Convex optimization

Project 5

Pierre-Gabriel Meyrignac

December 20, 2025

Description. Large language models (LLMs) (e.g., ChatGPT) are artificial intelligence (AI) systems that generate text. To make them useful, we need to *align* them with human preferences. Alignment is the process of fine-tuning them so that their outputs are helpful, honest, and harmless, based on human preferences. Learning those preferences is the task of *preference elicitation*. Since we cannot ask a human to rate every possible AI response (expensive), we must choose a small subset of responses to show to humans. The process of selecting which data to collect is known as an *optimal experimental design* (OED) problem. Here, we use convex optimization to choose the most informative prompts.

We will assume that for any given user prompt $i \in [L] := \{1, \dots, L\}$ (i.e., a question to the AI), there are K candidate answers. Each candidate answer k to the prompt i is represented by a vector $a_{i,k} \in \mathbb{R}^d$ called an *embedding*, which captures its semantic meaning. We assume that a human's rating $r_{i,k} \in \mathbb{R}$ of this answer is a linear function

$$r_{i,k} = a_{i,k}^\top \theta + \varepsilon_{i,k},$$

of its embedding $a_{i,k}$, where $\theta \in \mathbb{R}^d$ is an unknown parameter vector that defines human preferences, and $\varepsilon_{i,k} \sim \mathcal{N}(0, \sigma^2)$ is random noise following a Gaussian distribution with mean 0 and variance σ^2. Our goal is to estimate θ accurately. The core OED problem is: *with a limited budget of N total human ratings, which prompts (and their associated lists of answers) should we show to humans to learn θ most effectively?*

To formalize this problem, we concatenate the embedding vectors for each prompt i into a matrix

$$A_i = [a_{i,1}, a_{i,2}, \dots, a_{i,K}] \in \mathbb{R}^{d \times K},$$

and similarly stack their ratings

$$r_i = [r_{i,1}, r_{i,2}, \dots, r_{i,K}]^\top \in \mathbb{R}^K.$$

Throughout this project, we assume that the matrix

$$X = [A_1 \ \cdots \ A_L] \in \mathbb{R}^{d \times KL}$$

has full row rank, i.e.,

$$d = \operatorname{rank}(X) = \operatorname{rank}(XX^\top) = \operatorname{rank}\left(\sum_{i=1}^{L} A_i A_i^\top\right).$$

The matrix $A_i A_i^\top = \sum_{k=1}^{K} a_{i,k} a_{i,k}^\top$ can be viewed as the information gained from entering prompt i, that is, this matrix captures the diversity of the candidate answers for this prompt. For instance, if all answers are very similar (e.g., all short and polite), they correspond to similar semantic directions. Hence, $A_i A_i^\top$ is a matrix with a few large eigenvalues, meaning the prompt only explores a narrow subspace of the embedding space.

From now on we use

$$\Delta_L = \left\{\pi \in \mathbb{R}_+^L : \sum_{i=1}^{L} \pi_i = 1\right\}$$

to denote the set of sampling distributions. Given any $\pi \in \Delta_L$, we sample the prompts $i_1, \ldots, i_N \in [L]$ independently from π. We can then construct the sampled design matrix

$$\hat{X} = [A_{i_1} \ \cdots \ A_{i_N}] \in \mathbb{R}^{d \times KN}$$

by stacking the individual answer embeddings. We can also construct the corresponding sampled rating vector

$$\hat{r} = [\hat{r}_1^\top, \ldots, \hat{r}_N^\top]^\top \in \mathbb{R}^{KN}.$$

Assuming that $\operatorname{rank}(\hat{X}) = d$ (i.e., our sampled prompts are sufficiently diverse), it follows that $d = \operatorname{rank}(\hat{X})$ so that $\hat{X}\hat{X}^\top \in \mathbb{S}_+^d$ is invertible. Hence, the least-squares estimator

$$\hat{\theta} = (\hat{X}\hat{X}^\top)^{-1}\hat{X}\hat{r}$$

solves the regression problem $\min_{\theta \in \mathbb{R}^d} \|\hat{r} - \hat{X}^\top\theta\|_2^2$, is well-defined and has covariance matrix

$$\operatorname{Cov}(\hat{\theta}) = \sigma^2(\hat{X}\hat{X}^\top)^{-1}.$$

The law of large numbers implies that

$$\lim_{N\to\infty} \frac{1}{N}\hat{X}\hat{X}^\top = \lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{N} A_{i_n} A_{i_n}^\top = \sum_{i=1}^{L} \pi_i A_i A_i^\top,$$

and thus we have

$$\operatorname{Cov}(\hat{\theta}) \approx \frac{\sigma^2}{N}\left(\sum_{i=1}^{L} \pi_i A_i A_i^\top\right)^{-1} \quad \text{when } N \text{ is large.}$$

This observation motivates us to define

$$V(\pi) = \sum_{i=1}^{L} \pi_i A_i A_i^\top$$

as a weighted information matrix. Intuitively, $V(\pi)$ captures the *average information gained* if

many prompts are sampled under design π, and it scales inversely with the covariance matrix of $\hat{\theta}$.

