HW1 solutions

Problem 1 (a) Consider the following optimization problem:

$$\begin{aligned}
\max_f \ & \mathrm{Ent}(f) := \int_{-\infty}^{+\infty} f(x)\ln f(x)\mathrm{d}x, \\
\text{s.t.} \ & \int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1, \\
& \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x = 0, \\
& \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x = 1.
\end{aligned}$$

We set the Lagrangian as follows:

$$\mathcal{L}(f,\lambda,\mu,\nu) = \int_{-\infty}^{+\infty} f(x)\ln f(x)\mathrm{d}x + \lambda\left(1 - \int_{-\infty}^{+\infty} f(x)\mathrm{d}x\right) + \mu\left(-\int_{-\infty}^{+\infty} xf(x)\mathrm{d}x\right) + \nu\left(1 - \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x\right).$$

The functional derivative $\frac{\partial\mathcal{L}}{\partial f}$is defined as: for any test function h, the follwing equality holds:

$$\int_{-\infty}^{+\infty} \frac{\partial\mathcal{L}}{\partial f} h\mathrm{d}x = \frac{\partial}{\partial\varepsilon}\bigg|_{\varepsilon=0} \mathcal{L}(f+\varepsilon h,\lambda,\mu,\nu).$$

Hence

$$\int_{-\infty}^{+\infty} \frac{\partial\mathcal{L}}{\partial f} h\mathrm{d}x = \int_{-\infty}^{+\infty} (h\ln f + h)\mathrm{d}x - \lambda\int_{-\infty}^{+\infty} h\mathrm{d}x - \mu\int_{-\infty}^{+\infty} xh\mathrm{d}x - \nu\int_{-\infty}^{+\infty} x^2 h\mathrm{d}x.$$

Which implies that

$$0 = \frac{\partial\mathcal{L}}{\partial f} = \ln f + 1 - \lambda - \mu x - \nu x^2 \Rightarrow f(x) = e^{\lambda-1-\mu^2/(4\nu)} e^{\frac{(x+\mu/(2\nu))^2}{2/(2\nu)}}.$$

Since f is a probability density function, we can see that X follows normal distribution with mean 0 and variance 1.

(b) Similar to the previous problem, we consider the following optimization problem:

$$\begin{aligned}
\max_f \ & \mathrm{Ent}(f) := \int_{-\infty}^{+\infty} f(x)\ln f(x)\mathrm{d}x, \\
\text{s.t.} \ & \int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1, \\
& \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x = m_1, \\
& \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x = m_2, \\
& \cdots, \\
& \int_{-\infty}^{+\infty} x^k f(x)\mathrm{d}x = m_k,
\end{aligned}$$

We set the Lagrangian as follows:

$$\begin{aligned}
\mathcal{L}(f,\lambda,\mu_1,\mu_2,\cdots,\mu_k) = & \int_{-\infty}^{+\infty} f(x)\ln f(x)\mathrm{d}x + \lambda\left(1 - \int_{-\infty}^{+\infty} f(x)\mathrm{d}x\right) \\
& + \mu_1\left(m_1 - \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x\right) + \cdots + \mu_k\left(m_k - \int_{-\infty}^{+\infty} x^k f(x)\mathrm{d}x\right).
\end{aligned}$$

Similarily, we maximize $\mathrm{Ent}(f)$ by setting $\frac{\partial\mathcal{L}}{\partial f}$ to zero:

$$0 = \frac{\partial\mathcal{L}}{\partial f} = \ln f + 1 - \lambda - \mu_1 x - \cdots - \mu_k x^k \Rightarrow f = e^{\lambda-1+\mu_1 x+\cdots+\mu_k x^k}.$$

By plugging this into the constraint condition, we can solve $(\lambda,\mu_1,\cdots,\mu_k)$.

