Topics in Machine Learning : Mock Exam

Instructor: Lénaïc Chizat
TA: Guillaume Wang

December 2025

- No documents are permitted. Duration: 3 hours.
- Most questions can be solved independently; within an exercise, you may assume results from earlier parts.
- It is not necessary to answer all questions to obtain the maximum grade (the exam is intentionally long).

1 Questions on the course material (9 points)

Answer each of the following questions *with no justification.*

1.1. (2 points) For a classification dataset $(x_i, y_i)_{i\le n}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, write down the (unregularized) logistic regression objective. Under what conditions does a minimizer exist?

1.2. (2 points) Recall that the $\mathcal{F}_1$ and $\mathcal{F}_2$ norms of a function $f : \mathbb{R}^d \to \mathbb{R}$ are defined by

$$\|f\|_{\mathcal{F}_2} = \inf\left\{ \|\eta\|_{L_2(\tau)};\ \forall x, f(x) = \int_{\mathbb{R}^d} \eta(\theta)\sigma(\theta^\top x)\,\mathrm{d}\tau(\theta) \right\}$$
$$\|f\|_{\mathcal{F}_1} = \inf\left\{ \|\mu\|_{\mathrm{TV}};\ \forall x, f(x) = \int \sigma(\theta^\top x)\,\mathrm{d}\mu(\theta) \right\}$$

for an activation function σ and a reference probability measure τ, where $\|\mu\|_{\mathrm{TV}} = \int_{\mathbb{R}^d} |\mathrm{d}\mu(\theta)|$. Does it hold that $\|f\|_{\mathcal{F}_1} \le \|f\|_{\mathcal{F}_2}$ for all f? or that $\|f\|_{\mathcal{F}_2} \le \|f\|_{\mathcal{F}_1}$ for all f?

1.3. (2 points) Consider a 1D regression dataset $(x_i, y_i)_{i\le n}$. Denote by $X \in \mathbb{R}^n$ the vector of covariates and by $Y \in \mathbb{R}^n$ the vector of labels. Write the expression of the estimator for linear regression, including an intercept term.

1.4. (2 points) Recall that a random variable X is called sub-Gaussian with parameter σ^2 if

$$\forall \lambda \in \mathbb{R},\ \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \le e^{\lambda^2\sigma^2/2}.$$

If $(X_1, ..., X_n)$ is a collection of independent random variables and X_i is sub-Gaussian with parameter σ_i^2 for each i, give a sub-Gaussianity parameter for $S_n = \frac{1}{n}\sum_{i=1}^n X_i$.

1.5. (1 point) Write down the definition of α-strongly convexity for a differentiable function on $\mathbb{R}^d$.

2 Mean estimation (20 points)

Let $\mathcal{P}(\mathbb{R})$ be the set of all distributions supported on $\mathbb{R}$. For $\sigma^2 < \infty$, let $\mathcal{P}_{\sigma^2} = \{\rho \in \mathcal{P}(\mathbb{R}) : \mathbf{E}_{Z\sim\rho}[(Z-\mathbf{E}_{Z\sim\rho}[Z])^2] \leq \sigma^2\}$ be the subset of $\mathcal{P}(\mathbb{R})$ containing all distributions with variance bounded by σ^2. For any $\rho \in \mathcal{P}_{\sigma^2}$ and any real number θ define its population risk by

$$\mathcal{R}_\rho(\theta) = \mathbf{E}_{Z\sim\rho}[(\theta - Z)^2].$$

Let $S_n = (Z_1, \ldots, Z_n)$ be a collection of n i.i.d. samples from the same distribution $\rho \in \mathcal{P}_{\sigma^2}$, denoted in what follows by $S_n \sim \rho^{\otimes n}$. Then, for any real number θ, its empirical risk with respect to the sample S_n is defined by

$$\widehat{\mathcal{R}}_{S_n}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\theta - Z_i)^2.$$

2.1. (2 points) For a given $\rho \in \mathcal{P}_{\sigma^2}$, give the expressions of $\theta^*_\rho = \operatorname{argmin}_{\theta\in\mathbb{R}}\mathcal{R}_\rho(\theta)$ and of $\mathcal{R}_\rho(\theta^*_\rho)$.

2.2. (1 point) For a given data sample $S_n = (Z_1, \ldots, Z_n)$, give the expression of $\widehat{\theta}^{\mathrm{ERM}}(S_n) = \operatorname{argmin}_{\theta\in\mathbb{R}}\widehat{\mathcal{R}}_{S_n}(\theta)$.

