta\left(\lambda_i\{A\}\right)|$ associated with the largest and smallest singular values of A equal, i.e.,

$$|1 - \eta(\lambda_{\max}\{A\})| = |1 - \eta(\lambda_{\min}\{A\})|$$

Use this property to show that

$$\eta^\star = \frac{2}{\lambda_{\max}\{A\} + \lambda_{\min}\{A\}} \tag{14}$$

where $\lambda_{\min}\{A\} = \lambda_n\{A\}$ is the n^{th} largest singular value of A and similar for the maximum.

Solution: We have

$$\begin{aligned}
\left|1-\eta^{\star}\left(\lambda_{\min}\{A\}\right)\right| &= \left|1-\eta^{\star}\left(\lambda_{\max}\{A\}\right)\right| && (15)\\
1-\eta^{\star}\left(\lambda_{\min}\{A\}\right) &= -\left(1-\eta^{\star}\left(\lambda_{\max}\{A\}\right)\right) && (16)\\
1-\eta^{\star}\left(\lambda_{\min}\{A\}\right) &= \eta^{\star}\left(\lambda_{\max}\{A\}\right)-1 && (17)\\
2 &= \eta^{\star}\left(\lambda_{\max}\{A\}+\lambda_{\min}\{A\}\right) && (18)\\
\eta^{\star} &= \frac{2}{\lambda_{\max}\{A\}+\lambda_{\min}\{A\}}. && (19)
\end{aligned}$$

Here the second equality is the most challenging to derive. It follows from the first inequality by the following reasoning:

- If $1-\eta^{\star}\left(\lambda_{\min}\{A\}\right)$ and $1-\eta^{\star}\left(\lambda_{\max}\{A\}\right)$ have the same sign, then by the first equality, they must be equal. This means that $\lambda_{\max}\{A\}=\lambda_1\{A\}=\lambda_2\{A\}=\cdots=\lambda_n\{A\}=\lambda_{\min}\{A\}$ and the optimal learning rate $\eta^{\star}$ sets each rate $1-\eta^{\star}\left(\lambda_i\{A\}\right)$ to 0 simultaneously, ensuring convergence in one step. If both sides are 0 then the second equality holds (because $0=-0$).
- Otherwise, $1-\eta^{\star}\left(\lambda_{\min}\{A\}\right)$ and $1-\eta^{\star}\left(\lambda_{\max}\{A\}\right)$ have opposite signs. Since $\lambda_{\max}\{A\}>\lambda_{\min}\{A\}$ (since if they were equal we would be in the first case), we have $1-\eta^{\star}\left(\lambda_{\min}\{A\}\right)>1-\eta^{\star}\left(\lambda_{\max}\{A\}\right)$. Thus $1-\eta^{\star}\left(\lambda_{\min}\{A\}\right)$ must be positive and $1-\eta^{\star}\left(\lambda_{\max}\{A\}\right)$ must be negative. The absolute value of a negative number is its negative, so the second equality follows directly from the first equality.

(e) Finally, for the objective function (8), write an equation relating $\vec{x}_t$ to $\vec{x}_0$ if $\vec{x}_0=\begin{bmatrix}1\\-1\end{bmatrix}$. Assume for this part that $\sigma>1$ and reason about how quickly gradient descent converges when σ grows large. *HINT: What is the optimal step size for gradient descent, using the previous part? HINT: Also note that f is given by:*

$$f(\vec{x})=\vec{x}^{\top}A\vec{x} \text{ where } A=2\left(\sigma\begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}+\begin{bmatrix}\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{bmatrix}\right). \tag{20}$$

Solution: We first note that f is given by:

$$f(\vec{x})=\vec{x}^{\top}A\vec{x} \text{ where } A=2\left(\sigma\begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}+\begin{bmatrix}\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{bmatrix}\right). \tag{21}$$

Therefore, we have that $\lambda_{\max}$ of A is 2σ and $\lambda_{\min}$ is 2. Using the result from the previous part (and dividing by 2 since the optimal learning rate was computed for $\frac{1}{2}\vec{x}^{\top}A\vec{x}+\vec{x}^{\top}\vec{b}+c$ and not $\vec{x}^{\top}A\vec{x}-2\vec{b}^{\top}\vec{x}+c$), the step size for gradient descent is set to $1/(2\sigma+2)$. Now, we have that:

$$\nabla f((1,-1))=\begin{bmatrix}4\\-4\end{bmatrix}. \tag{22}$$

Therefore, we have that:

$$\vec{x}_1=\vec{x}_0-\eta\nabla f((1,-1))=\left(1-\frac{2}{\sigma+1}\right)\begin{bmatrix}1\\-1\end{bmatrix}. \tag{23}$$

By iterating the above procedure we see that:

$$\vec{x}_t=\left(1-\frac{2}{\sigma+1}\right)^t\begin{bmatrix}1\\-1\end{bmatrix}. \tag{24}$$

Therefore, when σ is large, the convergence rate of gradient descent is really slow. However, Newton's method would find the optimum in one step.

