AMATH 515 Homework #3 - Vinsensius
Due: Friday, March 3rd, by 11:59 pm

Number 1

Let f be a closed proper convex function. The convex conjugate of f, called f^*, is defined by

$$f^*(z) = \sup_x \left\{ z^T x - f(x) \right\}.$$

Compute the conjugates of the following functions.

(a) $f(x) = \delta_{\mathbb{B}_\infty}(x)$.

(b) $f(x) = \delta_{\mathbb{B}_2}(x)$.

(c) $f(x) = \exp(x)$.

(d) $f(x) = \log(1 + \exp(x))$

(e) $f(x) = x \log(x)$

(a) We can begin with scalar case

$$f^*(z) = \sup_x zx - \delta_{\mathbb{B}_\infty}(x)$$

We know that the function will be maximized when $x \in \delta_{\mathbb{B}_\infty}(x)$ or $x \in [-1, 1]$. So the function reduces to

$$f^*(z) = \sup_{x \in [-1,1]} zx$$

Clearly, the function will be maximized when x is either -1 or 1 depending on the sign of z. If $z > 0$, the sup is when x=1, while if $z < 0$, the sup is when $x = -1$. Thus, we can rewrite the result as

$$f^*(z) = \sup_{x \in [-1,1]} zx = |z|$$

For the vector case,

$$f^*(z) = \sup_x z^T x - \delta_{\mathbb{B}_\infty}(x)$$

Using similar argument we can set $x \in [-1.1]^n$ such that the indicator function goes away. Thus, we only need to figure out the rest of the function

$$f^*(z) = \sup_{x \in [-1,1]^n} z^T x$$

We can rewrite the dot product as summation,

$$\begin{aligned} f^*(z) &= \sup_{x \in [-1,1]^n} z^T x \\ &= \sup_{x_i \in [-1,1]} \sum_{i=1}^{n} z_i x_i \\ &= \sum_{i=1}^{n} \sup_{x_i \in [-1,1]} z_i x_i \end{aligned}$$

We know from the scalar case, $\sup_{x_i \in [-1,1]} z_i x_i = |z_i|$. So, the equation becomes

$$\begin{aligned} f^*(z) &= \sum_{i=1}^{n} \sup_{x_i \in [-1,1]} z_i x_i \\ &= \sum_{i=1}^{n} |z_i| \\ &= \|z\|_1 \quad \text{(By definition of 1-norm)} \end{aligned}$$

(b) The convex conjugate of $f(x) = \delta_{\mathbb{B}_2}(x)$

$$f^*(z) = \sup_x z^T x - \delta_{\mathbb{B}_2}(x)$$

By same argument as previous part, the expression is maximized when $x \in \mathbb{B}_2$ or $\|x\|_2 \leq 1$.So, the equation becomes

$$\begin{aligned} f^*(z) &= \sup_{\|x\| \leq 1} z^T x \\ &\leq \sup_{\|x\| \leq 1} \|z\| \|x\| \quad \text{(By Cauchy-Schwartz inequality)} \\ &= \|z\|_2 \end{aligned}$$

So,

$$f^*(z) = \sup_{\|x\| \leq 1} z^T x$$

$$= \|z\|_2$$

(c) The convex conjugate of f(x)

$$f^*(z) = \sup_x zx - \exp(x)$$

We will assume that $z, x \in \mathbb{R}$. Thus, the problem is a scalar optimization. Then, we can take the gradient of the objective function and set it to 0 since it is continuous.

$$\begin{aligned} \nabla(zx - \exp(x)) &= 0 \\ z - \exp(x) &= 0 \\ x^* &= \ln(z) \end{aligned}$$

Where x^* is the optimal x. We can see that the optimal x only valid for $z > 0$. So, we need to consider cases where $z = 0$ and $z < 0$. For case $z > 0$, $f^*(z) = z\ln(z) - z$.

For the case $z = 0$, the expression $f^*(z)$ becomes $f^*(z) = \sup_x -\exp(x)$, so the we can set $x \to -\infty$. Thus, for $z = 0$, $f^*(z) = 0$.

For the case $z < 0$, the term $zx < 0$ in the $f^*(z)$. To make the expression be as positive as possible, set $x \to -\infty$. So, $f^*(z) = \infty$.