2.3. (1 point) Prove that for any $\rho \in \mathcal{P}_{\sigma^2}$ and any $\theta \in \mathbb{R}$ we have $\mathcal{R}_\rho(\theta) - \mathcal{R}_\rho(\theta^*_\rho) = (\theta - \theta^*_\rho)^2$.

2.4. (2 points) Prove that for any σ^2-sub-Gaussian random variable X and any $0 < \delta < 1$, it holds

$$\mathbb{P}\left(|X - \mathbb{E}X| \geq \sqrt{2\sigma^2\log(2/\delta)}\right) \leq \delta.$$

2.5. (2 points) Let $\mathcal{P}^{\mathrm{sg}}_{\sigma^2}$ be a subset of sub-Gaussian measures in $\mathcal{P}_{\sigma^2}$ defined as

$$\mathcal{P}^{\mathrm{sg}}_{\sigma^2} = \left\{\rho \in \mathcal{P}_{\sigma^2} : \forall\lambda \in \mathbb{R},\, \mathbf{E}_{Z\sim\rho}[\exp(\lambda(Z - \mathbf{E}_{Z\sim\rho}[Z])] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)\right\}.$$

Prove that for any $\rho \in \mathcal{P}^{\mathrm{sg}}_{\sigma^2}$, any $\delta \in (0,1)$, and any integer $n \geq 1$, it holds that

$$\mathbf{P}_{S_n\sim\rho^{\otimes n}}\left(\mathcal{R}_\rho(\widehat{\theta}^{\mathrm{ERM}}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{2\sigma^2\log(2/\delta)}{n}\right) \leq \delta.$$

2.6. (2 points) Prove that for any $\rho \in \mathcal{P}_{\sigma^2}$ it holds that

$$\lim_{n\to\infty}\mathbf{P}_{S_n\sim\rho^{\otimes n}}\left(\mathcal{R}_\rho(\widehat{\theta}^{\mathrm{ERM}}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{2\sigma^2\log(2/\delta)}{n}\right) \leq \delta.$$

Hint: by the Central Limit Theorem, the random variable $G_n = n^{-1/2}\sum_{i=1}^{n}(Z_i - \mathbf{E}_{Z\sim\rho}[Z])$ converges in distribution to a Gaussian with variance σ^2.

2.7. (2 points) Prove that for any $\rho \in \mathcal{P}_{\sigma^2}$, any $\delta \in (0,1)$, and any integer $n \geq 1$, it holds that

$$\mathbf{P}_{S_n\sim\rho^{\otimes n}}\left(\mathcal{R}_\rho(\widehat{\theta}^{\mathrm{ERM}}_{S_n}) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{\sigma^2}{n\delta}\right) \leq \delta.$$

2.8. (3 points) Prove that there exists a universal constant[1] $c > 0$ and a natural number n_0 such that for any integer $n \geq n_0$ and any confidence level $\delta \in (0,1)$ there exists a distribution $\rho_{n,\delta} \in \mathcal{P}_{\sigma^2}$ such that

$$\mathbf{P}_{S_n \sim \rho_{n,\delta}^{\otimes n}} \left(\mathcal{R}_{\rho_{n,\delta}}(\widehat{\theta}^{\mathrm{ERM}}_{S_n}) - \mathcal{R}_{\rho_{n,\delta}}(\theta^*_{\rho_{n,\delta}}) \geq c \frac{\sigma^2}{n\delta} \right) \geq \delta.$$

2.9. (3 points) Let m, k be positive integers and suppose that $n = mk$. Fix any $\rho \in \mathcal{P}_{\sigma^2}$ and let $S_n \sim \rho^{\otimes n}$. Divide the sample S_n into m datasets of size k each, such that for $l \in \{1, \ldots, m\}$ we define $S_n^l = (Z_{k(l-1)+1}, \ldots, Z_{kl})$. Prove that

$$\mathbf{P}_{S_n \sim \rho^{\otimes n}} \left(\left| \mathrm{median}[\widehat{\theta}^{\mathrm{ERM}}(S_n^1), \ldots, \widehat{\theta}^{\mathrm{ERM}}(S_n^m)] - \mathbf{E}_{Z \sim \rho}[Z] \right| \geq \frac{2\sigma}{\sqrt{k}} \right) \leq \exp\left(-m/8\right).$$

Hint 1: first prove that a random variable $U \sim \mathrm{Binomial}(m, 1/4)$ satisfies $\mathbf{P}(U \geq m/2) \leq \exp(-m/8)$.