We now describe a principled approach to OED called *G-optimal design*, which aims to select π that minimizes prediction variance. However, we want to avoid a situation where the model is very confident. on some prompts but extremely uncertain on others. Instead, we aim to control the *worst prediction variance* over all prompts. To formalize this, consider the predicted rating $a_{i,k}^\top \hat{\theta}$ corresponding to the candidate answer $a_{i,k}$. The variance of this prediction is

$$\mathrm{Var}(a_{i,k}^\top \hat{\theta}) = a_{i,k}^\top \,\mathrm{Cov}(\hat{\theta})\, a_{i,k} \approx \frac{\sigma^2}{N}\, a_{i,k}^\top V(\pi)^{-1} a_{i,k}.$$

This reflects how uncertain we are about the rating of that specific answer. For an entire prompt i, the total prediction variance over all its K candidate answers is

$$\sum_{k=1}^{K} \mathrm{Var}(a_{i,k}^\top \hat{\theta}) \approx \frac{\sigma^2}{N} \sum_{k=1}^{K} a_{i,k}^\top V(\pi)^{-1} a_{i,k} = \frac{\sigma^2}{N}\, \mathrm{tr}\big(A_i^\top V(\pi)^{-1} A_i\big).$$

The G-optimal design problem

$$\min_{\pi \in \Delta_L} \max_{i \in [L]} \mathrm{tr}\big(A_i^\top V(\pi)^{-1} A_i\big) \tag{G}$$

minimizes the worst-case total prediction variance across all prompts $i \in [L]$. Here, if $\det(V(\pi)) = 0$ ($V(\pi)$ is singular) for some $\pi \in \Delta_L$, then we interpret $\max_{i\in[L]} \mathrm{tr}\big(A_i^\top V(\pi)^{-1} A_i\big)$ as ∞.

1 Reformulation of the Robust problem

1.1 Reformulation of the Robust problem

We start with:

$$\min_{\pi \in \Delta_L} \max_{i \in [L]} \mathrm{tr}(A_i^\top V(\pi)^{-1} A_i).$$

Let

$$g_i(\pi) = \mathrm{tr}(A_i^\top V(\pi)^{-1} A_i).$$

Then (G) is equivalent to

$$\min_{\pi \in \Delta_L} \max_{i \in [L]} g_i(\pi).$$

We do an epigraphical reformulation using an auxiliary variable s and we obtain:

$$(G) \equiv \begin{array}{ll} \min\limits_{\pi \in \Delta_L,\, s \in \mathbb{R}} & s \\ \text{s.t.} & \mathrm{tr}(A_i^\top V(\pi)^{-1} A_i) \le s, \quad \forall i \in [L]. \end{array} \tag{6}$$

We know that

$$\mathrm{tr}(A_i^\top V(\pi)^{-1} A_i) = \mathrm{tr}\Big(V(\pi)^{-1} \sum_{k=1}^{K} a_{i,k} a_{i,k}^\top\Big) = \sum_{k=1}^{K} a_{i,k}^\top V(\pi)^{-1} a_{i,k}.$$

Hence

$$\operatorname{tr}(A_i^\top V(\pi)^{-1} A_i) = \sum_{k=1}^{K} a_{i,k}^\top V(\pi)^{-1} a_{i,k}.$$

Therefore (G) becomes

$$\min_{\pi \in \Delta_L,\, s \in \mathbb{R}} \quad s \quad \text{s.t.} \quad \sum_{k=1}^{K} a_{i,k}^\top V(\pi)^{-1} a_{i,k} \le s, \quad \forall i \in [L]. \tag{6}$$

We then use auxiliary variables $s_{i,k}$ and we can do an epigraphical reformulation of the inequality constraint. Thus (6) is equivalent to:

$$\min_{\pi \in \Delta_L,\, s \in \mathbb{R},\, s_{i,k} \in \mathbb{R}} \quad s \quad \text{s.t.} \quad \sum_{k=1}^{K} s_{i,k} \le s, \quad \forall i \in [L],$$

$$a_{i,k}^\top V(\pi)^{-1} a_{i,k} \le s_{i,k}, \quad \forall\, i \in [L],\ \forall\, k \in [K].$$

We know that $V(\pi)$ is positive semidefinite, since

$$v^\top A_i A_i^\top v = (A_i^\top v)^\top (A_i^\top v) = \|A_i^\top v\|_2^2 \ge 0, \quad \forall\, v \in \mathbb{R}^d.$$

Therefore we can use the Schur lemma:

$$a_{i,k}^\top V(\pi)^{-1} a_{i,k} \le s_{i,k} \quad \Longleftrightarrow \quad \begin{pmatrix} s_{i,k} & a_{i,k}^\top \\ a_{i,k} & V(\pi) \end{pmatrix} \succeq 0.$$

We therefore obtain:

$$(G) \equiv \begin{array}{ll} \min\limits_{\pi \in \Delta_L,\, s \in \mathbb{R},\, s_{i,k} \in \mathbb{R}} & s \\ \text{s.t.} & \sum\limits_{k=1}^{K} s_{i,k} \le s, \quad \forall i \in [L], \\ & \begin{pmatrix} s_{i,k} & a_{i,k}^\top \\ a_{i,k} & V(\pi) \end{pmatrix} \succeq 0, \quad \forall\, i \in [L],\ \forall\, k \in [K]. \end{array} \tag{6}$$

Finally, since $V(\pi)$ must be invertible, we must have

$$\operatorname{rank}\big(V(\pi)\big) = d.$$

This gives us the final form of problem (6).

1.2 Relaxation of condition

Given that $\max_{i \in [L]} \operatorname{tr}\big(A_i^\top V(\pi)^{-1} A_i\big) = \infty$ when $\det(V(\pi)) = 0$, by minimizing over π we will land on π such that $\max_{i \in [L]} \operatorname{tr}\big(A_i^\top V(\pi)^{-1} A_i\big) < \infty \Leftrightarrow \det(V(\pi)) \neq 0$ so we dont need a rank constraint.

