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 $(x_i, y_i)_{i\leq n}$ 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?

Solution: From lecture 9,

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i x_i^\top \theta}).$$

A minimizer exists if and only if the dataset is not separable, i.e., if for all $\theta \in \mathbb{R}^d$, there exists i such that $y_i \theta^\top x_i \leq 0$ (if the negation of this condition holds, then the infimum of the objective is 0 and is not attained).

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 $\|f\|_{\mathcal{F}_1} \leq \|f\|_{\mathcal{F}_2}$ for all f? or that $\|f\|_{\mathcal{F}_2} \leq \|f\|_{\mathcal{F}_1}$ for all f?

Solution: It holds $\|f\|_{\mathcal{F}_1} \leq \|f\|_{\mathcal{F}_2}$ for all f, and the converse inequality is not true in general, (this can be shown using Jensen's inequality, see lecture 6).

1.3. (2 points) Consider a 1D regression dataset $(x_i, y_i)_{i \leq n}$. 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.

Solution: In one dimension, the estimator of linear regression solves the following optimization problem:

$$\hat{\beta}_n \in \operatorname*{argmin}_{\beta \in \mathbb{R}^2} \left\| Y - \beta_0 + X\beta_1 \right\|^2.$$

Solving the gradients yields: $\hat{\beta}_0 = \bar{Y}$ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\hat{\beta}_1 = (X^\top X)^{-1} X^\top (Y - \bar{Y})$. Another solution is to add the intercept in the input matrix, writing $\tilde{X} := \left[1, X\right]$ the $(n \times 2)$ matrix where the first column is $(1, \dots, 1)^\top \in \mathbb{R}^n$, we have $\hat{\beta}_n = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y$.

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

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

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 $S_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Solution: S_n is sub-Gaussian with parameter $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2$, see the proof of Proposition 2.2 in lecture 3.

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

Solution: A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called α-strongly convex if

$$\forall x, y \in \mathbb{R}^d,\ f(x) - f(y) - \nabla f(y)^\top (x - y) \geq \frac{\alpha}{2}\|x - y\|^2.$$

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 $S_n = (Z_1, \dots, Z_n)$ 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 $\mathcal{R}_\rho(\theta_\rho^*)$.

Solution: To simplify the notation, in what follows we shall write $\mathcal{R}_\rho = \mathcal{R}$, $\mathbf{E}_{Z\sim\rho}[Z] = \mu$, $\theta^*_\rho = \theta^*$, $\widehat{\theta}^{\mathrm{ERM}} = \widehat{\theta}$.

For any $\theta \in \mathbb{R}$ we have

$$\begin{aligned}\mathcal{R}(\theta) &= \mathbf{E}(\theta - Z)^2 \\ &= \mathbf{E}(\theta - \mu + \mu - Z)^2 \\ &= (\theta - \mu)^2 + \mathrm{Var}_{Z\sim\rho}(Z). \end{aligned} \tag{1}$$

The above expression is minimized at $\theta = \mu$. Hence, we have $\theta^* = \mu$ and $\mathcal{R}(\theta^*) = \mathrm{Var}_{Z\sim\rho}(Z)$.

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

Solution: By the previous part applied to the empirical measure $\rho_{S_n} = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ we have $\widehat{\theta}(S_n) = \frac{1}{n}\sum_{i=1}^n Z_i$. That is, $\widehat{\theta}(S_n)$ is the sample mean.

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

Solution: The result is immediate by (1).

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

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

Solution: See exercise sheet 3, problem 1.3.

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

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

Solution: By exercise sheet 3, problem 1.4, the random variable

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n Z_i$$

is σ^2/n-sub-Gaussian. Hence, the result follows by exercise sheet 3, problem 1.3.

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.

