Solution Sheet.

Problem 1

(a) Let X be a random variable with $\mathrm{E}[X] = 0$. Suppose that the moment-generating function of X^2 is bounded at some point, that is,

$$\mathrm{E}\left[e^{X^2}\right] \leq 2.$$

Prove that X satisfies the two-sided tail bound

$$\mathrm{P}(|X| > t) \leq 2e^{(-t^2)} \text{ for all } t \geq 0.$$

(b) Prove that if X is a non-negative random variable with expectation $\mathrm{E}[X]$, then for all $t > 0$, we have $\mathrm{P}[X \geq t] \leq \mathrm{E}[X]/t$.

(c) Recall Chernoff's inequality: Let X_i be independent Bernoulli random variables with success probability p_i. Consider their sum $S_N = \sum_{i=1}^{N} X_i$ and denote its mean by $\mu = \mathrm{E}[S_N]$. Then, for any $t > \mu$, we have

$$\mathrm{P}(S_N \geq t) \leq e^{t-\mu}\left(\frac{\mu}{t}\right)^t.$$

Consider 200 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 150. Use Chernoff's inequality.

(d) Let X_i, for $I = 1, \ldots, n$, be a random sample of a random variable X. Let X have mean μ and variance σ^2. Find the size of the sample (n), such that the probability that the difference between sample mean and true mean is smaller that $\frac{\sigma}{10}$ is at least 0.95. Hint: Derive a version of the Chebyshev inequality for $\mathrm{P}(|X - \mu| \geq a)$ using Markov inequality.

Solution.

(a) It holds that

$$\mathrm{P}(|X| > t) = \mathrm{P}(e^{X^2} > e^{t^2}) \leq \frac{2}{e^{t^2}}.$$

(we just apply Markov in the last step)

(b) It holds that

$$\mathrm{P}(X \geq t) = \mathrm{P}(t\mathbb{1}_{X\geq t} \geq t) = \mathrm{E}\left[\mathbb{1}_{X\geq t}\right] \leq \mathrm{E}\left[\frac{X}{t}\right] = \frac{\mathrm{E}[X]}{t}.$$

Alternative way:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} xf(x)dx \geq \int_{t}^{\infty} xf(x)dx \geq \int_{t}^{\infty} tf(x)dx = t\,\mathrm{P}(X \geq t)$$

$$\implies \mathrm{E}[X] \geq t\,\mathrm{P}(X \geq t) \implies \frac{\mathrm{E}[X]}{t} \geq \mathrm{P}(X \geq t)$$

(c) Chernoff gives $e^{50}\left(\frac{2}{3}\right)^{150} = \left(\frac{8e}{27}\right)^{50}$. It is not necessary to simplify this further.

(d) Let $\hat{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, $\mathrm{E}\left[\hat{X}\right] = \mu$ and $\mathrm{Var}\left[\hat{X}\right] = \frac{\sigma^2}{n}$. Now, we need to determine n such that

$$\mathrm{P}(|\hat{X}-\mu| \leq \frac{\sigma}{10}) \geq 0.95 \implies \mathrm{P}(|\hat{X}-\mu| \geq \frac{\sigma}{10}) \leq 0.05$$

We can write the probability as:

$$\mathrm{P}(\sqrt{(\hat{X}-\mu)^2} \geq \frac{\sigma}{10}) = \mathrm{P}((\hat{X}-\mu)^2 \geq \frac{\sigma^2}{100}) \leq \frac{\mathrm{Var}\left[\hat{X}\right]}{\frac{\sigma^2}{100}} = \frac{\sigma^2}{n}\frac{100}{\sigma^2} = \frac{100}{n} \leq 0.05$$

$$\implies \frac{100}{0.05} \leq n$$

Therefore, we need a sample size of $n \geq 2000$.

Problem 2

1. Estimation of diagonal covariances: Let $(X_i)_{i=1,\ldots,n}$ be an i.i.d. sequence of d-dimensional vectors, drawn from a zero-mean distribution with diagonal covariance matrix $\Sigma = D$. Consider the estimate $\widehat{D} = \mathrm{diag}(\widehat{\Sigma})$, where $\widehat{\Sigma}$ is the usual sample covariance matrix. Suppose further that each component X_{ij} is sub-Gaussian with parameter at most $\sigma = 1$. Show the following:

   (a) X_{ij}^2 is sub-exponential with parameters (2,4).

   (b) $\sum_{i=1}^{n} X_{ij}^2$ is sub-exponential with parameters $(2\sqrt{n}, 4)$

   (c) For each $i = 1, \ldots, d$, we get

   $$\mathrm{P}\left(|\widehat{D}_{ii} - D_{ii}| \geq t\right) \leq 2e^{-\frac{n}{8}\min\left\{t,t^2\right\}}.$$

2. Suppose that the random vector $X \in \mathrm{R}^n$ has a $N_n(\mu, \Sigma)$ distribution, where Σ is positive. Show the the random variable $Y = (X-\mu)^T \Sigma (X-\mu)$ is sub-exponential.

