EE5121: Convex Optimization

Assignment #2

Name: Swathi Shree Narashiman — Marks: 60

Roll No: EE22B149 — Due: Nov 11, 23:59

Problem 1 — 22 marks

Let $\{\phi_i\}_{i=1}^n \subset \mathbb{R}^k$ be feature vectors and let $\Phi \in \mathbb{R}^{n\times k}$ collect them by rows. Let $\mu \in \mathbb{R}^k$ be a target moment vector. Define the probability simplex

$$\Delta_n := \left\{ p \in \mathbb{R}^n_+ : \sum_{i=1}^n p_i = 1 \right\}.$$

Consider the following problem:

$$\begin{aligned} \max_{p\in\mathbb{R}^n} \quad & H(p) := -\sum_{i=1}^n p_i \log p_i \\ \text{s.t.} \quad & \sum_{i=1}^n p_i \phi_i = \mu, \\ & \sum_{i=1}^n p_i = 1, \\ & p_i \geq 0, \quad \forall i, \end{aligned}$$

with the convention $0 \log 0 := 0$. For subquestions (e) and (f), refer to the uploaded Excel sheet for input data, $\Phi \in \mathbb{R}^{n\times k}$ and $\mu \in \mathbb{R}^k$.

(a) [4 marks] Using dual variables θ (for the moment constraint) and ν (for the simplex constraint), derive the dual function $g(\theta, \nu)$ and the associated dual optimization problem.

Solution: Let us add a new coefficient in the Lagrangian θ_2 to account for the constraint $\sum_{i=1}^n p_i = 1$.

$$p_i \geq 0 \implies -p_i \leq 0$$

$$L(p,\theta,\theta_2,\nu) = H(p) + \sum_{i=1}^n \nu_i(-p_i) + \theta^\top\left(\sum_{i=1}^n p_i\phi_i - \mu\right) + \theta_2\left(\sum_{i=1}^n p_i - 1\right)$$

$$L(p,\theta,\theta_2,\nu) = H(p) - \nu^\top p + \theta^\top\left(\sum_{i=1}^n p_i\phi_i - \mu\right) + \theta_2\left(\sum_{i=1}^n p_i - 1\right)$$

Using the definition of entropy, we get

$$L(p,\theta,\theta_2,\nu) = -\sum_{i=1}^n p_i \log p_i + \sum_{i=1}^n \nu_i(-p_i) + \theta^\top\left(\sum_{i=1}^n p_i\phi_i - \mu\right) + \theta_2\left(\sum_{i=1}^n p_i - 1\right)$$

To find the dual, we differentiate the Lagrangian. For each i, we get

$$\frac{\partial L}{\partial p_i}(p,\theta,\theta_2,\nu) = -(1+\log p_i) - \nu_i + \theta^\top\phi_i + \theta_2$$

$$\implies p_i^* = e^{\theta^\top \phi_i + \theta_2 - 1 - \nu_i}.$$

The dual function is

$$g(\theta, \theta_2, \nu) = \inf_p L(p, \theta, \theta_2, \nu),$$

which simplifies to

$$g(\theta, \theta_2, \nu) = \sum_{i=1}^n p_i^* - \theta^\top \mu - \theta_2.$$

Substituting p_i^*,

$$g(\theta, \theta_2, \nu) = e^{\theta_2 - 1} \sum_{i=1}^n e^{\theta^\top \phi_i - \nu_i} - \theta^\top \mu - \theta_2 = e^{\theta_2 - 1} S - \theta^\top \mu - \theta_2,$$

where $S = \sum_{i=1}^n e^{\theta^\top \phi_i - \nu_i}$.
We can remove the dependency on θ_2 by minimizing with respect to it:

$$\frac{\partial g(\theta, \theta_2, \nu)}{\partial \theta_2} = 0 \implies e^{\theta_2 - 1} S - 1 = 0 \implies e^{\theta_2 - 1} = \frac{1}{S}.$$

Substituting this back,

$$g(\theta, \nu) = \log S - \theta^\top \mu = \log\left(\sum_{i=1}^n e^{\theta^\top \phi_i - \nu_i} \right) - \theta^\top \mu.$$

Finally, minimize over $\nu \geq 0$. Since each ν_i appears only by **decreasing** the i-th exponent, the minimizer is $\nu_i^* = 0$ for all i (consistent with complementary slackness: if $p_i^* > 0$, then $\nu_i^* = 0$). Thus, the ν-variables can be removed and the dual reduces to the standard form:

Dual:

$$\min_{\theta \in \mathbb{R}^d} \left\{ \log\left(\sum_{i=1}^n e^{\theta^\top \phi_i} \right) - \theta^\top \mu \right\}$$

(b) [3 marks] State precise conditions under which strong duality holds with respect to (Φ, μ).

