期中報告　0751231 曾揚

題目：

Bias-Corrected Q-Learning With Multistate Extension

研究動機：

Reward 的隨機性在很多 stochastic optimization 的問題中是很具有影響力的，這可能導致在估計一個 state 的 value 時具有一個很高的 variance 跟不確定性。Q-Learning 做為一個廣泛被使用的 sample-based, value-based, model-free 的一種強化學習演算法，受隨機性的影響也是不容忽視。Q-Learning 的演算法主要的 value iteration 及 stochastic approximation 如下:

$$\hat{Q}^n \leftarrow \hat{C}(s^n, a^n) + \gamma \max_{a' \in \boldsymbol{A}(s^{n+1})} \bar{Q}^{n-1}(s^{n+1}, a') \quad .$$

$$\bar{Q}^n(s^n, a^n) \leftarrow (1 - \alpha(s^n, a^n))\ \bar{Q}^{n-1}\ (s^n, a^n) + \alpha(s^n, a^n)\hat{Q}^n$$

雖然身為一個相當廣泛被使用的演算法，其的收斂性質已經從理論上被證明及被接受，但由於其本身演算法的特性是透過估計 Q-Value 去選擇 policy，若 Q-Value 的估計含有很高的 stochastic noise，則其的 transition function 也會含有很高的隨機性。而 Q-Learning 在產生一個 value estimate 時，含有 maximum operator，這會導致在 sample 樣本數較低時，產生一個過估計(overestimate)的現象，這個現象可能會持續幾十或幾百萬次的 sample 才會漸漸收斂至理想上的值，收斂速度非常的慢，這會使得在實驗時我們會誤以為其已經抵達收斂值，而停止繼續讓 model 學習及獲得一個有偏差的估計值。

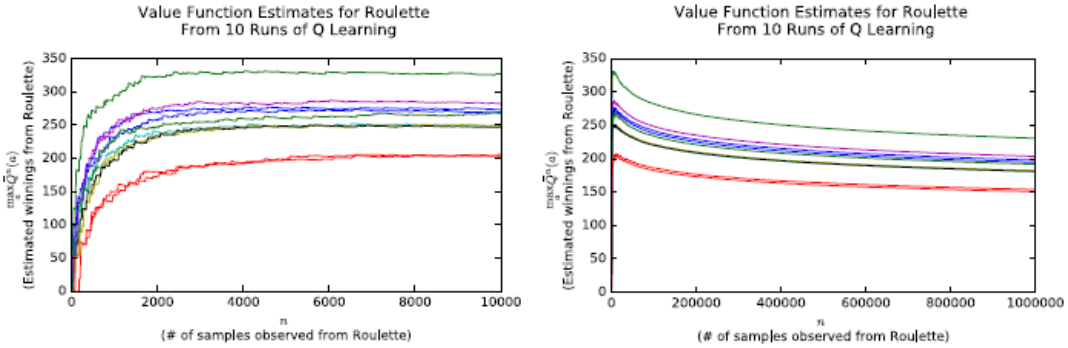
此篇 paper 旨透過在 value iteration 的式中加入一個作者提出的偏差修正項，使得 model 在學習的過程中得偏差可以大幅度得縮小，解決此過估計得問題。

實驗環境：

作者使用的實驗環境為 gym 套件中的俄羅斯輪盤遊戲(Roulette)，轉盤上有 1~37 共 37 個數字，每一輪遊戲，輪盤會隨機開出一個數字做為中獎號碼，其每個號碼被開出的機率相同。而每輪 Agent 可以選擇一個數字做為賭注或選擇不賭，共計有 38 種 action 可以選擇。為了簡化問題及方便觀察期望值，每一次的賭注只能投注 1 塊錢，若賭中則獲得的 reward 為 1，反之則 reward 為-1，且每一輪賭注結束之後即重置輪盤。在選擇賭其中一個數字的情況下，獲得的 reward 之期望值為-0.052 元，而選擇不賭的期望值為 0 元，為 38 個 action 中期望值最高的 action，亦為理想上 Q-Learning 於完成學習時應收斂到的 value 值。

Q-Learning Bias 之確立與定義：

如上的研究動機所述，Q-Learning 在學習時，由於環境的隨機性以及 maximum operator 的關係，會產生估計值偏差，在未做偏差修正的情況下，其學習的狀況如下圖所示。下圖之實驗中使用得衰減係數 $\gamma = 0.99$，為了使 transition function 造成的偏差更加明顯。