1.3 Optimal points candidate

Let

$$f(\pi) = \log\det V(\pi).$$

To compute

$$\frac{\partial f(\pi)}{\partial \pi_i},$$

we use the chain rule:

$$\frac{\partial f(\pi)}{\partial \pi_i} = \left\langle \frac{\partial f(\pi)}{\partial V(\pi)}, \frac{\partial V(\pi)}{\partial \pi_i} \right\rangle$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product defined as:

$$\langle X, Y \rangle = \mathrm{tr}(X^\top Y) \quad \text{for } X, Y \in \mathbb{R}^{d\times d}.$$

Hence we have:

$$\frac{\partial f(\pi)}{\partial V(\pi)} = V(\pi)^{-1}$$

and using

$$V(\pi) = \sum_{i=1}^{L} \pi_i A_i A_i^\top, \quad \text{we have} \quad \frac{\partial V(\pi)}{\partial \pi_i} = A_i A_i^\top.$$

Therefore,

$$\frac{\partial f(\pi)}{\partial \pi_i} = \left\langle V(\pi)^{-1},\ A_i A_i^\top \right\rangle = \mathrm{tr}\Big((V(\pi)^{-1})^\top A_i A_i^\top\Big) = \mathrm{tr}\Big(A_i^\top V(\pi)^{-1} A_i\Big),$$

since $V(\pi)$ is symmetric and therefore $V(\pi)^{-1}$ is also symmetric.

Now coming back to our problem:

$$\max_{\pi\in\Delta_L} f(\pi) \quad \text{s.t.} \quad \pi_i \geq 0 \ \ \forall i \in [L], \qquad \sum_{i=1}^{L} \pi_i = 1.$$

We define the Lagrangian operator:

$$L(\pi, \lambda, \mu) = -\log\det\big(V(\pi)\big) - \lambda^\top \pi + \mu\Big(\sum_{i=1}^{L} \pi_i - 1\Big),$$

where $\lambda \in \mathbb{R}^L$ and $\mu \in \mathbb{R}$ are dual variables.

We then compute

$$\frac{\partial L}{\partial \pi_i} = -\,\mathrm{tr}\left(A_i^\top V(\pi)^{-1} A_i\right) - \lambda_i + \mu.$$

We can therefore state the KKT conditions at an optimal solution π^*, with associated dual variables λ^* and μ^*:

(PF) Primal feasibility:

$$\pi_i^* \geq 0 \ \ \forall i \in [L], \qquad \sum_{i=1}^{L} \pi_i^* = 1.$$

(DF) Dual feasibility:

$$\lambda_i^* \geq 0 \quad \forall i \in [L].$$

(CS) Complementary slackness:

$$\lambda_i^* \pi_i^* = 0 \quad \forall i \in [L].$$

(S) Stationarity:

$$-\operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) - \lambda_i^* + \mu^* = 0, \quad \forall i \in [L].$$

From (S) we have

$$\operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) = \mu^* - \lambda_i^*.$$

From (DF), we know that $\lambda_i^* \geq 0$ for all $i \in [L]$. Therefore

$$\operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) \leq \mu^*. \tag{i}$$

Now multiplying each side of the stationarity condition by π_i^*, we obtain

$$\pi_i^* \operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) = \pi_i^* \mu^* - \lambda_i^* \pi_i^*.$$

By (CS), $\lambda_i^* \pi_i^* = 0$, hence

$$\pi_i^* \operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) = \pi_i^* \mu^*.$$

Summing this over $i \in [L]$, we have

$$\sum_{i=1}^{L} \pi_i^* \operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) = \mu^* \sum_{i=1}^{L} \pi_i^*.$$

Using (PF), $\sum_{i=1}^{L} \pi_i^* = 1$, so the right-hand side is μ^*.

On the left-hand side, we can write

$$\begin{aligned}
\sum_{i=1}^{L} \pi_i^* \operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) &= \operatorname{tr}\left(V(\pi^*)^{-1} \sum_{i=1}^{L} \pi_i^* A_i A_i^\top\right) \\
&= \operatorname{tr}\left(V(\pi^*)^{-1} V(\pi^*)\right) \\
&= \operatorname{tr}(I_d) = d.
\end{aligned}$$

Therefore we have

$$\mu^* = d.$$

Hence, from (i) we obtain

$$\operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) \leq d, \quad \forall i \in [L].$$

Finally, if $\pi_i^* > 0$, then by (CS) we have $\lambda_i^* = 0$, and the stationarity condition becomes

$$\operatorname{tr}\left(A_i^\top V(\pi^*)^{-1} A_i\right) = \mu^* = d.$$

Thus,

$$\boxed{\operatorname{tr}\left(A_i^\top V(\pi^*)^{-1}A_i\right) \le d \quad \forall i \in [L], \qquad \operatorname{tr}\left(A_i^\top V(\pi^*)^{-1}A_i\right) = d \text{ if } \pi_i^* > 0.}$$

1.4 Key equivalence between two ways of choosing prompts (i.e., designing the sampling distribution π) for the LLM preference-elicitation problem.

