Solutions to Zach Furman's Singular Learning Theory Exercises

These are a set of solutions I have written to these released exercises on LessWrong from Zach Furman. I have found them as a great introduction to the field of *Singular* Learning Theory.

I am newbie to singular learning theory so it is best not to over-index on these solutions. Please email anish.lakkapragada@yale.edu for any questions or errors.

Exercise: 1

a)

$$I(\mu) = -\mathbb{E}[\frac{\partial^2}{\partial \mu^2} \log p(x \mid \mu)] = -\mathbb{E}[\frac{\partial^2}{\partial \mu^2}(-\log\left(\sqrt{2\pi}\right) - \frac{(x-\mu)^2}{2})] = -\mathbb{E}[-1] = 1$$

b) Note that true distribution $q(x) = p(x \mid \mu_0) \implies$ our true distribution $\sim \mathcal{N}(\mu_0, 1)$. Using the formula for the KL divergence between two normals, we get:

$$\text{KL}(q(x) \mid\mid p(x \mid \mu)) = \frac{(\mu_0 - \mu)^2}{2}$$

c) Note that $\forall \epsilon > 0, K(\mu) < \epsilon \iff (\mu_0 - \mu)^2 < 2\epsilon \iff \mu \in (\mu_0 - \sqrt{2\epsilon}, \mu_0 + \sqrt{2\epsilon})$. Thus,

$$V(\epsilon) = \int_{\{\mu | K(\mu) < \epsilon\}} \varphi(\mu) d\mu = \int_{\mu_0 - \sqrt{2\epsilon}}^{\mu_0 + \sqrt{2\epsilon}} d\mu = 2\sqrt{2\epsilon}$$

d) We compute λ below, choosing $a \in \mathbb{R}/\{1\}$:

$$\lambda = \lim_{\epsilon \to 0} \frac{\log(V(a\epsilon)/V(\epsilon))}{\log(a)} = \lim_{\epsilon \to 0} \frac{\log\left(\frac{2\sqrt{2a\epsilon}}{2\sqrt{2\epsilon}}\right)}{\log(a)} = \lim_{\epsilon \to 0} \frac{\log \sqrt{a}}{\log a} = \frac{1}{2}\frac{\log a}{\log a} = \boxed{\frac{1}{2}}$$

Exercise: 2

a) Say we have some mean $\mu \in \mathbb{R}$ in an ordinary normal model, then in our cubic model we can just choose $\mu^{1/3}$.

b) Note that in our cubicly-parameterized normal model we have that $\mathbb{E}[x] = \mu^3$:

$$I(\mu) = -\mathbb{E}[\frac{\partial^2}{\partial \mu^2} \log p(x \mid \mu)] = -\mathbb{E}[\frac{\partial^2}{\partial \mu^2}(-\log\left(\sqrt{2\pi}\right) - \frac{(x-\mu^3)^2}{2})] = -\mathbb{E}[6\mu(x-\mu^3) - 9\mu^4]$$
$$= -6\mu\mathbb{E}[x] + 6\mu^4 + 9\mu^4 = -6\mu^4 + 6\mu^4 + 9\mu^4 = 9\mu^4$$

So the model is singular in the case of $\mu = 0$.

c) The KL divergence between the true distribution $q(x) = p(x \mid \mu_0)$ and $p(x \mid \mu)$ can again be given based on the divergence between two normals:

$$\mathrm{KL}(q(x) \,||\, p(x \mid \mu)) = \frac{(\mu_0^3 - \mu^3)^2}{2}$$

Then using our improper prior $\varphi(\mu) = 1$ again:

$$V(\epsilon) = \int_{\{\ \mu | K(\mu) < \epsilon\}} \varphi(\mu) d\mu = \int_{(\mu_0^3 - \sqrt{2}\epsilon)^{1/3}}^{(\mu_0^3 + \sqrt{2\epsilon})^{1/3}} d\mu = (\mu_0^3 + \sqrt{2\epsilon})^{1/3} - (\mu_0^3 - \sqrt{2\epsilon})^{1/3}$$