Combining the results, the $f*(z)$ of $\exp(x)$ is

$$f^*(z) = \begin{cases} z\ln(z) - z & z > 0 \\ 0 & z = 0 \\ \infty & z < 0 \end{cases}$$

(d) The convex conjugate of f(x) is

$$f^*(z) = \sup_x z^T x - \ln(1 + \exp(x))$$

Similar to previous problem, we will assume z,x are scalars.

We can apply gradient on the function to see the restriction on z and consider edge cases from there.

So, when we take the gradient on the objective and set it to 0, we get

$$\begin{aligned} \nabla(zx - \ln(1 + \exp(x))) &= 0 \\ z - \frac{\exp(x)}{1 + \exp(x)} &= 0 \end{aligned}$$

$$x^* = \ln\left(\frac{z}{1-z}\right)$$

So, the optimal point x^* exists when $z \in (0, 1)$.

For z= 0, we can set $x \to -\infty$, so $f^*(z) = 0$.

For z=1, we can rewrite the equation $f^*(z)$

$$\begin{aligned} f^*(z) &= \sup_x x - \ln(1 + \exp(x)) \\ &= \sup_x x - \ln(\exp(x)(1 + \exp(-x))) \\ &= \sup_x x - \ln(1 + \exp(-x)) - x \\ &= \sup_x - \ln(1 + \exp(-x)) \end{aligned}$$

So, $f^*(z) = 0$ by setting $x \to \infty$ for the case z=1.

For $z < 0$, set $x \to -\infty$, meaning that $f^*(z) = \infty$ since it is unbounded.

For $z > 1$, set $x \to \infty$, meaning that $f^*(z) = \infty$ since it is already unbounded.

Combining the results we get

$$f^*(z) = \begin{cases} z \ln\left(\frac{z}{1-z}\right) - \ln\left(\frac{1}{1-z}\right) & z \in (0, 1) \\ 0 & z = 0, 1 \\ \infty & \text{else} \end{cases}$$

(e) The convex conjugate of f(x)

$$f^*(z) = \sup_x z^T x - x \ln(x)$$

We will again assume that $z, x \in \mathbb{R}$.

Similar to previous problem, we will take the gradient of the function and set it to 0.

$$\begin{aligned} \nabla(zx - x\ln(x)) &= 0 \\ z - 1 - \ln(x) &= 0 \\ x^* &= \exp(z - 1) \end{aligned}$$

where z is continuous so, we can just substitute the optimal x^* into the $f^*(z)$ equation

and we get

$$f^*(z) = z^T \exp(z-1) - (z-1)\exp(z-1)$$

Number 2

Let g be any convex function; f is formed using g. Compute f^* in terms of g^*.

(a) $f(x) = \lambda g(x)$.

(b) $f(x) = g(x-a) + \langle x, b\rangle$.

(c) $f(x) = \inf_z \{g(x,z)\}$.

(d) $f(x) = \inf_z \left\{\frac{1}{2}\|x-z\|^2 + g(z)\right\}$

(a) The convex conjugate of f(x) is

$$\begin{aligned} f^*(z) &= \sup_x z^T x - \lambda g(x) \\ &= \lambda \sup_x \frac{z^T}{\lambda} x - g(x) \\ &= \lambda g^*\left(\frac{z}{\lambda}\right) \quad \text{(By definition of convex conjugate)} \end{aligned}$$

(b) The convex conjugate of f(x) is

$$f^*(z) = \sup_x z^T x - g(x-a) - x^T b$$

We can do change of variable $w = x - a$, and will maximized over variable w. The equation becomes

$$\begin{aligned} f^*(z) &= \sup_w z^T(w+a) - g(w) - (w+a)^T b \\ &= \sup_w w^T(z-b) - g(w) + a^T(z-b) \\ &= \sup_w w^T(z-b) - g^{**}(w) + a^T(z-b) && \text{(convexity of g)} \\ &= g^*(z-b) + a^T(z-b) && \text{(definition of } g^*\text{)} \end{aligned}$$