Solution:

(1) Feasibility (necessary). A necessary condition for the primal to be feasible is

$$\mu \in \operatorname{conv}\{\phi_1, \ldots, \phi_n\},$$

because any feasible p realizes $\mu = \sum_i p_i \phi_i$, a convex combination of the ϕ_i.

(2) Sufficient condition for strong duality (Slater). If there exists a strictly feasible point

$$p^\circ \in \mathbb{R}^n \quad \text{with} \quad p_i^\circ > 0 \ \forall i, \qquad \sum_{i=1}^n p_i^\circ \phi_i = \mu, \qquad \sum_{i=1}^n p_i^\circ = 1,$$

(i.e. a distribution with *all* entries strictly positive that satisfies the equalities), then Slater's condition holds for the convex formulation (minimize $-H(p)$ subject to linear constraints). Hence **strong duality holds**: the duality gap is zero ($p^* = d^*$) and the dual optimum is attained.

Equivalently, the Slater condition can be stated geometrically as

$$\mu \in \operatorname{relint}\big(\operatorname{conv}\{\phi_1, \dots, \phi_n\}\big),$$

i.e. μ belongs to the *relative interior* of the convex hull of the ϕ_i. This is exactly the condition guaranteeing the existence of a strictly positive probability vector attaining the moment μ.

(3) Boundary case and remarks.

- If $\mu \notin \operatorname{conv}\{\phi_i\}$ then the primal is infeasible and there is no meaningful strong duality statement.
- If $\mu \in \operatorname{conv}\{\phi_i\}$ but lies on the boundary (so every feasible p has at least one zero entry), Slater's condition fails. In that situation:
  - A duality gap *may* still be zero, but it is no longer guaranteed by Slater's theorem.
  - The dual optimum may fail to be attained, and nonzero Lagrange multipliers ν_i (for $p_i \geq 0$) can be active for indices with $p_i^* = 0$ (complementary slackness).
- In practice, for the finite discrete max-entropy problem the usual clean sufficient condition used to guarantee the Gibbs form

  $$p_i^* \propto e^{\theta^\top \phi_i}$$

  and zero gap is $\mu \in \operatorname{relint}(\operatorname{conv}\{\phi_i\})$.

(c) [2 marks] Prove that any primal maximizer has the Gibbs form

$$p_i^* \propto \exp(\theta^{*\top} \phi_i),$$

together with $\sum_i p_i^* \phi_i = \mu$ and $\sum_i p_i^* = 1$, where p^* and θ^* are primal and dual optimizers.

Solution:

From part (a), in the process of finding the dual we minimized the Lagrangian to get the optimal p_i^* given by,

$$\implies p_i^* = e^{\theta^\top \phi_i + \theta_2 - 1 - \nu_i}$$

Using the law of total probability,

$$\sum_i p_i^* = 1 \implies \sum_{i=1}^{n} e^{\theta^\top \phi_i + \theta_2 - 1 - \nu_i} = 1$$

$$\implies e^{\theta_2 - 1} \sum_{i=1}^{n} e^{\theta^\top \phi_i - \nu_i} = 1$$

$$\implies e^{\theta_2 - 1} = \frac{1}{\sum_{i=1}^{n} e^{\theta^\top \phi_i - \nu_i}}$$

Therefore,

$$p_i^* = \frac{e^{\theta^\top \phi_i - \nu_i}}{\sum_{i=1}^{n} e^{\theta^\top \phi_i - \nu_i}}$$

Hence we proved that,

$$p_i^* \propto e^{\theta^\top \phi_i}$$

(d) [3 marks] Write the dual function exclusively in terms of θ and explain the role of θ^* in the optimal solution p^*.

Solution:

From part (a), we can see already that we can eliminate ν from the dual problem as shown below.