Part 1

Let

$$s_i(\pi) := \operatorname{tr}\big(A_i^\top V(\pi)^{-1}A_i\big), \qquad i \in [L].$$

From the optimality candidate part, we know that

$$\sum_{i=1}^{L} \pi_i s_i(\pi) = d. \tag{1}$$

We use the fact that a weighted average is always bounded above by the maximum:

$$\sum_{i=1}^{L} \pi_i s_i(\pi) \;\le\; \sum_{i=1}^{L} \pi_i \max_{j\in[L]} s_j(\pi) = \max_{j\in[L]} s_j(\pi) \sum_{i=1}^{L} \pi_i = \max_{j\in[L]} s_j(\pi),$$

since $\sum_{i=1}^{L} \pi_i = 1$. Therefore, using (1), we obtain:

$$\max_{i\in[L]} s_i(\pi) \ge d. \tag{2}$$

In particular,

$$\min_{\pi\in\Delta_L} \max_{i\in[L]} s_i(\pi) \;\ge\; d. \tag{3}$$

Now let π_D^* be an optimal solution of (D). From Question 1.3 we have:

$$s_i(\pi_D^*) = \operatorname{tr}\big(A_i^\top V(\pi_D^*)^{-1}A_i\big) \le d, \qquad \forall i \in [L]. \tag{4}$$

Hence, in particular,

$$\max_{i\in[L]} s_i(\pi_D^*) = \max_{i\in[L]} \operatorname{tr}\big(A_i^\top V(\pi_D^*)^{-1}A_i\big) \le d. \tag{5}$$

Combining (3) and (5), we get:

$$d \;\le\; \min_{\pi\in\Delta_L} \max_{i\in[L]} \operatorname{tr}\big(A_i^\top V(\pi)^{-1}A_i\big) \;\le\; \max_{i\in[L]} \operatorname{tr}\big(A_i^\top V(\pi_D^*)^{-1}A_i\big) \;\le\; d.$$

Therefore,

$$\min_{\pi\in\Delta_L} \max_{i\in[L]} \operatorname{tr}\big(A_i^\top V(\pi)^{-1}A_i\big) = \max_{i\in[L]} \operatorname{tr}\big(A_i^\top V(\pi_D^*)^{-1}A_i\big) = d.$$

Thus, any optimal solution π_D^* of (D) is also optimal for (G), and the optimal value of (G) is

$$\boxed{d.}$$

Part 2

We now prove the converse statement: any optimal solution of (G) is also optimal for (D).

Recall the notation

$$s_i(\pi) := \mathrm{tr}\big(A_i^\top V(\pi)^{-1} A_i\big), \qquad i \in [L],$$

so that the objective of (G) is

$$g(\pi) = \max_{i\in[L]} s_i(\pi).$$

From Question 1.3, for any $\pi \in \Delta_L$ such that $V(\pi) \succ 0$ we have

$$\sum_{i=1}^{L} \pi_i \, s_i(\pi) = d. \tag{1}$$

From the first part of Question 1.4 we already know that the optimal value of (G) is d:

$$\min_{\pi\in\Delta_L} \max_{i\in[L]} s_i(\pi) = d.$$

Let π_G^* be an optimal solution of (G). Then

$$\max_{i\in[L]} s_i(\pi_G^*) = d. \tag{2}$$

Since this value is finite, by the convention in the statement we must have $V(\pi_G^*) \succ 0$, and therefore identity (1) applies at $\pi = \pi_G^*$:

$$\sum_{i=1}^{L} \pi_{G,i}^* \, s_i(\pi_G^*) = d. \tag{3}$$

Since $\max_{i\in[L]} s_i(\pi_G^*) = d$, we have

$$s_i(\pi_G^*) \le d \quad \text{for all } i \in [L].$$

We now show that if $\pi_{G,i}^* > 0$ then necessarily $s_i(\pi_G^*) = d$. Indeed, suppose by contradiction that there exists an index i such that

$$\pi_{G,i}^* > 0 \quad \text{and} \quad s_i(\pi_G^*) < d.$$

Then we can write

$$\sum_{j=1}^{L} \pi_{G,j}^* \, s_j(\pi_G^*) = \pi_{G,i}^* s_i(\pi_G^*) + \sum_{j\neq i} \pi_{G,j}^* s_j(\pi_G^*).$$

Using $s_i(\pi_G^*) < d$ and $s_j(\pi_G^*) \le d$ for $j \neq i$, we obtain the upper bound

$$\sum_{j=1}^{L} \pi_{G,j}^* \, s_j(\pi_G^*) \le \pi_{G,i}^* s_i(\pi_G^*) + \sum_{j\neq i} \pi_{G,j}^* d.$$

Since $s_i(\pi_G^*) < d$, it follows that

$$\pi_{G,i}^* s_i(\pi_G^*) < \pi_{G,i}^* d,$$

and hence

$$\sum_{j=1}^{L} \pi_{G,j}^*\, s_j(\pi_G^*) < \pi_{G,i}^* d + \sum_{j\neq i} \pi_{G,j}^* d = d\Big(\pi_{G,i}^* + \sum_{j\neq i} \pi_{G,j}^*\Big) = d\sum_{j=1}^{L} \pi_{G,j}^* = d,$$

where we used $\sum_{j=1}^{L} \pi_{G,j}^* = 1$. This contradicts (3), which states that $\sum_{j=1}^{L} \pi_{G,j}^* s_j(\pi_G^*) = d$. Therefore our assumption was false, and we must have

$$s_i(\pi_G^*) = d \quad \text{whenever } \pi_{G,i}^* > 0.$$

Thus we have shown:

$$s_i(\pi_G^*) = \operatorname{tr}(A_i^\top V(\pi_G^*)^{-1} A_i) \leq d \quad \forall i,$$

and

$$\operatorname{tr}\left(A_i^\top V(\pi_G^*)^{-1} A_i\right) = d \qquad \text{whenever } \pi_{G,i}^* > 0. \tag{4}$$

We now show that π_G^* satisfies the KKT conditions of (D).

