Contents

HW1 — Computational Statistics, ISYE 6416

Ashish Dhiman — adhiman9@gatech.edu

May 7, 2023

1 Question 1: Moment Generating Function (MGF)

We are given, MGF as

$$G(\theta) = E[e^{\theta X}]$$

Now by Taylor expansion we have

$$e^{\theta X} = \sum_{i=0}^{\infty} \frac{(\theta X)^r}{r!}$$

This can be expanded as:

$$\begin{aligned} e^{\theta X} &= 1 + \frac{\theta X}{1} + \frac{(\theta X)^2}{2} + \frac{(\theta X)^3}{6} + ... \\ &\text{Taking partial derivative wrt theta} \\ \frac{\partial e^{\theta X}}{\partial \theta} &= 0 + X + (\theta X).X + \theta(...) \\ \text{similarly, } \frac{\partial^2 e^{\theta X}}{\partial \theta^2} &= 0 + 0 + (X).X + \theta(...) \end{aligned} \tag{1}$$

Now

$$\frac{\partial G(\theta)}{\partial \theta} = E[\frac{\partial e^{\theta X}}{\partial \theta}] \quad \text{By definition of MGF}$$

$$\begin{aligned} \implies \frac{\partial G(\theta)}{\partial \theta} &= E[0 + X + (\theta X).X + \theta(...)] \\ \implies \frac{\partial G(\theta)}{\partial \theta}|_{\theta=0} &= E[X] \\ &\text{Similarly} \\ \implies \frac{\partial^2 G(\theta)}{\partial \theta^2}|_{\theta=0} &= E[X^2] \end{aligned} \tag{2}$$

2 Question 2: Maximum likelihood estimator

2.1 part a: MLE Estimate

We are given $\{X_i\}$ as IIDs RVs sampled from pdf:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

For MLE of a and b we need to find the Likelihood of $X_1, X_2...X_n$ and maximise it.

$$a_{MLE}, b_{MLE} = arg\max_{(a,b)} = Lik(X_1, X_2...X_n|a, b)$$

$$\begin{aligned} Lik(X_1, X_2...X_n|a, b) &= p(X_1, X_2...X_n|a, b) \\ = p(X_1|a, b) * p(X_1|a, b)...p(X_1|a, b) &= \prod_{i=1}^{n} p(X_i|a, b) \\ &= \prod_{i=1}^{n} f(x_i) \quad \text{(From pdf)} \end{aligned} \tag{3}$$

Now the above $Lik > 0$ only if all product terms are greater than 0. This is true if $\forall X_i \in [a, b]$, otherwise Lik becomes 0. In other words:

$$b >= \max(X_1, X_2...X_n) \quad \text{and } a <= \min(X_1, X_2...X_n)$$

Keeping the above constraints in mind Lik is maximised if $b - a$ is as small as possible. Now minimising $b - a \implies$ pick the smallest b and largest a.

$$\begin{aligned} a_{MLE} &= \min(X_1, X_2...X_n), \quad \text{and} \\ b_{MLE} &= \max(X_1, X_2...X_n) \end{aligned} \tag{4}$$

2.2 part b: Unbiased Estimate

A estimator θ^* is unbiased for parameter θ iff $E[\theta^*] = \theta$. Hence for a_{MLE}, b_{MLE} to be unbiased, $E[a_{MLE}] = a$ and $E[b_{MLE}] = b$.

Let's find $P(b_{MLE} <= b)$ or the CDF of b_{MLE}.

$$\begin{aligned} P(b_{MLE} <= b) = P(b_{MLE} < b) \quad &\text{(for continuous distribution point prob. = 0)} \\ &\text{From part a we also have} \\ b >= \max(X_1, X_2...X_n) \text{ and } b_{MLE} &= \max(X_1, X_2...X_n) \\ \implies b >= b_{MLE} \\ \implies P(b_{MLE} <= b) = P(b_{MLE} < b) &= 1 \\ \implies E[b_{MLE}] &< b \end{aligned} \tag{5}$$