Since each ν_i appears only by **decreasing** the i-th exponent, the minimizer is $\nu_i^* = 0$ for all i (consistent with complementary slackness: if $p_i^* > 0$, then $\nu_i^* = 0$). Thus, the ν-variables can be removed and the dual reduces to the standard form:

Dual:

$$\min_{\theta \in \mathbb{R}^d} \left\{ \log\left(\sum_{i=1}^{n} e^{\theta^\top \phi_i} \right) - \theta^\top \mu \right\}$$

Role of θ^* in the optimal primal p^*. Let θ^* be a minimizer of $g(\theta)$. From the Gibbs form result in part (c) we can write:

$$p_i^* \propto \exp(\theta^{*\top} \phi_i), \qquad p_i^* = \frac{\exp(\theta^{*\top} \phi_i)}{\sum_{j=1}^{n} \exp(\theta^{*\top} \phi_j)}.$$

Differentiating the dual function,

$$\nabla_\theta g(\theta) = \frac{\sum_{i=1}^{n} \phi_i e^{\theta^\top \phi_i}}{\sum_{j=1}^{n} e^{\theta^\top \phi_j}} - \mu$$

$$\nabla_\theta g(\theta) = \sum_{i=1}^{n} p_i^* \phi_i - \mu,$$

where p_θ denotes the Gibbs distribution with parameter θ. At optimality, $\nabla_\theta g(\theta^*) = 0$, so

$$\sum_{i=1}^{n} p_i^* \phi_i = \mu.$$

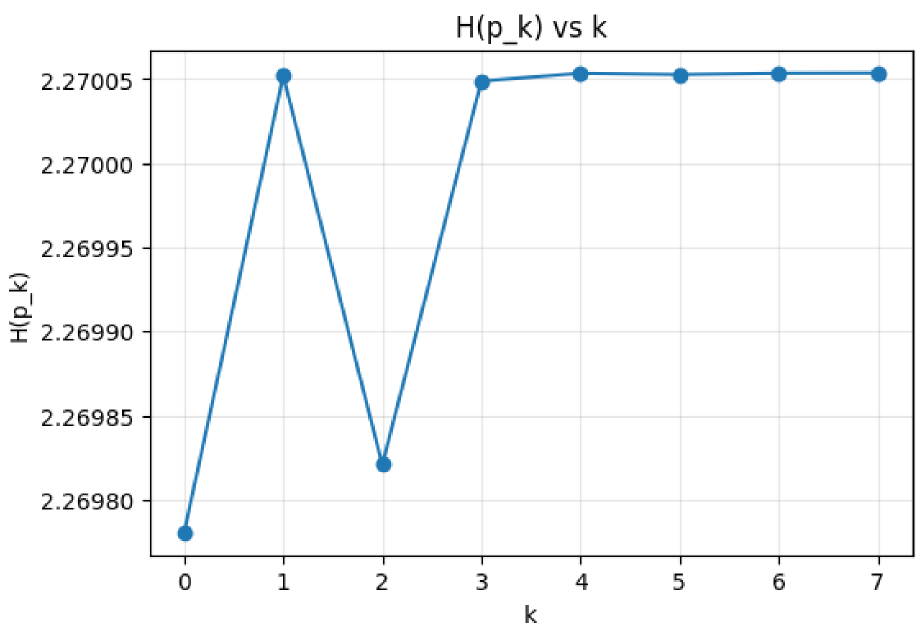
Thus θ^* is the Lagrange multiplier (dual variable) that enforces the moment constraint: it parameterizes the Gibbs family, and its optimal value is chosen so that the Gibbs distribution p_{θ^*} satisfies the required moment μ (after normalization, $\sum_i p_i^* = 1$).

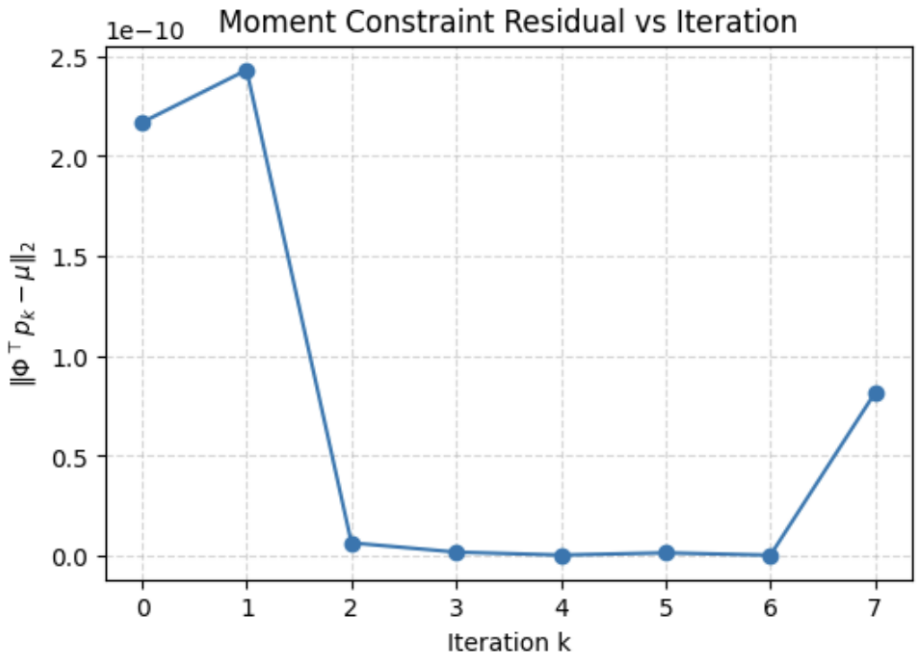
(e) [5 marks] Solve the primal problem using any library to obtain p^*. Generate an iterative sequence $\{p_k\}$ over Δ_n and plot:

