Recap of Lecture 3

- general structure of convex optimization
- properties that make them attractice
  - local minima are global minima
  - even if solution is non unique, the set of solutions is a convex set
- First order optimality condition
  - $\nabla f(x^*)(y - x^*) \geqslant 0$
  - negative gradient has to push you out of the domain set
  - if an interior point is optimal, then the gradient is zero
- Rewriting convex optimization problems
  - eliminating constraints
  - partial optimization

A. Relaxations

- expand the set over which you search for solutions
- you are guaranteed at least as good a solution as to the original problem
- $min\ f(x)\ subject\ to\ x \in C = min\ f(x)\ subject\ to\ x \in C_2,\ C_2 \supseteq C$
  - this can be a simpler problem to solve!
- Many times, this solution can be tight i.e. it is the same solution as you would have gotten to the original problem

A1. Relaxing non-affine equalities

Problem set up:

$min\ f(x)$
$subject\ to$
$g_i(x) = 0,\ g_i\ are\ convex\ but\ not\ affine$
$Ax = b$
$h_i(x) \leqslant 0$

This is not a convex opt problem because the set of solutions is not convex any longer

- Q: why? Proof?

The classic relaxation here is to relax the equalities to become inequalities, i.e.

$min\ f(x)$
$subject\ to$
$g_i(x) \leqslant 0,\ g_i\ are\ convex\ but\ not\ affine$
$Ax = b$
$h_i(x) \leqslant 0$

- In general this relaxation is not tight, but there are cases when it is

Example 1: Investment/Expense optimization
(problem from BV)

$$max \sum \alpha_i u(x_i)$$
$$such\ that$$
$$x_i \geqslant 0$$
$$b_{t+1} = b_t + f(b_t) - x_t$$

You have an initial budget b_0
In every cycle you invest your budget for the cycle and spend some amount
x is the expense
f() is the return on investment, this is increasing and concave
u() is the utility function, also concave

Fact 1: This is not a convex program

- The criterion function is concave, but we are maximizing it, so that is ok
- But the constraint $b_{t+1} = b_t + f(b_t) - x_t$ is not convex!

So we can try relaxing this constraint to
$b_{t+1} \leqslant b_t + f(b_t) - x_t$

- This is the same as $b_{t+1} - b_t - f(b_t) + x_t \leqslant 0$
  - negative of concave function is convex
- argue that optimum there wont be any wastage, i.e. there will be equality, so we are solving the original problem as well - so this relaxation is tight

Example 2: Low rank matrix approximation
Frobenius norm of a matrix = 2norm if you turned the matrix into a long vector

- https://www.youtube.com/watch?v=Gt56YxMBlVA&ab_channel=SteveBrunton

Problem setup:
Given fixed matrix $X \in R^{n\ x\ d}$
Find a low rank representation R
$min_R\ ||X - R||_F^2\ such\ that\ rank(R) = k$

This is a non convex problem. Why?

- The objective is convex
- But the rank(R) function is not a convex function!
- Why?
  - Find examples to the contrary.

  - For example: matrix A can have rank r. But A - A will have rank 0
  - Or you can add two matrices, each having rank 1, and get a matrix of rank 2

Arguably one of the most important non convex problems that we know the solution to!
- The solution is the SVD.
- The optimal $R^* = U_k D_k V_k^T$ i.e. the truncated SVD.
  - When we compute the SVD of a variance covariance matrix, we call it principal components analysis

Take another approach to this problem.
Let $R = XZ$ where Z is a projection matrix into some low rank subspace

For a projection matrix like Z:
- ZZ^T is by definition idempotent i.e. $ZZ^T = Z$
- Therefore, the eigenvalues of a projection matrix are 0 and 1
  - https://math.stackexchange.com/questions/1157589/find-the-eigenvalues-of-a-projection-operator
- Because of this, the tr(Z) = rank(Z)
  - the sum of the eigenvalues is hte same as the number of non zero eigenvalues, since every eigenvalue is either 0 or 1

So the problem is

$$min_Z \, ||X - XZ||_F^2, \; Z \; is \; a \; projection, \; rank(Z) = k$$
$$= min_Z \, Tr\left[(X - XZ)^T (X - XZ)\right] \; by \; definition$$
$$= min_Z \, Tr\left[X^T X - X^T XZ - Z^T X^T X + Z^T X^T XZ\right]$$
$$\equiv max_Z \, Tr\left[X^T XZ\right]$$

- Idk how to make that last step work? Presumably it ends with something of the form $min \; -Tr\left[X^T XZ\right]$

Subject to constraints $\{Z \in Symmetric, \; rank(Z) = tr(Z) = k, \; \lambda_i(Z) = \{0, 1\}\}$
- Only the last constraint is non convex!
- So we relax it to be $0 \leqslant \lambda_i(Z) \leqslant 1, \; or \; 0 \preccurlyeq Z \preccurlyeq I$

This constraint set is called the Fantope.