Solution: Let

$$G_n = n^{-1/2} \sum_{i=1}^n (Z_i - \mu).$$

By the Central Limit Theorem, G_n converges in distribution to $N(0, \sigma^2)$ as $n \to \infty$. Let U be a random variable distributed as $N(0, \sigma^2)$. Hence,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}_{S_n\sim\rho^{\otimes n}} \left(\mathcal{R}_\rho(\widehat{\theta}^{\mathrm{ERM}}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{2\sigma^2 \log(2/\delta)}{n} \right) \\
&= \lim_{n\to\infty} \mathbf{P}_{S_n\sim\rho^{\otimes n}} \left(n \left(\frac{1}{n} \sum_{i=1}^n (Z_i - \mu) \right)^2 \geq 2\sigma^2 \log(2/\delta) \right) \\
&= \lim_{n\to\infty} \mathbf{P}_{S_n\sim\rho^{\otimes n}} \left(|G_n| \geq \sqrt{2\sigma^2 \log(2/\delta)} \right) \\
&= \mathbf{P}(|U| \geq \sqrt{2\sigma^2 \log(2/\delta)}) \leq \delta,
\end{aligned}$$

where the final line follows by noting that U is σ^2-sub-Gaussian and applying the result of exercise sheet 3 problem 1.3.

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

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

Solution: By Markov's inequality we have

$$\begin{aligned}
&\mathbf{P}_{S_n\sim\rho^{\otimes n}} \left(\mathcal{R}_\rho(\widehat{\theta}^{\mathrm{ERM}}_{S_n}) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{\sigma^2}{n\delta} \right) \\
&= \mathbf{P}_{S_n\sim\rho^{\otimes n}} \left(\left(\frac{1}{n} \sum_{i=1}^n (Z_i - \mu) \right)^2 \geq \frac{\sigma^2}{n\delta} \right) \\
&\leq \frac{\frac{1}{n}\mathrm{Var}(Z_1)}{\frac{\sigma^2}{n\delta}} \leq \delta.
\end{aligned}$$

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

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

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

Solution: Let $\rho_{n,\delta}$ be the distribution of a random variable Z such that

$$Z = \begin{cases} 0 & \text{with probability } 1-2p, \\ \frac{\sigma}{\sqrt{2p}} & \text{with probability } p, \\ -\frac{\sigma}{\sqrt{2p}} & \text{with probability } p, \end{cases}$$

where $p \in [0, 1/2]$ will be chosen later. Then $\mathbf{E}[Z] = 0$ and $\mathrm{Var}(Z) = \sigma^2$; hence, $\rho_{n,\delta} \in \mathcal{P}_{\sigma^2}$. Observer that

$$\begin{aligned}
&\mathbf{P}_{S_n \sim \rho_{n,\delta}^{\otimes n}} \left(\mathcal{R}_{\rho_{n,\delta}}(\widehat{\theta}^{\mathrm{ERM}}_{S_n}) - \mathcal{R}_{\rho_{n,\delta}}(\theta^*_{\rho_{n,\delta}}) \geq \frac{\sigma^2}{n^2 2p} \right) \\
&= \mathbf{P}_{S_n \sim \rho_{n,\delta}^{\otimes n}} \left(\left(\frac{1}{n} \sum_{i=1}^n Z_i \right)^2 \geq \frac{\sigma^2}{n^2 2p} \right) \\
&= \mathbf{P}_{S_n \sim \rho_{n,\delta}^{\otimes n}} \left(\left| \sum_{i=1}^n Z_i \right| \geq \frac{\sigma}{\sqrt{2p}} \right) \\
&\geq \mathbf{P}_{S_n \sim \rho_{n,\delta}^{\otimes n}} \left(|\{i : Z_i \neq 0\}| = 1 \right) \\
&= n \cdot 2p \cdot (1-2p)^{n-1}.
\end{aligned}$$

It remains to choose p such that for some absolute constant $c > 0$ and large enough n the following two inequalities hold:

$$\begin{aligned}
\frac{\sigma^2}{n^2 2p} &\geq c \frac{\sigma^2}{n\delta}, \\
n \cdot 2p \cdot (1-2p)^{n-1} &\geq \delta.
\end{aligned}$$