Hence b_{MLE} is biased. Similar proof for a_{MLE}

3 Question 3: Bayesian Inference

We are given: $x \sim \mathcal{N}(\mu, \sigma^2)$

3.1 part a: Normal-Normal

$\mu \sim \mathcal{N}(\theta, \tau^2)$ We need to find posterior for μ, or

$$p(\mu|x) = \frac{p(x|\mu) * p(\mu)}{p(x)} \propto p(x|\mu) * p(\mu)$$

Hence plugging in pdf function we have:

$$\begin{aligned} p(\mu|x) &\propto \big[\frac{1}{\sqrt{2\pi\sigma^2}} * exp(\frac{-(x-\mu)^2}{2\sigma^2})\big] * \big[\frac{1}{\sqrt{2\pi\tau^2}} * exp(\frac{-(\mu-\theta)^2}{2\tau^2})\big] \\ &\propto \exp\big(-0.5 * [\frac{-(x-\mu)^2}{2\sigma^2} + \frac{-(\mu-\theta)^2}{2\tau^2}]\big) \end{aligned}$$

Looking on terms only inside exponenent we have

$$\begin{aligned} [\frac{-(x-\mu)^2}{2\sigma^2} + \frac{-(\mu-\theta)^2}{2\tau^2}] &= \frac{\tau^2 * (x-\mu)^2) + \sigma^2 * (\mu-\theta)^2}{\sigma^2\tau^2} \\ &= \frac{\mu^2(\sigma^2+\tau^2) - 2\mu(\tau^2 x + \sigma^2\theta) + \tau^2 x^2 + \sigma^2\theta^2}{\sigma^2\tau^2} \end{aligned} \tag{6}$$

Dividing num and deonom by $(\tau^2 + \sigma^2)$

$$\begin{aligned} \frac{num}{(\tau^2+\sigma^2)} &= \mu^2 - 2\mu(\frac{\tau^2 x}{\tau^2+\sigma^2} + \frac{\sigma^2\theta}{\tau^2+\sigma^2}) + \frac{(\tau^2 x^2 + \sigma^2\theta^2)}{\tau^2+\sigma^2} \\ &= \ldots \pm (\frac{2\tau^2\sigma^2 x\theta}{(\tau^2+\sigma^2)^2}) \quad \text{completing squares} \\ &= (\frac{\tau^2 x}{\tau^2+\sigma^2} + \frac{\sigma^2\theta}{\tau^2+\sigma^2} - \mu)^2 + \text{constant} \end{aligned} \tag{7}$$

Hence,

$$\implies p(\mu|x) \propto \exp -0.5 * \big(\frac{(\frac{\tau^2 x}{\tau^2+\sigma^2} + \frac{\sigma^2\theta}{\tau^2+\sigma^2} - \mu)^2}{\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}}\big) \tag{8}$$

Hence

$$\mu|x \sim \mathcal{N}(\frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\theta, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$$

3.2 part b: Normal-Uniform

Here $\mu \sim \mathcal{U}(0, 1)$

$$\implies p(\mu|x) \propto \begin{cases} \big[\frac{1}{\sqrt{2\pi\sigma^2}} * exp(\frac{-(x-\mu)^2}{2\sigma^2})\big] * \big[\frac{1}{1-0}\big], & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Hence

$$\mu|x \sim \begin{cases} \mathcal{N}(\mu, \sigma^2), & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

4 Question 4: Basic Optimisation

4.1 part a: Examples

1. First Order: A classic example of first order optimisation method is the Gradient Descent method. The core idea of this method is to pick a starting point and perturb the point in the direction of negative gradient, which is the direction of steepest descent of the function.

2. Accelerated First Order: One example of the Accelerated First Order Method is the Momentum Gradient Descent. It is supposed to be better than Gradient Descent, because it sort of has this implicit memory about the updates. It does better than GD in cases where along some variables the descent is steep, while along the others it is plateauing.

3. Second Order method: An example of second order numerical method is the Newton method. It leverages not only gradient but also information about the curvature (or Hessian) to make updates to the point. It is derived from applying Newton's Iterative rule to find the roots of the equation, and in this context we want to find roots of gradient of a function.

   The question doesn't ask for exact update rules so I haven't written those out. But they can be easily found in the slides.