If V_k is unique, Z will be tight.

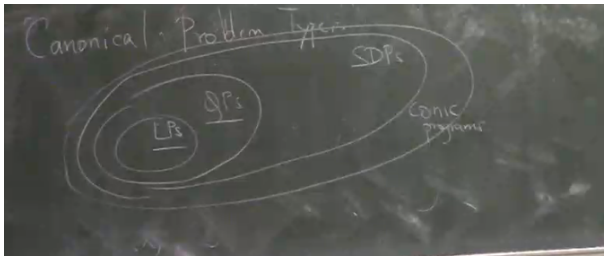
Why do you do this?
- The convex nature of this problem makes it easier to think about regularization
  - for example to solve a sparse PCA

- Sparse PCA: solve $max\ Tr\left[X^T XZ\right]\ +\ \lambda|Z|_1,\ Z\ \in Fantope$
    - * we want the V coming out of the SVD to have many zeros
- in the unconstrained case, the relaxation is tight

B. Canonical Classes of Convex Programs

some types of convex programs are easier to solve than others

Hierarchy from easiest to hardest to solve:
LP = objective linear constraints linear
QP = objective quadratic constraints linear
SDP = Semi definite programs
Conic Programs



B1. Linear Programs

- Criterion is linear and constraints are linear

General form of LP

$$min\ c^T x$$
$$subject\ to$$
$$Ax\ =\ b$$
$$Cx\ \leqslant\ d$$

- The inequality constraints define a *polytope*
- Each constraint says you need to be on some side of a plane
- In the end the feasible set are i the intersection of a bunch of half spaces = polytope
- iThe optimum is usually a vertex of the polytope

Standard form of the LP

$$min\ c^T x$$
$$subject\ to$$
$$Ax\ =\ b$$
$$x\ \geqslant\ 0$$

The general and standard forms are equivalent.

You go from G to S by introducing some slack variables.

LPs are used for a lot of OR/planning problems.

Example: Diet problem

d possible things you can eat
cost of each is c_i
D_{ij} = amount of nutrient i in food j
d_i = min amount on nutrient i you need

$$min\ c^T x$$
$$subject\ to\ Dx\ \geqslant\ d$$
$$x\ \geqslant\ 0$$

Example 2: Basis pursuit
Solve a system of linear equations that are under determined.
$y\ \in R^n,\ X \in R^{n\ x\ m},\ p\ \gg n$

model is $y\ =\ XB$ but there are many B that will solve this system
You want to find some solution that you like - for example the sparsest solution
Problem:

$$min\ ||B||_0\ =\ \sum 1\{B_j\ \neq 0\}$$
$$subject\ to$$
$$y\ =\ XB$$

This is a non convex program because the L0 norm is non convex.

- same reason that the rank() is non convex.
- You can take two vectors that have for e.g. L0 = 1, add them, and get a vector that has L0 > 1

Can we relax this problem to make it a convex program?
Instead of minimizing L0 norm, can you minimize the convex hull of the L0 norm?

- This is the same as L1 norm!

New problem:

$$min \sum_{j=1}^{p} ||B_j||_1$$
$$such\ that\ y\ =\ XB$$

This problem is called basis pursuit.

The claim is that this problem is an LP even if it doesn't look like it!
How do we see this?
Introduce proxies for the absolute variables to form an LP with linear cost and constraints!

$$min_{B,\,Z} \sum_{i=1}^{p} z_i$$
$$such\ that$$
$$Z_i \geqslant B_i$$
$$Z_i \geqslant -B_i\ [sic]$$
$$y = XB$$

Even though Z can be $\gg$ B, at optimum it shuold be on the boundary i.e. equal

- Questions
  - should the second constraint not be -Z?

Example 3: Dantzig selector
The objective is noise reduction
$y = XB + noise$, you want to remove said noise
One way to do this is LASSO - how?
This is another.