(c) The convex conjugate of f(x) is

$$\begin{aligned} f^*(z) &= \sup_x z^T x - \inf_w g(x,w) \\ &= \sup_x z^T x + \sup_w -g(x,w) \\ &= \sup_{x,w} z^T x + 0^T w - g(x,w) \end{aligned}$$

$$= g^*(z, 0) \quad \text{(convex conjugate definition)}$$

(d) The convex conjugate of f(x) is

$$\begin{aligned} f^*(z) &= \sup_x z^T x - \inf_w \left(\frac{1}{2}\|x - w\|^2 + g(w) \right) \\ &= \sup_{x,w} z^T x - \frac{1}{2}\|x - w\|^2 - g(w) \end{aligned}$$

Define $\tilde{z} = x - w$, so the equation becomes

$$\begin{aligned} f^*(z) &= \sup_{x,\tilde{z}} z^T(\tilde{z} + w) - \frac{1}{2}\|\tilde{z}\|^2 - g(w) \\ &= \sup_{x,\tilde{z}} z^T\tilde{z} - \frac{1}{2}\|\tilde{z}\|^2 + z^T w - g(w) \\ &= \frac{1}{2}\|z\|^2 + g^*(z) \quad \text{(Definition of } g^*) \end{aligned}$$

Number 3

Moreau Identities.

(a) Derive the Moreau Identity:

$$\text{prox}_f(z) + \text{prox}_{f^*}(z) = z.$$

(b) Use the Moreau identity and 1a, 1b to check your formulas for

$$\text{prox}_{\|\cdot\|_1}, \quad \text{prox}_{\|\cdot\|_2}$$

from last week's homework.

(a) For this problem, we will assume that f is convex.

Using the definition of $\text{prox}_f(z)$

$$\text{prox}_f(z) = \arg\min_x \frac{1}{2}\|z - x\|^2 + f(x)$$

Let $z^* = \text{prox}_f(z)$Applying the optimality condition, we have

$$\begin{aligned} 0 &\in -(z - z^*) + \partial f(z^*) \\ y = z - z^* &\in \partial f(z^*) \end{aligned}$$

Since f is convex, we can rewrite the optimality condition as f^*

$$z^* \in \partial f^*(y) = \partial f^*(z - z^*)$$

So, $z - z^*$ is $\text{prox}_{f^*}(z)$.

Thus,

$$\text{prox}_f(z) + \text{prox}_{f^*}(z) = z^* + z - z^* = z$$

(b)

- For $f(x) = \|x\|_1$ and $f^*(x) = \delta_{\mathbb{B}_\infty}(x)$ From the previous homework, the definition of $\text{prox}_{\|\|_1}(z)$

$$\text{prox}_{\|\|_1}(z_i) = \begin{cases} z_i - 1 & z_i > 1 \\ 0 & z_i \in (-1, 1) \\ z_i + 1 & z < -1 \end{cases}$$

From the lecture notes, the definition of $\text{prox}_{\delta_{\mathbb{B}_\infty}}(z)$

$$\text{prox}_{\|\|_1^*}(z_i) = \begin{cases} 1 & z_i > 1 \\ z_i & z_i \in (-1, 1) \\ -1 & z < -1 \end{cases}$$

For each cases, we see that $\text{prox}_{\|\|_1}(z) + \text{prox}_{\delta_{\mathbb{B}_\infty}}(z) = z_i$

- For $f(x) = \|x\|_2$ and $f^*(x) = \delta_{\mathbb{B}_2}(x)$ From the previous homework, the definition of $\text{prox}_{\|\|_2}(z)$

$$\text{prox}_{\|\|_2}(z) = \begin{cases} 0 & \|z\| \leq 1 \\ z - \frac{z}{\|z\|} & \|z\| > 1 \end{cases}$$

From the lecture notes, the definition of $\text{prox}_{\delta_{\mathbb{B}_2}}(z)$

$$\text{prox}_{\|\|_2^*}(z_i) = \begin{cases} z & \|z\| \leq 1 \\ \frac{z}{\|z\|} & \|z\| > 1 \end{cases}$$

For each cases, we see that $\text{prox}_{\|\|_2}(z) + \text{prox}_{\delta_{\mathbb{B}_2}}(z) = z$

Number 4

Duals of regularized GLM. Consider the Generalized Linear Model family:

$$\min_x \sum_{i=1}^n g(\langle a_i, x\rangle) - b^T Ax + R(x),$$

Where g is convex and R is any regularizer.