With the choice $p = \frac{e\delta}{n}$ the first inequality holds with $c = 1/(2e)$ and the second inequality holds for large enough n because

$$\begin{aligned}
&n \cdot \frac{2e\delta}{n} \cdot (1 - \frac{2e\delta}{n})^{n-1} \\
&\geq 2e\delta \cdot (1 - \frac{2e\delta}{n})^{n} \\
&\to 2e\delta \exp(-2e\delta) \geq 2\delta \text{ as } n \to \infty.
\end{aligned}$$

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

Solution: Define the event

$$E = \left\{ \left| \text{median}[\widehat{\theta}^{\text{ERM}}(S_n^1), \ldots, \widehat{\theta}^{\text{ERM}}(S_n^m)] - \mathbf{E}_{Z\sim\rho}[Z] \right| \geq \frac{2\sigma}{\sqrt{k}} \right\}.$$

For $l \in \{1, \ldots, m\}$ define the events

$$E_l = \left\{ \left| \widehat{\theta}^{\text{ERM}}(S_n^l) - \mathbf{E}_{Z\sim\rho}[Z] \right| \geq \frac{2\sigma}{\sqrt{k}} \right\}.$$

If the event E happens, then at least $m/2$ of the events $E_1, \ldots, E_m$ must happen. By Chebyshev's inequality, $\mathbf{P}(E_l) \leq 1/4$, and moreover, the events $E_1, \ldots, E_m$ are independent. Hence,

$$\mathbf{P}(E) \leq \mathbf{P}(U \geq m/2),$$

where $U \sim \text{Binomial}(m, 1/4)$. The bound stated in the hint follows via Hoeffding's inequality, which yields the desired result.

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.

Solution: Take $m = \lceil 8 \log(1/\delta) \rceil$ in the previous question.

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 $x_1, \ldots, x_n$

Solution: Take $m = n = d = x_1 = w^* = 1$. Then, $U, V \in \mathbb{R}$ and we have

$$F(\theta) = F((U, V)) = \frac{1}{2}(UV - 1)^2.$$

Let $\theta_1 = (U_1, V_1) = (1, 3)$ and $\theta_2 = (U_2, V_2) = (3, 1)$. We have

$$F(\theta_1) = F(\theta_2) = \frac{1}{2}(3-1)^2 = 2. \quad \text{and} \quad F((\theta_1+\theta_2)/2) = F((2,2)) = \frac{1}{2}(4-1)^2 = 9/2.$$

It follows that

$$F((\theta_1+\theta_2)/2) = \frac{9}{2} > 2 = \frac{1}{2}(F(\theta_1) + F(\theta_2)).$$

The above inequality shows that F is non-convex.

(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 $\theta_t = (U_t, V_t)$.

Solution: Observe that for any x_i we have

$$\nabla_U \frac{1}{2}(f_\theta(x_i) - x_i^\top w^*) = \nabla_U \frac{1}{2}(x_i^\top V^\top U - x_i^\top w^*)^2 = V x_i (f_\theta(x_i) - x_i^\top w^*)$$

and

$$\nabla_V F(\theta) = \nabla_V \frac{1}{2}(\mathrm{trace}(V^\top U x_i^\top) - x_i^\top w^*)^2 = U x_i^\top (f_\theta(x_i) - x_i^\top w^*).$$

Let $X \in \mathbb{R}^{n\times d}$ be the matrix whose i-th row equal to $x_i^\top$ and let $w_t = V_t^\top U_t$ so that $f_{\theta_t}(x) = w_t^\top x$. Then $\Sigma = X^\top X/n$ and from the above equations we have

$$\begin{aligned}\frac{d}{dt}U_t &= -\nabla_{U_t} F((U_t, V_t)) \\ &= -\frac{1}{n} V_t \left(\sum_{i=1}^n (f_{\theta_t}(x_i) - x_i^\top w^*) x_i \right) \\ &= -V_t \Sigma (w_t - w^*)\end{aligned} \tag{2}$$

and similarly

$$\begin{aligned}\frac{d}{dt}V_t &= -\nabla_{V_t} F((U_t, V_t)) \\ &= -\frac{1}{n} \sum_{i=1}^n (f_{\theta_t}(x_i) - x_i^\top w^*) U_t x_i^\top \\ &= -U_t (w_t - w^*)^\top \Sigma.\end{aligned} \tag{3}$$

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

$$\frac{d}{dt}w_t = -K_t \Sigma (w_t - w^*), \tag{4}$$

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.