Hint 2: consider the events $E_\ell = \left\{ \left| \widehat{\theta}^{\mathrm{ERM}}(S_n^\ell) - \mathbf{E}_{Z \sim \rho}[Z] \right| \geq \frac{2\sigma}{\sqrt{k}} \right\}$.

2.10. (2 points) Fix any $\delta \in (0,1)$. Using the result of the previous question, design an estimator $\widehat{\theta}_\delta$ (i.e., the estimator is allowed to depend on δ) such that for any distribution $\rho \in \mathcal{P}_{\sigma^2}$ and any integer $n \geq 1$ it holds that

$$\mathbf{P}_{S_n \sim \rho^{\otimes n}} \left(\mathcal{R}_\rho(\widehat{\theta}_{\sigma^2,\delta}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) \geq c \frac{\sigma^2 \log(1/\delta)}{n} \right) \leq \delta,$$

where $c > 0$ is a universal constant.

3 Linear neural networks (25 points)

Let $\theta = (U, V) \in \mathbb{R}^m \times \mathbb{R}^{m \times d}$ be the weights of the neural network $f_\theta(x) = U^\top V x$, where $x \in \mathbb{R}^d$. In particular, V is the weight matrix of the first layer, m is the width of the network, the activation function is the identity and the second layer weights are given by U. Consider the function $F : \mathbb{R}^m \times \mathbb{R}^{m \times d} \to [0, \infty)$ defined by

$$F(\theta) = \frac{1}{2n} \sum_{i=1}^n \left(f_\theta(x_i) - x_i^\top w^* \right)^2,$$

where $x_1, \ldots, x_n \in \mathbb{R}^d$ are arbitrary fixed points and $w^* \in \mathbb{R}^d$ is some fixed vector. In what follows, denote $\Sigma = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top$.

3.1. (Gradient flow dynamics)

(a) (2 points) Is the function $\theta \mapsto F(\theta)$ convex? Either give a proof or provide a counterexample for some specific choice of problem parameters m, d, n, w^* and $x_1, \ldots, x_n$

(b) (2 points) Let $\theta_0 = (U_0, V_0)$ and consider the gradient flow dynamics

$$\frac{d}{dt}\theta_t = -\nabla F(\theta_t).$$

Compute $\frac{d}{dt}U_t$ and $\frac{d}{dt}V_t$, where $\theta_t = (U_t, V_t)$.

[1] That is, a constant that does not depend on the problem parameters $n, \delta, \sigma^2, \rho$.

(c) (2 points) Let $w_t = V_t^\top U_t$. Show that

$$\frac{d}{dt} w_t = -K_t \Sigma (w_t - w^*), \tag{1}$$

where $K_t = V_t^\top V_t + I_d U_t^\top U_t$ and I_d is the $d \times d$ identity matrix.

(d) (2 points) Compute $\frac{d}{dt} F(\theta_t)$ in terms of w_t, w^*, Σ and K_t.

3.2. (Finite width analysis with deterministic initialization) In this part, let $m \geq d+1$. Consider the deterministic initialization

$$U_0 = \begin{pmatrix} 1 \\ 0_{(m-1)\times 1} \end{pmatrix} \in \mathbb{R}^m \text{ and } V_0 = \begin{pmatrix} 0_{1\times d} \\ I_d \\ 0_{(m-d-1)\times d} \end{pmatrix} \in \mathbb{R}^{m\times d},$$

where $0_{a\times b}$ is an $a \times b$ matrix with all entries equal to zero and I_d is the $d \times d$ identity matrix.

(a) (4 points) Let $z \in \mathbb{R}^d$ be a vector such that $z^\top x_i = 0$ for all $i = 1, \ldots, n$. In this question, we aim to show that $z^\top w_t = 0$ for all $t \geq 0$.

i. Let $N_t = z^\top w_t$. Compute $\frac{d}{dt} N_t$.

ii. Show that for all $t \geq 0$ we have $V_t^\top V_t z = V_t^\top V_0 z$.

iii. Let $A_t = U_t^\top V_0 z$ and $B_t = V_t^\top V_0 z$. Let $C_t = (A_t, B_t)$. Compute $\frac{d}{dt} C_t$ to obtain an ordinary differential equation (ODE) that must be satisfied by C_t for all $t \geq 0$. Show that $C_t = (A_t, B_t) = (0, B_0)$ is a particular solution of the obtained ODE.

iv. The ODE for C_t obtained in the previous question has a unique solution for the initial value $C_0 = (A_0, B_0)$ (you do not need to prove this fact). Hence, the particular solution obtained in the previous question is the unique solution. Use this fact to deduce that $z^\top w_t = 0$ for all $t \geq 0$.

