Instructions.

• Everyone must submit an individual write-up.

• This is a calibration homework; please work alone and don’t hunt for solutions.

• Homework is due Thursday, September 5, at 3:00pm; no late homework accepted.

• Please consider using the provided \LaTeX file as a template.
1. **Ordinary least squares.**

Recall the least squares problem

\[
\arg \min_{w \in \mathbb{R}^d} \frac{1}{n} \|Xw - y\|_2^2,
\]

where \(X \in \mathbb{R}^{n \times d}\) and \(y \in \mathbb{R}^n\).

(a) Prove that \(\hat{w} \in \mathbb{R}^d\) is optimal iff

\[
X^T X \hat{w} = X^T y.
\]

(b) Let \(X^\dagger\) denote the pseudoinverse of \(X\), and \(\ker(X) = \{v \in \mathbb{R}^d : Xv = 0\}\) the right nullspace of \(X\). Prove that \(\hat{w}\) is optimal iff

\[
\hat{w} - X^\dagger y \in \ker(X).
\]

**Hint.** One direction of the “iff” is harder than the other; the SVD gives one short solution.

(c) Construct an \(X\) and \(y\) so that \(X^T X\) is not invertible, but \(\hat{w} := X^\dagger y\) exists and is optimal; include a complete verification of optimality of \(\hat{w}\).

**Solution.**
2. Logistic regression.

Consider a logistic regression problem, where the goal is now to minimize the logistic risk

\[ R(w) := \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-(Xw)_{iy_i})) , \]

where as before \( X \in \mathbb{R}^{n \times d} \) and \( y \in \mathbb{R}^n \) and \( w \in \mathbb{R}^d \).

(a) Prove \( R \) is convex.
(b) Suppose there exists \( w \) so that \( \min_i (Xw)_{iy_i} > 0 \). Prove \( \inf_{w \in \mathbb{R}^d} R(w) = 0 \).
(c) Suppose we rewrite \( w \in \mathbb{R}^d \) as a linear network of width \( m \) and depth 2, meaning we replace \( w \) with \( AB \), where \( A \in \mathbb{R}^{d \times m} \) and \( B \in \mathbb{R}^{m \times 1} \) are the new parameters, and the new risk is

\[ \tilde{R}(A, B) := R(AB) = \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-(XAB)_{iy_i})) . \]

Is \( \tilde{R} \) convex in the joint variable \((A, B)\)?

Simply saying “yes” or “no” suffices for this part. Feel free to provide proofs/counterexamples, though.

**Hint.** If “joint” seems ambiguous but you’d still like to attempt the problem, note that this problem is convex in \( A \) and \( B \) separately, meaning \( A \mapsto \tilde{R}(A, B) \) is convex for all \( B \), and \( B \mapsto \tilde{R}(A, B) \) is convex for all \( A \). This problem is asking about something else, about the joint variable \((A, B)\)...

**Solution.**
3. Random initialization.

Initialization is an essential part of the deep learning story. In this problem, we will study two quantities,

\[ E\|Wz\|^2 \quad \text{and} \quad E\|\sigma_r(Wz)\|^2, \]

where \( W \in \mathbb{R}^{m \times p} \) is a matrix with iid Gaussian entries, specifically \( W_{ij} \sim \mathcal{N}(0, \frac{1}{m}) \), and \( z \in \mathbb{R}^p \), and \( \sigma_r: \mathbb{R}^m \rightarrow \mathbb{R}^m \) applies the ReLU nonlinearity \( x \mapsto \max\{0, x\} \) coordinate-wise. \( W \) is the only random variable here, and \( E \) is an expectation over \( W \). Effectively, this problem studies how norms are altered via the application of a single layer of a network; \( z \) need not be the input, it can be the representation produced by some later layer.

(a) Prove that \( E\|Wz\|^2 = \|z\|^2 \).

**Hint.** Try the case \( p = 1 \) first.

(b) Prove that \( E\|\sigma_r(Wz)\|^2 = \frac{1}{2}\|z\|^2 \).

**Hint.** Try using the symmetry of the Gaussian distribution.

(c) The standard concentration inequality for Gaussians is: if \( X \in \mathbb{R}^d \) has standard Gaussian coordinates \( X_i \sim \mathcal{N}(0, 1) \), and if \( f: \mathbb{R}^d \rightarrow \mathbb{R} \) is \( L \)-Lipschitz in \( X \), meaning

\[ |f(X) - f(X')| \leq L \cdot \|X - X'\|, \]

then with probability at least \( 1 - \delta \),

\[ |f(X) - E f(X)| < L \sqrt{2 \ln \frac{2}{\delta}}. \]

Use this inequality (and perhaps Jensen’s inequality…) to prove that for any fixed \( z \in \mathbb{R}^p \), with probability at least \( 1 - \delta \),

\[ \|\sigma_r(Wz)\| \leq \|z\| \left( 1 + \sqrt{\frac{2 \ln(2/\delta)}{m}} \right). \]

**Remark.** We can’t (solely) use Hoeffding’s inequality for this problem… why not?

**Solution.**