2. **Gradient Descent vs Newton Method**

   Run the Jupyter notebook `gradient_vs_newton.ipynb` which demonstrates differences between gradient descent and Newton's method.

3. LASSO vs. Ridge

Say that we have the data set $\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1,\ldots,n}$ of samples $\vec{x}^{(i)} \in \mathbb{R}^d$ and values $y^{(i)} \in \mathbb{R}$.

Define $X = \begin{bmatrix} \vec{x}^{(1)} & \ldots & \vec{x}^{(n)} \end{bmatrix}^\top$ and $y = \begin{bmatrix} y^{(1)} & \ldots & y^{(n)} \end{bmatrix}^\top$.

For the sake of simplicity, assume that the data has been centered and whitened so that each feature has mean 0 and variance 1 and the features are uncorrelated, i.e. $X^\top X = nI$. Consider the linear least squares regression with regularization in the ℓ_1-norm, also known as LASSO:

$$\vec{w}^\star = \underset{\vec{w}\in\mathbb{R}^d}{\operatorname{argmin}} \; \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_1 . \tag{25}$$

This problem will compare ℓ_1-regularization with ℓ_2-regularization (ridge regression) to understand their similarities and differences. We will do this by looking at the elements of $\vec{w}^\star$ in the solution to each problem.

(a) First, we decompose this optimization problem into d univariate optimization problems over each element of $\vec{w}$. Let $X = \begin{bmatrix} \vec{x}_1 & \ldots & \vec{x}_d \end{bmatrix}$ and recall that $X^\top X = nI$.

Solution:

$$\|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_1 = \sum_{i=1}^{d} \left[n w_i^2 - 2\vec{y}^\top \vec{x}_i w_i + \lambda |w_i| \right] + \vec{y}^\top \vec{y}. \tag{26}$$

Hence the original problem becomes

$$\min_{\vec{w}\in\mathbb{R}^d} \sum_{i=1}^{d} \left[n w_i^2 - 2\vec{y}^\top \vec{x}_i w_i + \lambda |w_i| \right] + \vec{y}^\top \vec{y}. \tag{27}$$

Since the objective is separable in w_i the problem decomposes into d univariate optimization problems and hence we have

$$\sum_{i=1}^{d} \min_{w_i\in\mathbb{R}} \left[n w_i^2 - 2\vec{y}^\top \vec{x}_i w_i + \lambda |w_i| \right] + \vec{y}^\top \vec{y}. \tag{28}$$

(b) If $w_i^\star > 0$, then what is the value of $w_i^\star$? What is the condition on $\vec{y}^\top \vec{x}_i$ for this to be possible?

Solution: If $w_i^\star > 0$, then the first order optimality conditions for $w_i^\star$ write

$$2n w_i^\star - 2\vec{y}^\top \vec{x}_i + \lambda = 0, \tag{29}$$

from which we obtain

$$w_i^\star = \frac{2\vec{y}^\top \vec{x}_i - \lambda}{2n}, \tag{30}$$

which is positive when

$$\vec{y}^\top \vec{x}_i > \frac{\lambda}{2}. \tag{31}$$

(c) If $w_i^\star < 0$, then what is the value of $w_i^\star$? What is the condition on $\vec{y}^\top \vec{x}_i$ for this to be possible?

Solution: If $w_i^\star < 0$, then the first order optimality conditions for $w_i^\star$ write

$$2n w_i^\star - 2\vec{y}^\top \vec{x}_i - \lambda = 0, \tag{32}$$

from which we obtain

$$w_i^\star = \frac{2\vec{y}^\top \vec{x}_i + \lambda}{2n}, \tag{33}$$

which is negative when

$$\vec{y}^\top \vec{x}_i < -\frac{\lambda}{2}. \tag{34}$$

(d) What can we conclude about $w_i^\star$ if $\left|\vec{y}^\top \vec{x}_i\right| \leq \frac{\lambda}{2}$? How does the value of λ impact the individual entries $w_i^\star$?

Solution: From the previous parts we have $w_i^\star \neq 0 \Rightarrow \left|\vec{y}^\top \vec{x}_i\right| > \frac{\lambda}{2}$. Hence if $\left|\vec{y}^\top \vec{x}_i\right| \leq \frac{\lambda}{2}$ then we must have $w_i^\star = 0$. This means that a larger value of λ will force more entries of $\vec{w}$ to be zero — i.e. larger λ will imply higher sparsity.

(e) Now consider the case of ridge regression, which uses the the ℓ_2 regularization $\lambda \left\|\vec{w}\right\|_2^2$.

$$\vec{w}^\star = \operatorname*{argmin}_{\vec{w} \in \mathbb{R}^d} \left\|X\vec{w} - \vec{y}\right\|_2^2 + \lambda \left\|\vec{w}\right\|_2^2. \tag{35}$$

Write down the new condition for $\vec{w}_i^\star$ to be 0. How does this differ from the condition obtained in part (4) and what does this suggest about LASSO?