Fix $a \in \mathbb{R}/\{1\}$. We now analyze λ in the two cases below:

(i) Singularity Case: $\mu_0 = 0$
In this case, $V(\epsilon) = 2(2\epsilon)^{1/6}$ so

$$\lambda = \lim_{\epsilon \to 0} \frac{\log V(a\epsilon) - \log V(\epsilon)}{\log(a)} = \lim_{\epsilon \to 0} \frac{\frac{1}{6}\log(2a\epsilon) - \frac{1}{6}\log(2\epsilon)}{\log(a)} = \frac{1}{6} \lim_{\epsilon \to 0} \frac{\log(a)}{\log(a)} = \frac{1}{6}$$

(ii) Regular Case: $\mu_0 \neq 0$
We can define $\delta := \sqrt{2\epsilon}$ and $f(x) := (\mu_0^3 + t)^{1/3}$ so we have:

$$V(\epsilon) = f(\delta) - f(-\delta)$$

For clarity, we first give the derivative of f:

$$f'(t) = \frac{1}{3}(\mu_0^3 + t)^{-2/3}$$

So $f'(0) = \frac{1}{3}\mu_0^{-2}$. Then because $\mu_0 \neq 0$, we can use identical Taylor Expansions around $t = 0$ on both $f(\delta)$ and $f(-\delta)$ below so only the odd-powered terms remain:

$$V(\epsilon) = f(\delta) - f(-\delta) \approx 2f'(0)\delta + O(\delta^3) = \frac{2}{3}\mu_0^{-2}\sqrt{2\epsilon} + O(\epsilon^{3/2})$$

Note that when $\epsilon \to 0$, $O(\epsilon^{3/2})$ can be expressed as $o(\epsilon^{1/2})$. So for any $a \in \mathbb{R}/\{1\}$:

$$\frac{V(a\epsilon)}{V(\epsilon)} = \frac{(\frac{2}{3}\mu_0^{-2}\sqrt{2}(a\epsilon)^{1/2} + o((a\epsilon)^{1/2}))}{(\frac{2}{3}\mu_0^{-2}\sqrt{2}(\epsilon)^{1/2} + o(\epsilon^{1/2}))} = \frac{a^{1/2} + \underbrace{\frac{o((a\epsilon)^{1/2})}{\frac{2}{3}\mu_0^{-2}\sqrt{2}\epsilon^{1/2}}}_{\to 0}}{1 + \underbrace{\frac{o(\epsilon^{1/2})}{\frac{2}{3}\mu_0^{-2}\sqrt{2}\epsilon^{1/2}}}_{\to 0}} = a^{1/2}$$

So $\lambda = \lim_{\epsilon \to 0} \frac{\log \frac{V(a\epsilon)}{V(\epsilon)}}{\log(a)} = \frac{1}{2}$, as is expected in a regular model.

For the cubicly-parameterized normal model, we can give $V(\epsilon) = C\epsilon^{1/2} + O(\epsilon^{3/2})$ for some constant $C \in \mathbb{R}$ whereas in our normal model in Exercise 1 we simply had $V(\epsilon) \propto \epsilon^{1/2}$. What this means is that for all the regular cases of models $\mu_0 \neq 0$, the learning coefficient $(\frac{1}{2})$ is identical as the dominating term is $\epsilon^{1/2}$. However, at the singularity $\mu_0 = 0$, we analytically can give $V(\epsilon) \propto \epsilon^{1/6}$ which is a lower learning coefficient.

(d) Omitted.