圖中橫軸為 sample 之次數，縱軸為 state 之估計值，理想上其應收斂至 0。總共進行了 100 萬次的 sample，左圖為次數前 1%之估計狀況，其由於 sample 的次數尚不夠多，故產生了極大的估計偏差。右圖為完整 100 萬次的 sample 的估計值趨勢圖，可以發現其在 sample 前段會有相當高的偏差，而後隨著 sample 次數的增多，估計值逐漸下降，但即便已經 sample 了 100 萬次，其估計值仍與實際值有相當大的差距。

作者在 paper 中首先定義了偏差值的大小如何計算，如下式:

$$B^n := \hat{Q}^n - E[\hat{Q}^n | \tau^{n-1}] \tag{1}$$

在此環境中，後者之值為理想上收斂之值，即為 0。而後作者證明了只要 reward function 以及 transition function 為獨立之隨機變數，則偏差一定有機會發生。如下 lemma 所示:

$$P[B^n > 0 | \tau^{n-1}] > 0$$

透過此 lemma，作者確立了問題確實是存在的，因此嘗試提出一個修正項去修正此問題。

偏差值之分項劃分：

透過(1)式及 Q 值的定義式，將偏差值展開，如下:

$$\boldsymbol{B^n := \hat{Q}^n - E\left[\hat{Q}^n \middle| \tau^{n-1}\right]}$$

$$= \hat{C}(s^n, a^n) + \gamma \max_{a' \in \boldsymbol{A}(s^{n+1})} \bar{Q}^{n-1}(s^{n+1}, a') - \left(E\left[\hat{C}(s^n, a^n) | \tau^{n-1}\right] + \gamma E\left[\max_{a' \in \boldsymbol{A}(T^n)} \{\bar{Q}^{n-1}(T(s,a), a') | \tau^{n-1}\} \right] \right)$$

$= \left(\hat{C}(s^n, a^n) - E\left[\hat{C}(s^n, a^n)|\tau^{n-1}\right]\right) + \gamma\left(\max_{a' \in \boldsymbol{A}(s^{n+1})} \bar{Q}^{n-1}(s^{n+1}, a') - E\left[\max_{a' \in \boldsymbol{A}(T^n)} \{\bar{Q}^{n-1}(T(s,a), a')|\tau^{n-1}\}\right]\right)$

$$= \boldsymbol{B_C^n} + \boldsymbol{B_T^n}$$

作者將偏差劃分為來自於 reward 項之偏差以及 transition 項之偏差，個別做處理。並且由此式可以看出:

$$Var^{n-1}\hat{C}(s^n, a^n) = 0 \rightarrow \boldsymbol{B_C^n} = 0$$
$$\gamma = 0 \text{ or } |\boldsymbol{S}| = 1 \quad \rightarrow \boldsymbol{B_T^n} = 0$$

同時也是在一次映證了問題的存在。

$\boldsymbol{B_C^n}$ 之偏差消除：

為了簡化問題，作者先將環境設定為一個 MDP-SS(single state)，亦即一個只有一輪的俄羅斯輪盤遊戲，此種環境使得$|\boldsymbol{S}| = 1$ ，藉此先消除$\boldsymbol{B_T^n}$帶來的影響，單純先探討$\boldsymbol{B_C^n}$。作者在 paper 中企圖透過推導找到偏差的機率分布，並將其消除。他提出了一個偏差修正項，如下:

$$\tilde{B}_C^{n-1}(s^n, a^n) = \left(\frac{\xi}{b_{|\boldsymbol{A}(s^n)|}} + b_{|\boldsymbol{A}(s^n)|}\right)\sigma(s^n, a^n)$$

- Where ξ is Euler-Mascheroni constant. $\xi \approx 0.5774$
- $b_M := (2\log(M+7) - \log\log(M+7) - \log 4\pi)^{\frac{1}{2}}$
- $\sigma(s^n, a^n) := \sqrt{Var[C(s,a)]}$