(a) Write down the dual obtained by dualizing g.

(b) Specify your formula to Ridge-regularized logistic regression:

$$\min_x \sum_{i=1}^n \log(1 + \exp(\langle a_i, x\rangle)) - b^T Ax + \frac{\lambda}{2}\|x\|^2.$$

(c) Specify your formula to 1-norm regularized Poisson regression:

$$\min_x \sum_{i=1}^n \exp(\langle a_i, x\rangle) - b^T Ax + \lambda\|x\|_1.$$

(a) We started with primal problem and want to derive the dual problem of GLM

$$\begin{aligned}
&= \inf_x \sum_{i=1}^n g(a_i^T x) - b^T Ax + R(x) \\
&= \inf_x \sum_{i=1}^n g(a_i^T x) - b^T Ax + R(x) \\
&= \inf_x \sum_{i=1}^n \sup_{w_i} w_i(a_i^T x) - g^*(a_i^T x) - b^T Ax + R(x) \quad \text{(Def of } g^*\text{)} \\
&= \inf_x \sup_{w_1,\ldots,w_n} \sum_{i=1}^n w_i(a_i^T x) - g^*(w_i) - b^T Ax + R(x) \\
&\leq \sup_{w_1,\ldots,w_n} \inf_x w^T Ax - b^T Ax + R(x) - \sum_{i=1}^n g^*(w_i) \\
&= \sup_{w_1,\ldots,w_n} \inf_x x^T(Aw - A^T b) + R(x) - \sum_{i=1}^n g^*(w_i) \\
&= \sup_{w_1,\ldots,w_n} -\sup_x x^T(A^T b - A^T w) + R(x) - \sum_{i=1}^n g^*(w_i)
\end{aligned}$$

$$= \sup_{w_1,\ldots,w_n} -R^*(A^Tb - A^Tw) + \sum_{i=1}^{n} g^*(w_i) \quad \text{(Def of } R^*\text{)}$$

So, the dual problem is

$$\sup_{w_1,\ldots,w_n} -R^*(A^Tb - A^Tw) + \sum_{i=1}^{n} g^*(w_i)$$

(b) Using the result from the previous part, we can calculate the dual problem of the regularized Ridge logistic regression. The $g(z) = \ln(1 + \exp(z))$ and $R(z) = \frac{\lambda}{2}\|z\|^2$.

We know that $f(z) = \frac{1}{2}\|z\|^2$ is self-conjugate so, $f^*(y) = \frac{1}{2}\|z\|^2$. Thus, $R^*(y) = \frac{\lambda}{2}\|\frac{y}{\lambda}\|^2 = \frac{1}{2\lambda}\|y\|^2$ from previous part. We can also find $g^*(y)$ from previous part.

Combining the results, the dual problem is

$$\sup_y \sum_{i=1}^{n} \begin{cases} y_i \ln\left(\frac{y_i}{1-y_i}\right) - \ln\left(\frac{1}{1-y_i}\right) - \frac{1}{2\lambda}\left\|A^Tb - A^Ty\right\|^2 & y_i \in (0,1) \\ -\frac{1}{2\lambda}\left\|A^Tb - A^Ty\right\|^2 & y_i = 0, 1 \\ \infty & \text{else} \end{cases}$$

where $y_i = a_i x$

(c) Using the result from the previous part, we can calculate the dual of 1-norm regularized Poisson regression problem. The $g(z) = \exp(z)$ and $R(z) = \lambda\|z\|_1$. From previous part we know what is $R^*(y)$ and $g^*(y)$.

$$\sup_y \sum_{i=1}^{n} \begin{cases} y_i \ln(y_i) - y_i - \lambda\delta_{\mathbb{B}_\infty}\left(\frac{A^Tb - A^Ty}{\lambda}\right) & y_i > 0 \\ -\lambda\delta_{\mathbb{B}_\infty}\left(\frac{A^Tb - A^Ty}{\lambda}\right) & y_i = 0 \\ \infty & y_i < 0 \end{cases}$$

where $y_i = a_i x$

Number 5

Coding Assignment

In this problem you will write a routine to project onto the capped simplex.
The Capped Simplex Δ_k is defined as follows:

$$\Delta_k := \{x : 1^T x = k, \quad 0 \leq x_i \leq 1 \quad \forall i.\}$$

This is the intersection of the k-simplex with the unit box.
The projection problem is given by

$$\text{proj}_{\Delta_k}(z) = \arg\min_{x \in \Delta_k} \frac{1}{2}\|x - z\|^2.$$

(a) Derive the (1-dimensional) dual problem by focusing on the $1^T x = k$ constraint.