4.2 part b: Stopping Criteria

The possible stopping criteria are:

1. Based on change in x^k

$$||x^{k+1} - x^k||_2 < \epsilon$$

2. Based on change in f

$$|f(x^{k+1}) - f(x^k)| < \epsilon$$

3. Based on plateuing of ∇f:

$$|\nabla f| < \epsilon$$

4.

$$\frac{R^2 + G^2 \sum_{i=1}^{k} t_i^2}{2 \sum_{i=1}^{k} t_i}$$

Now if the function is strongly convex with constant μ:

$$f(y) >= f(x) + < \nabla f(x), y - x > + \frac{m||y-x||_2^2}{2}$$

Let x^* be the optimal point then

$$f(x) - f(x^*) <= \frac{||\nabla f(x)||^2}{2m}$$

If we are at x_k in GD iteration, the above eqn becomes

$$f(x_k) - f(x^*) <= \frac{||\nabla f(x_k)||^2}{2m} \tag{10}$$

Also let's say our stopping criteria is met, i.e. no further iteration

$$||\nabla f(x)||^2 <= \epsilon \text{In this case:}$$

$$\implies f(x_k) - f(x^*) <= \frac{1}{2m} * \epsilon$$

Hence we have a upper bound on difference between current function value from GD and true optimal value.

4.3 part c: Step size and Global Convergence for Convex function

Possible step size:

1. constant step size say, 0.01
2. $t_k = \frac{p}{|\nabla f(x_k)|}$
3. $t_k = 1/k$

Convergence:

$$\text{For a Convex function, let } x_* \text{ be minimiser}$$

$$||x_{k+1} - x_*||^2 <= ||(x_k - t_k g_k) - x_*||^2$$

$$= ||((x_{k-1} - t_{k-1} g_{k-1}) - t_k g_k) - x_*||_2$$

$$= ||x_1 - x_*||^2 - 2 \sum_{i=1tok} t_i (f(x_i - f_*) + \sum_{i=1tok} t_i^2 ||g_i||^2$$

$$\text{Let } G >= max(\{g_k | \forall k\})$$

$$\text{Also } f_k - f_* >= f_b - f_*, f_b = min(\{f_k | \forall k\})$$

$$\implies f_b - f_* <= \frac{R^2 + G^2 \sum_{i=1}^{k} t_i^2}{2 \sum_{i=1}^{k} t_i} \tag{11}$$

$$\text{Now for step size } t_k = 1/k$$

$$\lim_{n \to \text{inf}} \sum_{k=1}^{n} t_k = \text{inf}, \lim_{n \to \text{inf}} \sum_{k=1}^{n} t_k^2 = 0$$

$$\text{Hence for large k}$$

$$f_b - f_* <= \frac{finite}{\text{inf}} = 0$$

$$\text{Thus } f_b \to f_*$$

5 Question 5: Weighted Regression

Given loss of Regression function:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n} w_i(\theta^T x_i - y_i)^2.$$

5.1 part a: Weighted Loss Proof

Let X is matrix of observations with dimensions $(n*p)$ i.e. n obs. and p features, and theta be the weights vector of dimensions $(p*1)$.

In other words

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{1p} \\ x_{21} & x_{22} & x_{2p} \\ x_{n1} & x_{n2} & x_{np} \end{bmatrix}_{n*p} = \begin{bmatrix} {x_1}^T \\ {x_2}^T \\ {x_n}^T \end{bmatrix}_{n*p}$$

$$\implies \hat{y} = X\theta \quad \text{then,}$$

$$X\theta - y = \hat{y} - y = \begin{bmatrix} x_1\theta - y1 \\ x_2\theta - y2 \\ x_n\theta - yn \end{bmatrix}_{n*1}$$

$$\text{Also let diagonal weights matrix} W = \begin{bmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_n \end{bmatrix}_{n*n}$$

Now we can see that

$$J(\theta) = (X\theta - y)^T W(X\theta - y) =$$

$$\begin{bmatrix} x_1\theta - y1 & x_2\theta - y2 & x_n\theta - yn \end{bmatrix}_{1*n} * \begin{bmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_n \end{bmatrix}_{n*n} * \begin{bmatrix} x_1\theta - y1 \\ x_2\theta - y2 \\ x_n\theta - yn \end{bmatrix}_{n*1}$$

$$= \begin{bmatrix} w_1(x_1\theta - y1) & w_2(x_2\theta - y2) & w_n(x_n\theta - yn) \end{bmatrix}_{1*n} * \begin{bmatrix} x_1\theta - y1 \\ x_2\theta - y2 \\ x_n\theta - yn \end{bmatrix}_{n*1}$$

$$\implies J(\theta) = w1(x_1\theta - y1)^2 + ... + w_n(x_n\theta - yn)^2 \tag{12}$$