Exercise: 9

1. This should be second nature now:

$$K(\mu) = \int_{-\infty}^{\infty} p(x \mid \mu_0) \log \frac{p(x \mid \mu_0)}{p(x \mid \mu)} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \underbrace{[-\frac{1}{2}x^2 + \frac{1}{2}(x - \mu(\mu-2)^2)^2]}_{\frac{\mu^2(\mu-2)^4}{2} - x\mu(\mu-2)^2}$$

This is just an expectation over r.v. $X \sim \mathcal{N}(0,1)$ so:

$$K(\mu) = \mathbb{E}[\frac{\mu^2(\mu-2)^4}{2} - x\mu(\mu-2)^2] = \frac{\mu^2(\mu-2)^4}{2}$$

Plotting $K(\mu)$, we can see that it has two minimas: $\mu = 0$ and $\mu = 2$, where the latter is wider.

b) *It would be helpful to first read through Exercise 5(c)'s description of the Watanabe's normal crossing method to find the learning coefficient. We will apply that here for our 1D KL divergence function $K(\mu)$ to find the LLCs at $\mu^* = 0$ and $\mu^* = 2$.*

(a) **Near $\mu^* = 0$**
We are essentially trying to find $k \in \mathbb{Z}$ where $K(\mu) \asymp c\mu^{2k}$ for μ near μ^*. Observe then that:

$$K(\mu) = \frac{\mu^2(\mu-2)^4}{2} \approx 8\mu^2 + O(\mu^{>2})$$

and so for close values near μ^* we can give $K(\mu) \approx 8\mu^{2k}$ for $k = 1 \implies$ the LLC of $\mu^* = 0$ can be given as $\frac{1}{2}$.

(b) **Near $\mu^* = 2$**
For convenience, we will reparametrize $K(\mu)$ with $t := \mu - 2$. Then note that:

$$K(\mu) = \frac{\mu^2(\mu-2)^4}{2} = \frac{(t+2)^2t^4}{2} \approx 2t^4 + O(t^{>4})$$

and so the LLC of $\mu^* = 2$ can be given as $\frac{1}{4}$.

This aligns with our qualitative description of $K(\mu)$ in part (a) – the minima in the KL divergence for $\mu^* = 2$ is much wider than $\mu^* = 0$ so it makes sense that the LLC of $\mu^* = 2$ will be less (as this implies more volume for an arbitrarily close solution.)

c) We now compute the *global* learning coefficient using the full method we have used before. First note that for any small $\epsilon > 0$,

$$\{\mu : K(\mu) < \epsilon\} \approx \{8\mu^2 + O(\mu^{>2}) < \epsilon\} \cup \{2t^4 + O(t^{>4}) < \epsilon\}$$
$$\approx \{\mu : |\mu| < \sqrt{\frac{\epsilon}{8}}\} \cup \{\mu : |t| < (\frac{\epsilon}{2})^{1/4}\}$$

Note that the first set is for all the μ close to $\mu^* = 0$ and the second set is for all μ close to $\mu^* = 2$. Recall that using our prior $\varphi(\mu)$ we have $V(\epsilon) = \int_{\{\mu:K(\mu)<\epsilon\}} \varphi(\mu)d\mu$. Using our improper prior $\varphi(\mu) = 1$ we have:

$$V(\epsilon) \approx 2\sqrt{\frac{\epsilon}{8}} + 2(\frac{\epsilon}{2})^{1/4} = \frac{\epsilon^{1/2}}{\sqrt{2}} + 2^{3/4}\epsilon^{1/4} \implies V(\epsilon) \propto \epsilon^{1/4}$$

The last statement is due to the fact that the dominating term in $V(\epsilon)$ is $\epsilon^{1/4}$ as ϵ is small. So for any $a \in \mathbb{R}/\{1\}$:

$$\lambda = \lim_{\epsilon\to 0} \frac{\log(V(a\epsilon)/V(\epsilon))}{\log(a)} = \lim_{\epsilon\to 0} \frac{\log[\frac{(a\epsilon)^{1/4}}{\epsilon^{1/4}}]}{\log(a)} = \frac{1}{4}$$

So the global learning coefficient is $\lambda = \frac{1}{4}$, which matches the lower LLC computed at minimizer $\mu^* = 2$ in part (b).

d) Watanabe's free energy asymptotic tells us that as $n \to \infty$, we will arrive at a minimizer of the KL divergence. However, when training models we are using finite data ($n < \infty$) and thus it might happen that we are in simple but less optimal minimas. The LLC helps us compare these minimas we might encounter throughout training and understand when a phase transition might occur; the global learning coefficient is useful insofar as understanding the final result of training.