(b) Implement a routine to solve this dual. It's a scalar root finding problem, so you can use the root-finding algorithm provided in the code.

(c) Using the dual solution, write down a closed form formula for the projection. Use this formula, along with your dual solver, to implement the projection. You can use the unit test provided to check if your code id working correctly.

The primal problem is

$$\text{proj}_{\Delta_k}(z) = \arg\min_{x \in \Delta_k} \frac{1}{2}\|x - z\|^2$$

We can rewrite the constraint $1^T x = k$ as

$$\begin{aligned} 1^T x &= k \\ 1^T x - k = 0 &\implies \delta_0(1^T x - k) \end{aligned}$$

Similarly,we can rewrite the constraint $0 \leq x \leq 1$ as

$$\delta_{[0,1]^n}(x)$$

So, the primal problem becomes

$$\arg\min_x \frac{1}{2}\|x-z\|^2 + \delta_0(1^T x - k) + \delta_{[0,1]^n}(x)$$

We can rewrite the $\delta_0(1^T x - k)$ in terms of its conjugate.
Consider a simple scalar case $\delta_0(x)$ and we want to find $\delta_0^*(z)$.

$$\delta_0^*(z) = \sup_x zx - \delta_0(x)$$

Clearly, the expression above is maximize when $x = 0$. Thus, $\delta_0^*(z) = 0$.
So, $\delta_0(1^T x - k)$ in terms of its conjugate is

$$\delta_0(1^T x - k) = \sup_\lambda \lambda(1^T x - k)$$

So the primal problem becomes

$$\begin{aligned}
&= \arg\min_x \frac{1}{2}\|x-z\|^2 + \sup_\lambda \lambda(1^T x - k) + \delta_{[0,1]^n}(x) \\
&\geq \sup_\lambda \arg\inf_x \frac{1}{2}\|x-z\|^2 + \lambda(1^T x - k) + \delta_{[0,1]^n}(x) \\
&= \sup_\lambda \arg\inf_x \frac{1}{2}\langle x,x\rangle - \langle x,z\rangle + \langle \lambda 1, x\rangle + \delta_{[0,1]^n}(x) + \frac{1}{2}\langle z,z\rangle - \lambda k \\
&= \sup_\lambda \arg\inf_x \frac{1}{2}\langle x,x\rangle - \langle z - 1\lambda, x\rangle + \frac{1}{2}\langle z-\lambda 1, z-\lambda 1\rangle + \delta_{[0,1]^n}(x) + \frac{1}{2}\langle z,z\rangle - \lambda k - \\
&\quad \frac{1}{2}\langle z-\lambda 1, z-\lambda 1\rangle \\
&= \sup_\lambda \arg\inf_x \frac{1}{2}\|x-(z-\lambda 1)\|^2 + \delta_{[0,1]^n}(x) + \frac{1}{2}\|z\|^2 - \lambda k - \frac{1}{2}\|z-\lambda 1\|^2
\end{aligned}$$

So, the dual problem is

$$\sup_\lambda \arg\inf_x \frac{1}{2}\|x-(z-\lambda 1)\|^2 + \delta_{[0,1]^n}(x) + \frac{1}{2}\|z\|^2 - \lambda k - \frac{1}{2}\|z-\lambda 1\|^2$$

We can rewrite the minimization of x part as prox operator

$$x^* = \text{prox}_{\delta_{[0,1]^n}}(z-\lambda 1) = \arg\inf_x \frac{1}{2}\|x-(z-\lambda 1)\|^2 + \delta_{[0,1]^n}(x)$$

$$\text{proj}_{[0,1]^n}(z-\lambda 1) = \max(\min(z-\lambda 1, 1), 0) \quad \text{(From notes projection onto box)}$$

So, x^* is the optimal condition for the primal problem.
Assuming strong duality, we can solve the dual problem through the primal constraint

$1^T x = k$ using the primal solution.