Problem (D) is

$$\max_{\pi\in\Delta_L} f(\pi) = \log\det V(\pi).$$

From Question 1.3 we know that

$$\frac{\partial f(\pi)}{\partial \pi_i} = \operatorname{tr}\big(A_i^\top V(\pi)^{-1} A_i\big) = s_i(\pi).$$

Introduce dual variables $\lambda_i \geq 0$ for the constraints $\pi_i \geq 0$ and $\mu \in \mathbb{R}$ for the equality $\sum_{i=1}^{L} \pi_i = 1$. The KKT conditions for (D) are:

$$\pi \in \Delta_L, \quad \lambda_i \geq 0, \quad \lambda_i \pi_i = 0, \quad s_i(\pi) - \mu + \lambda_i = 0.$$

We now define dual variables corresponding to π_G^*:

$$\mu^* := d, \qquad \lambda_i^* := d - s_i(\pi_G^*).$$

By (4), $s_i(\pi_G^*) \leq d$ for all i, thus $\lambda_i^* \geq 0$ and dual feasibility holds.

Check stationarity at $\pi = \pi_G^*$:

$$s_i(\pi_G^*) - \mu^* + \lambda_i^* = s_i(\pi_G^*) - d + (d - s_i(\pi_G^*)) = 0, \qquad \forall i.$$

Complementary slackness:

$$\lambda_i^* \pi_{G,i}^* = \big(d - s_i(\pi_G^*)\big)\, \pi_{G,i}^* = 0.$$

Indeed, if $\pi_{G,i}^* > 0$, then $s_i(\pi_G^*) = d$ by (4), so $\lambda_i^* = 0$. If $\pi_{G,i}^* = 0$, the product is trivially zero.

Primal feasibility holds since $\pi_G^* \in \Delta_L$.

Thus π_G^*, together with (λ^*, μ^*), satisfies all KKT conditions of (D). Since (D) is a concave maximization over the convex compact set Δ_L, these conditions are sufficient for optimality. Hence π_G^* is an optimal solution of (D).

We conclude that any optimal solution of (G) is also optimal for (D).

Question 2

Question 2.1

We consider the function

$$h(X) = -\log\det(X), \qquad dom(h) = \mathbb{S}_{++}^d,$$

with respect to the inner product $\langle X, Y\rangle = \mathrm{tr}(XY)$. By definition, its convex conjugate is

$$\begin{aligned} h^*(Y) &= \sup_{X\in\mathbb{S}_{++}^d} \big\{\langle X, Y\rangle - h(X)\big\} \\ &= \sup_{X\in\mathbb{S}_{++}^d} \big\{\mathrm{tr}(XY) + \log\det X\big\}. \end{aligned}$$

Define

$$\phi_Y(X) := \mathrm{tr}(XY) + \log\det X.$$

We first compute the stationary point and candidate maximizer. On $\mathbb{S}_{++}^d$, the gradient of $\log\det X$ is $\nabla_X \log\det X = X^{-1}$. Hence

$$\nabla_X \phi_Y(X) = Y + X^{-1}.$$

A stationary point (if it exists) must satisfy

$$Y + X^{-1} = 0 \quad\Longleftrightarrow\quad X^{-1} = -Y \quad\Longleftrightarrow\quad X = -Y^{-1}.$$

For $X = -Y^{-1}$ to lie in $\mathbb{S}_{++}^d$, we need $-Y^{-1} \succ 0$, i.e. $Y \in -\mathbb{S}_{++}^d$.

Moreover, $\log\det X$ is strictly concave on $\mathbb{S}_{++}^d$ and $\mathrm{tr}(XY)$ is linear in X, so ϕ_Y is strictly concave in X. Therefore any stationary point is the unique maximizer of ϕ_Y over $\mathbb{S}_{++}^d$.

Thus, for $Y \in -\mathbb{S}_{++}^d$, the unique maximizer is

$$X^*(Y) = -Y^{-1},$$

and

$$\begin{aligned} h^*(Y) &= \phi_Y(X^*) \\ &= \mathrm{tr}(-Y^{-1}Y) + \log\det(-Y^{-1}) \\ &= -\mathrm{tr}(I) + \log\det(-Y^{-1}) \\ &= \log\det(-Y^{-1}) - d. \end{aligned}$$

Now we look at the domain of h^*. We now show that if $Y \notin -\mathbb{S}^d_{++}$, then $h^*(Y) = +\infty$.

If $Y \not\succ 0$, there exists a unit vector v such that

$$\lambda := v^\top Y v \geq 0.$$

Consider the family of matrices

$$X_t := I + t\, vv^\top, \qquad t > 0.$$

Each X_t is positive definite. We compute

$$\operatorname{tr}(X_t Y) = t\, v^\top Y v + \operatorname{tr}(Y) = t\lambda + \operatorname{tr}(Y),$$

and, since $vv^\top$ has eigenvalues $1, 0, \ldots, 0$,

$$\log\det X_t = \log\det(I + t\, vv^\top) = \log(1 + t).$$

Therefore

$$\phi_Y(X_t) = t\lambda + \operatorname{tr}(Y) + \log(1 + t).$$

If $\lambda > 0$, then $\phi_Y(X_t) \to +\infty$ as $t \to \infty$ because of the linear term $t\lambda$. If $\lambda = 0$, then $\phi_Y(X_t) = \operatorname{tr}(Y) + \log(1 + t) \to +\infty$ as $t \to \infty$. In both cases,

$$\sup_{X \in \mathbb{S}^d_{++}} \phi_Y(X) = +\infty,$$

so $h^*(Y) = +\infty$ and Y is not in the effective domain of h^*.