Solution: In the case of ridge regression the optimal weight vector $\vec{w}$ is given by

$$w_i^\star = \frac{\vec{y}^\top \vec{x}_i}{n + \lambda}, \ i = 1, \ldots, d. \tag{36}$$

So $w_i^\star$ is only zero when $\vec{y}^\top \vec{x}_i = 0$, in contrast to LASSO where $w_i^\star$ is zero when $\vec{y}^\top \vec{x}_i \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right]$. This suggest that LASSO forces a lot of coordinates to be zero, i.e. induces sparsity to the optimal weight vector.

4. More Fun with Lasso and Ridge

Complete the Jupyter notebook `ridge_vs_lasso.ipynb` which demonstrates differences between ridge regression and LASSO.

5. Safe feature elimination in LASSO

Safe feature elimination is a technique that applies to problems with ℓ_1-norm penalties. It allows one to remove features (rows of a data matrix) before solving the problem, based on a quick computation, leading to a great reduction in overall computational effort. In this exercise we illustrate the approach with the so-called LASSO model:

$$\min_{\vec{w}} \ P(\vec{w}) := \frac{1}{2}\|X^\top \vec{w} - \vec{y}\|_2^2 + \lambda\|\vec{w}\|_1$$

where $X \in \mathbb{R}^{n\times m}$ a data matrix, with each column a data point in $\mathbb{R}^n$, and $\lambda \geq 0$ is a hyper-parameter. We denote by $\vec{x}_i^T$ the i-th row of X.

(a) Show that the dual can be written as

$$\max_{\vec{u}} \ D(\vec{u}) := \begin{cases} \vec{y}^T\vec{u} - \frac{1}{2}\vec{u}^T\vec{u} & \text{if } \vec{u} \in \mathcal{U} \\ -\infty & \text{otherwise,} \end{cases}$$

where $\mathcal{U} := \{\vec{u} \,:\, \|X\vec{u}\|_\infty \leq \lambda\}$ is the dual feasible set.

HINT:

$$\min_{\vec{a}} \ \left(\lambda\|\vec{a}\|_1 - \vec{b}^\top\vec{a}\right) = \begin{cases} 0 & \textit{if } \|\vec{b}\|_\infty \leq \lambda \\ -\infty & \textit{otherwise} \end{cases}$$

where $a, b \in \mathbb{R}^n$.

Solution: To derive the dual, first introduce an auxiliary variable $\vec{z} \in \mathbb{R}^m$ and constraint $\vec{z} = X^\top\vec{w}$, which gives the equivalent problem

$$\min_{\vec{w},\vec{z}} \ \frac{1}{2}\|\vec{z} - \vec{y}\|_2^2 + \lambda\|\vec{w}\|_1 \quad \text{subject to} \quad \vec{z} = X^\top\vec{w}.$$

Introducing Lagrange multiplier $\vec{u} \in \mathbb{R}^m$ for the equality constraint, the Lagrangian is

$$\mathcal{L}(\vec{w}, \vec{z}, \vec{u}) = \frac{1}{2}\|\vec{z} - \vec{y}\|_2^2 + \lambda\|\vec{w}\|_1 + \vec{u}^\top(\vec{z} - X^\top\vec{w}),$$

and the dual function is

$$D(\vec{u}) = \min_{\vec{w},\vec{z}} \mathcal{L}(\vec{w}, \vec{z}, \vec{u}) = \min_{\vec{w},\vec{z}} \ \frac{1}{2}\|\vec{z} - \vec{y}\|_2^2 + \lambda\|\vec{w}\|_1 + \vec{u}^\top(\vec{z} - X^\top\vec{w}).$$

Observe the dual function can be minimized separately in $\vec{w}, \vec{z}$

$$\min_{\vec{w},\vec{z}} \ \frac{1}{2}\|\vec{z} - \vec{y}\|_2^2 + \lambda\|\vec{w}\|_1 + \vec{u}^\top(\vec{z} - X^\top\vec{w}) = \min_{\vec{z}} \ \left(\frac{1}{2}\|\vec{z} - \vec{y}\|_2^2 + \vec{u}^\top\vec{z}\right) + \min_{\vec{w}} \ \left(\lambda\|\vec{w}\|_1 - (X\vec{u})^\top\vec{w}\right).$$

To perform the minimization over $\vec{z}$, notice the function is convex; taking the derivative and setting it equal to zero gives $\vec{z}^\star - \vec{y} + \vec{u} = 0$, so $\vec{z}^\star = \vec{y} - \vec{u}$, and thus

$$\min_{\vec{z}} \ \left(\frac{1}{2}\|\vec{z} - \vec{y}\|_2^2 + \vec{u}^\top\vec{z}\right) = \vec{y}^\top\vec{u} - \frac{1}{2}\vec{u}^\top\vec{u}.$$

To perform the minimization over $\vec{w}$, notice the objective can be minimized separately for each coordinate

$$\min_{\vec{w}} \ \left(\lambda\|\vec{w}\|_1 - (X\vec{u})^\top\vec{w}\right) = \min_{\vec{w}} \ \sum_{i=1}^{n} \lambda|w_i| - (\vec{x}_i^\top\vec{u})w_i = \sum_{i=1}^{n} \min_{w_i}(\lambda - \text{sign}(w_i)\vec{x}_i^\top)\vec{u}|w_i|.$$