Problem 2 The log-likelihood can be written as:

$$\ell(\theta|y) = yb(\theta) + c(\theta) + d(y).$$

The expectation of score function is 0:

$$0 = \mathbb{E}\left[\frac{\partial \ell(\theta|y)}{\partial \theta}\right] = \mathbb{E}\left[yb'(\theta) + c'(\theta)\right] \Rightarrow \mathbb{E}\left[y\right] = \frac{-c'(\theta)}{b'(\theta)}.$$

The property of Fisher information yields:

$$\begin{aligned}
0 =& \mathbb{E}\left[\frac{\partial^2 \ell(\theta|y)}{\partial \theta^2}\right] + \mathrm{Var}\left[\frac{\partial \ell(\theta|y)}{\partial \theta}\right] = \mathbb{E}\left[yb''(\theta) + c''(\theta)\right] + (b'(\theta))^2 \,\mathrm{Var}\left[y\right] \\
\Rightarrow \mathrm{Var}\left[y\right] =& \frac{-b''(\theta)\mathbb{E}\left[y\right] - c''(\theta)}{(b'(\theta))^2} = \frac{\frac{c'(\theta)b''(\theta)}{b'(\theta)} - c''(\theta)}{(b'(\theta))^2} = \frac{\left(\frac{-c'(\theta)}{b'(\theta)}\right)'}{b'(\theta)}.
\end{aligned}$$

(a) Denote $\boldsymbol{Y} = (y_1, \cdots, y_n)^\top, \boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^\top$. By the chain rule:

$$\begin{aligned}
\boldsymbol{s}^\top = \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{Y})}{\partial \boldsymbol{\beta}^\top} =& \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{Y})}{\partial \boldsymbol{\theta}^\top} \cdot \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}^\top} \cdot \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}^\top} \cdot \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}^\top} \\
=& \begin{pmatrix} y_1 b'(\theta_1) + c'(\theta_1) & \cdots & y_n b'(\theta_n) + c'(\theta_n) \end{pmatrix} \begin{pmatrix} \left(\frac{-c'(\theta_1)}{b'(\theta_1)}\right)' & & \\ & \ddots & \\ & & \left(\frac{-c'(\theta_n)}{b'(\theta_n)}\right)' \end{pmatrix}^{-1} \\
& \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & & \\ & \ddots & \\ & & \frac{\partial \mu_n}{\partial \eta_n} \end{pmatrix} \boldsymbol{x} \\
=& \begin{pmatrix} \frac{y_1 - \frac{-c'(\theta_1)}{b'(\theta_1)}}{\frac{\left(\frac{-c'(\theta_1)}{b'(\theta_1)}\right)'}{b'(\theta_1)}} \cdot \frac{\partial \mu_1}{\partial \eta_1} & \cdots & \frac{y_n - \frac{-c'(\theta_n)}{b'(\theta_n)}}{\frac{\left(\frac{-c'(\theta_n)}{b'(\theta_n)}\right)'}{b'(\theta_n)}} \cdot \frac{\partial \mu_n}{\partial \eta_n} \end{pmatrix} \boldsymbol{x} \\
=& \begin{pmatrix} \sum_{i=1}^n \frac{y_i - \mu_i}{\mathrm{Var}[Y_i]} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{i1} & \cdots & \sum_{i=1}^n \frac{y_i - \mu_i}{\mathrm{Var}[Y_i]} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{in} \end{pmatrix}
\end{aligned}$$

Which is the desired result.

(b) From (1)

$$\begin{aligned}
\mathcal{I}_{jk} =& \mathbb{E}\left[s_j s_k\right] = \mathbb{E}\left[\left(\sum_{i=1}^n \frac{y_i - \mu_i}{\mathrm{Var}\left[Y_i\right]} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij}\right)\left(\sum_{i=1}^n \frac{y_i - \mu_i}{\mathrm{Var}\left[Y_i\right]} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ik}\right)\right] \\
=& \sum_{1 \le i, l \le n} \mathbb{E}\left[\frac{(y_i - \mu_i)(y_l - \mu_l)}{\mathrm{Var}\left[Y_i\right] \mathrm{Var}\left[Y_l\right]} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \mu_l}{\partial \eta_l} \cdot x_{ij} x_{lk}\right] \\
=& \sum_{i=1}^n \mathbb{E}\left[\frac{(y_i - \mu_i)(y_i - \mu_i)}{\mathrm{Var}\left[Y_i\right] \mathrm{Var}\left[Y_i\right]} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij} x_{ik}\right] \qquad \text{(since } y_i\text{s are i.i.d. sample)} \\
=& \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\mathrm{Var}\left[Y_i\right]} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.
\end{aligned}$$