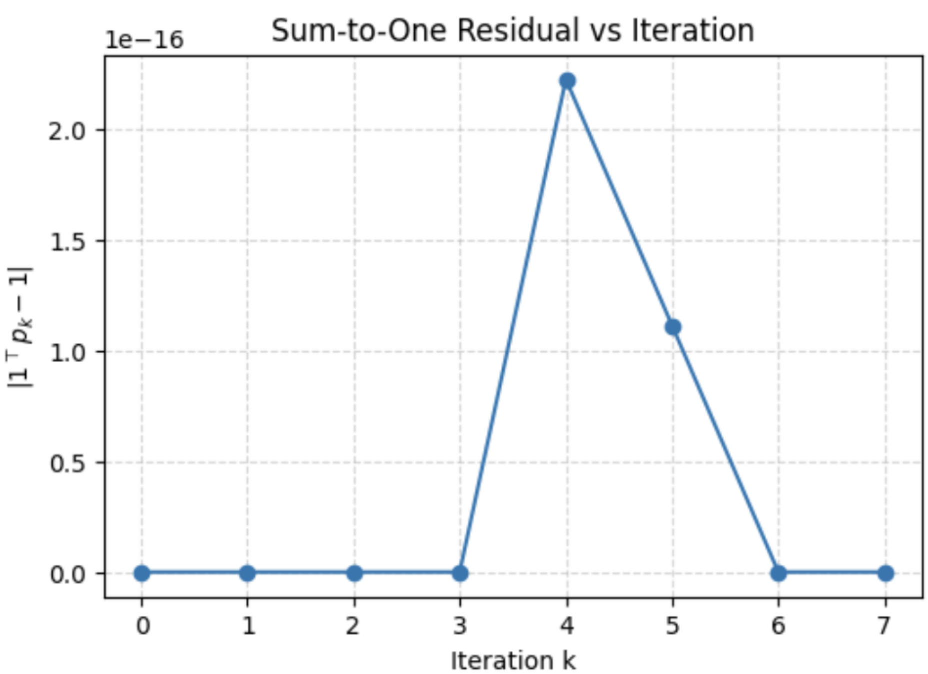
- $H(p_k)$ versus iteration k, indicating the optimal value $H(p^*)$;
- feasibility residuals $\|\Phi^\top p_k - \mu\|_2$ and $|1^\top p_k - 1|$ versus k.

Solution: The code files for the given question is available here

The plots for the same are attached here:




Moment Constraint Residual vs Iteration
1e−10
2.5
2.0
1.5
1.0
0.5
0.0
$\|\Phi^\top p_k - \mu\|_2$
0
1
2
3
4
5
6
7
Iteration k


Sum-to-One Residual vs Iteration
1e−16
2.0
1.5
1.0
0.5
0.0
$|1^\top p_k - 1|$
0
1
2
3
4
5
6
7
Iteration k

(f) [5 marks] Formulate and solve the dual problem directly using any library to obtain (θ^*, ν^*) and $g(\theta^*, \nu^*)$. Compute the optimal point

$$\tilde{p}_i \propto \exp(\theta^{*\top} \phi_i),$$

and report $\|p^* - \tilde{p}\|_\infty$. Report the duality gap

$$|H(p^*) - g(\theta^*, \nu^*)|.$$

Solution: The code files for the same is attached here
The Optimal point is:

$$\theta^* = \begin{bmatrix} 0.06249744 \\ 0.32719376 \end{bmatrix}$$

The dual function value at this point is:

$$g(\theta^*) = 2.270053591693747$$

The primal optimal value was:

$$H(p^*) = 2.270053591654.$$

Thus, the duality gap is:

$$|H(p^*) - g(\theta^*)| = 3.930456 \times 10^{-11},$$

which is essentially zero up to numerical precision, confirming that strong duality holds and the computed solution is effectively optimal.