For each coordinate i, we can argue by cases. If $|\vec{x}_i^\top \vec{u}| > \lambda$, we can pick $\mathrm{sign}(w_i)$ so that $\lambda - \mathrm{sign}(w_i)\vec{x}_i^\top \vec{u} < 0$ and send $|w_i| \to \infty$, so the optimal value is $-\infty$. Otherwise, if $\lambda \geq |\vec{x}_i^\top \vec{u}|$, then for any w_i, $\lambda - \mathrm{sign}(w_i)\vec{x}_i^\top \vec{u} \geq 0$, so the objective is minimized for $w_i = 0$. Summarizing this discussion, we've shown

$$\min_{w_i}(\lambda - \mathrm{sign}(w_i)\vec{x}_i^\top \vec{u})|w_i| = \begin{cases} 0 & \text{if } |\vec{x}_i^\top \vec{u}| \leq \lambda \\ -\infty & \text{otherwise.} \end{cases}$$

Since this is true for all i,

$$\min_{\vec{w}} \ \left(\lambda\|\vec{w}\|_1 - (X\vec{u})^\top \vec{w}\right) = \begin{cases} 0 & \text{if } \|X\vec{u}\|_\infty \leq \lambda \\ -\infty & \text{otherwise.} \end{cases}$$

Combined with our previous work, we've shown the dual function is

$$D(\vec{u}) = \min_{\vec{w},\vec{z}} \mathcal{L}(\vec{w},\vec{z},\vec{u}) = \begin{cases} \vec{y}^\top \vec{u} - \frac{1}{2}\vec{u}^\top \vec{u} & \text{if } \|X\vec{u}\|_\infty \leq \lambda \\ -\infty & \text{otherwise,} \end{cases}$$

and the associated dual problem is then $\max_{\vec{u}} D(\vec{u})$.

(b) Explain why the so-called dual gap $G(\vec{w},\vec{u}) := P(\vec{w}) - D(\vec{u})$ is always non-negative, for any $\vec{w}$ and any $\vec{u}$ with $\|X\vec{u}\|_\infty \leq \lambda$.

Solution: This is a consequence of weak-duality. In particular, for any $\vec{w}$ and feasible $\vec{u}$, $P(\vec{w}) \geq \min_{\vec{w}'} P(\vec{w}') \geq \max_{\vec{u}'} D(\vec{u}') \geq D(\vec{u})$, where the second inequality uses weak duality. Rearranging, $G(\vec{w},\vec{u}) = P(\vec{w}) - D(\vec{u}) \geq 0$.

(c) Let $\vec{u}^\star$ be an optimal dual point. Show that, if for a given i, $|\vec{x}_i^T \vec{u}^\star| < \lambda$ then we can safely remove the i-th feature.

HINT: You may use a lemma we proved in the last homework. Namely, if $\vec{u}^\star$ is a optimal for the original dual, then $\vec{u}^\star$ is also optimal for the problem with the inactive constraint removed:

$$\max_{\vec{u}} D(\vec{u}) \ : \ \forall\, j \neq i, \ \ |\vec{x}_j^\top \vec{u}| \leq \lambda.$$

Solution: Since the dual problem is convex, if $\vec{u}^\star$ is optimal, it is also optimal for the problem with the inactive constraint removed:

$$\max_{\vec{u}} D(\vec{u}) \ : \ \forall\, j \neq i, \ \ |\vec{x}_j^\top \vec{u}| \leq \lambda.$$

Let X' denote the data matrix with the features i removed. Consider the modified LASSO problem,

$$\min_{\vec{w}} \ P'(\vec{w}) := \frac{1}{2}\|X'^\top \vec{w} - \vec{y}\|_2^2 + \lambda\|\vec{w}\|_1.$$

Then, from part (a), the dual problem is

$$D'(\vec{u}) = \begin{cases} \vec{y}^\top \vec{u} - \frac{1}{2}\vec{u}^\top \vec{u} & \text{if } \|X'\vec{u}\|_\infty \leq \lambda \\ -\infty & \text{otherwise.} \end{cases}$$

Since the only constraints in the primal problem are affine constraints, Slater's condition holds, and strong duality obtains. Thus, $D'(\vec{u}^\star) = \min_{\vec{w}} P'(\vec{w})$. By the above argument (Prop 8.1), $\vec{u}^\star$ is also dual optimal for D', so $D'(\vec{u}^\star) = D(\vec{u}^\star)$. Putting all of these pieces together, we've shown

$$\min_{\vec{w}} P'(\vec{w}) = D'(\vec{u}^\star) = D(\vec{u}^\star) = \min_{\vec{w}} P(\vec{w}).$$

and thus removing the i-th feature does not change the optimum of the LASSO problem.

