Modern Paradigms in Learning and Decision-Making (CS-GY 9223), Fall 2026
Homework 1. Due at the start of class on September 16.

Problem 0.

Read or review Chapter 1 and Sections 2.1–2.2 of Vershynin.

Problem 1.

A function $f : K \to \mathbb{R}$ is convex if for every $x, y \in K$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y). \tag{1}$$

Using this definition prove the following statements. Assume throughout that $K \subset \mathbb{R}^d$ is a convex set.

1. Assume $f_1, \ldots, f_m : K \to \mathbb{R}$ be convex. Show that $g(x) = \max_{i \le m} f_i(x)$ is convex.
2. Prove that a function $f : K \to \mathbb{R}$ is convex if and only if for every $m \in \mathbb{N}$, any vectors $x_i \in K$, and weights $\lambda_i \ge 0$ with $\sum_{i=1}^m \lambda_i = 1$,
$$f\Big(\sum_{i=1}^m \lambda_i x_i\Big) \le \sum_{i=1}^m \lambda_i f(x_i).$$
3. Let X be a random vector in $\mathbb{R}^d$ taking finitely many values and let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Deduce that $f(\mathbb{E}X) \le \mathbb{E} f(X)$. This is known as Jensen's inequality.

Problem 2.

1. Assume that f is differentiable at a point x. Prove directly from the definition of convexity in Equation (1) that for all $y \in \mathbb{R}^d$,
$$f(y) \;\ge\; f(x) + \langle \nabla f(x), y - x\rangle. \tag{2}$$
Hint: use the fact that
$$\lim_{t\to 0} \frac{f(x+tz) - f(x)}{t} = \nabla f(x)^\top z$$
2. Let $x^\star \in \operatorname{argmin}_{x\in K} f(x)$ be a minimizer of a convex function f.
Prove that $x^\star \in \operatorname{argmin}_{x\in K} f(x)$ if and only if
$$\langle \nabla f(x^\star), y - x^\star\rangle \;\ge\; 0 \qquad \text{for all } y \in K.$$
3. Assume that $f : K \to \mathbb{R}$ is γ-strongly convex, that is
$$f(y) \;\ge\; f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\gamma \|y - x\|_2^2$$
where K is a compact and convex set. Prove that f has a unique minimizer.

Problem 3.

Let K be a compact convex set and let $\Pi_K(x) = \text{argmin}_{y\in K}\|y-x\|_2^2$ be the projection operator onto K.

1. Prove that $\Pi_x(y) = \|y-x\|_2^2$ is a convex function in y for every x.
2. Using Equation (2) show that

$$\langle x - \Pi_K(x), z - \Pi_K(x)\rangle \le 0 \qquad \text{for all } z \in K. \tag{3}$$

3. Now prove using Equation (3) that the projection operator is non-expansive. For all $x, y \in \mathbb{R}^d$,

$$\|\Pi_K(x) - \Pi_K(y)\|_2 \le \|x-y\|_2.$$

   This fact regarding non-expansiveness of the projection operator will come up repeatedly when analyzing the convergence behavior of algorithms like stochastic gradient descent for constrained optimization problems.

Problem 4.

Throughout the following problem, let $\mathcal{F}$ be a finite collection of functions $f : \mathcal{Z} \to [0,1]$, let $\mathcal{D}$ be a distribution over $\mathcal{Z}$, and let $z_1, \dots, z_n$ be n i.i.d samples drawn from $\mathcal{D}$.

1. Using Hoeffding's inequality (HDP Theorem 2.2.6) show that if,

$$n \;\ge\; \frac{\log|\mathcal{F}| + \log(2/\delta)}{2\varepsilon^2},$$

   then with probability at least $1-\delta$ over the draws from the distribution:

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}_{z\sim D}[f(z)]\right| \le \varepsilon.$$

2. Let

$$f^\star = \text{argmin}_{f\in\mathcal{F}}\,\frac{1}{n}\sum_{i=1}^{n} f(z_i)$$

   be the empirical risk minimizer (assume it is unique). Show that with probability $1-\delta$,

$$\mathbb{E}_{z\sim\mathcal{D}}[f^\star(z)] \le \min_{f\in\mathcal{F}}\mathbb{E}[f(z)] + 2\sqrt{\frac{\log|\mathcal{F}| + \log(2/\delta)}{n}}$$

   This idea is known as a uniform convergence argument. It is a very common tool in learning theory that we use when trying to fit functions to data

Problem 5.

Let P and Q be probability distributions on a finite set $\mathcal{Z}$, and define the total variation distance

$$\mathrm{TV}(P,Q) := \max_{A \subseteq \mathcal{Z}} |P(A) - Q(A)|.$$

1. Show that $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_{z\in\mathcal{Z}} |P(z) - Q(z)|$.
2. Show that $\mathrm{TV}(P,Q) = \sup_{f:\mathcal{Z}\to[0,1]} |\mathbb{E}_{z\sim P} f(z) - \mathbb{E}_{z\sim Q} f(z)|$

Problem 6.

A loss function $\ell(p,y) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_{\geq 0}$ is a proper scoring rule if for every distribution $\mathcal{D}$ over outcomes y, the mean of the distribution $p = \mathbb{E}[Y]$ is in $\mathrm{argmin}_{t\in\mathbb{R}}\mathbb{E}_{y\sim\mathcal{D}}[\ell(q,y)]$.

1. Show that the squared loss $\ell(t,y) = (y-t)^2$ is a a proper scoring rule.
2. Assume y is binary. In this case, $\mathbb{E}[Y] = \Pr[Y=1]$. Show that the log loss

   $$\ell(t,y) = -y\log t - (1-y)\log(1-t)$$

   is a proper scoring rule .
3. Show that absolute loss $\ell(y,p) = |y-p|$ is not proper. What is its minimizer if y is binary? What is its minimizer if y is real-valued?