Convex Optimization

HW6

---

1. *Designing face masks*, In response to the dramatic increase in face mask demand due to the COVID- 19 pandemic, a company wants to mass-produce face masks with the minimum cost possible. Although not to scale, the following diagram shows a proposed face mask design where the filter is shown by light green and the elastic cord is shown by dark green.

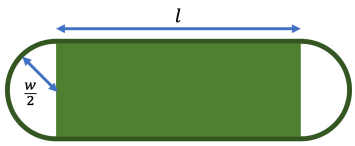

Figure 1: face mask

The basic idea is that the elastic cord loop goes through the top and bottom edge of the filter and forms two half-circles at its opposing ends. The filter is of width w and length l and must have a minimum area of $300cm^2$ according to the standards. Moreover, they impose the following upper and lower bounds on the aspect ratio,

$$2 \geq \frac{l}{w} \geq 1$$

The minimum and maximum accepted width are $10cm$ and $20cm$ and the minimum and maximum accepted length are $20cm$ and $30cm$. Suppose that each cm^2 of filter costs twice each cm of elastic band and that the company wants to minimize the production cost of the design which is equal to the price that must be paid for the required filter and elastic band per mask (manufacturing cost is negligible).

(a) Reformulate this face mask design problem as a convex optimization problem.

(b) Using $CVXPY$ find the size of the filter in the optimal design for mass production.

2. A.6.5

3. A.3.32

4. Suppose predictors (columns of the design matrix $X \in R^{n\times(p+1)}$) in a regression problem split up into J groups:

$$X[\mathbf{1}\ \mathbf{X_{(1)}}...\mathbf{X_{(J)}}]$$

where $\mathbf{1} = (\mathbf{11}...\mathbf{1}) \in \mathbf{R^n}$. To achieve sparsity over non-overlapping groups rather than individual predictors, we may write $\beta = (\beta_0, \beta_{(1)}, ..., \beta_{(J)})$, where β_0 is an intercept term and each $\beta_{(j)}$ is an appropriate coefficient block of β corresponding to X_j, and solve the regularized regression problem:

$$min_{\beta\in R^{p+1}} \quad g(\beta) + h(\beta)$$

In the following problems, we will use linear regression to predict the Parkinson's disease (PD) symptom score on the `Parkinsons` dataset. The PD symptom score is measured on the unified Parkinson's disease rating scale (UPDRS). This data contains $5,785$ observations, 18 predictors (in `X_train.csv`), and an outcome—the toal UPDRS (in `y_train.csv`). The data were collected at the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation. The 18 columns in the predictor matrix have the following groupings (in column ordering):

- `age`: Subject age in years
- `sex`: Subject gender, '0'–male, '1'–female
- `Jitter(%), Jitter(Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP`: Several measures of variation in fundamental frequency of voice
- `Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA`: Several measures of variation in amplitude of voice
- `NHR, HNR`: Two measures of ratio of noise to tonal components in the voice
- `RPDE`: A nonlinear dynamical complexity measure
- `DFA`: Signal fractal scaling exponent
- `PPE`: A nonlinear measure of fundamental frequency variation

(a) We first consider the ridge regression problem, where $h(\beta) = \frac{\lambda}{2}\|\beta\|_2^2$:

$$\min_{\beta \in R^{p+1}} \frac{1}{2N}\|X\beta - y\|^2 + \frac{\lambda}{2}\|\beta\|_2^2$$

where N is the number of samples. Note: in your implementation for this problem, if you added a ones vector to X ($X = \begin{bmatrix}\mathbf{1} & X_{(1)} & X_{(2)} & \dots & X_{(J)}\end{bmatrix}$), you should not include the bias term β_0 associated with the ones vector in the penalty.

i. Derive the stochastic gradient update w.r.t. a batch-size B and a step-size t. Hint: you will need to a separate update for β_0 since it should not be penalized.

ii. Implement the stochastic gradient descent algorithm to solve the ridge regression problem. Initialize β with random normal values. Fit the model parameters on the training data (`X_train.csv`, `Y_train.csv`) and evaluate the objective function after each epoch (you will need to plot these values later). Set $\lambda = 1$. Try different batch-sizes from $\{10, 20, 50, 100\}$ and different step-sizes from $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. Train for 500 epochs (an epoch is one iteration though the dataset).

iii. Plot $f^k - f^\star$ versus k ($k = 1, \dots, 500$) on a semi-log scale (i.e. where the y-axis is in log scale) for all setting combinations, where f^k denotes the objective value averaged over all samples at epoch k, and the optimal objective value is $f^\star = 57.0410$. What do you find? How do the different step sizes and batch sizes affect the learning curves (i.e. convergence rate, final convergence value, etc.)?

(b) Next, we consider the least squares group LASSO problem, where $h(\beta) = \lambda \sum_j w_j \|\beta_{(j)}\|_2$:

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{2N}\|X\beta - y\|^2 + \lambda \sum_j w_j \|\beta_{(j)}\|_2$$

A common choice for weights on groups w_j is $\sqrt{p_j}$, where p_j is number of predictors that belong to the jth group, to adjust for the group sizes.

We will solve the problem using proximal gradient descent algorithm (over the whole dataset).

i. Derive the proximal operator $\text{prox}_{h,t}(x)$ for the non-smooth component $h(\beta) = \lambda \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_2$.

ii. Derive the proximal gradient update for the objective.

iii. Implement proximal gradient descent to solve the least squares group lasso problem on the `Parkinsons` dataset. Set $\lambda = 0.02$. Use a fixed step-size $t = 0.005$ and run for 10000 steps.

iv. Plot $f^k - f^\star$ versus k for the first 10000 iterations ($k = 1, \ldots, 10000$) on a semi-log scale (i.e. where the y-axis is in log scale) for both train and test data, where f^k denotes the objective value averaged over all samples at step k, and the optimal objective value is $f^\star = 49.9649$. Print the components of the solutions numerically. What are the selected groups?

v. Now implement the LASSO (hint: you shouldn't have to do any additional coding), with fixed step-size $t = 0.005$ and $\lambda = 0.02$. Compare the LASSO solution with your group lasso solutions.

vi. Implement accelerated proximal gradient descent with fixed step-size under the same setting in part (c). Hint: be sure to exclude the bias term β_0 from the proximal update, just use a regular accelerated gradient update. Plot $f^k - f^\star$ versus k for both methods (unaccelerated and accelerated proximal gradient) for $k = 1, \ldots, 10000$ on a semi-log scale and compare the selected groups. What do you find?

References

[A] Boyd, S., Vandenberghe, L, "Additional Exercises for Convex Optimization," , Jan 2021.