Problem 3 (a) We generate observations of Y following

$$\begin{aligned}
p &= \frac{\exp(X^T \beta_0)}{1 + \exp(X^T \beta_0)}; \\
Y &\sim \mathrm{Bernoulli}(p)
\end{aligned}$$

The generated observations of Y are

```
array([0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1,
                                                                  1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0,
                                                                   0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1
                                                                  , 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1,
                                                                  0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0,
                                                                   0, 0])
```

(b) IRLS algorithm for logistic regression can be found in lecture 2, thus you can use the update rules on the slides directly. The IRLS update rule is

$$\beta^{(k+1)} = \beta^{(k)} + \left(X^T W^{(k)} X\right)^{-1} \left(X^T (Y - p^{(k)})\right)$$

where

$$\begin{aligned} p^{(k)} &= \frac{1}{1+\exp(-X^T\beta^{(k)})} \\ W^{(k)} &= \operatorname{diag}\{p_1^{(k)}(1-p_1^{(k)}), p_2^{(k)}(1-p_2^{(k)}), \ldots, p_n^{(k)}(1-p_n^{(k)})\} \end{aligned}$$

The MLE found by IRLS algorithm may differ with different initial value and stopping criterion. An example is $\hat{\beta} = (1.37086595; 0.66987777)^T$.

(c) The theoretical distribution of β is $N(\beta_0, \Sigma)$ where

$$\beta_0 = (-2, 1)^T, \quad \Sigma = (X^T W X)^{-1} = \begin{pmatrix} 0.19350715 & -0.0440933 \\ -0.0440933 & 0.10205911 \end{pmatrix}$$

The empirical mean and variance of 100 estimations of β is:

$$\hat{\mu} = \frac{1}{100}\sum_{i=1}^{100} \hat{\beta}^{(i)} = (2.19250614, 1.07360059)^T$$

$$\hat{\Sigma} = \frac{1}{99}\sum_{i=1}^{100} (\hat{\beta}^{(i)} - \hat{\mu})((\hat{\beta}^{(i)} - \hat{\mu}))^T = \begin{pmatrix} 0.24955422 & -0.040046 \\ -0.040046 & 0.12947053 \end{pmatrix}$$

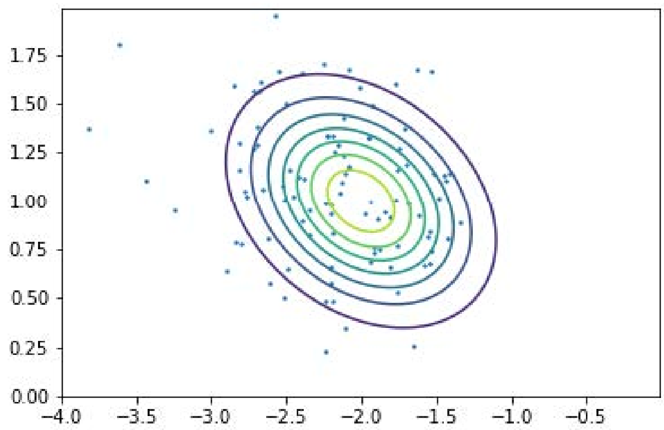

Figure 1: Contour plot for $n = 100$

(d) We repeat the procedure in (c) for $n = 10000$ and obtain the theoretical mean and variance

$$\beta_0 = (-2, 1)^T, \quad \Sigma = (X^T W X)^{-1} = \begin{pmatrix} 0.00174174 & -0.00053308 \\ -0.00053308 & 0.00096404 \end{pmatrix}$$

and the empirical mean and variance

$$\hat{\mu} = \frac{1}{100}\sum_{i=1}^{100}\hat{\beta}^{(i)} = (-2.00174993, 0.99698396)^T$$

$$\hat{\Sigma} = \frac{1}{99}\sum_{i=1}^{100}(\hat{\beta}^{(i)} - \hat{\mu})((\hat{\beta}^{(i)} - \hat{\mu}))^T = \begin{pmatrix} 0.00186519 & -0.00062197 \\ -0.00062197 & 0.00114749 \end{pmatrix}$$

(All these results may differ with different settings.)