Problem 2 — 17 marks

You are given binary labels $y_i \in \{-1, +1\}$ and features $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$. Consider the (unregularized) logistic regression problem:

$$\min_{w \in \mathbb{R}^d} L(w) := \sum_{i=1}^{n} \log\big(1 + \exp(-y_i x_i^\top w)\big). \tag{1}$$

(a) [2 marks] Explain briefly (no more than two lines) why (1) corresponds to a probabilistic binary classifier that separates the data into two classes. asses via the sigmoid decision rule sign$(x_i^\top w)$.

Solution: Because the model assumes each label $y_i \in \{0, 1\}$ is drawn from a Bernoulli distribution with

$$P(y_i = 1 \mid x_i; w) = \sigma(x_i^\top w),$$

minimizing (1) is equivalent to maximizing the likelihood of a probabilistic binary classifier. Thus, the learned parameter w separates the data into two classes through the sigmoid-based decision rule.

(b) [5 marks] Implement any smooth solver in CVXPY or a hand-coded optimizer (e.g., gradient descent or L-BFGS) for solving (1) using data provided in the Excel sheet. Mention the solver and record the iterates $\{w_k\}$, the objective values $L(w_k)$, and the norms $\|w_k\|_2$. Produce two plots (logarithmically scaled):

$$k \mapsto L(w_k), \qquad k \mapsto \|w_k\|_2.$$

Solution: The hand-coded optimizer using gradient descent is implemented in the code file attached here

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \log\big(1 + \exp(-y_i x_i^\top w)\big),$$

which is the correct loss for labels $y_i \in \{\pm 1\}$. Before optimization the features were standardized (zero mean, unit variance) and an intercept (bias) column was appended.

Hyper-parameters and recording.

- Initialization: $w_0 = 0$.
- Step-size: $\alpha = 0.1$ (chosen after simple tuning for smooth decay).
- Iterations: $K = 2000$.
- At every iteration k, I recorded the iterate w_k, the objective $L(w_k)$, and the Euclidean norm $\|w_k\|_2$.