Exercise: 10

For this question, note that $\int$ *refers to* $\int_{-\infty}^{\infty}$. *We also define* $L_n(w) := -\frac{1}{n}\sum_i \log p(X_i \mid w)$.

1. We first compute the RHS:

$$K(w) + S = \int q(x) \log \frac{q(x)}{p(x \mid w)} - \int q(x) \log q(x) dx = -\int q(x) \log p(x \mid w) dx$$

and then the LHS:

$$\mathbb{E}_{D_n}[L_n(w)] = \mathbb{E}_{D_n}[-\frac{1}{n}\sum_i \log p(X_i \mid w)] = -\frac{1}{n}\sum_i \mathbb{E}_{X_i\sim q(x)}[\log p(X_i \mid w)] =$$
$$-\frac{1}{n}\sum_i \int \log p(X_i \mid w) q(X_i) dX_i = -\int q(x) \log p(x \mid w) dx$$

and so we are finished.

b) We put the PDF below for convenience:

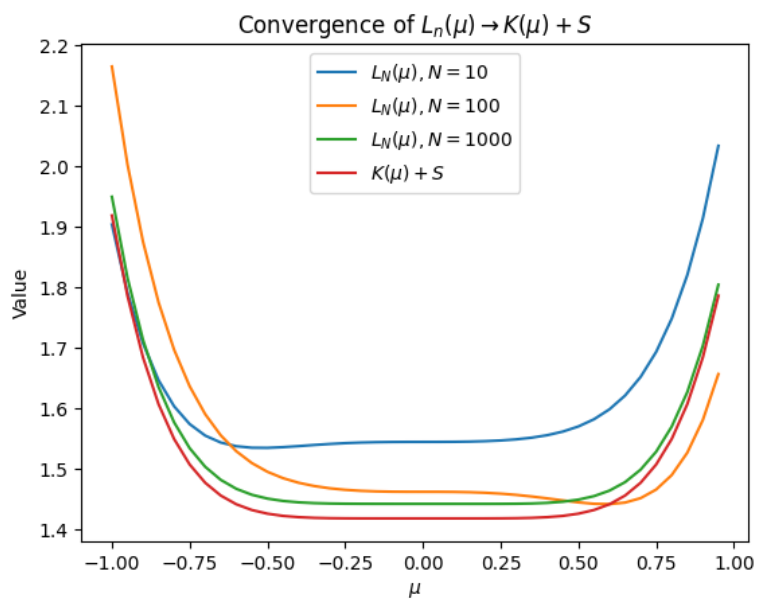
$$p(x \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu^3)^2\right)$$

Note that with the true parameter $\mu_0 = 0$, $p(x \mid \mu_0)$ gives the $\mathcal{N}(0,1)$ PDF. We also have:

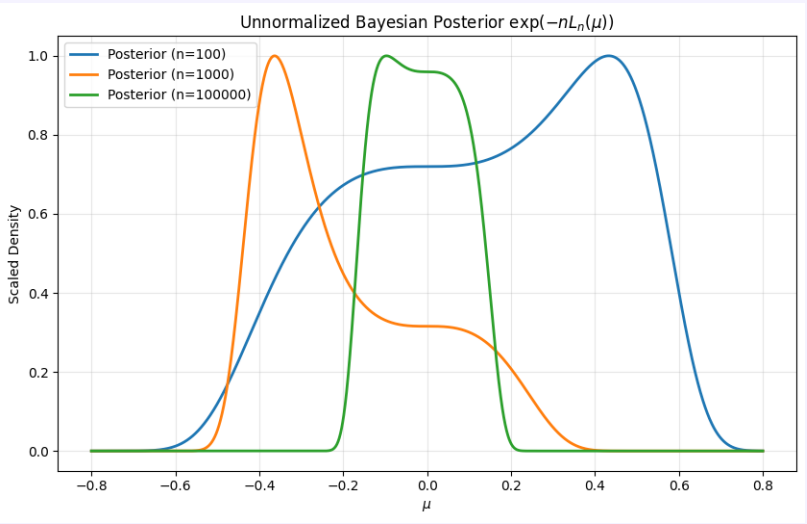
$$K(\mu) = \frac{\mu^6}{2}, \quad S = -\mathbb{E}[\log[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\chi_1^2\right)]] = \log\sqrt{2\pi} + \mathbb{E}[\frac{1}{2}\chi_1^2] = \log\sqrt{2\pi} + \frac{1}{2}$$