Thus

$$dom(h^*) = -\mathbb{S}^d_{++}.$$

To Conclude: For $h(X) = -\log\det(X)$ with $dom(h) = \mathbb{S}^d_{++}$, the convex conjugate is

$$h^*(Y) = \begin{cases} \log\det(-Y^{-1}) - d, & \text{if } Y \in -\mathbb{S}^d_{++}, \\ +\infty, & \text{otherwise,} \end{cases}$$

so on its domain we have

$$h^*(Y) = \log\det(-Y^{-1}) - d \quad \text{with} \quad dom(h^*) = -\mathbb{S}^d_{++}.$$

Question 2.2

We know from Question 2.1 that for

$$h(X) = -\log\det(X), \quad \operatorname{dom}(h) = S^d_{++},$$

its convex conjugate is

$$h^*(Y) = \log\det(-Y^{-1}) - d, \quad \mathrm{dom}(h^*) = -S^d_{++}.$$

By the fact that h is closed and convex (closed because continuous), we get:

$$h(X) = h^{**}(X) = \sup_{Y \in -S^d_{++}} \big\{\langle X, Y\rangle - h^*(Y)\big\},$$

where $\langle X, Y\rangle = \mathrm{tr}(XY)$. Taking $X = V(\pi) \succ 0$ gives

$$-\log\det V(\pi) = \sup_{Y \in -S^d_{++}} \Big\{ \mathrm{tr}\big(V(\pi)Y\big) - \big(\log\det(-Y^{-1}) - d\big)\Big\}.$$

Hence

$$\log\det V(\pi) = \inf_{Y \in -S^d_{++}} \Big\{ -\mathrm{tr}\big(V(\pi)Y\big) + \log\det(-Y^{-1}) - d\Big\}.$$

Now perform the change of variable $Y = -W$ with $W \in S^d_{++}$. Then $-Y^{-1} = W^{-1}$ and

$$-\mathrm{tr}(V(\pi)Y) = -\mathrm{tr}\big(V(\pi)(-W)\big) = \mathrm{tr}\big(V(\pi)W\big).$$

Therefore

$$\begin{aligned}\log\det V(\pi) &= \inf_{W \in S^d_{++}} \Big\{ \mathrm{tr}\big(V(\pi)W\big) + \log\det(W^{-1}) - d\Big\} \\ &= \inf_{W \in S^d_{++}} \Big\{ \mathrm{tr}\big(V(\pi)W\big) - \log\det W - d\Big\},\end{aligned}$$

since $\log\det(W^{-1}) = -\log\det W$.

By definition of $V(\pi)$,

$$V(\pi) = \sum_{i=1}^{L} \pi_i A_i A_i^\top,$$

so

$$\begin{aligned}\mathrm{tr}\big(V(\pi)W\big) &= \mathrm{tr}\Big(W \sum_{i=1}^{L} \pi_i A_i A_i^\top\Big) \\ &= \sum_{i=1}^{L} \pi_i \,\mathrm{tr}\big(W A_i A_i^\top\big) \\ &= \sum_{i=1}^{L} \pi_i \,\mathrm{tr}\big(A_i^\top W A_i\big).\end{aligned}$$

Plugging this into the previous expression yields

$$\boxed{\log\det V(\pi) = \inf_{W \in S^d_{++}} \left\{ \sum_{i=1}^{L} \pi_i \,\mathrm{tr}\big(A_i^\top W A_i\big) - \log\det W - d \right\}.}$$

This proves the desired representation.

Question 2.3

From Question 2.2 we know that for every $\pi \in \mathbb{R}^L$ with $V(\pi) \succ 0$

$$\log\det V(\pi) = \inf_{W \in S^d_{++}} \left\{ \sum_{i=1}^{L} \pi_i \operatorname{tr}\Big(A_i^\top W A_i\Big) - \log\det W - d \right\}.$$

Define

$$g(\pi, W) := \sum_{i=1}^{L} \pi_i \operatorname{tr}\Big(A_i^\top W A_i\Big) - \log\det W - d, \qquad \pi \in \Delta_L,\ W \in S^d_{++}.$$

Then problem (D) can be written as

$$(D) \qquad \max_{\pi \in \Delta_L} \log\det V(\pi) = \max_{\pi \in \Delta_L} \inf_{W \succ 0} g(\pi, W).$$

We can Use of Sion's minimax theorem.
Indeed, for fixed π, $g(\pi, \cdot)$ is convex in W because it is the sum of a linear function of W and the convex function $-\log\det W$. For fixed W, $g(\cdot, W)$ is linear (hence concave) in π. The sets Δ_L and S^d_{++} are convex, and Δ_L is compact. Moreover g is continuous in (π, W). Hence the assumptions of Sion's minimax theorem hold, and we can swap max and inf:

$$\max_{\pi \in \Delta_L} \inf_{W \succ 0} g(\pi, W) = \inf_{W \succ 0} \max_{\pi \in \Delta_L} g(\pi, W).$$

For fixed W, the dependence on π is only through

$$\sum_{i=1}^{L} \pi_i \operatorname{tr}\Big(A_i^\top W A_i\Big),$$

which is a linear functional of π. Since Δ_L is the simplex, we have

$$\max_{\pi \in \Delta_L} \sum_{i=1}^{L} \pi_i \operatorname{tr}\Big(A_i^\top W A_i\Big) = \max_{i \in [L]} \operatorname{tr}\Big(A_i^\top W A_i\Big).$$

Therefore

$$\max_{\pi \in \Delta_L} g(\pi, W) = \max_{i \in [L]} \operatorname{tr}\Big(A_i^\top W A_i\Big) - \log\det W - d.$$

Hence

$$(D) \equiv \inf_{W \succ 0} \left\{ \max_{i \in [L]} \operatorname{tr}\Big(A_i^\top W A_i\Big) - \log\det W - d \right\}.$$

Introduce an epigraph variable t:

$$\inf_{W \succ 0,\, t} \left\{ t - \log\det W - d \ :\ \operatorname{tr}\Big(A_i^\top W A_i\Big) \le t,\ \forall i \in [L] \right\}.$$

