# Collection of Problems that I think are Cool

Benjamin Chu

September 2, 2019

## 1 Math

**Problem 1.1**

Let $\mathbf{X} = \mathbb{R}^{n\times n}$ random matrix. Show that probability that $\det(\mathbf{X}) = 0$ is 0. That is, almost all $n \times n$ random matrices are invertible.

**Problem 1.2**

For $n \in \mathbb{R}$, prove that the following optimization problem

$$\begin{aligned} \max \quad & xy \\ \text{s.t.} \quad & x + y = n. \end{aligned}$$

has optimal point $x = y = \frac{n}{2}$. Then show that, with the additional constraint that $x, y, n \in \mathbb{Z}$, the solution is achieved by $x = \lceil n/2 \rceil, y = \lfloor n/2 \rfloor$.

*Proof.* Since $y = n - x$, the problem is equivalent to maximizing $x(n-x)$ with no constraint. Completing the square, we have

$$x(n-x) = -(x^2 - nx) = -\left(x - \frac{n}{2}\right)^2 + \frac{n}{4}.$$

Since $n$ is fixed, the objectice is maximized when $-\left(x - \frac{n}{2}\right)^2 = 0 \iff x = n/2 = y$. If we seek integer solutions, minimizing $x - n/2$ is achieved by rounding $n/2$ to the nearest integer. Thus $y$ is just $n - \lceil n/2 \rceil = \lfloor n/2 \rfloor$.

**Note to self:** the intuitive method of differentiation natural to all calculus students fails for the integer case, whereas completing the square method natural to middle school students is straightforward. $\square$

**Problem 1.3**

Continuing the previous problem, for $n \in \mathbb{R}$, prove that the following optimization problem

$$\begin{aligned} \max \quad & x_1 x_2 \cdots x_m \\ \text{s.t.} \quad & \sum_{i=1}^{m} x_i = n. \end{aligned}$$

has optimal point $x_i = \frac{n}{m}$. What would the solution look like with the additional constraint $x_i, n \in \mathbb{Z}$?

**Problem 1.4**

Given a line of length $l$, show that the maximum area it can enclose is achieved by a circle of radius $\frac{l}{2\pi}$.

# 2 Statistics

**Problem 2.1**

Consider a multiple regression where $n > p$ and $rank(\mathbf{X}) = p$. Let

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2$$

where $\mathbf{e} = (e_1, ..., e_n)^t = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ are the regression residuals and $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. Show that $\hat{\sigma}^2$ is an unibased estimator of $\sigma^2$.

*Proof.* We have

$$\hat{\sigma^2} = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-p} (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}).$$

Also, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$. Repeatedly applying cyclic permuation and linearity of trace operator, we have

$$\begin{aligned}
&\mathrm{E}\left((\mathbf{y} - \mathbf{H}\mathbf{y})^T(\mathbf{y} - \mathbf{H}\mathbf{y})\right) = \mathrm{E}(\mathbf{y}^T(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\mathbf{y}) = \mathrm{E}(\mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}) \\
&= \mathrm{tr}\left(\mathrm{E}(\mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y})\right) = \mathrm{E}\left(\mathrm{tr}((\mathbf{X}\boldsymbol{\beta} + \varepsilon)^T(\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \varepsilon))\right) \\
&= \mathrm{E}\left(\mathrm{tr}((\mathbf{X}\boldsymbol{\beta} + \varepsilon)^T(\mathbf{X}\boldsymbol{\beta} + \varepsilon - \mathbf{H}\mathbf{X}\boldsymbol{\beta} - \mathbf{H}\varepsilon))\right) = \mathrm{E}\left(\mathrm{tr}((\mathbf{X}\boldsymbol{\beta} + \varepsilon)^T(\mathbf{I} - \mathbf{H})\varepsilon)\right) \\
&= \mathrm{E}\left(\mathrm{tr}(\varepsilon^T(\mathbf{I} - \mathbf{H})\varepsilon)\right) = \mathrm{tr}\left((\mathbf{I} - \mathbf{H})\,\mathrm{E}(\varepsilon\varepsilon^T)\right) = \mathrm{tr}\left((\mathbf{I} - \mathbf{H})\,\mathrm{Var}(\varepsilon)\right) \\
&= \sigma^2\,\mathrm{tr}(\mathbf{I} - \mathbf{H}) = \sigma^2\left(\mathrm{tr}(\mathbf{I}_{n\times n}) - \mathrm{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\right) = \sigma^2(n-p).
\end{aligned}$$

□

**Problem 2.2**

Show that sample mean and sample variance are 2 independent statistics.