To motivate the Dantzig selector, we will first look at unconstrained least squares.
The unconstrained LS problem is $min_B\ ||y - XB||_2^2$
You can solve this by setting the gradient of the objective wrt B to zero (from first-order characterization)

$$||y - XB||_2^2 = (y - XB)^T(y - XB)$$
$$= y^Ty - y^TXB - B^TX^Ty + B^TX^TXB$$
$$So\ grad\ wrt\ B = -2X^Ty + 2X^TXB = 2\left(X^TXB - X^Ty\right) = 0\ for\ optimum$$
$$so\ \nabla f(B) = X^T(y - XB) = 0$$

In the Dantzig selector, we don't set this gradient to zero.
Instead, we say it can be non-zero, but we want B to have small L1 norm.
So we trade off the gradient quantity for the B L1 norm.

Problem:
$$min\ ||B||_1$$
$$such\ that$$

$$||X^T(y - XB)||_{inf} \leqslant \lambda = max\ absolute\ value\ in\ the\ vector$$

where λ is a tuning parameter.
One way of thinking about is is: every predictor has a small correlation with the residual.

- 'if any column had a large correlation with the residual, you shuold include more of that column in your regressor'
- Stop including columns in your regression when columns are not predictive of the residual
  - i dont understand this intuition

LASSO is a QP, this is an LP.

B2. Quadratic programs
Quadratic convex objective + linear constraints

$$min\ c^T x + \frac{1}{2} x^T Q x$$
$$such\ that$$
$$Ax = b$$
$$Cx \leqslant d$$
$$Q \succcurlyeq 0\ i.e.\ Q\ is\ psd$$

- As a reminder, if Q is not psd then this is not a convex objectice function - the Hessian of the objective must be psd, and Q is the Hessian

- these problems can be solved fairly quickly

Example 1: Least squares
Example 2: LASSO
Example 3: SVM
Example 4: Markowitz Portfolio Optimization

- some mean and covariance between stocks
- Σ = covariance between stocks
- μ = mean return of each stock
- want to max return and minimize volatility
- x = money you put into each stock

$$max_x\ \mu^T x - \frac{\gamma}{2} x^T \Sigma x$$
$$such\ that$$
$$x \succcurlyeq 0\ (no\ short\ selling)$$
$$\sum x_i = B\ (budget)$$

Q: what does it intuitively mean to do $x^T \Sigma x$?

- https://stats.stackexchange.com/questions/326508/intuitive-meaning-of-vector-multiplication-with-covariance-matrix
- However I already have an understanding of the term xTAx in the following sense: I see it as the dot product between a vector and it's image according to a linear map A. Therefore it can be interpreted as the correlation of x and Ax. If its absolute value is small (or 0), then Ax is (almost) orthogonal to x. A positive definite matrix therefore "upholds the orientation" (i.e. it doesn't change direction). However this doesn't help me for my question (as I don't know what r should be). Nevertheless I will have a look at your provided links

Answer in the link:

Your answer is good. Note that since Σ is symmetric and square so is Σ^{-1}. The matrix, its transpose, or inverse all project your vector Σr in the same space.

Since Σ and Σ^{-1} are positive definite, all eigenvalues are positive. Thus a multiplication with a vector always ends up in the same halfplane of the space.

Now if Σ or Σ^{-1} would be a diagonal matrix, then the multiplication would reweigh (or undo the reweigh) of only the lengths of the target vector in each dimension (as you noticed). If they are full matrices, then indeed the matrix is full rank as it is PSD, the eigendecomposition exists and $\Sigma=V\Lambda V^{-1}$, here V is an orthonormal eigenvector matrix by the virtue of Σ being PSD, and Λ the diagonal with eigenvalues. Thus r is first rotated by V^{-1}, and then reweighed by Λ, then rotated back by V. The same thing goes for Σ^{-1}, but then r is rotated the other way around and the scaled by the diagonal of reciprocals Λ^{-1} and rotated back with V^{-1}. It is easy to see they are opposite processes.

Additionally, you may think of
$r^T\Sigma^{-1}r=(\Lambda^{-1/2}V^Tr)^T(\Lambda^{-1/2}V^Tr)=\|\|\Lambda^{-1/2}V^Tr\|\|^2$
as the length of your vector r reweighed by the "standard deviations" after correction for cross-correlations.

Hope that helps.