   Note: The question has a typo. It should be $Y = (X-\mu)^T \Sigma^{-1} (X-\mu)$

Solution.

1. (a) X_{ij}^2 is sub-exponential with parameters (2,4).

Approach 1 For this, we consider that a sub-gaussian variable of parameter at most σ will be bounded for above by a gaussian variable. Let $X \sim N(0, \sigma^2)$, and further assume $\sigma = 1$. Now, consider that X^2 follows a chi-squared distribution, and its moment generating function is defined as.

$$\mathrm{E}\left[e^{\lambda(X^2-1)}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda x^2} e^{-\frac{x^2}{2}} dx = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}, \text{ for } \lambda < \frac{1}{2}$$

The moment generating function is also obtain by using the gaussian distribution and considering $\mathrm{E}\left[X^2\right] = 1$. Following the definition of sub-exponential we have:

$$\mathrm{E}\left[e^{\lambda(X^2-1)}\right] \le e^{\frac{\nu^2\lambda^2}{2}} \text{ for all } \lambda^2 < \frac{1}{\alpha^2}$$

Now, considering $\nu = 2$, and $\alpha = 4$, we have that $\lambda^2 < \frac{1}{16} \implies \lambda \in (-\frac{1}{4}, \frac{1}{4})$. Therefore, the moment generation function previously calculated is bounded for these values of λ. With this it should hold that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{\frac{2^2\lambda^2}{2}} = e^{2\lambda^2} \text{ for all } \lambda^2 < \frac{1}{16}$$

Let's focus for $\lambda \in (-\frac{1}{4}, \frac{1}{4})$. Given that all terms are positive, we can square and reorder the inequality:

$$e^{-4\lambda^2-2\lambda} \le 1-2\lambda$$

$$\implies -4\lambda^2 - 2\lambda \le \log 1 - 2\lambda \implies 0 \le ln(1-2\lambda) + 4\lambda^2 + 2\lambda = f(\lambda)$$

It easy to show that $f(x)$ is a convex function in the domain of lambda. Therefore we can calculate its minimum with first and second order condition. If $\min f(\lambda) \ge 0$ for $|\lambda| < \frac{1}{4}$, the inequality holds and the variable is sub-exponential(2,4).

$$FOC: \frac{df(\lambda)}{d\lambda} = -\frac{2}{1-2\lambda} + 8\lambda + 2 = -2 + 8\lambda - 16\lambda^2 + 2 - 4\lambda = 0$$

$$4\lambda(1-4\lambda) = 0 \implies \lambda = 0 \lor \lambda = 1/4, \text{ we can see that 1/4 is not a minimizer.}$$

The second derivative evaluated in $\lambda = 0$ has a value of 4, therefore $\lambda = 0$ is a proper minimizer of $f(\lambda)$, and $f(0) = 0$. Thus the inequality holds and the variable X^2 is sub-exponential of parameter (2,4).

Possible Alternative Another way to approach the problem will be: Let $Z = X^2 - \mathrm{E}[X^2]$. Then, we calculated its moment generation function using the Taylor expansion of the exponential,

$$\mathrm{E}\left[e^{\lambda Z}\right] \le \mathrm{E}\left[1 + \sum_{k=1}^{\infty} \frac{\lambda^k Z^k}{k!}\right]$$

Following this, we can keep bounding using Jensen's inequality and the bounds available given that X is sub-gaussian the parameter at most 1.

(b) Let $Z_{ij} = X_{ij}^2$, and therefore be sub-exponential with parameters $(2, 4)$. Now we compute the moment generating function:

$$\mathrm{E}\left[e^{\lambda \sum_{i=1}^{n}(Z_{ij} - \mathrm{E}[Z_{ij}])}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda (Z_{ij} - \mathrm{E}[Z_{ij}])}\right]$$

Now, following the bounds obtained beforehand:

$$\prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda (Z_{ij} - \mathrm{E}[Z_{ij}])}\right] \leq \prod_{i=1}^{n} e^{\nu^2 \frac{\lambda^2}{2}} = e^{\sum_{i=1}^{n}(\nu^2 \frac{\lambda^2}{2})} \qquad \forall |\lambda| \leq \frac{1}{4}$$

$$\implies E\left[e^{\lambda \sum_{i=1}^{n}(Z_{ij} - \mathrm{E}[Z_{ij}])}\right] \leq e^{(\sqrt{n}\nu)^2 \frac{\lambda^2}{2}} \qquad \forall |\lambda| \leq \frac{1}{4}$$

Finally, we can conclude that $\sum_{i=1}^{n} X_{ij}^2$ is sub-exponential with parameters $(2\sqrt{n}, 4)$.

