1. Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, consider the lasso problem

$$
\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \| y - X \beta \|_2^2 + \lambda \| \beta \|_1.
$$

Rewrite this as

$$
\min_{\beta \in \mathbb{R}_p, z \in \mathbb{R}^n} \frac{1}{2} \| y - z \|_2^2 + \lambda \| \beta \|_1 \quad \text{subject to} \quad z = X \beta, \tag{1}
$$

and now, starting from the above representation, derive its dual. Show that the dual is

$$
\max_{u \in \mathbb{R}^n} \frac{1}{2} \| y \|_2^2 - \frac{1}{2} \| y - u \|_2^2 \quad \text{subject to} \quad \| X^T u \|_\infty \leq \lambda.
$$

Do we have strong duality here? And, what is the relationship between a primal solution $\hat{\beta}$ and a dual solution $\hat{u}$? (Hint: look at the KKT conditions for the problem (1).)

2. Let $\Sigma \in \mathbb{R}^{p \times p}$ an unknown covariance matrix, assumed to be strictly positive definite. Let $S \in \mathbb{R}^{n \times n}$ be the sample covariance matrix constructed from $n$ i.i.d. observations from $N(0, \Sigma)$. Recall that we can draw a conditional independence graph between the nodes $1, \ldots, p$ based on the entries of the inverse covariance matrix $\Sigma^{-1}$: if $X \sim N(0, \Sigma)$, then $(\Sigma^{-1})_{i,j} = 0$ if and only if $X_i, X_j$ are conditionally independent given $X_\ell, \ell \neq i,j$. Hence, in our conditional independence graph, we only connect nodes $i$ and $j$ for which $(\Sigma^{-1})_{i,j} \neq 0$.

To form an estimate $\hat{\Theta}$ of the underlying inverse covariance matrix $\Sigma^{-1}$, consider solving the graphical lasso problem

$$
\min_{\Theta \in S_{++}^p} -\log \det \Theta + \text{tr}(S \Theta) + \lambda \| \Theta \|_1.
$$

Here, recall, $S_{++}^p$ is the set of strictly positive definite $p \times p$ matrices, tr$(\cdot)$ denotes the trace operator (sum of diagonal elements), and $\| \Theta \|_1 = \sum_{i,j=1}^p |\Theta_{i,j}|$. In short, the graphical lasso performs $\ell_1$ penalized maximum likelihood estimation.

(a) Argue that the graphical lasso problem is convex, in fact strictly convex, so that it has a unique solution $\hat{\Theta}$.

(b) Show that the KKT conditions for this problem are

$$
-\hat{\Theta}^{-1} + S + \lambda \Gamma = 0,
$$

where $\Gamma_{i,j} \in \partial |\hat{\Theta}_{i,j}|$ a subgradient of the absolute value function evaluated at $\hat{\Theta}_{i,j}$, for each $i,j = 1, \ldots, p$. (You do not need to explicitly derive the gradient of $-\log \det \Theta$, just state what it is.)

(c) Recall that we think of each nonzero entry $\hat{\Theta}_{i,j}$ as an indication that we will join nodes $i,j$ by an edge. Prove, using the KKT conditions from part (b), that for any tuning parameter value $\lambda \geq 0$, the connected components of the graph defined by the solution $\hat{\Theta}$ are simply given by the connected components of graph defined by an appropriately thresholded covariance matrix $\hat{S}$, defined as

$$
\hat{S}_{i,j} = S_{i,j} \cdot 1(|S_{i,j}| > \lambda), \quad i,j = 1, \ldots, p.
$$

Hint: if we use the nonzero entries of a matrix $M$ to define a graph, then the connected components in this graph exactly correspond to a block diagonal structure in $M$ (subject to rearranging the row/column order).
3. Recall from lecture that we say points $X_1, \ldots, X_p \in \mathbb{R}^n$ are in general position provided that the following property holds: any affine subspace $L \subseteq \mathbb{R}^n$ of dimension $k < \min(n, p)$ contains at most $k + 1$ points among $\pm X_1, \ldots, \pm X_n$, excluding antipodal pairs. E.g., if $L$ contains both $+X_1$ and $-X_1$, we don’t count this as containing 2 separate points.

(a) Prove that this is equivalent to the following statement: for any $k < \min(n, p)$, any indices $i_1, \ldots, i_{k+1} \in \{1, \ldots, p\}$, and any signs $\sigma_1, \ldots, \sigma_{k+1} \in \{-1, +1\}$, the affine subspace spanned by $\sigma_1 X_{i_1}, \ldots, \sigma_{k+1} X_{i_{k+1}}$ does not contain any element of the set $\{\pm X_i : i \neq i_1, \ldots, i_{k+1}\}$.

(b) Prove that if $X_1, \ldots, X_p$ are jointly drawn from a continuous probability distribution on $\mathbb{R}^n$, then $X_1, \ldots, X_p$ are in general position with probability 1.

(c) Consider the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p} f(X\beta) + \lambda \|\beta\|_1,$$

where $f$ is a strictly convex, differentiable function, and $\lambda > 0$. You may assume that a solution $\hat{\beta}$ to the above problem exists. By recreating the arguments from lecture for the special case $f(u) = \|y - u\|^2_2$, prove that the solution $\hat{\beta}$ to the above problem is unique whenever the columns of $X$ are in general position. (Hence, from what you’ve shown in (b): if the columns $X_1, \ldots, X_p$ come from a continuous probability distribution, the solution is unique almost surely, regardless of the sizes of $n$ and $p$.)

(d) Prove that if the solution $\hat{\beta}$ to the problem in part (c) is unique, then it can have at most $\min(n, p)$ nonzero entries.

4. Let $X_1, \ldots, X_n \sim \text{Unif}(0, \theta)$ where $\theta > 0$. Let $\hat{\theta}$ be the maximum likelihood estimator. Show that $n(\hat{\theta} - \theta)$ converges in distribution to a nondegenerate distribution $F$. Show that the bootstrap is inconsistent, i.e. $\hat{F}(t) = \mathbb{P}(n(\hat{\theta} - \theta) \leq t | X_1, \ldots, X_n)$ does not converge to $F$.

5. Let $X_1, \ldots, X_n \in \mathbb{R}^d$. Let $\Sigma$ be the $d \times d$ covariance matrix for $X_i$. The covariance graph $G$ puts an edge between $(j, k)$ if $\Sigma_{jk} \neq 0$. Here we will use the bootstrap to estimate the covariance graph.

Let $\Sigma$ have the following form: $\Sigma_{jj} = 1$, $\Sigma_{j,k} = a$ if $|j - k| = 1$ and $\Sigma_{j,k} = 0$ otherwise. Here, $0 < a < 1$.

Let $d = 100$ and $n = 50$. Generate $n$ observations. Compute a 95 percent bootstrap confidence set for $\Sigma$ using the bootstrap distribution

$$\mathbb{P}\left[\max_{j,k} \sqrt{n}|\hat{\Sigma}_{jk} - \tilde{\Sigma}_{jk}| \leq t \mid X_1, \ldots, X_n\right].$$

This gives (uniform) confidence intervals for all the elements of $\Sigma_{jk}$. For each $(j, k)$, put an edge if the confidence interval for $\Sigma_{jk}$ excludes 0. Plot your graph. Try this for different values of $a$. 

2