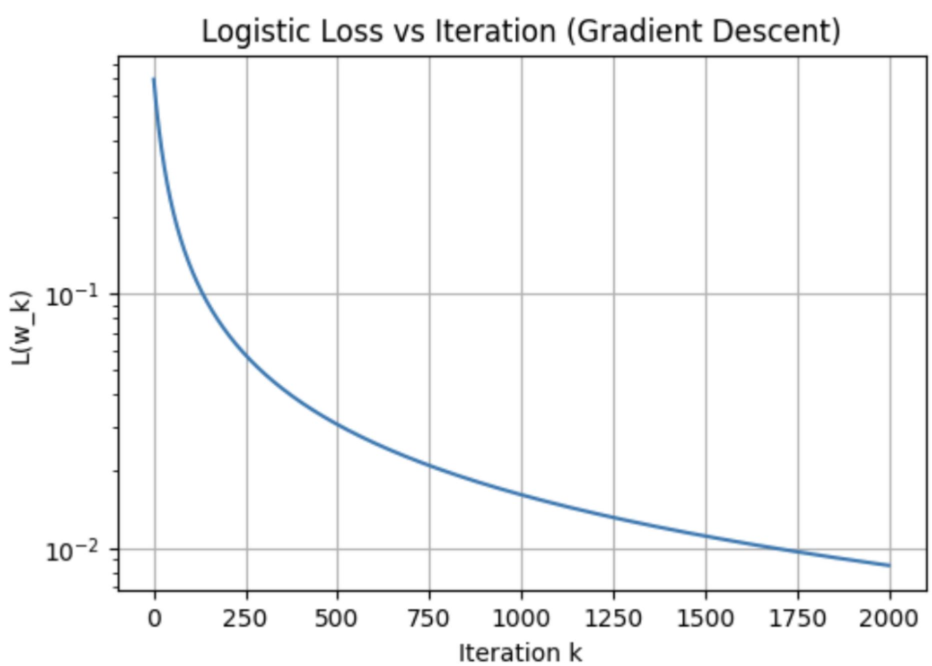


Figure 1: Loss vs k

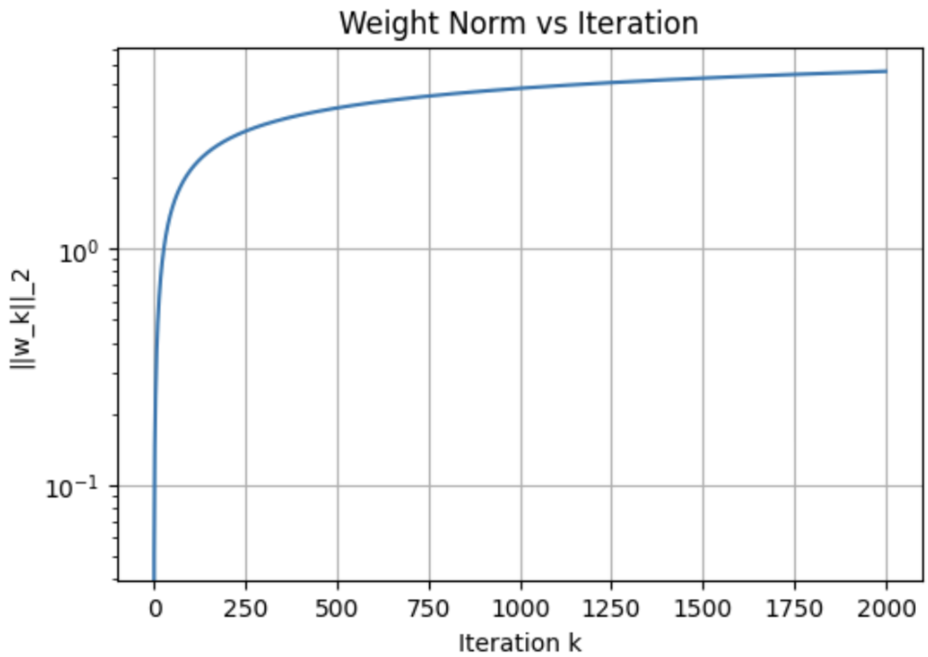


Figure 2: Weight norm vs k

(c) [5 marks] From your plots in (b), justify why the algorithm you used in (b) fails to deliver a finite optimizer on the dataset.

Solution: The plots show that the logistic loss $L(w_k)$ decreases steadily while the parameter norm $\|w_k\|_2$ grows without bound. This combination is the tell-tale signature of a separable dataset: if there exists w with $y_i(x_i^\top w) > 0$ for all i, then for any $\lambda > 0$ the scaled vector λw makes the margins $y_i x_i^\top(\lambda w) = \lambda y_i x_i^\top w$ arbitrarily large, and hence

$$L(\lambda w) = \frac{1}{n}\sum_i \log\bigl(1 + \exp(-\lambda y_i x_i^\top w)\bigr) \to 0 \quad \text{as } \lambda \to \infty.$$

Thus the unregularized logistic objective has no finite minimizer (its infimum is attained only at infinity).

The empirical evidence (loss $\downarrow$ while $\|w_k\| \uparrow$ with no plateau) confirms gradient descent is driving the iterates to infinity to reduce the loss, so the algorithm cannot produce a finite optimizer on this dataset.

(d) [5 marks] Propose a method to solve the above problem by changing the optimization formulation. Justify using appropriate plots and results that your proposal works.

Solution: The code files for the same can be found here

Proposed function Add an ℓ_2 regularizer (ridge) to the logistic objective and solve

$$\min_{w\in\mathbb{R}^d} L_\lambda(w) = \frac{1}{n}\sum_{i=1}^{n} \log\bigl(1 + \exp(-y_i x_i^\top w)\bigr) + \frac{\lambda}{2}\|w\|_2^2,$$

with $\lambda > 0$.

This works because, the regularizer makes L_λ strongly convex and coercive, hence it has a unique finite minimizer w_λ^*. Intuitively the $\frac{\lambda}{2}\|w\|_2^2$ term prevents the optimizer from sending $\|w\| \to \infty$ to drive the logistic loss to zero. This instead the trade-off between loss and norm yields a finite solution.

Final Results :
Unregularized final loss: 0.048710117993198916
Unregularized final $\|w\|$: 3.335903816908147
Regularized final loss: 0.10944739400401371
Regularized final $\|w\|$: 2.955711471899615

Conclusions : The un-regularized model achieves a lower loss but only by increasing the weight norm without bound, confirming that logistic regression on separable data does not admit a finite minimizer. Introducing L2 regularization stabilizes the optimization and yields a finite solution with a controlled weight magnitude, at the cost of a slightly higher loss. The comparison demonstrates the necessity of regularization for well-posed optimization in separable settings.

