Statistical Models and Computing Methods, Problem Set 1

September 30, 2021 Due 10/14/2021

Problem 1.

(1) Show that $X \sim \mathcal{N}(0,1)$ is the maximum entropy distribution such that $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$. *(10 points)*

(2) Generalize the result in (1) for the maximum entropy distribution given the first k moments, *i.e.*, $\mathbb{E}X^i = m_i,\ i = 1, \ldots, k$. *(5 points)*

Problem 2.

Let $Y_1, \ldots, Y_n$ be a set of independent random variables with the following pdfs

$$p(y_i|\theta_i) = \exp(y_i b(\theta_i) + c(\theta_i) + d(y_i)), \quad i = 1, \ldots, n$$

Let $\mathbb{E}(Y_i) = \mu_i(\theta_i),\ g(\mu_i) = x_i^T\beta$, where g is the link function and $\beta \in \mathbb{R}^d$ is the vector of model parameters.

(1) Denote $g(\mu_i)$ as η_i, and let s be the score function of β. Show that *(15 points)*

$$s_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}\mathrm{ar}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}, \quad j = 1, \ldots, d$$

(2) Let $\mathcal{I}$ be the Fisher information matrix. Show that *(5 points)*

$$\mathcal{I}_{jk} = \mathbb{E}(s_j s_k) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\mathbb{V}\mathrm{ar}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \quad \forall\ 1 \le j, k \le d.$$

Problem 3.

Use the following code to generate covariate matrices X

```
import numpy as np
np.random.seed(1234)

n = 100
X = np.random.normal(size=(n,2))
```

(1) Generate $n = 100$ observations Y following the logistic regression model with true parameter $\beta_0 = (-2, 1)$. *(5 points)*

(2) Find the MLE using the iteratively reweighted least square algorithm. *(10 points)*

(3) Repeat (1) and (2) for 100 instances. Compare the MLEs with the asymptotical distribution $\hat{\beta} \sim \mathcal{N}(\beta_0, \mathcal{I}^{-1}(\beta_0))$. Present your result with a scatter plot for MLEs with contours for the pdf of the asymptotical distribution. *(10 points)*

(4) Try the same for $n = 10000$. Does the asymptotical distribution provide a better fit to the MLEs? You can use the empirical covariance matrix of the MLEs for comparison. *(5 points)*

Problem 4.
Consider minimizing the following Beale's function

$$f(x, y) = (1.5 - x + xy)^2 + (2.25 - x + xy^2)^2 + (2.625 - x + xy^3)^2$$

with an initial position at $(x_0, y_0) = (0.5, 2)$.
(1) Find the global minima and show the contour plot over $[-4, 4] \times [-4, 4]$. *(5 points)*
(2) Compare gradient descent, gradient descent with momentum and nesterov's accelerated gradient descent. Plot $f - f^*$ as a function of the number of iterations. *(15 points)*
(3) Compare vanilla stochastic gradient descent with different adaptive stochastic gradient descent methods, including AdaGrad, RMSprop, and Adam. Plot $f - f^*$ as a function of the number of iterations. You can add some random noise (e.g., $\mathcal{N}(0, 0.01)$) to the exact gradient to form the stochastic gradient. *(15 points)*