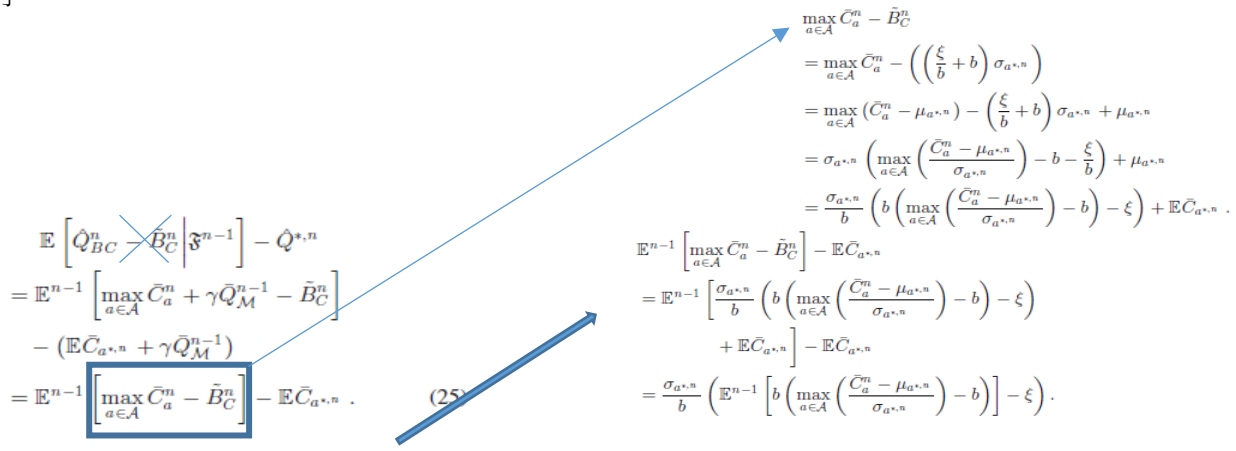
在每一輪做 value iteration 時將此偏差修正項計算後帶入修正 Q value 之值，如下圖演算法所示 ：

Algorithm 1: BCQ for SS-MDP.

Require: $\bar{Q}^0(s,a)$, s^0, γ, access to MDP $(\mathcal{S}, \mathcal{A}, C, T)$, N, stepsize rule α_n

1: **for** $n = 0, 1, \ldots, N-1$ **do**
2: Decide a^n
3: Observe $\hat{C}^{n+1} \sim C(s^n, a^n)$
4: Observe $s^{n+1} = \hat{T}^{n+1} \sim T(s^n, a^n)$
5: **if** $\tilde{B}_C^{n+1}$ is computable **then**
6: Compute $\tilde{B}_C^{n+1} \leftarrow \left(\frac{\xi}{b_{|\mathcal{A}(s^n)|}} + b_{|\mathcal{A}(s^n)|}\right)\bar{\sigma}_{n+1}(s^n, a^n)$
7: Compute $\hat{Q}_{BC}^{n+1} \leftarrow \bar{C}^{n+1}(s^n, a^n) + \gamma \max_{a \in \mathcal{A}(s^{n+1})} \bar{Q}^n(s^{n+1}, a) - \tilde{B}_C^{n+1}$
8: Update $\bar{Q}^{n+1}(s^n, a^n) \leftarrow (1 - \alpha_n(s^n, a^n))\bar{Q}^n(s^n, a^n) + \alpha_n(s^n, a^n)\hat{Q}_{BC}^{n+1}$
9: **end if**
10: **end for**
11: **return** $\bar{Q}^N(s,a)$

透過將提出修正項帶入原先之偏差項之定義式，推導後得到下圖右下角之式子。



$$\mathbb{E}\left[\hat{Q}^n_{BC} - \tilde{B}^n_C \middle| \mathfrak{F}^{n-1}\right] - \hat{Q}^{*,n}$$
$$= \mathbb{E}^{n-1}\left[\max_{a\in A} \bar{C}^n_a + \gamma \bar{Q}^{n-1}_M - \tilde{B}^n_C\right]$$
$$- \left(\mathbb{E}\bar{C}_{a^{*,n}} + \gamma \bar{Q}^{n-1}_M\right)$$
$$= \mathbb{E}^{n-1}\left[\max_{a\in A} \bar{C}^n_a - \tilde{B}^n_C\right] - \mathbb{E}\bar{C}_{a^{*,n}} \,. \quad (25)$$