Working through this myself:

$$\Sigma x = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \sigma_1\sigma_3 \\ \sigma_2\sigma_1 & \sigma^2 & \sigma_2\sigma_3 \\ \sigma_3\sigma_1 & \sigma_3\sigma_2 & \sigma_3^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \sigma_1 x_1^2 + \sigma_1\sigma_2 x_2 + \sigma_1\sigma_3 x_3 \\ \dots \\ \dots \end{bmatrix}$$

every element of x weighted by covariance
Then

$$x^T\Sigma x = [x_1\ x_2\ x_3]\begin{bmatrix}\sigma_1 x_1^2 + \sigma_1\sigma_2 x_2 + \sigma_1\sigma_3 x_3\\ \dots\\ \dots\end{bmatrix} = \begin{bmatrix}x_1\left(\sigma_1 x_1^2 + \sigma_1\sigma_2 x_2 + \sigma_1\sigma_3 x_3\right)\\ \dots\\ \dots\end{bmatrix}$$

Σ is the variance covariance matrix

The eigendecomposition will be $\Sigma = Q\Lambda Q^{-1}$

so $\Sigma r = Q\Lambda Q^{-1} r$

i.e. rotate r in the direction of the principal components of x, stretch r by the variance, then rotate it back to the original orientation

Then $r^T\Sigma r$ is ?

$r^T\Sigma r = r^T\left(Q\Lambda Q^{-1}\right)r = r^T Q\Lambda^{0.5}\Lambda^{0.5}Q^{-1}r = \left(\Lambda^{0.5T}Q^T r\right)^T\left(\Lambda^{0.5T}Q^T r\right) = ||\Lambda^{0.5T}Q^T r||_2^2$

= length of vector r reweighed by standard deviations after correction for cross correlations

B3. Semi definite programs

Cones

<u>Standard form:</u>

optimize over some matrix variable, solve something that looks like an LP

For some matrix X

$min\ C.X$

such that

$A_i.X = B_i$

$X \succcurlyeq 0$

$C.X$

where $= \sum C_{ij}X_{ij}$ (*string out the two matrices as vectors and get the inner product*)

$= tr\left(C^T X\right)$

Why?

$C = \begin{bmatrix}a_1 & b_1\\ c_1 & d_1\end{bmatrix}, X = \begin{bmatrix}a_2 & b_2\\ c_2 & d_2\end{bmatrix}.$ *Then* $C.X = a_1a_2 + b_1b_2 + \dots$

And $C^T X = \begin{bmatrix}a_1 & c_1\\ b_1 & d_1\end{bmatrix}\begin{bmatrix}a_2 & b_2\\ c_2 & d_2\end{bmatrix} = \begin{bmatrix}a_1a_2 + c_1c_2 & \dots\\ \dots & b_1b_2 + d_1d_2\end{bmatrix}$

so $tr\left(C^T X\right) = a_1a_2 + c_1c_2 + \dots$

so the problem becomes a convex problem

$min\ tr\left(C^T X\right)$

such that

$tr\left(A_i^T B\right) = b_i$

$X \succcurlyeq 0$

i.e. the matrix that you are optimizing has to be psd

Example 1: Fantope optimization
Example 2: Matrix completion/Trace norm

Unknown matrix X
take a bunch of linear measurements of it
Observe $(y_i,\ A_i)$
Think of matrix completion where you see a few values only of a matrix and need to fill in the rest.
e.g. if you only observe

$$X_{obs} = \begin{bmatrix} & 7 & \\ & & 17 \end{bmatrix}$$

model this as

$$y_i = 7 = tr\left(A_i^T X\right)$$

$$where\ A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

The matrix completion problem is: given these $y_i$observations, fill in the rest of the numbers in X!
There are many ways to do this.
One way us

$$min\ rank(X)$$

$$such\ that$$

$$y = tr\left(A_i^T X\right)$$

But rank is not a convex function! So this problem is non-convex.
So we relax/reform the objective of the problem to

$min \sum 1\{\sigma_i(X) \neq 0\}$ (rank is sum of non zero singluar values)

= $min \sum \sigma_i(X)$ (minimise the L1 norm on the singular values)
But this can be written as an SDP.
This is the trace norm of X.
When X is sdp, this is the trace of X.

$$= min\ tr(X)$$

We can relax min rank completion to min tracenorm completion, which is a convex SDP.

- this is proved using duality - we will see this later

(subject to you have observed the values you saw)