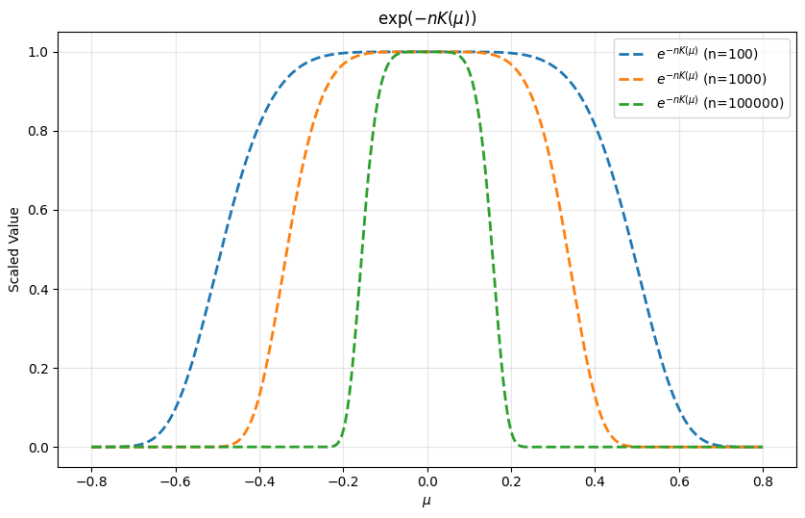
where χ_1^2 is the standard chi-squared distribution with one degree of freedom.

We attach a plot of $L_n(\mu)$ and $K(\mu) + S$ below for part (i). We can see that as n increases, $L_n(\mu) \to K(\mu) + S$.



We move onto part (ii) now. We give plots of $\exp(-nL_n(\mu))$ and $\exp(-nK(\mu))$ below:





We can see that as n increases, $\exp(-nK(\mu))$ looks more Gaussian around $\mu = 0$. Furthermore, more of the posterior mass (i.e. $\exp(-nL_n(\mu))$) shifts near $\mu = 0$.

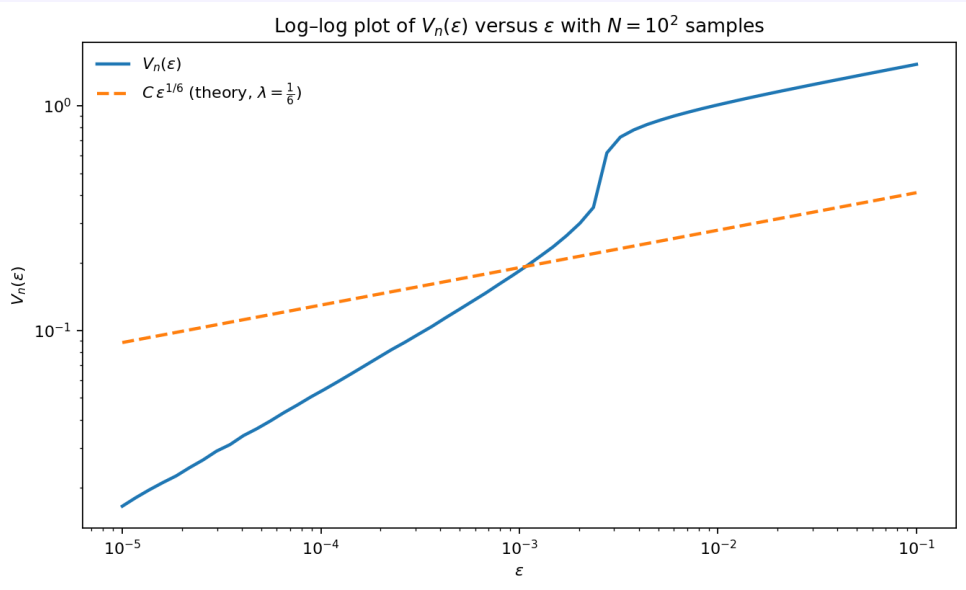
It can be shown that $\exp(-nK(\mu))$ drives changes in the posterior away from the prior. Thus, $\exp(-nK(\mu))$ can give us information on the posterior contraction rate.