(c) $\widehat{D}_{ii}$ is the usual sample covariance matrix, and it's defined as $\widehat{D}_{ii} = \dfrac{1}{n}\sum_{i=1}^{n} x_{ij}^2$. This estimator is unbiased, i.e., $\mathrm{E}\left[\widehat{D}_{ii}\right] = D$. Also, following the previous exercises we have that $\widehat{D}_{ii}$ is sub-exponential de parameters $(\frac{2}{\sqrt{n}}, \frac{4}{n})$. Now, given sub-exponential concentration

$$\mathrm{P}\left(|\widehat{D}_{ii} - D_{ii}| \geq t\right) \leq 2 \exp^{-\dfrac{1}{2}\min\left\{\dfrac{t}{\alpha}, \dfrac{t^2}{\nu^2}\right\}}.$$

Replacing nu and α for their respective values, we get:

$$\mathrm{P}\left(|\widehat{D}_{ii} - D_{ii}| \geq t\right) \leq 2e^{-\dfrac{n}{8}\min\{t, t^2\}}.$$

2. $Y = (X - \mu)^T \Sigma^{-1} (X - \mu)$. Let's consider the spectral decomposition of $\Sigma = Q\Lambda Q^T$, where $Q^T Q = I$. Now, $\Sigma^{-\frac{1}{2}}$ is ten defined as $Q\Lambda^{\frac{1}{2}}Q^T$, and therefore $\Sigma^{-1} = \Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}$. Let $Z = \Sigma^{-\frac{1}{2}}(X - \mu)$, and corresponds to random variable that follows a normal standard distribution. Then,

$$Y = Z^T Z = \sum_i Z_i^2$$

Therefore, as presented in the lecture, $Y \sim \chi^2(n)$ is a sub-exponential variable.

Problem 3

For the orthogonal case, i.e. when $X'X = I$, derive the following explicit forms for estimators,

(a) For ridge:

$$\widehat{\beta}^{Ridge} = \widehat{\beta}^{OLS}/(1+\lambda).$$

(b) For lasso:

$$\widehat{\beta}_i^{Lasso} = \text{sign}(\widehat{\beta}_i^{OLS})(|\widehat{\beta}_i^{OLS}| - \lambda)_+,$$

where $\widehat{\beta}^{OLS}$ is the regular OLS estimator and $\widehat{\beta}_i^{OLS}$ its ith component. Note that the results can differ depending on how one chooses the multiplicative constants. The solutions in Problem 3 are based on the following objective functions:

$$\widehat{\beta}^{Ridge} = \text{argmin}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right\}.$$

$$\widehat{\beta}^{LASSO} = \text{argmin}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right\}.$$

Solution.

(a) For ridge regression, we know

$$\widehat{\beta}^{Ridge} = \left(X^T X + \lambda I\right)^{-1} X^T y$$

$$= \frac{1}{1+\lambda} X^T y = \frac{\widehat{\beta}^{OLS}}{(1+\lambda)}.$$

(b) For lasso let us write the objective with matrices:

$$\widehat{\beta}^{LASSO} = \text{argmin}\left\{\frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1\right\}$$

$$= \text{argmin}\left\{\frac{1}{2}(y^t y - 2y^T X\beta + \beta^T X^T X\beta) + \lambda||\beta||_1\right\} \equiv \text{argmin}\left\{-y^T X\beta + \frac{1}{2}\beta^T\beta + \lambda||\beta||_1\right\}$$

$$= \text{argmin}\left\{\lambda||\beta||_1 - \beta^T\widehat{\beta}^{OLS} + \frac{1}{2}\beta^T\beta\right\}$$

$$= \text{argmin}\left\{\sum_i \lambda|\beta_i| - \beta_i\widehat{\beta}_i^{OLS} + \frac{1}{2}\beta_i^2\right\}.$$

We can see that the problem is separable, thus it can be solve for each individual i separately. We have two cases:

- When $\widehat{\beta}_i^{OLS} \geq 0$, we have that the optimal solution follows $\beta_i^\star \geq 0$. It can be show that if $\beta^\star < 0$, the exist a new solution within an ε-neighborhood of $\beta^\star$ with better objective, contradicting the optimality of $\beta^\star$. Thus, the problem to solve is reduced to

$$\min_{\beta \geq 0} \left\{ \beta_i(\lambda - \widehat{\beta}_i^{OLS}) + \frac{1}{2}\beta_i^2 \right\}.$$

  And it's optimal value is achieved when $\beta^\star = \widehat{\beta}_i^{OLS} - \lambda$. However, as $\beta \geq 0$ we need to define the solution fo only when $\beta^\star$ is non-negative, i.e., $\beta^\star = (\widehat{\beta}_i^{OLS} - \lambda)_+$.

- Analogously, when $\widehat{\beta}_i^{OLS} \geq 0$, we have that the optimal solution follows $\beta_i^\star \leq 0$. Then, we now solve

$$\min_{\beta \leq 0} \left\{ -\beta_i(\lambda + \widehat{\beta}_i^{OLS}) + \frac{1}{2}\beta_i^2 \right\}.$$

  which has solution $\beta^\star = (\widehat{\beta}_i^{OLS} + \lambda)_-$.

In both cases the solution ca be written as:

$$\widehat{\beta}_i^{Lasso} = \text{sign}(\widehat{\beta}_i^{OLS})(|\widehat{\beta}_i^{OLS}| - \lambda)_+,$$