Now perform the change of variables

$$W = t\,\widetilde{W}, \qquad t > 0,\ \widetilde{W} \succ 0.$$

Then for each i,

$$\operatorname{tr}\Big(A_i^\top W A_i\Big) = t\,\operatorname{tr}\Big(A_i^\top \widetilde{W} A_i\Big) \le t \quad \Longleftrightarrow \quad \operatorname{tr}\Big(A_i^\top \widetilde{W} A_i\Big) \le 1.$$

The objective becomes

$$t - \log\det(t\widetilde{W}) - d = t - d\log t - \log\det\widetilde{W} - d.$$

For fixed $\widetilde{W}$, minimize over $t > 0$ the function $\phi(t) := t - d\log t$. We have

$$\phi'(t) = 1 - \frac{d}{t} = 0 \quad \Longrightarrow \quad t^\star = d,$$

and

$$\phi(t^\star) = d - d\log d.$$

Thus, the problem is equivalent to:

$$(P) \qquad \inf_{\widetilde{W}\succ 0} \ -d\log d - \log\det\widetilde{W} \quad \text{s.t.} \quad \operatorname{tr}\Big(A_i^\top \widetilde{W} A_i\Big) \le 1,\ \forall i \in [L],$$

which is the desired reformulation.

We now give the smallest K-ellipsoid interpretation.

Let each design list $A_i \in \mathbb{R}^{d\times K}$ be seen as a point in $\mathbb{R}^{dK}$ via $\operatorname{vec}(A_i)$. For $\widetilde{W} \succ 0$, define the block matrix

$$\widetilde{W}_K := I_K \otimes \widetilde{W} \in S^{dK}_{++}. \ \text{(concatenation)}$$

Then

$$\operatorname{vec}(A_i)^\top \widetilde{W}_K \operatorname{vec}(A_i) = \operatorname{tr}\Big(A_i^\top \widetilde{W} A_i\Big).$$

Hence the constraints $\operatorname{tr}\Big(A_i^\top \widetilde{W} A_i\Big) \le 1$ are exactly

$$\operatorname{vec}(A_i)^\top \widetilde{W}_K \operatorname{vec}(A_i) \le 1, \quad \forall i \in [L],$$

which means that all points $\operatorname{vec}(A_i)$ lie in the ellipsoid

$$\mathcal{E}(\widetilde{W}_K) := \Big\{ z \in \mathbb{R}^{dK} \ :\ z^\top \widetilde{W}_K z \le 1 \Big\}.$$

The volume of this ellipsoid satisfies

$$\operatorname{vol}\big(\mathcal{E}(\widetilde{W}_K)\big) = c_{dK}\,(\det\widetilde{W}_K)^{-1/2} = c_{dK}\,(\det\widetilde{W})^{-K/2},$$

for a constant c_{dK} depending only on the dimension. Thus minimizing $-\log\det\widetilde{W}$ is equivalent to minimizing $\operatorname{vol}(\mathcal{E}(\widetilde{W}_K))$.

Therefore problem (P) chooses $\widetilde{W}$ (equivalently, $\widetilde{W}_K$) so that the ellipsoid $\mathcal{E}(\widetilde{W}_K)$ has minimum volume among all ellipsoids that contain all design lists $\{A_i\}_{i=1}^L$. In other words, (P) finds the smallest K-ellipsoid containing all design lists.

Question 3

Question 3.1

The Frank–Wolfe (FW) algorithm from Algorithm 1 was implemented to solve the D-optimal design problem using the dataset `q31.npy` ($L = 400$, $d = 5$, $K = 2$). FW was executed with tolerance 10^{-3} and $T = 1000$ iterations for $L = 100, 200, 300, 400$. The resulting design vectors were compared with the D-optimal and G-optimal solutions obtained via MOSEK.

L	**Support(D)**	**Support(G)**	**Support(FW)**
100	100	100	91
200	200	200	159
300	300	300	204
400	400	400	233

Table 1: Support size (threshold 10^{-4}) for D-optimal, G-optimal, and FW solutions.

The D-optimal and G-optimal solutions have full support (support size = L). The FW solution uses a reduced subset of indices depending on L. This is expected, since the Frank-Wolfe algorithm selects extreme points of the feasible set and therefore concentrates the weight on fewer indices. The slow and non-linear growth of the FW support also contributes to its low computational cost. Only a limited number of new indices are activated as L increases. So each iteration remains inexpensive, and the overall runtime scales almost linearly with L as it can be seen in the runtime comparison table2.

L	**Runtime D (s)**	**Runtime G (s)**	**Runtime FW (s)**
100	0.1550	16.9621	0.0975
200	0.1434	68.6294	0.1339
300	0.2009	149.4846	0.1742
400	0.2942	204.8496	0.2425

Table 2: Runtime comparison between D-optimal, G-optimal, and Frank–Wolfe algorithms.

The SDP-based G-optimal problem is the most computationally demanding, reaching over 200 seconds at $L = 400$. The D-optimal problem is considerably faster but still slower than FW. FW consistently achieves the lowest runtime, remaining below 0.25 seconds for all values of L, and exhibits near-linear scaling, making it highly suitable for large-scale instances.

L	$\|\pi_{\mathrm{FW}} - \pi_{\mathrm{D}}\|_1$	$\|\pi_{\mathrm{G}} - \pi_{\mathrm{D}}\|_1$	$\|\pi_{\mathrm{FW}} - \pi_{\mathrm{G}}\|_1$
100	8.4459×10^{-1}	6.1169×10^{-2}	8.6601×10^{-1}
200	1.0003	1.9483×10^{-1}	1.1035
300	1.0546	8.1096×10^{-3}	1.0541
400	1.1265	1.1184×10^{-2}	1.1294

Table 3: ℓ_1 distances between FW, D-optimal, and G-optimal solutions.