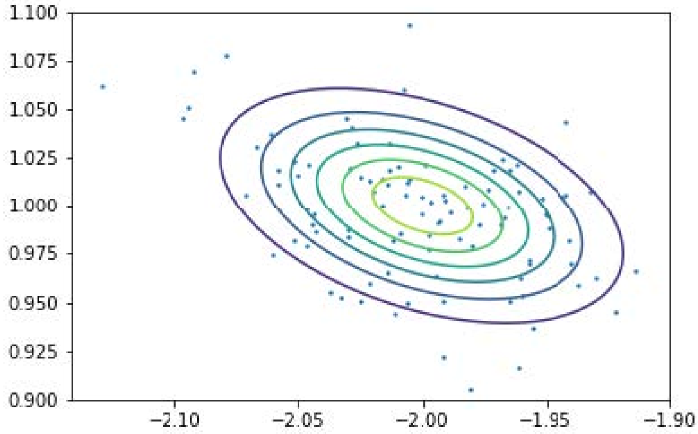

Figure 2: Contour plot for $n = 10000$

Problem 4 (a) It's obvious that $f(x,y) \geq 0$ and $f(x,y) = 0 \Rightarrow 1.5 - x + xy = 2.25 - x + xy^2 = 2.625 - x + xy^3 = 0 \Rightarrow (x,y) = (3, 0.5)$. Hence $(x^*, y^*) = (3, 0.5)$ is the unique point who reaches the global minimum.

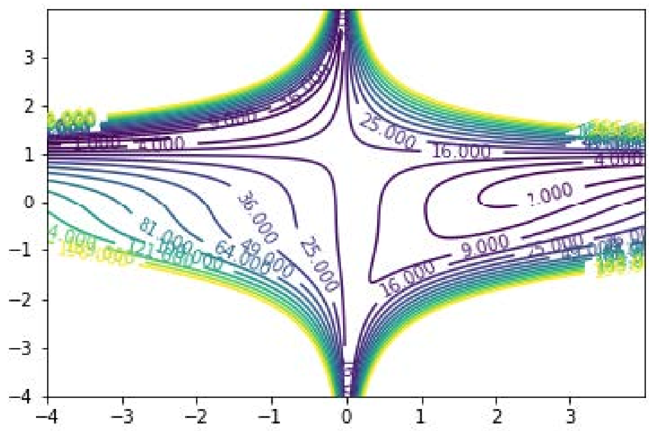

Figure 3: Contour plot for f

(b) Let $z = (x,y)^T$. The gradient of Beal function is

$$\frac{\partial f}{\partial x} = (2y-2)(1.5 - x + xy) + (2y^2 - 2)(2.25 - x + xy^2) + (2y^3 - 2)(2.625 - x + xy^3)$$
$$\frac{\partial f}{\partial y} = 2x(1.5 - x + xy) + 4xy(2.25 - x + xy^2) + 6xy^2(2.625 - x + xy^3)$$

For different optimzation algorithms, the update schemes are as follows.

- gradient descent:

$$z^{(k+1)} = z^{(k)} - \alpha \nabla f(z^{(k)}).$$

- gradient descent with momentum:

$$\begin{aligned} m^{(k)} &= \mu m^{(k-1)} + (1-\mu)\nabla f(z^{(k)}) \\ z^{(k+1)} &= z^{(k)} - \alpha m^{(k)} \end{aligned}$$

- Nesterov's acceleration:

$$\begin{aligned} \eta &= z^{(k)} + \frac{k-1}{k+2}\left(z^{(k)} - z^{(k-1)}\right) \\ z^{(k+1)} &= \eta - \alpha)\nabla f(\eta) \end{aligned}$$

