MATH 547: HOMEWORK 5
DUE ON: MONDAY, DECEMBER 2, 9AM.

Problem 1: Hoeffding's inequality in Hilbert spaces, 15 points:
Let $X_1, \dots, X_n$ be independent random vectors in a $\mathbb{R}^p$ such that $\mathbb{E}X_i = \vec{0},\ i = 1, \dots, n$ and $\|X_i\|_2 \le c/2$ (almost surely) for some $c > 0$, and set $B^2 := \frac{nc^2}{4}$. Then for all $t \ge B$,

$$\Pr\left(\left\|\sum_{i=1}^n X_i\right\|_2 \ge t\right) \le \exp\left(-\frac{(t-B)^2}{8B^2}\right).$$

[Hint: apply the bounded difference inequality to $\left\|\sum_{i=1}^n X_i\right\|_2$. Also, note that the result is valid for any separable Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle)$, not only $\mathbb{R}^p$.]

Problem 2: Linear regression, 25pts:
Let $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. training data. The Ridge regression solves the following problem:

$$\frac{1}{2}\sum_{j=1}^n \left(Y_j - \beta^T X_j\right)^2 + \lambda \sum_{j=1}^d \beta_j^2 \to \text{minimize over } \beta = (\beta_1, \dots, \beta_d)^T \in \mathbb{R}^d,$$

where $\lambda > 0$ is called the "regularization parameter" and the term $\lambda\sum_{j=1}^d \beta_j^2$ is called a "penalty."

(1) **(10 pts)** Show that the solution of the Ridge regression is given by

$$\hat{\beta}_\lambda = \left(\mathbf{X}^T\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

Here, I is the $d \times d$ identity matrix, $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ and $\mathbf{X}$ is a matrix with rows $X_1, \dots, X_n$. Does the inverse matrix $(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}$ always exist? One possible approach is to define

$$F(\beta) = \frac{1}{2}\sum_{j=1}^n \left(Y_j - \beta^T X_j\right)^2 + \lambda\|\beta\|_2^2$$

and to compute the directional derivative of $F(\beta)$ in direction u:

$$DF(\beta; u) := \lim_{t\to 0}\frac{F(\beta + tu) - F(\beta)}{t}.$$

Then use the fact that $DF(\beta; u) = \langle \nabla F(\beta), u\rangle$ to get the expression for the gradient $\nabla F(\beta)$.

(2) **(15 pts)** If the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible, then $\mathbf{X}^\dagger = \lim_{\lambda\to 0}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T$ is called the Moore-Penrose pseudo-inverse of $\mathbf{X}$. If $\mathbf{X^TX}$ is invertible, then the pseudo-inverse coincides with the usual inverse matrix. Prove that the solution $\hat{\beta}_0$ of the problem

$$\sum_{j=1}^n \left(Y_j - \beta^T X_j\right)^2 \to \text{minimize over } \beta = (\beta_1, \dots, \beta_d)^T \in \mathbb{R}^d$$

given by $\hat{\beta} = \mathbf{X}^\dagger\mathbf{Y}$ has the smallest $\|\cdot\|_2$ - norm among all solutions.

Problem 3, Minkowski functional, 15 points:

Let $K \subset \mathbb{R}^p$ be closed, bounded, convex, symmetric (meaning that $K = -K$), and have non-empty interior (so that K contains some Euclidean ball). Show that the Minkowski functional (gauge)

$$\|x\|_K = \inf\left\{t > 0 : \ \frac{x}{t} \in K\right\}$$

is a norm.

Problem 4, properties of the Gaussian mean width, 20 points):

Required background will be covered on Monday, November 18. Recall the definition of the Gaussian mean width of a bounded set $K \subset \mathbb{R}^p$:

$$w(K) = \mathbb{E} \sup_{z \in K-K} \langle z, g \rangle ,$$

where g has $N(0, I_p)$ distribution and $K - K = \{u - v, \ u, v \in K\}$. Show that

(a) $w(K) = 2\,\mathbb{E}\sup_{z\in K} \langle z, g\rangle$.

(b) $w(K)$ is invariant under affine transformations, meaning that for any $y \in \mathbb{R}^p$ and any $Q \in \mathbb{R}^{p\times p}$ such that $Q^{-1} = Q^T$, $w(QK + y) = w(K)$.

(c) $w(K)$ is invariant with respect to taking the convex hull: if $co(K)$ is the convex hull of K, then $w(\mathrm{co}(K)) = w(K)$.

(d) Let $\mathrm{diam}(K)$ be the diameter of K. Show that

$$\sqrt{\frac{2}{\pi}}\,\mathrm{diam}(K) \le w(K) \le \sqrt{p}\,\mathrm{diam}(K).$$

Problem 5, bonus: matrix inverse, 10 points: Let A be a positive definite matrix. Prove that the map $A \mapsto A^{-1}$ is convex, that is, for any positive definite A, B and any $0 \le \lambda \le 1$,

$$(\lambda A + (1-\lambda)B)^{-1} \preceq \lambda A^{-1} + (1-\lambda)B^{-1}.$$

(Hint – use the following fact without the proof: the block matrix $\begin{pmatrix} T & C \\ C^T & M \end{pmatrix}$ is nonnegative definite if and only if $C^T T^1 C \preceq M$; the matrix $C^T T^{-1} C - M$ is known as the Schur complement. Now apply it (how?) with $T = A$, $M = A^{-1}$, $C = I$ and $T = B$, $M = B^{-1}$, $C = I$)