Solution: By the chain rule and equations (2) and (3) we have

$$\frac{d}{dt}w_t$$

$$
\begin{aligned}
&= \frac{d}{dt}(V_t^\top U_t) \\
&= \left(\frac{d}{dt}V_t^\top\right) U_t + V_t^\top \left(\frac{d}{dt}U_t\right) \\
&= \left(-U_t(w_t - w^*)^\top \Sigma\right)^\top U_t + V_t^\top \left(-V_t \Sigma (w_t - w^*)\right) \\
&= -\Sigma(w_t - w^*)\|U_t\|_2^2 - V_t^\top V_t \Sigma (w_t - w^*) \\
&= -\left(I_d \|U_t\|_2^2 + V_t^\top V_t\right) \Sigma (w_t - w^*),
\end{aligned}
$$

which is what we wanted to show.

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

Solution: Observe that

$$
F(\theta_t) = G(V_t^\top U_t) = G(w_t), \quad \text{where} \quad G(w) = \frac{1}{2n}\sum_{i=1}^n \left(x_i^\top w - x_i^\top w^*\right).
$$

Letting $X \in \mathbb{R}^{n\times d}$ be the matrix whose i-th row equals $x_i^\top$ we have $G(w) = \frac{1}{2n}\|Xw - Xw^*\|_2^2$. In particular,

$$
\nabla G(w) = \frac{1}{n} X^\top (Xw - Xw^*) = \Sigma(w - w^*).
$$

Hence,

$$
\frac{d}{dt}F(\theta_t) = \frac{d}{dt}G(w_t) = \langle \nabla G(w_t), \frac{d}{dt}w_t \rangle = -(w_t - w^*)^\top \Sigma K_t \Sigma (w_t - w^*). \tag{5}
$$

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 $t \geq 0$.

Solution: Define

$$
N_t = z^\top w_t
$$

and observe that by (4) we have

$$\frac{d}{dt}N_t = -z^\top I_d \|U_t\|_2^2 \underbrace{z^\top X^\top}_{(Xz)^\top=0} X(w_t - w^*) - z^\top V_t^\top V_t X^\top X(w_t - w^*)$$
$$= -z^\top V_t^\top V_t X^\top X(w_t - w^*).$$

To simplify the above derivative further, notice using (3) that

$$(\frac{d}{dt}V_t)z = -\frac{1}{n}U_t(w_t - w^*)^\top X^\top \underbrace{Xz}_{=0} = 0.$$

In particular, the above implies that for all $t \geq 0$ we have $V_t z = V_0 z$. Therefore, we have

$$\frac{d}{dt}N_t = -z^\top V_0^\top V_t X^\top X(w_t - w^*).$$

We now aim to show that for all $t \geq 0$ we have $V_t^\top V_0 z = z$, which, using the above identity, implies that $\frac{d}{dt}N_t = 0$, thus proving the claim of this exercise.
To prove that $V_t^\top V_0 z = z$ for all $t \geq 0$ consider

$$A_t = U_t^\top V_0 z,$$
$$B_t = V_t^\top V_0 z.$$

We will show that $B_t = B_0 = z$ for all $t \geq 0$. Indeed, observe that

$$\frac{d}{dt}A_t = -(w_t - w^*)^\top \Sigma V_t^\top V_0 z = -(w_t - w^*)^\top \Sigma B_t,$$
$$\frac{d}{dt}B_t = -\Sigma(w_t - w^*)^\top U_t^\top V_0 z = -\Sigma(w_t - w^*)A_t.$$

The above ODE has a unique solution. We can verify that $(A_t, B_t) = (0, B_0)$ is a solution. Hence, $B_t = B_0$ for all $t \geq 0$, which implies that $\frac{d}{dt}N_t = 0$ for all $t \geq 0$. Hence, $N_t = N_0 = 0$ for all $t \geq 0$, which is what we wanted to show.