5.2 part b: Varying variance

We are given (x_i, y_i), $i = 1, \ldots, n$ of n independent examples, and $y_i = \theta^T x_i + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, {\sigma_i}^2)$

This means,

$$p(y_i|x_i, \theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2})$$

Now we want to find θ_{MLE} s.t.:

$$\theta_{MLE} = arg\max_{\theta} Lik(y_1, y_2, ..., y_n|x_1, x_2, ..., x_n, \theta)$$

$$
\begin{aligned}
Lik &= p(y_1, y_2, ..., y_n | x_1, x_2, ..., x_n, \theta) \\
&= \prod_{i=1}^{n} p(y_i | x_i, \theta) \quad \text{Independent examples} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2}\right)\right] \\
&\text{ignoring } \pi, \sigma_i \text{ terms as they won't impact optimisation wrt } \theta \qquad (13) \\
&\propto \exp\Big(-0.5 * \sum_{i=1}^{n} \frac{(y_i - \theta^T x_i)^2}{\sigma_i^2})\Big) \\
&\text{Let } w_i = 1/\sigma_i^2, \text{ then:} \\
Lik &\propto \exp\Big(-0.5 * \sum_{i=1}^{n} w_i (y_i - \theta^T x_i)^2\Big)
\end{aligned}
$$

Now maximising above likelihood is equivalent to minimising the terms in exponent since if x goes to -inf exp(-x) goes to inf

$$
\implies \theta_{MLE} = arg\max_{\theta} Lik = arg\min_{\theta} \sum_{i=1}^{n} w_i (y_i - \theta^T x_i)^2 \qquad (14)
$$

The above problem is convex b'cos from part 1 above

$$
J(\theta) = \frac{1}{2} \sum_{i=1}^{n} w_i (\theta^T x_i - y_i)^2 = (X\theta - y)^T W (X\theta - y)
$$

In such form Hessian of J is W, and since W is diagonal with positive entries, its semi definite hence problem is convex.

5.3 part c: Fit unweighted regression

The data looks like:

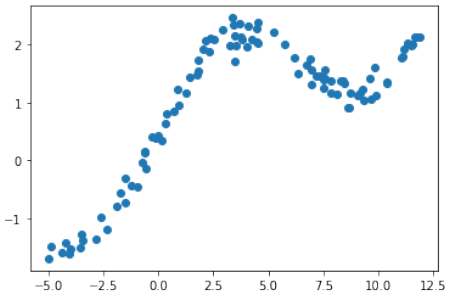

Figure 1: Scatter plot of Data

The gradient is:

$$
\nabla = 2X^T(X\beta - y)
$$

The Regression fit looks as below:

With 100 iterations of gradient descent the gradient descent fit line has almost converged to true fit $(X^T X)^{-1}(Xy)$. Since the data is quadratic, the linear fit is not enough to accommodate it.

$$
\beta_{true} \begin{bmatrix} 0.17531122 \\ 0.32767539 \end{bmatrix} \quad \beta_{gradient} = \begin{bmatrix} 0.15631765 \\ 0.50857532 \end{bmatrix}
$$

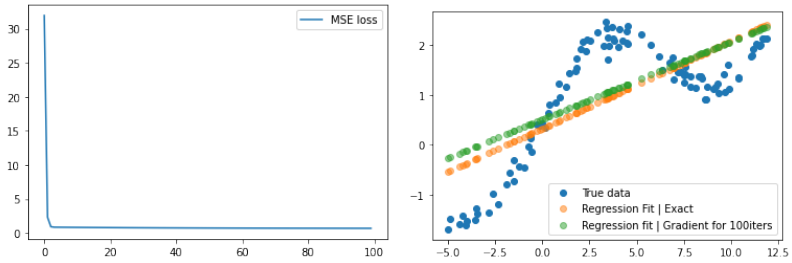

Figure 2: Loss (left) and Regression fits (right)