The D-optimal and G-optimal solutions are extremely close, with ℓ_1-distances below 0.02 for $L \geq 300$. So D-optimal and G-optimal designs coincide at optimality. The FW solution differs more noticeably from both D-optimal and G-optimal solutions. This behavior can result from early stopping due to the tolerance 10^{-3}, the limited iteration $T = 1000$, and the inherent sparsity of FW iterates. Despite this, FW remains computationally attractive and provides an efficient approximate solution.

Question 3.2

The Frank-Wolfe algorithm is applied to the transformed LLM dataset in `q32.pkl` with tolerance 10^{-3} and $T = 2000$ iterations.

Verification of the optimality condition.

The resulting design vector π has support size

$$\#\{\, i : \pi_i > 10^{-4} \,\} = 829 \quad \text{out of } L = 1000.$$

For the active indices, the computed values

$$s_i = \mathrm{tr}\Big(A_i^\top V(\pi)^{-1} A_i\Big)$$

have an empirical mean of

$$\mathrm{mean}(s_i \ : \ \pi_i > 10^{-4}) \ \approx \ 390.30,$$

Since the dimension of the transformed LLM training dataset is $d = 384$ and the active values having an empirical mean of ≈ 390.30, the quantities s_i for i with $\pi_i > 10^{-4}$ are all close to d on average. This confirms that the Frank-Wolfe solution is close to satisfying the optimality condition for active indices, such that $s_i \approx d$..

This behavior is expected. At each iteration, the Frank-Wolfe algorithm selects the index

$$i^\star \ \leftarrow \ \arg\max_{i \in [L]} \mathrm{Tr}\Big(A_i^\top V_t^{-1} A_i\Big)\,,$$

This corresponds to the prompt that is currently the most informative. The algorithm updates the design through

$$\pi^{(t+1)} \ \leftarrow \ (1-\gamma_t)\,\pi^{(t)} + \gamma_t\, e_{i^\star}.$$

Therefore,the mass assigned to informative prompts gradually increases. As the algorithm converges, the active prompts tend to contribute equally to the information matrix, which makes the quantities

$$s_i \ = \ \mathrm{tr}\Big(A_i^\top V(\pi)^{-1} A_i\Big)$$

nearly constant across all indices with $\pi_i > 10^{-4}$. Since these contributions must collectively balance the model dimension, the values s_i naturally settle close to d.

Geometry of the top 25 selected prompts.

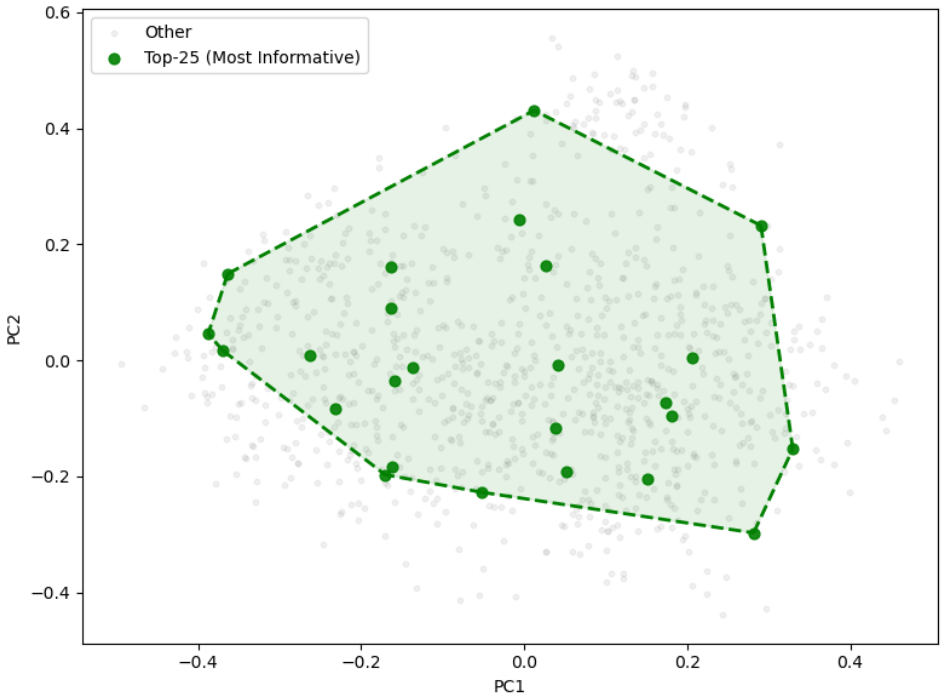

Figure 1: PCA embedding of all prompts with the top 25 most informative prompts selected by the FW algorithm.

The top 25 prompts, ranked by decreasing values of s_i, can be visualized in the provided PCA embedding (Figure 1). The prompts do not form a tight cluster. Instead, they are spread widely across the embedding. They approximately trace the convex hull of the dataset. In other words, the selected points originate from multiple distinct regions of the representation space.

This behavior is expected. Indeed, D-optimal design aims to maximize the determinant of the information matrix. It requires selecting prompts whose feature directions span the space as broadly as possible. Thus, the most informative prompts tend to lie on the geometric "extremes" of the dataset, covering diverse clusters and forming a broad convex region in the PCA projection.

The top 25 selected prompts do span diverse clusters. Intuitively, the wide spatial dispersion of the top prompts indicates that the FW solution identifies directions of maximal variability in the data. By selecting prompts that span diverse clusters, the algorithm constructs a design that is well-conditioned and highly informative for estimating the underlying model parameters.