(d) Let $\vec{u}^\star$ be an optimal dual point. Show that, for any $\vec{u}$ with $\|X\vec{u}\|_\infty \leq \lambda$:

$$\nabla D(\vec{u}^\star)^T(\vec{u}-\vec{u}^\star) = (\vec{y}-\vec{u}^\star)^T(\vec{u}-\vec{u}^\star) \leq 0$$

HINT: By first-order optimality conditions, if f be a convex function and C be a closed convex set on which f is differentiable. Then $x^\star \in \arg\min_{x\in C} f(x)$ if and only if $\langle \nabla f(x^\star), \vec{y}-x^\star\rangle \geq 0$ for all $\vec{y} \in C$.

Solution: This follows from the first-order optimality condition for the concave function D on the set $\mathcal{U}$. Fix $\vec{u} \in \mathcal{U}$, and suppose $\nabla D(\vec{u}^\star)^\top(\vec{u}-\vec{u}^\star) > 0$. Since $\vec{u}^\star$ is an optimal dual point and strong duality obtains, $D(\vec{u}^\star) > -\infty$ and therefore $\|X\vec{u}^\star\|_\infty \leq \lambda$. This implies both $\vec{u}, \vec{u}^\star \in \mathcal{U}$, which is convex (since it's the sublevel set of a convex function). By assumption, D is locally increasing around $\vec{u}^\star$ on the line to $\vec{u}$, which contradicts the optimality of $\vec{u}^\star$.

Slightly more formally, define $h(t) = D(\vec{u}^\star + t(\vec{u}-\vec{u}^\star))$, and note for all $t \in [0,1]$, $\vec{u}^\star + t(\vec{u}-\vec{u}^\star) \in \mathcal{U}$. Now, observe $h'(0) = D(\vec{u}^\star)^\top(\vec{u}-\vec{u}^\star) > 0$, so for sufficiently small t, $D(\vec{u}^\star + t(\vec{u}-\vec{u}^\star)) = h(t) > h(0) = D(\vec{u}^\star)$, a contradiction.

(e) Prove that, for any $\vec{u} \in \mathcal{U}$:

$$\frac{1}{2}\|\vec{u}-\vec{u}^\star\|_2^2 \leq P(\vec{w}) - D(\vec{u}) = G(\vec{w},\vec{u}).$$

HINT: A twice-differentiable convex function f is α-strongly convex if $\nabla^2 f(x) \succeq \alpha I$. An interesting consequence of strong-convexity if that for any $x, \vec{y} \in \mathrm{dom}(f)$,

$$f(\vec{y}) \geq f(x) + \nabla f(x)^\top(\vec{y}-x) + \frac{\alpha}{2}\|\vec{y}-x\|_2^2.$$

Start by showing the function $-D$ is strongly convex, and then combine this inequality with the previous parts.

Solution: On the set $\mathcal{U}$, $-D$ is 1-strongly convex (since $\nabla^2(-D) = I$). Therefore, using an equivalent definition of strong convexity, for any $\vec{u} \in \mathcal{U}$,

$$(-D)(\vec{u}) \geq (-D)(\vec{u}^\star) + \nabla(-D)(\vec{u}^\star)^\top(\vec{u}-\vec{u}^\star) + \frac{1}{2}\|\vec{u}-\vec{u}^\star\|_2^2.$$

Rearranging this inequality and applying the previous part,

$$\frac{1}{2}\|\vec{u}-\vec{u}^\star\|_2^2 \leq D(\vec{u}^\star) - D(\vec{u}) + \nabla D(\vec{u}^\star)^\top(\vec{u}-\vec{u}^\star) \leq D(\vec{u}^\star) - D(\vec{u}).$$

By weak-duality, $D(\vec{u}^\star) \leq \min_{\vec{w}'} P(\vec{w}') \leq P(\vec{w})$, so

$$\frac{1}{2}\|\vec{u}-\vec{u}^\star\|_2^2 \leq D(\vec{u}^\star) - D(\vec{u}) \leq P(\vec{w}) - D(\vec{u}) = G(\vec{w},\vec{u}),$$

as required.

(f) The above sphere of confidence for an optimal dual point $\vec{u}^\star$ allows to safely eliminate features. Recall from part c that if $|\vec{x}_i^T\vec{u}^\star| < \lambda$ then the i-th feature can be removed. Show that if

$$\lambda > \max_{\xi\,:\,\|\xi\|_2\leq 1} |\vec{x}_i^T(\vec{u} + \sqrt{2G(\vec{w},\vec{u})}\xi)| = |\vec{x}_i^T\vec{u}| + \sqrt{2G(\vec{w},\vec{u})}\|\vec{x}_i\|_2 \tag{37}$$

then one can safely remove the i-th feature.