# 3 Useful Tricks and Identities

**Problem 3.1** [Dobson and Barnett, 2008, Chapter 3.4]

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\lambda_i \in \mathbb{R}$, and $\mathbf{x}_i^T \in \mathbb{R}^p$ be a row of $\mathbf{X}$. Show that

$$\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^T = \mathbf{X}^T \begin{bmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{bmatrix} \mathbf{X}$$

*Proof.* This is a definition problem. By definition we have

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & — & \mathbf{x}_n \\ | & & | \end{bmatrix} \begin{bmatrix} — & \mathbf{x}_1^T & — \\ & | & \\ — & \mathbf{x}_n^T & — \end{bmatrix} \equiv \begin{bmatrix} c_{11} & \cdots & c_{ij} \\ \vdots & & \vdots \\ & & c_{nn} \end{bmatrix}$$

Therefore $c_{11} = x_{11}x_{11} + x_{21}x_{21} + ... + x_{n1}x_{n1}$. Similarly,

$$\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \begin{bmatrix} d_{11} & \cdots & d_{ij} \\ \vdots & & \vdots \\ & & d_{nn} \end{bmatrix} \iff d_{11} = \left(\mathbf{x}_1\mathbf{x}_1^T\right)_{11} + \left(\mathbf{x}_2\mathbf{x}_2^T\right)_{11} ... + \left(\mathbf{x}_n\mathbf{x}_n^T\right)_{11} = c_{11}.$$

Therefore the entries match up judiciously. □

**Problem 3.2 Exact 2nd order Taylor's expansion**

Suppose $f \in C^2(\mathbb{R})$. Show that there exists $y \in (x_0, x)$ such that:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(y)(x - x_0)^2.$$

This motivates the quadratic upper bound principle, which is used ubiquitously in MM algorithms.

*Proof.* Applying fundamental theorem of calculus twice, we have

$$\begin{aligned} f(x) &= f(x_0) + \int_{x_0}^{x} f'(x_1)dx_1 \\ &= f(x_0) + \int_{x_0}^{x} \left( f'(x_0) + \int_{x_0}^{x_1} f''(x_2)dx_2 \right) dx_1 \\ &= f(x_0) + f'(x_0)(x-x_0) + \int_{x_0}^{x}\int_{x_0}^{x_1} f''(x_2)dx_2dx_1. \end{aligned}$$

By mean value theorem, there exists $y \in (x_0, x_1)$ such that $\int_{x_0}^{x_1} f''(x_2)dx_2 = f''(y)(x_1 - x_0)$. Thus

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x-x_0) + \int_{x_0}^{x} f''(y)(x_1-x_0)dx_1 \\ &= f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2} f''(y)(x-x_0). \end{aligned}$$

□

**Problem 3.3 Clever use of Cauchy-Schwarz** [Lange, 2016, Exercise 1.4.18]

Prove the majorization

$$\begin{aligned} (x+y-z)^2 \leq -(x_n+y_n-z_n)^2 + 2(x_n+y_n-z_n)(x+y-z) \\ +3[(x-x_n)^2+(y-y_n)^2+(z-z_n)^2] \end{aligned}$$

which separtes the variables $x, y$, and $z$. In examples 1.3.6 and 1.3.7 this would facilitate penalizing parameter curvature rather than changes in parameter values.

*Proof.* First, move the first two terms on the right to the left:

$$(x+y-z)^2 - 2(x_n+y_n-z_n)(x+y-z) + (x_n+y_n-z_n)^2 \leq 3[(x-x_n)^2+(y-y_n)^2+(z-z_n)^2]$$

The left can be factored cleanly as

$$\big((x+y-z)-(x_n+y_n-z_n)\big)^2 \leq 3[(x-x_n)^2+(y-y_n)^2+(z-z_n)^2]$$

$$\iff (a+b+c)^2 \leq 3a^2+3b^2+3c^2$$

where $a = x - x_n, b = y - y_n, c = z - z_n$. Now define $v = (1,1,1), u = (a,b,c)$. By Cauchy-Schwarz, we obtain the desired result:

$$(a+b+c)^2 \leq 3(a^2+b^2+c^2).$$

□

# 4 Real worl Application Problems

**Problem 4.1**

Suppose we have a huge number of samples and small number of covariate (large $n$ small $p$) problem and we wish to fit a linear regression. Furthremore, every day millions of new sample points are generated, i.e. $\hat{\boldsymbol{\beta}}$ must be updated continuously whenever new data arrives. How would one obtain $\hat{\boldsymbol{\beta}}$ without saving larger and larger matrices?