So, the dual solution λ^* is the solution of the equation

$$1^T \max(\min(z - \lambda 1, 1), 0) - k = 0$$

which is a scalar problem and can be found using root-finding algorithm such as bisection.

Once we found optimal λ^*, we can found x^*, which is the solution to the projection problem

$$x^* = \max(\min(z - \lambda^* 1, 1), 0)$$

Number 6

Coding Assignment

In this problem you will learn apply proximal operators to matrices to perform matrix completion. You'll find that you can recover most of the information in a low rank matrix despite only seeing a small percentage of the entries. At the end of this problem, you'll see something that is, in my opinion, **truly remarkable**.

Consider a matrix X, (for example, where it is a matrix of ratings, and $X_{i,j}$ is the rating that user i gave to item j). However, you only observed the entries $(i,j) \in \Omega$. However, we assume that the matrix X is low rank, and we will use this to try prior knowledge to try and recover it. Similar to how the lasso or ℓ_1 penalty promotes sparsity in regression, the nuclear norm $\|\cdot\|_*$ promotes low rank matrices. We will implement and experiment with a handful of approaches to the matrix completion problem.

(a) One natural approach to setting up an optimization problem that models this situation is to assume that X is rank k, and solve the optimization problem

$$\begin{aligned} &\text{minimize } \|P \odot (Y - X)\|_F^2 \\ &\text{subject to } \text{rank}(Y) \leq k \end{aligned} \tag{1}$$

where $\odot$ denotes the elementwise product, and P is a matrix with $P_{i,j} = 1$ when $(i,j) \in \Omega$ and 0 if $(i,j) \notin \Omega$ (essentially a matrix of which entries are measured). **Is this problem convex? Why or why not?**

(b) We relax the constraint above into a nuclear norm constraint

$$\begin{aligned} &\text{minimize } \|P \odot (Y - X)\|_F^2 \\ &\text{subject to } \|Y\|_* \leq K \end{aligned}$$

Is this problem convex? Why or why not? Rather than the problem above, we replace this with a penalized version,

$$\text{minimize } \|P \odot (Y - X)\|_F^2 + \lambda \|Y\|_* \tag{2}$$

because the projection onto the ℓ_1 ball requires a little bit of work to write (though it can be done exactly in $O(n \log(n))$, using a sort+O(n) operations to collect the result, and is faster if most of the entries are zero. This is even

better for us since we're already computing an SVD so the singular values will come sorted.) This approach is known as soft-impute in the literature.

(c) Derive formulas for (or procedures for computing) the following proximal and projection operators

$$\text{prox}_{t\|\cdot\|_*}(Y) = \arg\min_M \frac{1}{2t}\|Y - M\|_F^2 + \|M\|_*$$

$$\text{proj}_{\text{rank}_k}(Y) = \arg\min_{\text{rank}(M)\leq k} \frac{1}{2}\|Y - M\|_F^2$$

$$\text{proj}_{*K}(Y) = \arg\min_{\|M\|_*\leq K} \frac{1}{2}\|Y - M\|_F^2$$

For the third one, you may write the answer in terms of an ℓ_1 ball projection

$$\text{proj}_{\ell_1\leq R}(x) = \arg\min_{\|y\|_1\leq R} \frac{1}{2}\|x - y\|_2^2$$

(d) Suppose we want to solve the optimization problem in (b) for many values of λ to give ourselves the best chance of finding a good recovery (performing some kind of cross validation or scoring to pick the best one. In the coding section, we'll just compare the true reconstruction error $\|X - Y\|_F^2$ for simplicity in order to understand how changing the regularization parameter changes the performance of the model. This data wouldn't be available in practice since we only observe $P \odot X$ but there are ways to approximate that quantity. However, as you'll see, even for small examples, this problem can be somewhat time consuming. **What could we do to speed this up? Do you expect a larger value of λ to make the problem converge faster or slower?**

(e) Suppose that we rather than knowing the observed entries of X exactly, there were some large corrupted entries. **What could we change to make our problem robust to such corruptions?**