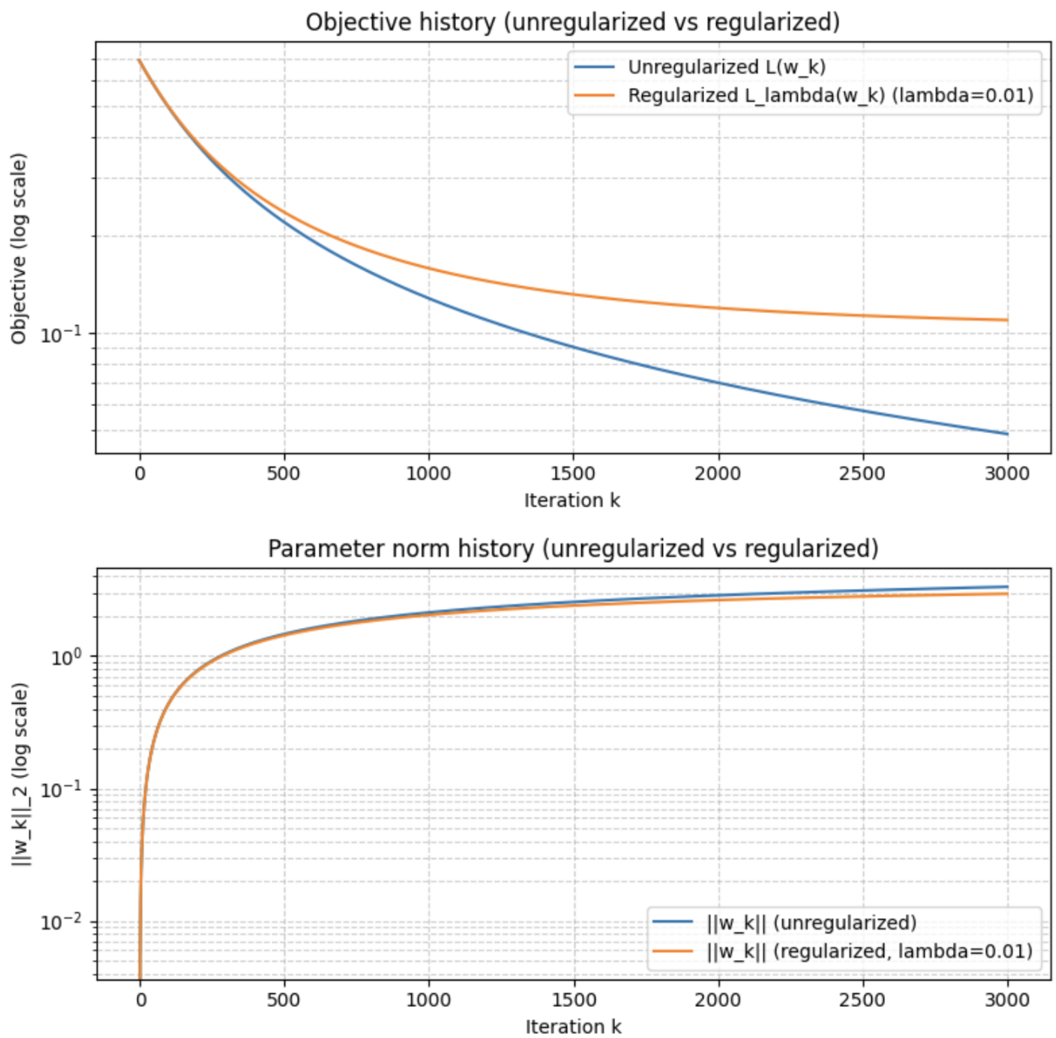


Figure 3: Comparison of Plain Gradient Descent and Gradient Descent with L2 regularizer

Problem 3 **21 marks**

Let $G=(V,E)$ be a simple undirected graph with $|V|=n$. Recall the vector-coloring SDP from the previous assignment formulated as the following optimization problem:

$$\begin{aligned} \min_{G,\rho} \quad & \rho \\ \text{s.t.} \quad & G \succeq 0, \\ & G_{vv}=1 \quad (\forall v \in V), \\ & G_{uv} \le \rho \quad (\forall (u,v) \in E), \end{aligned} \tag{2}$$

where $G \in \mathbb{S}^n$ is a Gram matrix and $\rho \in \mathbb{R}$ bounds edge-wise inner products. The K-color vector-feasibility threshold is $\rho \le -\frac{1}{K-1}$ (e.g., $-\frac{1}{2}$ for $K=3$).

Graphs for all computations in this problem:

1. $V=\{1,2,3,4\}$; $\quad E=\{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$
2. $V=\{1,2,3,4,5\}$; $\quad E=\{(1,2),(2,3),(3,4),(4,5),(5,1)\}$

(a) [5 marks] Treat (G,ρ) as primal variables. Introduce multipliers $\lambda \in \mathbb{R}^n$ for $G_{vv}=1$, nonnegative $\alpha_{uv} \ge 0$ for $G_{uv}-\rho \le 0$ on edges, and the PSD constraint via a dual slack $S \succeq 0$. Derive the dual problem.