*Proof.* Let $\mathbf{y}_i$ and $\mathbf{X}_i$ denote the samples and corresponding data of day $i$. Then up to day $n$, the concatenated full design matrix $\mathbf{X}$ and full sample vector $\mathbf{y}$ is

$$[\mathbf{X}\mathbf{y}] = \begin{bmatrix} [\mathbf{X}_1\mathbf{y}_1] \\ \vdots \\ [\mathbf{X}_n\mathbf{y}_n] \end{bmatrix}.$$

Of course we do not want to store this entire matrix because it gets bigger each day. Fortunately, the gram matrix of $[\mathbf{X}\mathbf{y}]$ is a small $p \times p$ matrix and can be readily computed:

$$\begin{aligned} [\mathbf{X}\mathbf{y}]^t[\mathbf{X}\mathbf{y}] &= [\mathbf{X}_1\mathbf{y}_1]^t[\mathbf{X}_1\mathbf{y}_1] + \ldots + [\mathbf{X}_n\mathbf{y}_n]^t[\mathbf{X}_n\mathbf{y}_n] \\ &= \begin{bmatrix} \mathbf{X}_1^t\mathbf{X}_1 & \mathbf{X}_1^t\mathbf{y} \\ \mathbf{y}_1^t\mathbf{X}_1 & \mathbf{y}_1^t\mathbf{y}_1 \end{bmatrix} + \ldots + \begin{bmatrix} \mathbf{X}_n^t\mathbf{X}_1 & \mathbf{X}_n^t\mathbf{y} \\ \mathbf{y}_n^t\mathbf{X}_1 & \mathbf{y}_n^t\mathbf{y}_1 \end{bmatrix}. \end{aligned}$$

By property of the sweep operator, we know that sweeping on this full gram matrix have the property:

$$\begin{aligned} \text{sweep}\left([\mathbf{X}\mathbf{y}]^t[\mathbf{X}\mathbf{y}]\right) &= \begin{bmatrix} -(\mathbf{X}^t\mathbf{X})^{-1} & (\mathbf{X}^t\mathbf{X})\mathbf{X}^t\mathbf{y} \\ \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X}) & \mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^y\mathbf{X}^t\mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} -\sigma^{-2}\operatorname{Cov}(\hat{\beta}) & \hat{\beta} \\ \hat{\beta}^t & ||\mathbf{y} - \hat{\mathbf{y}}||_2^2 \end{bmatrix} \end{aligned}$$

Therefore, we store the *sum* of all preceeding days of data in the form of a gram matrix. When new data arrives, we add the new data's gram matrix to the previous sum and sweep until the 2nd to last entry. Then the fitted model $\hat{\beta}$ will be on the top right column. Since $n \gg p$, the gram matrix is small and thus easy to store. □

**Problem 4.2 Modeling count data** [Dobson and Barnett, 2008, 3.5.b]

To model count data, one can choose among Poisson, Negative Binomial, and Binomial distributions. Given a set of observations $y_i$ and assuming a common rate parameter, how would one decide which of these distribution are more appropriate?

*Proof.* The 3 different models under consideration are:

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i) \\
y_i &\sim \text{NegBin}(r, p) \\
y_i &\sim \text{Binomial}(n, p).
\end{aligned}
$$

The simplest way is to use the relationship between mean and variance of $\mathbf{y}$. For Poisson, $\text{E}(by) = \text{Var}(by)$. For negative binomial, $\text{Var}(\mathbf{y}) > \text{E}(\mathbf{y})$. And for Binomial, $\text{E}(\mathbf{y}) > \text{Var}(\mathbf{y})$.

□

# References


[Dobson and Barnett, 2008] Dobson, A. J. and Barnett, A. G. (2008). *An introduction to generalized linear models*. Chapman and Hall/CRC.

[Lange, 2016] Lange, K. (2016). *MM optimization algorithms*, volume 147. SIAM.