$$\max_{a\in A} \bar{C}^n_a - \tilde{B}^n_C$$
$$= \max_{a\in A} \bar{C}^n_a - \left(\left(\frac{\xi}{b} + b\right)\sigma_{a^{*,n}}\right)$$
$$= \max_{a\in A} \left(\bar{C}^n_a - \mu_{a^{*,n}}\right) - \left(\frac{\xi}{b} + b\right)\sigma_{a^{*,n}} + \mu_{a^{*,n}}$$
$$= \sigma_{a^{*,n}}\left(\max_{a\in A}\left(\frac{\bar{C}^n_a - \mu_{a^{*,n}}}{\sigma_{a^{*,n}}}\right) - b - \frac{\xi}{b}\right) + \mu_{a^{*,n}}$$
$$= \frac{\sigma_{a^{*,n}}}{b}\left(b\left(\max_{a\in A}\left(\frac{\bar{C}^n_a - \mu_{a^{*,n}}}{\sigma_{a^{*,n}}}\right) - b\right) - \xi\right) + \mathbb{E}\bar{C}_{a^{*,n}} \,.$$

$$\mathbb{E}^{n-1}\left[\max_{a\in A} \bar{C}^n_a - \tilde{B}^n_C\right] - \mathbb{E}\bar{C}_{a^{*,n}}$$
$$= \mathbb{E}^{n-1}\left[\frac{\sigma_{a^{*,n}}}{b}\left(b\left(\max_{a\in A}\left(\frac{\bar{C}^n_a - \mu_{a^{*,n}}}{\sigma_{a^{*,n}}}\right) - b\right) - \xi\right)\right.$$
$$\left. + \mathbb{E}\bar{C}_{a^{*,n}}\right] - \mathbb{E}\bar{C}_{a^{*,n}}$$
$$= \frac{\sigma_{a^{*,n}}}{b}\left(\mathbb{E}^{n-1}\left[b\left(\max_{a\in A}\left(\frac{\bar{C}^n_a - \mu_{a^{*,n}}}{\sigma_{a^{*,n}}}\right) - b\right)\right] - \xi\right).$$

而後作者提出下式 lemma :

$$b_M \cdot \left(\max_i \left\{\frac{\hat{X}_i - E\hat{X}_i}{\sqrt{Var\hat{X}_i}} \middle| i \in \{1,2,\dots,M\}\right\} - b_M\right) \xrightarrow{M\to\infty} G(0,1)$$

透過此 lemma 我們可將其帶回上圖之右下角式中。由於得知括號中之期望值為一標準 gumbel 分布，故其期望值就等於標準 gumbel 分布時的 mean 值，即為 $\xi \approx 0.5774$ ，因此透過減掉 ξ 將整個式子，亦即 Bias，修正到 0。

B^n_T 之偏差消除：

B^n_T之偏差根據定義，如下式:

$$B^n_T := \gamma\left(\max_{a'\in A(s^{n+1})} \bar{Q}^{n-1}(s^{n+1}, a') - E\left[\max_{a'\in A(T^n)} \{\bar{Q}^{n-1}(T(s,a),a')|\tau^{n-1}\}\right]\right)$$

此處作者簡單的使用多個 Sample 取平均的方式去近似期望值，如下式:

$$\tilde{B}^n_T := \gamma\left(\max_{a'\in A(s^{n+1})} \bar{Q}^{n-1}(s^{n+1}, a') - \frac{1}{K}\sum_{k=1}^{K}\left[\max_{a'\in A(T^n)} \{\bar{Q}^{n-1}(T(s,a),a')\}\right]\right)$$

演算法的流程方面，與 BCQ-SS 時類似僅於修正時多加入此一$\tilde{B}^n_T$項，如下圖:

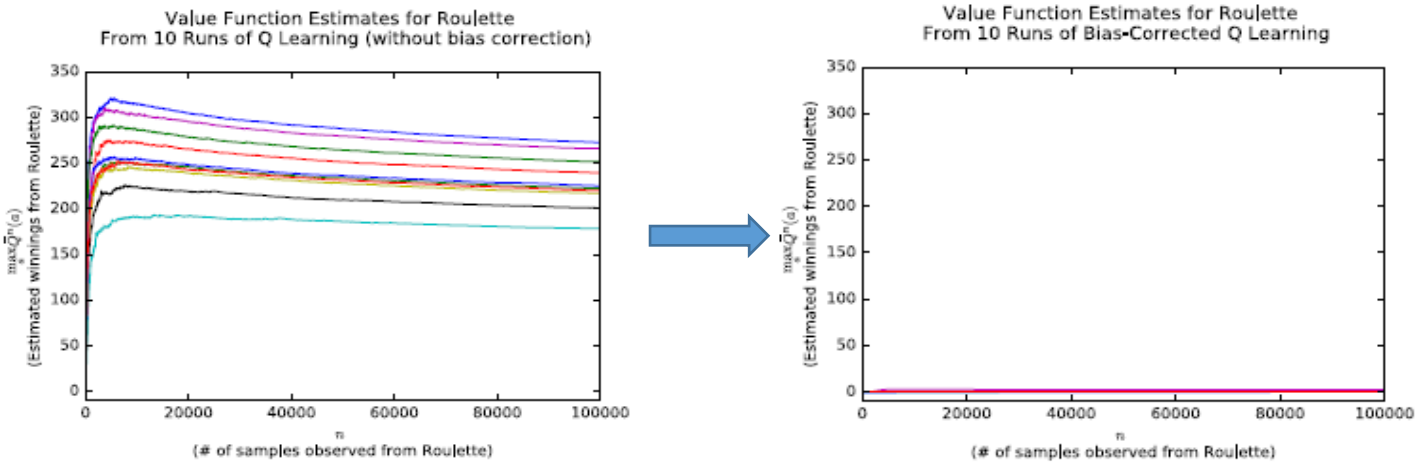
Algorithm 2: BCQ Algorithm With Multistate Extension.

Require: $\bar{Q}^0(s,a)$, s^0, γ, access to MDP $(\mathcal{S},\mathcal{A},C,T)$, N, stepsize rule α_n, Burn-in parameter K

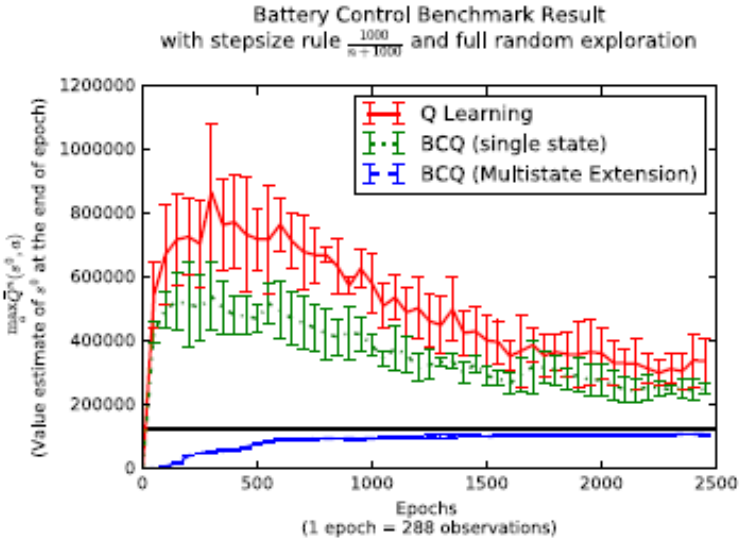
1: **for** $n = 0, 1, \ldots, N-1$ **do**
2: Decide a^n
3: Observe $\hat{C}^{n+1} \sim C(s^n, a^n)$
4: Observe $s^{n+1} = \hat{T}^{n+1} \sim T\left(s^{n-1}, a^{n-1}\right)$
5: **if** $\tilde{B}_C^{n+1}$ is computable and $\tilde{B}_T^{n+1}$ is computable **then**
6: Compute $\tilde{B}_C^{n+1} \leftarrow \left(\frac{\xi}{b_{|\mathcal{A}(s^n)|}} + b_{|\mathcal{A}(s^n)|}\right) \bar{\sigma}_n (s^n, a^n)$
7: Compute $\tilde{B}_T^{n+1} \leftarrow \gamma(\max_{a' \in \mathcal{A}(s^{n+1})}(\bar{Q}^n(s^{n+1}, a')) - \frac{1}{K}\sum_{k=1}^{K}(\max_{a' \in \mathcal{A}(\hat{T}^k)} \bar{Q}^n(\hat{T}^k, a')))$
8: Compute $\hat{Q}_{BC}^{n+1} \leftarrow \bar{C}^{n+1}(s^n, a^n) + \gamma \max_{a \in \mathcal{A}(s^{n+1})} \bar{Q}^n\left(s^{n+1}, a\right) - (\tilde{B}_C^n + \tilde{B}_T^n)$
9: Update $\bar{Q}^{n+1}\left(s^n, a^n\right) \leftarrow (1 - \alpha_n\left(s^n, a^n\right))\bar{Q}^n\left(s^n, a^n\right) + \alpha_n\left(s^n, a^n\right)\hat{Q}_{BC}^{n+1}$
10: **end if**
11: **end for**
12: **return** $\bar{Q}^N(s,a)$