HINT: Consider a specific $\xi = \frac{\vec{u}^\star - \vec{u}}{\sqrt{2G(\vec{w},\vec{u})}}$

Solution: We show the condition implies $|\vec{x}_i^\top\vec{u}^\star| < \lambda$. Let $\xi = \frac{\vec{u}^\star-\vec{u}}{\sqrt{2G(\vec{w},\vec{u})}}$. Then, using part e and non-negativity of $G(\vec{w},\vec{u})$, we have

$$\|\xi\|_2 = \left\|\frac{\vec{u}^\star - \vec{u}}{\sqrt{2G(\vec{w},\vec{u})}}\right\|_2 = \frac{1}{\sqrt{2G(\vec{w},\vec{u})}}\|\vec{u}^\star - \vec{u}\| \leq 1.$$

Further $\vec{u} + \sqrt{2G(\vec{w}, \vec{u})}\xi = \vec{u}^\star$. Hence, if $\lambda > |\vec{x}_i^\top(\vec{u} + \sqrt{2G(\vec{w}, \vec{u})}\xi)|$ for all ξ with $\|\xi\|_2 \leq 1$, then $\lambda > |\vec{x}_i^\top \vec{u}^\star|$, so it is safe to remove the i-th feature.

(g) Apply the technique in part f with the choice $\vec{w} = 0$ and dual feasible point $\vec{u} = \lambda\vec{y}/\lambda_{\max}$, where $0 \leq \lambda \leq \lambda_{\max} = \|X\vec{y}\|_\infty$, to show that the condition

$$\lambda > \frac{|\vec{x}_i^T\vec{y}| + \|\vec{x}_i\|_2\|\vec{y}\|_2}{1 + \frac{\|\vec{x}_i\|_2\|\vec{y}\|_2}{\lambda_{\max}}}$$

allows one to safely remove the i-th feature.

HINT: Since $0 \leq \lambda \leq \lambda_{\max}$, *use the upper bound* $(\lambda/\lambda_{\max})|\vec{x}_i^\top\vec{y}| \leq |\vec{x}_i^\top\vec{y}|$.

Solution: We explicitly evaluate equation (37) for the given data. First, $|\vec{x}_i^\top\vec{u}| = \frac{\lambda|\vec{x}_i^\top\vec{y}|}{\lambda_{\max}}$. Next, the duality gap is

$$\begin{aligned} G(\vec{w}, \vec{u}) &= \frac{1}{2}\|X^\top\vec{w} - \vec{y}\|_2^2 + \lambda\|\vec{w}\|_1 - \vec{y}^\top\vec{u} + \frac{1}{2}\vec{u}^\top\vec{u} \\ &= \frac{1}{2}\|\vec{y}\|_2^2 - \frac{\lambda\vec{y}^\top\vec{y}}{\lambda_{\max}} + \frac{\lambda^2}{2\lambda_{\max}^2}\vec{y}^\top\vec{y} \\ &= \frac{1}{2}\left(\frac{\lambda}{\lambda_{\max}} - 1\right)^2\|\vec{y}\|_2^2. \end{aligned}$$

Consequently,

$$\sqrt{2G(\vec{w}, \vec{u})}\|\vec{x}_i\|_2 = \left(\frac{\lambda_{\max} - \lambda}{\lambda_{\max}}\right)\|\vec{y}\|_2\|\vec{x}_i\|_2.$$

All together, this gives

$$\begin{aligned} |\vec{x}_i^\top\vec{u}| + \sqrt{2G(\vec{w}, \vec{u})}\|\vec{x}_i\|_2 &= \frac{\lambda}{\lambda_{\max}}|\vec{x}_i^\top\vec{y}| + \left(\frac{\lambda_{\max} - \lambda}{\lambda_{\max}}\right)\|\vec{x}_i\|_2\|\vec{y}\| \\ &\leq |\vec{x}_i^\top\vec{y}| + \|\vec{x}_i\|_2\|\vec{y}\| - \left(\frac{\lambda}{\lambda_{\max}}\right)\|\vec{x}_i\|_2\|\vec{y}\|. \end{aligned}$$

Rearranging, this last term is less than λ if and only if

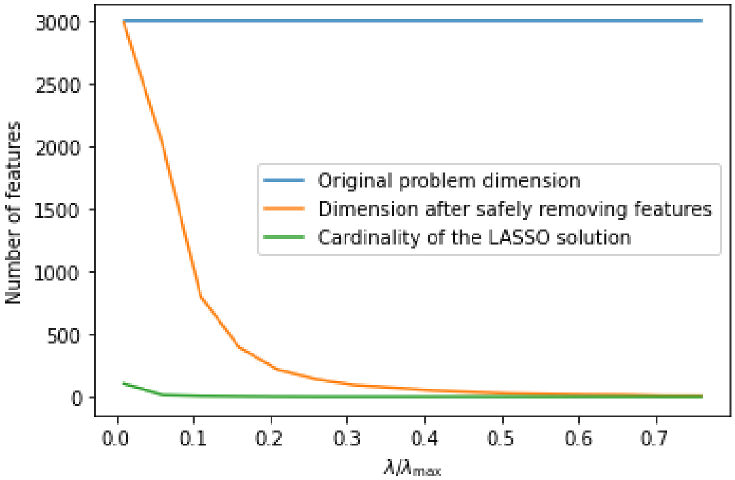
$$\lambda > \frac{|\vec{x}_i^\top\vec{y}| + \|\vec{x}_i\|\|\vec{y}\|}{1 + \frac{\|\vec{x}_i\|_2\|\vec{y}\|}{\lambda_{\max}}},$$

so if the above condition holds, it is safe to remove the i-th feature.