(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 $F(\theta_{t^*}) < F(0) - \varepsilon$.

Solution: Using the computation of $\frac{d}{dt}F(\theta_t)$ in (5) we have

$$\frac{d}{dt}F(\theta_t) = -(w_t - w^*)^\top \Sigma K_t \Sigma(w_t - w^*).$$

Because Σ is full rank, it has positive eigenvalues. Also, K_0 has only positive eigenvalues. Hence, the matrix $\Sigma K_0 \Sigma$ has only positive eigenvalues. Since $w_0 = 0$ and $w^* \neq 0$, we have

$$\frac{d}{dt}F(\theta_t)|_{t=0} = -(w^*)^\top \Sigma K_0 \Sigma w^* < 0.$$

The result follows.

(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 $\|V_t\|_F^2 - d = \|U_t\|_2^2 - 1$.

Solution: We have

$$\|V_t\|_F^2 = \text{trace}(V_t^\top V_t)$$

and

$$\frac{d}{dt}(V_t^\top V_t) = -\Sigma(w_t - w^*)U_t^\top V_t - V_t^\top U_t(w_t - w^*)^\top \Sigma.$$

It follows that

$$\begin{aligned}
&\frac{d}{dt}\|V_t\|_F^2 \\
&= \text{trace}\left(-\Sigma(w_t - w^*)U_t^\top V_t\right) + \text{trace}\left(-V_t^\top U_t(w_t - w^*)^\top \Sigma.\right) \\
&= \text{trace}\left(-U_t^\top V_t \Sigma(w_t - w^*)\right) + \text{trace}\left(-(w_t - w^*)^\top \Sigma V_t^\top U_t\right) \\
&= 2\langle U_t, \frac{d}{dt}U_t\rangle \\
&= \frac{d}{dt}\|U_t\|_2^2.
\end{aligned}$$

Hence, for all $t \geq 0$ we have $\|V_t\|_F^2 - \|U_t\|_2^2 = \|V_0\|_F^2 - \|U_0\|_2^2 = d - 1$. Rearranging yields the desired result.

(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 (4).

Solution: Let $G(w_t) = \frac{1}{2n}\|X(w_t - w^*)\|_2^2$. Then $G(w_t) = F(\theta_t)$. By question 2 (b), for all $t \geq t^*$ we have $G(w_t) < G(0) - \varepsilon$ for some $\varepsilon > 0$. Since all eigenvalues of Σ are strictly positive, we have, in particular, $\|w_t\|_2^2 \geq c'$ for all $t \geq 0$ and some constant c' (that depends on Σ and w^*). It follows that for all $t \geq t^*$ we have

$$\begin{aligned}
c' &\leq \|w_t\|_2^2 \|w_t\|_F^2 \\
&= \|V_t^\top U_t\|_F^2 \\
&= \text{trace}(U_t^\top V_t V_t^\top U_t) \\
&= \text{trace}(V_t V_t^\top U_t U_t^\top) \\
&= \text{trace}(V_t V_t^\top U_t U_t^\top) \\
&= \langle V_t V_t^\top, U_t U_t^\top \rangle_F \\
&\leq \|V_t V_t^\top\|_F \|U_t U_t^\top\|_F \\
&\leq \|V_t\|_F^2 \|U_t\|_2^2 \\
&= (d - 1 + \|U_t\|_2^2)\|U_t\|_2^2.
\end{aligned}$$

From the above, it follows that for all $t \geq t^*$ we have $\|U_t\|_2^2 \geq c'' > 0$ for some constant c''.

It follows that for all $t \geq t^*$ we have $K_t \succcurlyeq I_d \|U_t\|_2^2 \succcurlyeq c'' I_d$, which is what we wanted to show (the notation $A \succcurlyeq B$ means that $A - B$ is positive semi-definite).

(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 $\frac{d}{dt}F(\theta_t) \le -c'F(\theta_t)$ for all $t \ge t^*$ and some constant $c' > 0$ independent of t; conclude the proof using Grönwall's inequality.