實驗結果：

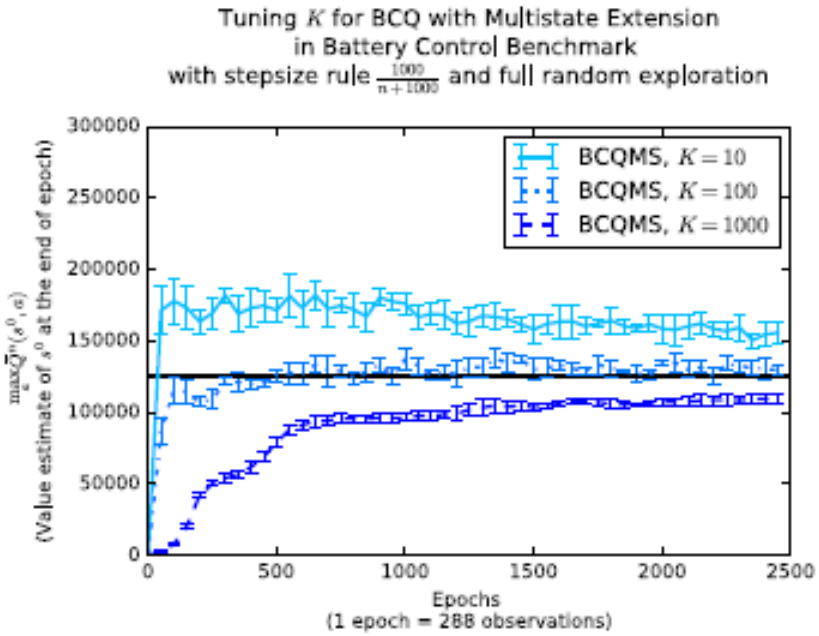
在 MDP-SS 中，沒有$\boldsymbol{B_T^n}$，故僅修正$\boldsymbol{B_C^n}$。其實驗結果如下圖所示，原本於一開始會造成之過估計值，在加入修正項後，幾乎沒有造成任何過估計的現象，估計值良好的在理想值 0 附近。



而在 BCQ-MS (multi-state)中，同時具有$\boldsymbol{B_T^n}$與$\boldsymbol{B_C^n}$，經過作者做偏差修正後，實驗結果如下圖所示，圖中也比較了僅修正$\boldsymbol{B_C^n}$與沒有做修正時的曲線。從圖中不難觀察出將偏差全數修正過後的曲線較其他兩者在收斂上更為平滑且穩定。



作者同時也做了 K 值大小的比較，如下圖。我們可以發現 K 值為 100 時收斂的表現就已經相當不錯，反而 K 值為 1000 時收斂的速度較為緩慢。推測原因為因為每 1000 個 sample 才做一次 value iteration，所以在相同 learning rate 的狀況下，估計值更新的速度較慢，但其仍然是以穩定的速度收斂到理想中的期望值。



結果討論：

作者表示使用此方法會使得計算的成本提高 40%，但可以良好得將偏差修正，在一些需要精準估計得狀況下，40%得計算成本是可以被接受的。

期末探討：

期末我主要會以實現此 Paper 的內容並確認其實驗結果為主要目標，並且對兩個點做深入探討。

1. 作者的所有推倒中都是基於 M 趨近於無限大的前提下做的推導，其於 paper 中表示 M 有沒有趨近於無限大的影響相較於 maximum operator 給予他的 Bias 小了許多，固可忽略，但其並沒有給予證明或實驗映證其論點。期末我會對此點做加以探索與比較。
2. 在作者提出的修正項中，

$$b_M := (2\log(M+7) - \log\log(M+7) - \log 4\pi)^{\frac{1}{2}}$$

在上式中，包含兩個(M+7)項，但於其證明過程中對於此常數 7 並未做過多解釋，且證明過程中其作用亦不明。故期末會探討此常數給予修正項的影響。