(h) Now, we're going to apply this technique to real data. The `safe_feature_selection.ipynb` python notebook downloads and featurizes a sentiment classification task for the IMDB movie review dataset[1]. As λ varies between $\lambda \in [0.01\lambda_{\max}, 0.8\lambda_{\max}]$, apply the condition derived in part 7 to safely eliminate features from the Lasso problem. Concretely, plot (1) the cardinality of the LASSO solution, (2) the number of features remaining after safely removing features using part 7, and (3) the original number of features as a function of the regularization parameter λ for $\lambda \in [0.01\lambda_{\max}, 0.8\lambda_{\max}]$.

Solution:

[1] https://ai.stanford.edu/ amaas/data/sentiment/



The code generating this plot is in `safe_feature_selection_solution.ipynb`

6. Connecting Ridge Regression, LASSO, and Constrained Least Squares

This question aims to help you develop an understanding of how a constraint in an optimization problem has the same effect as a penalty term in the objective.

(a) Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be strictly convex and such that $\lim_{\alpha\to\infty} f(\alpha\vec{v}) = \infty$ for any nonzero $\vec{v} \in \mathbb{R}^n$. Let $g\colon \mathbb{R}^n \to \mathbb{R}_+$ be convex and take non-negative values. Further, suppose that there exists $\vec{x}_0 \in \mathbb{R}^n$ such that $g(\vec{x}_0) = 0$.

For $\lambda \geq 0$ and $k \geq 0$, define the "penalty" and "constraint" programs

$$P(\lambda) \doteq \operatorname*{argmin}_{\vec{x}}\{f(\vec{x}) + \lambda g(\vec{x})\} \tag{38}$$

$$C(k) \doteq \operatorname*{argmin}_{\vec{x}\colon\, g(\vec{x})\leq k} f(\vec{x}). \tag{39}$$

Show that:

- for every $\lambda \geq 0$ there exists $k \geq 0$ such that $P(\lambda) = C(k)$, and
- for every $k > 0$ there exists $\lambda \geq 0$ such that $P(\lambda) = C(k)$.

HINT: First show using strict convexity that, for $k \geq 0$ and $\lambda \geq 0$, both $P(\lambda)$ and $C(k)$ have exactly one element, i.e., each problem has exactly one optimal solution. You may use without proof that $P(\lambda)$ and $C(k)$ have at least one element each (this is true from assumptions but requires some analysis to show).

To show the first direction (i.e. for all λ there exists k...), let $\vec{x}^\star \in P(\lambda)$ and show that $\vec{x}^\star \in C(k)$ for $k = g(\vec{x}^\star)$. You might need the fact that $P(\lambda)$ and $C(k)$ have exactly one element. To show the other direction (i.e. for all k there exists λ...), prove that strong duality holds for the constraint problem, let $\vec{x}^\star \in C(k)$ and $\mu^\star$ be optimal primal and dual variables for the constraint problem and show that $\vec{x}^\star \in P(\lambda)$ for $\lambda = \mu^\star$.

Solution: Assume $k \geq 0$ and $\lambda \geq 0$. Since f is strictly convex and g is convex, both problems are convex with strictly convex objective. Thus both problems have at most one solution. Since we know from the assumptions that both problems have at least one solution, $P(\lambda)$ and $C(k)$ have exactly one element each.

We first claim that for every $\lambda \geq 0$ there exists $k \geq 0$ such that $P(\lambda) = C(k)$. Let $\lambda \geq 0$ and let $\vec{x}^\star \in P(\lambda)$, so that $P(\lambda) = \{\vec{x}^\star\}$. Set $k = g(\vec{x}^\star)$. Certainly $\vec{x}^\star$ is feasible for the constraint problem with this k. We claim that $\vec{x}^\star \in C(k)$, so that $C(k) = \{\vec{x}^\star\}$. Suppose for the sake of contradiction that $C(k) = \{\vec{z}\}$ where $\vec{z} \neq \vec{x}$. Then $\vec{z}$ must be feasible for the constraint problem, so $g(\vec{z}) \leq k = g(\vec{x}^\star)$, and it must be better than $\vec{x}^\star$, so $f(\vec{z}) < f(\vec{x}^\star)$. Then we have

$$\begin{aligned} f(\vec{z}) + \lambda g(\vec{z}) &< f(\vec{x}^\star) + \lambda g(\vec{z}) && (40)\\ &\leq f(\vec{x}^\star) + \lambda g(\vec{x}^\star) && (41)\end{aligned}$$

so $\vec{x}^\star \notin P(\lambda)$, a contradiction. Thus $\vec{x}^\star \in C(k)$, i.e., $C(k) = \{\vec{x}^\star\}$.