Solution: Let

$$G(w_t) = \frac{1}{2n}\|X(w_t - w^*)\|_2^2 = \frac{1}{2}\|\sqrt{\Sigma}(w_t - w^*)\|_2^2.$$

Then, $G(w_t) = F(\theta_t)$ and we have from (5)

$$\begin{aligned}\frac{d}{dt}G(w_t) &= -\|\sqrt{K_t}\Sigma(w_t - w^*)\|_2^2 \le -c''\|\Sigma(w_t - w^*)\|_2^2 \\ &\le -c'''\|\sqrt{\Sigma}(w_t - w^*)\|_2^2 = -c'''G(w_t),\end{aligned}$$

where $c''' > 0$ because Σ is full rank. By Grönwall's inequality, for any $t \ge t^*$ we have

$$G(w_t) \le \exp(-c'''(t - t^*))G(w_t^*) \le \exp(-c'''(t - t^*))G(0).$$

Thus, $G(w_t) \to 0$ as $t \to \infty$ and 0 is a global optimum of G.
As we have shown before, w_t stays in the span of the data. Since as $t \to \infty$, the predictor $\langle w_t, \cdot\rangle$ interpolates the data, it follows that w_t converges to a minimum ℓ_2 norm solution.

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.

Solution: The result follows immediately by the law of large numbers.

(b) (2 points) Show that for all $t \ge 0$ 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 $t \ge 0$ and all $i \in \{1, \dots, d\}$ the vector $(\tilde{V}_t)_i^{\text{col}}$ is spanned by the columns of J^m.

Solution: Denote $\xi_t = \Sigma(\tilde{w}_t - w^*)$ Then, we have

$$\frac{d}{dt}\tilde{U}_t = -\tilde{V}_t\xi_t = -\sum_{i=1}^{d}(\tilde{V}_t)_i^{\text{col}}(\xi_t)_i. \tag{6}$$

Notice that as long as the columns of $\tilde{V}_t$ belong to the span of the columns of J_m, so does the time derivative $\frac{d}{dt}\tilde{U}_t$. Similarly, we have

$$\frac{d}{dt}\tilde{V}_t = -\tilde{U}_t\xi_t^\top. \tag{7}$$

If $\tilde{U}_t$ belongs to the span of the columns of J_m, then so does the columns of the derivative $\frac{d}{dt}\tilde{V}_t$ (each column is proportional to $\tilde{U}_t$).
Observing that $\tilde{V}_0$ and $\tilde{U}_0$ trivially satisfy the column span condition, the result follows.

(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.

Solution: By (6), we have

$$\begin{aligned}
\frac{d}{dt}\tilde{U}_t &= -\sum_{i=1}^{d} (\tilde{V}_t)_i^{\text{col}} (\xi_t)_i. \\
&= -\sum_{i=1}^{d} (\sum_{k=1}^{d+1} \beta(t)_{ki} J_k^m)(\xi_t)_i. \\
&= -\sum_{k=1}^{d+1} J_k^m \sum_{i=1}^{d} \beta(t)_{ki} (\xi_t)_i. \\
&= -\sum_{k=1}^{d+1} J_k^m (\beta(t)\xi_t)_k.
\end{aligned}$$

In particular, we have

$$\frac{d}{dt}\alpha(t) = -\beta(t)\xi_t.$$

Similarly, from (7) we have

$$\frac{d}{dt}\tilde{V}_t = -\tilde{U}_t \xi_t^\top = -\sum_{k=1}^{d+1} \alpha(t)_k J_k^m \xi_t^\top.$$

Hence

$$\frac{d}{dt}\beta(t) = -\alpha(t)\xi_t^\top.$$

Hence, $\alpha(t)$ and $\beta(t)$ evolve according to the same dynamics as U_t and V_t in (2) and (3). Observe that using the orthonomality of columns of J_m we have

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

Therefore,

$$\tilde{w}_t = \tilde{V}_t^\top \tilde{U}_t = \beta(t)^\top \alpha(t).$$

Noting that $\alpha(0)$ and $\beta(0)$ match the deterministic initialization of U_0, V_0 (with $m = d+1$) in the second part of this question concludes the proof.