5.4 part d: Fit weighted regression

The weighted data looks like below if we choose $\sigma = 2$

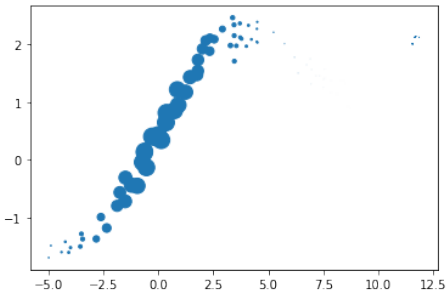

Figure 3: Scatter plot of Data

The gradient in this case is:

$$\nabla = 2X^T[W(X\beta - y)]$$

The Regression fit looks as below:

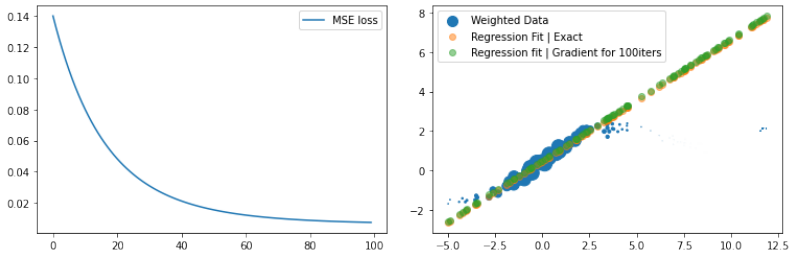

Figure 4: Loss (left) and Regression fits (right)

With points on the right being accorded less weight, we see the regression fit line is much more steeper now. True fit in this case is given by $(X^TWX)^{-1}(XWy)$.

$$\beta_{true} \begin{bmatrix} 0.61769085 \\ 0.43480531 \end{bmatrix} \quad \beta_{gradient} = \begin{bmatrix} 0.61865562 \\ 0.50510307 \end{bmatrix}$$

6 Question 6: Neural Networks and Back Propagation

6.1 part a: derivative wrt w

We have Loss as

$$L = \sum_{i=1}^{n} (y^i - \sigma(w^T z^i))^2 = \sum_{i=1}^{n} (y^i - \sigma(u^i))^2$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial u^i} * \frac{\partial u^i}{\partial w}$$

We have:

$$\frac{\partial L}{\partial u^i} = \sum_{i=1}^{n} 2(y^i - \sigma(u^i)) * (-\frac{\partial \sigma(u^i)}{\partial u^i}) * \frac{\partial u^i}{\partial w}$$

Also:

$$\frac{\partial \sigma(u^i)}{\partial u^i} = \frac{-1}{(1+\exp^{-u^i})^2} * (0 - \exp^{-u^i}) \tag{15}$$

$$= \frac{\exp^{-u^i}}{(1+\exp^{-u^i})^2}$$

$$= \frac{-1}{1-\exp^{-u^i}} * (1 - \frac{1}{1-\exp^{-u^i}})$$

$$= \sigma(u^i) * (1 - \sigma(u^i))$$

$$\frac{\partial u^i}{\partial w} = z^i$$

$$\implies \frac{\partial L}{\partial w} = -\sum_{i=1}^{n} 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i))z^i$$

6.2 part b: derivative wrt α, β

We want

$$\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial u^i} * \frac{\partial u^i}{\partial z^1} * \frac{\partial z^1}{\partial \alpha}$$

Similar to first part:

$$\implies \frac{\partial L}{\partial u^i} = -\sum_{i=1}^{n} 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i))$$

And:

$$\frac{\partial u^i}{\partial z^1} = w1$$

For $\frac{\partial z^1}{\partial \alpha}$

$$z^1 = \sigma(\alpha^T x_i) = \sigma(v_i)$$

$$\implies \frac{\partial z^1}{\partial \alpha} = \sigma(v_i) * (1 - \sigma(v_i)) * \frac{\partial v_i}{\partial \alpha}$$

$$= \frac{\partial z^1}{\partial \alpha} = \sigma(v_i) * (1 - \sigma(v_i)) * x_i$$

$$\implies \frac{\partial L}{\partial \alpha} = -\sum_{i=1}^{n} 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i)) * w1 * \sigma(v_i) * (1 - \sigma(v_i)) * x_i$$

By Symmetry:

$$\implies \frac{\partial L}{\partial \beta} = -\sum_{i=1}^{n} 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i)) * w2 * \sigma(p_i) * (1 - \sigma(p_i)) * x_i$$

where $p_i = \sigma(\beta^T x_i)$

(16)