c) Recall that we are using an improper prior $\varphi(\mu) = 1$ and so $V(\epsilon)$ by:

$$V(\epsilon) = \int_{\{\mu | K(\mu) < \epsilon\}} d\mu = (\mu_0^3 + \sqrt{2\epsilon})^{1/3} - (\mu_0^3 - \sqrt{2\epsilon})^{1/3} = 2(2\epsilon)^{1/6}$$

If we were to plot $\log[V(\epsilon)]$ versus ϵ, we would get a $\frac{1}{6}$ slope which gives the learning coefficient λ we derived in 2(c) for the case of $\mu_0 = 0$.

d) It is probably helpful to realize that $L_n(\mu) - \min_\mu L_n(\mu) \approx K(\mu) + S - S = K(\mu)$ as $\min_\mu L_n(\mu) \approx S$. Below is a log-log plot comparing $V_n(\epsilon)$ with ϵ using $N = 10^2$ samples:



We can see that the theoretical slope $\lambda = \frac{1}{6}$ matches better for larger values of ϵ. For smaller ϵ, $V_n(\epsilon)$ grows faster than $\epsilon^{1/6}$ relation we derived. This means that for smaller ϵ tolerances, there is actually a larger volume of possible solutions than we expected.

Exercise: 11

Note that because all $p(X_i \mid w) \in [0, 1], \beta$ is an *inverse* temperature as increasing β decreases the strength of the likelihood $\prod_i p(X_i \mid w)^\beta$ in moving the posterior $p(w \mid D_n)$. This differs from increasing n as increasing n will increase the strength of the likelihood in moving the posterior.

Exercise: 12

a) We provide the free energy $F_n(\beta)$ below:

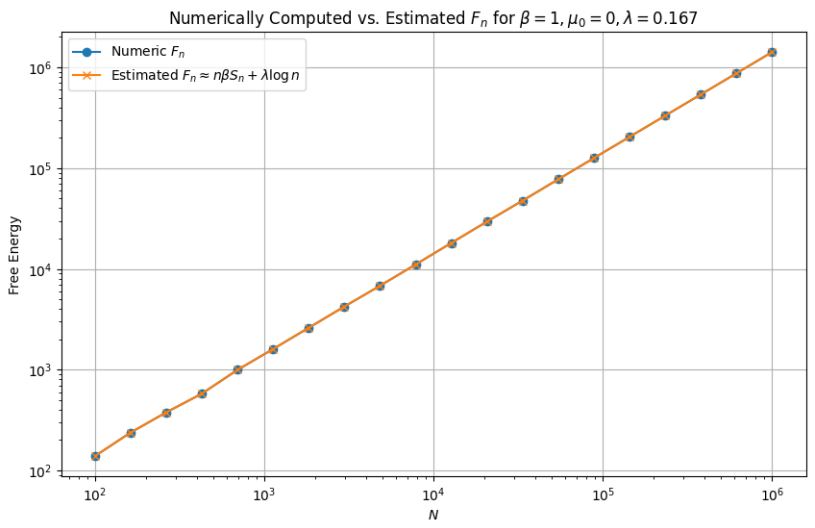
$$F_n(\beta) = -\log \int (\prod_i p(X_i \mid w)^\beta)\varphi(w)dw = -\log p(D_n)$$

We can think of $p(D_n)$ as a constant that is large when there exists a set of high-likelihood weights that strongly align with the prior. So lower values of $F_n(\beta)$ mean that there is a space of weights with good Bayesian strength, which is ideal compared to having all weights have low Bayesian strength (i.e. no weights fit our data well). Thus, we should always be aiming to choose models with lower free energy.