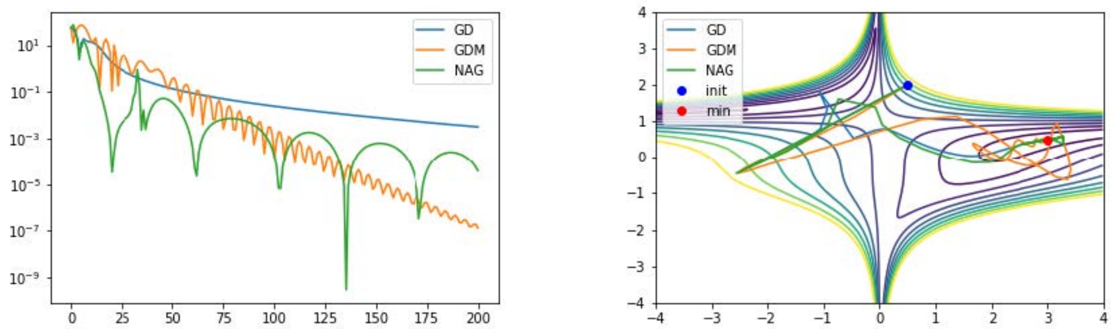

Figure 4: Stepsize $= 0.0213$ for gradient descent, Stepsize $= 0.0812, \mu = 0.9$ for gradient descent with momentum, Stepsize $= 0.0271$ for Nesterov's accelerated gradient descent.

(c) We can add random noise $N(0, 0.01)$ to obtain the stochastic gradient

$$g(z) = \nabla f(z) + \eta, \quad \eta \sim N(0, 1).$$

For vanilla SGD and its variants, the update schemes are:

- vanilla SGD:

$$z^{(k+1)} = z^{(k)} - \alpha g(z^{(k)}).$$

- AdaGrad:

$$\begin{aligned} G^{(k)} &= G^{(k-1)} + \nabla g(z^{(k)}) \odot g(z^{(k)}) \\ z^{(k+1)} &= z^{(k)} - \frac{\alpha}{\sqrt{G^{(k)} + \varepsilon}} \odot g(z^{(k)}). \end{aligned}$$

- RMSprop:

$$\begin{aligned} \mathbb{E}(g^2)_k &= \rho \mathbb{E}(g^2)_{k-1} + \rho g(z^{(k)}) \odot g(z^{(k)}) \\ z^{(k+1)} &= z^{(k)} - \frac{\alpha}{\sqrt{\mathbb{E}(g^2)_k + \varepsilon}} \odot g(z^{(k)}). \end{aligned}$$

- Adam:

$$
\begin{aligned}
m^{(k)} &= \beta_1 m^{(k-1)} + (1-\beta_1) g(z^{(k)}) \\
v^{(k)} &= \beta_2 v^{(t-1)} + (1-\beta_2) g(z^{(k)}) \odot g(z^{(k)}) \\
\hat{m}^{(k)} &= \frac{m^{(k)}}{1-\beta_1^k} \\
\hat{v}^{(k)} &= \frac{v^{(k)}}{1-\beta_2^k} \\
z^{(k+1)} &= z^{(k)} - \frac{\alpha}{\sqrt{\hat{v}^{(k)} + \varepsilon}} \odot \hat{m}^{(k)}.
\end{aligned}
$$

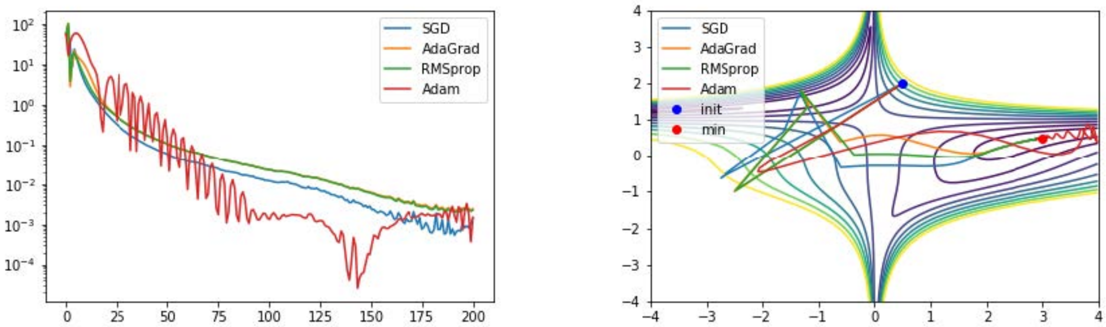

Figure 5: Stepsize = 0.029 for SGD, Stepsize = 2.99 for AdaGrad, Stepsize = 0.094, $\rho = 0.999$ for RMSprop, Stepsize = 1.152, $\beta_1 = 0.9, \beta_2 = 0.999$ for Adam.