(f) Fill in Problem 6 in the jupyter notebook, the nuclear norm and rank projection in proxes.py

There are a couple of other tricks to make this scalable to real world big data sets which we're not making you do in this class. The biggest speed up would come from using a "partial" or "truncated" SVD at each step since

we only care about the top components, along the lines of the algorithm in https://arxiv.org/pdf/1607.03463.pdf. Additionally, because they apply an iterative algorithm, you can both warm start the SVD itself with the previous SVD, and stop computing additional components once you know they will be "proxed" down to zero (either with the nonconvex projection or with nuclear norm prox).

For **extra credit** implement the LazySVD algorithm in this paper link with a warm start for the nuclear norm regularized problem, and adaptively quit computing components when you are sure that the rest will be sent to zero by the prox operator! You may have to write a custom accelerated proximal gradient descent function for this, as you'll need to save the decompositions between iterations!

Ultimately, for the best performance (assuming very low rank and very few observed entries), you would want to avoid multiplying the matrices together, in the SVD, making each matrix vector product $O(kn + |\Omega|)$, where the kn comes from the two factor matrices, and $|\Omega|$ comes from the sparsity of the gradient (the gradient of the smooth part is zero except for where you observed the entries of X).

(a) The problem is not convex because the constraint of $\text{rank}(Y) \leq K$ is not convex while the objective function is convex.

(b) The problem is convex because the constraint is convex since norm is convex and the objective function is convex.

(c) We will use the fact that

$$Y - M = U(\Sigma_Y - \Sigma_M)V^T$$

Where Σ is the diagonal matrix of the singular values.

- For $\text{prox}_{t\|\cdot\|_*}(Y) = \arg\min_M \frac{1}{2t}\|Y - M\|_F^2 + \|M\|_*$, we can rewrite the frobenius norm as $\|Y - M\|_F^2 = \|\sigma_Y - \sigma_M\|_2^2$. We can rewrite the nuclear norm $\|M\|_* = \|\sigma_M\|_1$. So, the prox problem becomes

$$\arg\min_{\sigma_M} \frac{1}{2t}\|\sigma_Y - \sigma_M\|_2^2 + \|\sigma_M\|_1$$

The problem just becomes a projection problem on L1-norm which we have done

in the previous homework. The optimal $\sigma^*_{i,M}$ is

$$\sigma^*_{i,M} = \begin{cases} \sigma_{i,Y} - t & \sigma_{i,Y} > t \\ 0 & |\sigma_{i,Y}| \leq t \\ \sigma_{i,Y} + t & \sigma_{i,Y} < -t \end{cases}$$

Then, optimal $M^* = U\text{diag}(\sigma^*_M)V^T$.

- For $\text{proj}_{\text{rank}_k}(Y) = \arg\min_{\text{rank}(M)\leq k} \frac{1}{2}\|Y - M\|_F^2$. We are essentially doing a low rank approximation of Y to create M such that the rank$(M) \leq K$ and still retain the same dimension.

  We know that rank of a matrix is the sum of nonzero singular values of matrix. Since singular values are arranged in descending order, we can zero out the singular values after k-th element.

  We can construct the matrix M^* by similar method as previous part.

- For the problem $\text{proj}_{*K}(Y) = \arg\min_{\|Y\|_*\leq K} \frac{1}{2}\|Y - M\|_F^2$, we can think of this problem as projection on the L1-norm ball since nuclear norm is a 1-norm of singular vector. So, we can rewrite the problem as

$$\text{proj}_{l_1\leq K}(Y) = \arg\min_{\|\sigma_{1,M}\|\leq K} \frac{1}{2}\|\sigma_Y - \sigma_M\|_F^2$$

(d) To speed the algorithm up, we can just compute the partial SVD for the nuclear norm for each iteration instead of doing full SVD because it is very expensive to do it in each iteration.

Larger value of λ will make the problem converge faster because of high importance of having low nuclear norm for the problem. As a result, the error increases. With low λ it will take the problem longer to solve because of low penalty of having large nuclear norm.

(e) Use non-convex penalty problem to deal with the corrupted elements (Enhanced Low-Rank Matrix Approximation).