(b) (1 point) By the previous question, by restricting our analysis to the range of Σ, we can assume for the remainder of this problem without loss of generality that the matrix Σ has full rank. Assume furthermore that $w^* \neq 0$.

Show that there exists some $t^* > 0$ and $\varepsilon > 0$ such that $F(\theta_{t^*}) < F(0) - \varepsilon$.

(c) (2 points) For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, let $\|A\|_F$ denote its Frobenius norm defined by $\|A\|_F^2 = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} A_{ij}^2$. Prove that for all $t \geq 0$ it holds that $\|V_t\|_F^2 - d = \|U_t\|_2^2 - 1$.

(d) (2 points) Show that there exists some $c > 0$ (allowed to depend on any problem parameters other than t) such that $\|U_t\|_2^2 \geq c$ for all $t \geq t^*$. Deduce that for all $t \geq t^*$ the matrix $K_t - cI_d$ is positive semi-definite (that is, all eigenvalues of $K_t - cI_d$ are non-negative), where recall that K_t is defined in (1).

(e) (2 points) Show that $F(\theta_t)$ converges to a global minimum. What can you say about the solution $w_\infty = \lim_{t\to\infty} w_t$ (in particular, think of the case where Σ is not full rank and there exists an infinite number of optimal solutions)?

Hint: for proving that θ_t converges to a minimizer first prove that $\frac{d}{dt} F(\theta_t) \leq -c' F(\theta_t)$ for all $t \geq t^*$ and some constant $c' > 0$ independent of t; conclude the proof using Grönwall's inequality.

3.3. (Random initialization and infinite width analysis) We now consider the random initialization where $(\tilde{U}_0)_i$ and $(\tilde{V}_0)_{ij}$ are i.i.d. $\mathcal{N}(0, 1/m)$ random variables. We aim to show that as $m \to \infty$, gradient flow with the above random initialization converges to the previous dynamics with deterministic initialization.

(a) (1 point) Define

$$J^m = \begin{pmatrix} \tilde{U}_0 & \tilde{V}_0 \end{pmatrix} \in \mathbb{R}^{m \times (d+1)}$$

and let $J_k^m \in \mathbb{R}^m$ be the k-th column of J^m. Show that for any k, k', $(J_k^m)^\top (J_{k'}^m)$ converges almost surely to $\delta_{k,k'}$, where $\delta_{k,k'} = 1$ if $k = k'$ and $\delta_{k,k'} = 0$ otherwise.

(b) (2 points) Show that for all $t \geq 0$ the vector $\tilde{U}_t$ is spanned by the columns of J^m.
Let $(\tilde{V}_t)_i^{\text{col}}$ be the i-th column of $\tilde{V}$. Show that for all $t \geq 0$ and all $i \in \{1, \ldots, d\}$ the vector $(\tilde{V}_t)_i^{\text{col}}$ is spanned by the columns of J^m.

(c) (3 points) To simplify the analysis, assume that the columns of J_m are orthonormal (this is only true approximately, but this assumption will considerably simplify our analysis that follows). Using the previous part of this question, we may write

$$\tilde{U}_t = \sum_{k=1}^{d+1} \alpha(t)_k J_k^m \quad \text{and} \quad (\tilde{V}_t)_i^{\text{col}} = \sum_{k=1}^{d+1} \beta(t)_{ki} J_k^m,$$

where $\alpha(t) \in \mathbb{R}^{d+1}$ and $\beta(t) \in \mathbb{R}^{(d+1)\times d}$.

Show that $(\alpha(t), \beta(t))$ evolves according to the same dynamics as (U_t, V_t) under the deterministic initialization of part 2 of this question with $m = d + 1$.

Deduce that evolution of the predictors $\tilde{w}_t = \tilde{U}_t^\top \tilde{V}_t$ and $w_t = V_t^\top U_t$ is the same.