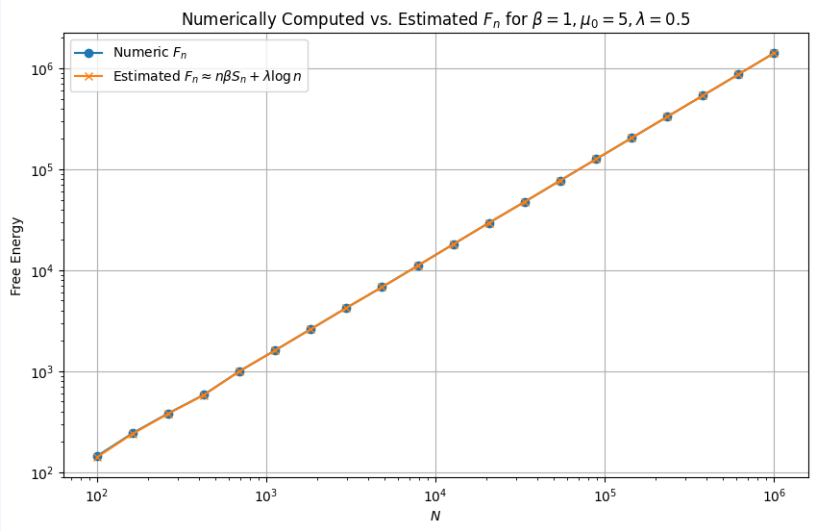
b) We again use the following setup from Question 2:

$$p(x \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu^3)^2\right), \quad \mu_0 = 0, \quad q(x) = p(x \mid \mu_0) \sim \mathcal{N}(0, 1), \quad \lambda = \frac{1}{6}$$

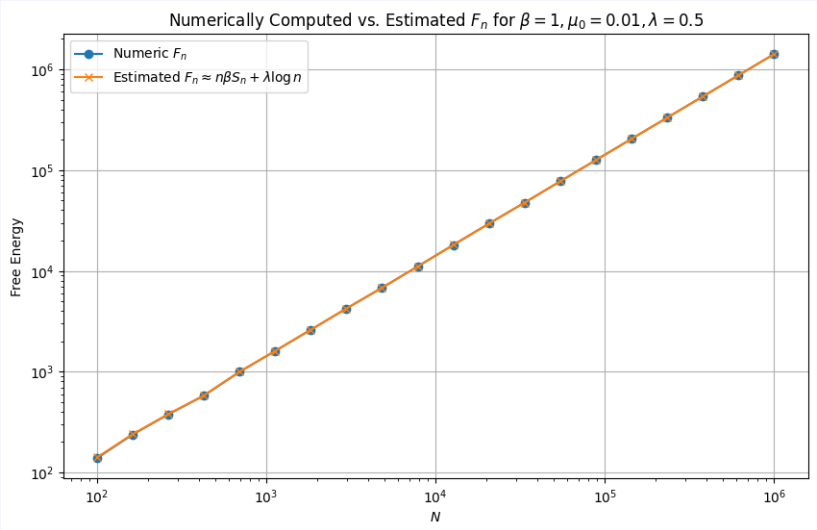
We provide a plot below of a numerically computed[a] $F_n(\beta = 1)$ free energy versus its estimate $n\beta S_n + \lambda \log n$ for $n = 10^2$ to $n = 10^6$:



c) We perform (b) again, this time for $\mu_0 = 0$ (recall then that our local learning coefficient is $\lambda = \frac{1}{2}$):



(d) We repeat (b) again with $\mu_0 = 0.01$. Even though this has a formal learning coefficient of $\frac{1}{2}$, its behavior matches more with the $\mu_0 = 0$ case in part (b) than in part (c).



[a] Note that naively integrating across μ to compute F_n is numerically unstable. I used GPT-5 when generating code to perform this integration. To view the code and generate the plots displayed in this question, please feel free to check out this https://gist.github.com/anish-lakkapragada/9f39da7f072e88d98cd56517e0193ce6 I made.