Now we claim that for every $k > 0$ there exists $\lambda \geq 0$ such that $P(\lambda) = C(k)$. We know that there is a point $\vec{x}_0$ such that $g(\vec{x}_0) = 0 < k$. Thus the constraint problem has a strictly feasible point. Since the constraint problem is a convex problem with a strictly feasible point, strong duality holds for it. The Lagrangian of the constraint problem is given by

$$L_k(\vec{x}, \mu) = f(\vec{x}) + \mu(g(\vec{x}) - k). \tag{42}$$

Since strong duality holds, let $(\vec{x}^\star, \mu^\star)$ be optimal primal and dual variables for the constraint problem (so that $\vec{x}^\star \in C(k)$, i.e., $C(k) = \{\vec{x}^\star\}$) such that

$$p^\star = L_k(\vec{x}^\star, \mu^\star) = d^\star. \tag{43}$$

Since the constraint problem is convex, and strong duality holds, we have

$$\begin{aligned}
\vec{x}^\star &\in \operatorname*{argmin}_{\vec{x}} L_k(\vec{x}, \mu^\star) && (44)\\
&= \operatorname*{argmin}_{\vec{x}}\{f(\vec{x}) + \mu^\star(g(\vec{x}) - k)\} && (45)\\
&= \operatorname*{argmin}_{\vec{x}}\{f(\vec{x}) + \mu^\star g(\vec{x}) - \mu^\star k\} && (46)\\
&= \operatorname*{argmin}_{\vec{x}}\{f(\vec{x}) + \mu^\star g(\vec{x})\} && (47)\\
&= P(\mu^\star), && (48)
\end{aligned}$$

so $\vec{x}^\star \in P(\lambda)$, i.e., $P(\lambda) = \{\vec{x}^\star\}$, with $\lambda = \mu^\star$.

Finally, we give some clarity on the fact that $P(\lambda)$ and $C(k)$ have at least one element each. This requires some analysis to prove, and so **it is all out of scope of the course**. We know that convex functions (hence also strictly convex functions) are continuous (see this MathStackExchange link). The assumption that $f(\vec{x}) \to \infty$ as $\|\vec{x}\|_2 \to \infty$ is called "coercivity". We know that continuous coercive functions on closed sets (such as $\mathbb{R}^n$ and $\{\vec{x} \in \mathbb{R}^n \mid g(\vec{x}) \leq 0\}$) attain their global minima, i.e., their problems have at least one solution (see this MathStackExchange link). This is how we are able to say that $P(\lambda)$ and $C(k)$ have at least one element each.

Let $A \in \mathbb{R}^{m \times n}$ have full column rank, and let $\vec{y} \in \mathbb{R}^m$. In the course, we have looked at LASSO:

$$\text{LASSO}(\lambda) \doteq \operatorname*{argmin}_{\vec{x}} \left\{\|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_1\right\} \tag{49}$$

and ridge regression:

$$\text{Ridge}(\lambda) \doteq \operatorname*{argmin}_{\vec{w}} \left\{\|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2\right\} \tag{50}$$

which add an ℓ^1 and ℓ^2 norm penalty to the least squares objective, respectively. The analogous constrant programs are the ℓ^1- and ℓ^2-constrained least squares problems:

$$\begin{aligned}
\ell^1\text{CLS}(k) &\doteq \operatorname*{argmin}_{\vec{x}\colon \|\vec{x}\|_1 \leq k} \|A\vec{x} - \vec{y}\|_2^2 && (51)\\
\ell^2\text{CLS}(k) &\doteq \operatorname*{argmin}_{\vec{x}\colon \|\vec{x}\|_2^2 \leq k} \|A\vec{x} - \vec{y}\|_2^2. && (52)
\end{aligned}$$

(b) Show that the result from part (a) can be used to show the equivalence of LASSO with ℓ^1CLS and the equivalence of ridge regression with ℓ^2CLS. Namely, for each pair of equivalent formulations, find f and g, prove that f is strictly convex, prove that g is convex, and prove that there is an $\vec{x}_0$ such that $g(\vec{x}_0) = 0$.

Solution: For all formulations,

$$f(\vec{x}) \doteq \|A\vec{x} - \vec{y}\|_2^2 \tag{53}$$

has Hessian

$$\nabla_{\vec{x}}^2 f(\vec{x}) = 2A^\top A \tag{54}$$

which is PD since A has full column rank. Thus f is strictly convex (and in fact is strongly convex), and in particular $f(\alpha\vec{v}) \to \infty$ as $\alpha \to \infty$ for any $\vec{v} \neq \vec{0}$.

In the LASSO-ℓ^1CLS formulations, $g(\vec{x}) = \|\vec{x}\|_1$, which is convex since it is a norm. In the LASSO-ℓ^2CLS formulations, $g(\vec{x}) = \|\vec{x}\|_2^2$, which is (strictly) convex since it has Hessian $2I$ which is PD.

For both g, $g(\vec{0}) = 0$.

(c) Complete the Jupyter notebook, which will use this equivalence to show geometrically why LASSO solutions tend to be sparse (i.e. have many zeros) while ridge regression doesn't, and attach a PDF printout of your answers.

Solution: See the Jupyter notebook solutions.

7. Homework Process

With whom did you work on this homework? List the names and SIDs of your group members.

NOTE: If you didn't work with anyone, you can put "none" as your answer.