一、填空题

1. 闭凸集$X \subset R^n$，任意一点$y \in R^n$,则点y在集合X上的投影是否是唯一的_是____。

2. 给定问题$\min\limits_{x\in R^2} f(x)$, $f(x) = -12x_2 + 4x_1^2 + 4x_2^2 + 4x_1x_2$,则Hessian矩阵$H =$ ________.

   若初始搜索方向$d_1 = \begin{pmatrix}1\\0\end{pmatrix}$,则该方向的共轭方向具有的特征为$d_2 =$ _______. 假设初始点为$x^{(1)} = \left(-\frac{1}{2}, 1\right)^T$,沿$d_1$最小化目标函数得到迭代点$x^{(2)} =$ __________. 然后从该点沿d_2方向最小化目标函数得$x^{(3)} =$ ______________.

3. 目标函数$f(x_1, x_2) = 3x_1^2 - 4x_1x_2 + x_2^2$在点$x^{(0)} = (0,1)^T$处的单位最速下降方向为: ___

4. $\boldsymbol{min\ x_1 + x_2, s.t.\ 2 - x_1^2 - x_2^2 \geq 0}$，**其由非线性乘子法构造的拉格朗日函数在最优点** $\boldsymbol{x^* =}$ __________**处对应的**$\boldsymbol{\nabla_x^2 L(x^*, \lambda^*) =}$ ____________________，**对应的乘子**$\boldsymbol{\lambda_1^* =}$ ___

5. 在 SVM 中的 Hinge 损失和L_2损失分别如下：

$$f_i(z) = \begin{cases} \max\{0, 1 - y_i z\}, \textit{Hinge损失} \\ \frac{1}{2}\max\{0, 1 - y_i z\}^2, \ L_2\textit{损失} \end{cases}$$

   其梯度$f_i'(z) = \begin{cases} (\qquad\qquad\qquad\qquad), \textit{Hinge损失} \\ (\qquad\qquad\qquad\qquad), \ L_2\textit{损失} \end{cases}$

6. 外点法与内点法中，只能对不等式约束问题进行求解是_内点_法。

二、计算证明题

7. 给定u, v，$x \equiv (w_1, w_2, b_1, w_3, w_4, b_2, \omega_1, \omega_2, c)$，令$h(u; x) := g(\omega_1 h_1(u; w_1, w_2, b_1) + \omega_2 h_2(u, w_3, w_4, b_2) + c) = g(\omega_1 g(w_1u_1 + w_2u_2 + b_1) + \omega_2 g(w_3u_1 + w_4u_2 + b_2) + c)$,这里：$g(z) = \frac{1}{1+\exp(-z)}$，并定义$f(x) = [h(u; x) - v]^2$

   (a) 计算函数f的 Hessian 矩阵，并证明函数f不一定是凸函数；
   (b) 计算函数f关于变量x的梯度∇f_x；
   (c) 讨论如何在计算机上有效计算函数f的梯度；
   (d) 推导在什么条件下，函数的梯度是 Lipschitz 连续的，即满足：$||\nabla f(x_1) - \nabla f(x_2)|| \leq L||x_1 - x_2||, \forall x_1, x_2 \in R^9$,常数$L > 0$.

8. **问题 P:** $\boldsymbol{min\ f(x); s.t.\ Ax \leq b; Ex = e}$; $\boldsymbol{f}$可微，$\boldsymbol{A_{m\times n}, E_{l\times n}, x \in R^n, b \in R^m, e \in R^l}$，证明：设$\boldsymbol{\bar{x}}$是上述问题的一个可行解，且在点$\boldsymbol{\bar{x}}$处有$\boldsymbol{A_1\bar{x} = b_1, A_2\bar{x} < b_2}$,其中$\boldsymbol{A = [A_1\ A_2]^T, b = [b_1\ b_2]^T}$则
   (a)向量$\boldsymbol{d}$($\boldsymbol{d \in R^n, d \neq 0}$)是点$\boldsymbol{\bar{x}}$处的可行方向的充要条件为$\boldsymbol{A_1 d \leq 0, Ed = 0}$;
   (b)若此时$\boldsymbol{d}$又满足:$\boldsymbol{\nabla f(\bar{x})^T d < 0}$,则$\boldsymbol{d}$是一个可行下降方向.

9. 考虑约束最优化问题$\min f(x) = 4x_1^2 + 5x_1x_2 + x_2^2, s.t. \begin{cases} x_1^2 - x_2 + 2 \leq 0 \\ x_1 + x_2 - 6 \leq 0 \\ x_1, x_2 \geq 0 \end{cases}$ 分别写出用罚函数法和闸函数法求解的辅助问题。

10. 给定问题(P)$\begin{cases} \min \left(x_1 - \frac{9}{4}\right)^2 + (x_2 - 2)^2 \\ s.t. \begin{cases} x_2 - x_1^2 \geq 0 \\ x_1 + x_2 \leq 6 \\ x_1, x_2 \geq 0 \end{cases} \end{cases}$，请验证$x^* = \left(\frac{3}{2}, \frac{9}{4}\right)^T$是KKT 点，并证明$x^*$是唯一

的最优解。

三、问答题

11. 请论述在深度学习优化算法中，是如何一步步对梯度下降法进行改进的。

12. 请论述线性规划算法在优化方法中的重要性。

11.批量梯度下降法（Batch Gradient Descent，BGD）：
BGD使用整个数据集来计算每个迭代步的梯度。这使得它非常准确，但是在处理大型数据集时速度会很慢，因为每次迭代都需要遍历整个数据集。
随机梯度下降法（Stochastic Gradient Descent，SGD）：
SGD在每次迭代中只随机选择一个样本来计算梯度，这样计算量大大减少，速度更快。但是，由于梯度的估计是基于单个样本，因此SGD的方差较大，导致更新过程更加波动。
小批量梯度下降法（Mini-batch Gradient Descent，MBGD）：
MBGD是BGD和SGD的折中方案，它在每次迭代中使用一小批样本（例如32或64个）来计算梯度。这样可以减少方差，同时比BGD更快地更新参数。
动量（Momentum）：
动量方法借鉴了物理学中的动量概念，它将之前梯度的指数衰减平均值考虑进来，以此来加速学习过程，并减少震荡。动量有助于跨越局部极小值和鞍点。
Nesterov 加速梯度（Nesterov Accelerated Gradient，NAG）：
NAG是动量方法的一个变种，它在计算梯度时考虑了参数更新后的近似位置，这通常可以提高优化的速度和性能。
自适应学习率算法：
自适应学习率算法根据参数的历史梯度来调整每个参数的学习率。这些方法包括：
AdaGrad：它为每个参数分配不同的学习率，通过累积梯度平方来更新学习率。
RMSprop：结合了动量和AdaGrad的思想，它使用梯度平方的移动平均来调整学习率。
Adam：结合了RMSprop和动量的特点，同时使用梯度的一阶矩估计（即动量）和二阶矩估计（即未中心化的方差）来调整学习率。