Solution:
Introduce Lagrange multipliers:

$$\lambda \in \mathbb{R}^n \text{ for } G_{ii}=1, \qquad \alpha_{uv} \ge 0 \text{ for } G_{uv}-\rho \le 0, \qquad S \succeq 0 \text{ for } G \succeq 0.$$

The Lagrangian is

$$\mathcal{L}(G,\rho;\lambda,\alpha,S)=\rho+\sum_{i=1}^{n}\lambda_i(G_{ii}-1)+\sum_{(u,v)\in E}\alpha_{uv}(G_{uv}-\rho)-\langle S,G\rangle.$$

Grouping terms:

$$\mathcal{L}=\rho\Big(1-\sum_{(u,v)\in E}\alpha_{uv}\Big)+\Big\langle(\lambda)+A_\alpha-S,\ G\Big\rangle-\sum_{i=1}^{n}\lambda_i,$$

where A_α is the symmetric matrix with (u,v)-entry equal to α_{uv} for edges and zero otherwise.
For the Lagrangian to have a finite infimum over G and ρ, we require

$$1-\sum_{(u,v)\in E}\alpha_{uv}=0, \qquad (\lambda)+A_\alpha-S=0.$$

Because $S \succeq 0$ and $\alpha \ge 0$, we obtain the dual:

$$\boxed{\begin{aligned} \max_{\lambda\in\mathbb{R}^n,\ \alpha\ge 0} \quad & -\sum_{i=1}^{n}\lambda_i \\ \text{s.t.} \quad & S:=(\lambda)+A_\alpha \succeq 0, \\ & \sum_{(u,v)\in E}\alpha_{uv}=1, \\ & \alpha_{uv}\ge 0 \qquad \forall (u,v)\in E. \end{aligned}}$$

Here the dual objective is $-\sum_i \lambda_i$, and weak duality ensures

$$\text{dual optimum} \ \le \ \rho^\star.$$

(b) [10 marks] For each of the graphs, solve the dual and primal problems, and report the dual optimal value, the primal optimal value $\rho^\star$, and the duality gap.

Solution: The code files for the same can be found here.

Graph (i):

$\rho^\star_{\text{primal}} = -0.33333333255748676$

Dual optimum $= -3.1716702746734213 \times 10^{-10}$

Duality gap $= 0.33333333224031975$

Graph (ii):

$\rho^\star_{\text{primal}} = -0.8090172334287339$

Dual optimum $= -3.4740024448663363 \times 10^{-10}$

Duality gap $= 0.8090172330813337$

(c) [6 marks] **Feasibility of 3-colorability**

(a) If your dual optimum satisfies $\sum_v \lambda_v > -\frac{1}{2}$, explain (in 3 lines) why this is a certificate that the vector 3-coloring relaxation is infeasible.

(b) For each of the graphs, interpret the outcomes to determine whether the graph is 3-colorable or not.

Solution:

(a) (3-line certificate.) Let $d^\star$ denote the optimal value of the dual derived in (a). By weak duality we have

$$d^\star \le \rho^\star,$$

where $\rho^\star$ is the primal optimum. If $d^\star > -\frac{1}{2}$ then $\rho^\star \ge d^\star > -\frac{1}{2}$, so the primal cannot attain $\rho \le -\frac{1}{2}$. Hence $d^\star > -\frac{1}{2}$ certifies that the vector 3-coloring relaxation (which requires $\rho \le -1/2$) is infeasible.

(b) (Interpretation for the two graphs.)

- **Graph (i)** K_4**.** The SDP primal optimum is $\rho^\star = -\frac{1}{3}$ (since distinct vertices of the 3-simplex have inner product $-1/3$). Because $-\frac{1}{3} > -\frac{1}{2}$, the relaxation cannot achieve the 3-color threshold; equivalently the graph is not 3-colorable (indeed $\chi(K_4) = 4$).

- **Graph (ii)** C_5**.** The 5-cycle admits a 3-coloring: one can assign the three equiangular unit vectors of an equilateral triangle to the three colors, yielding a feasible Gram with $\rho = -\frac{1}{2}$. Thus $\rho^\star \le -\frac{1}{2}$ (in fact $\rho^\star = -\frac{1}{2}$), so the vector 3-coloring relaxation is feasible and the graph is 3-colorable.

Collaboration

I would like to thank Aravind Ramana V (EP23B003) for the useful discussion and ideation on solving the problems in this assignment. However, the work presented in this submission is original to the best of my knowledge.