Mathematical Foundation of DNN : HW 12

Jeong Min Lee

June 14, 2024

1

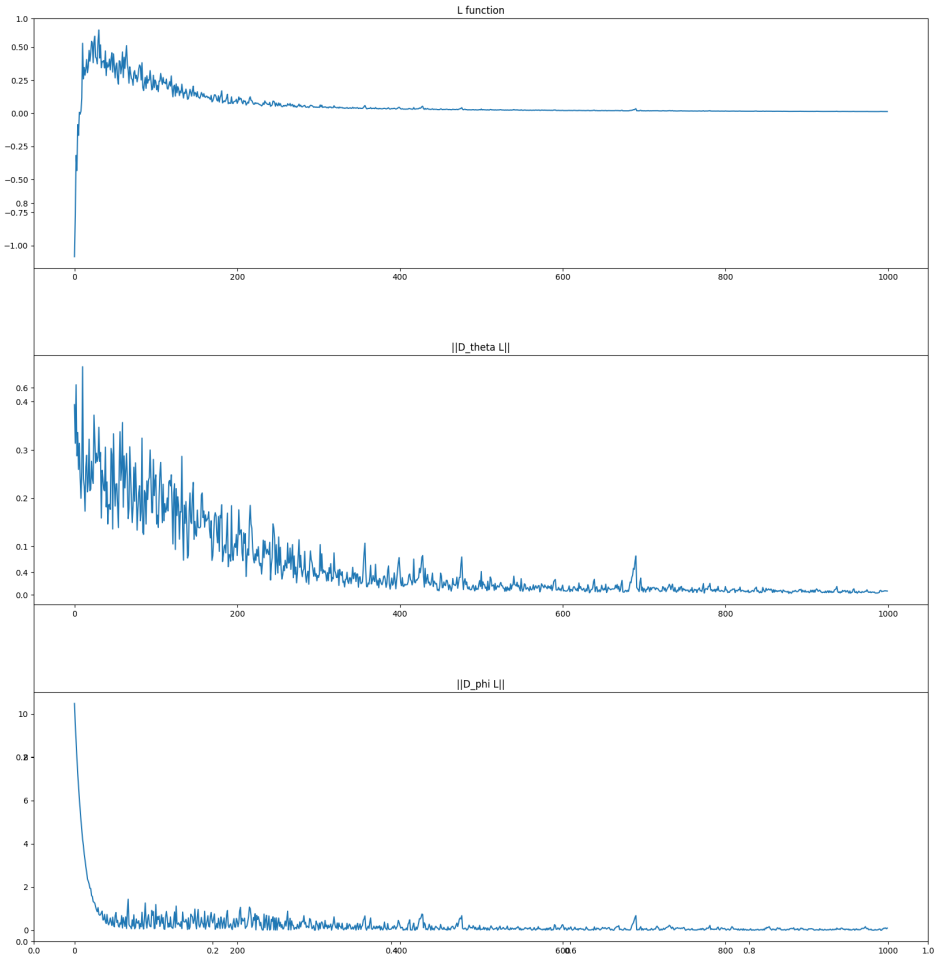


Figure 1: Loss and its gradient of the model's parameters.

As the Figure 1 shows, the loss and its gradient converge to the zero. This implies that the training of robust logistic model is successful. Note that the uneven and jittering curve is originated from the nature of SGD. Please, refer to the jupyter notebook file to see the implementation.

2

First of all, one have to verify that $f(u)$ is convex or not. To confirm it, it is necessary to find second derivative of $f(u)$ and its sign.

$$\frac{d^2}{du^2} f(u) = \frac{1}{u} - \frac{1}{u+\lambda} = \frac{\lambda}{u(u+\lambda)}$$

Since the domain of $f(u)$ is $\{u \in \mathbb{R} | u \geq 0\}$ and the regularization coefficient λ is positive, $f''(u) > 0$. This implies that $f(u)$ is convex function in its domain. Then, find the convex conjugate of $f(u)$, $f^*(t)$.

$$\begin{aligned} f^*(t) &= \sup_{u \in \mathbb{R}} \{tu - f(u)\} \\ &= \sup_{u \in \mathbb{R}} \left\{ tu - u \log \frac{u}{u+\lambda} - \lambda \log \frac{\lambda}{u+\lambda} - (1+\lambda)\log(1+\lambda) + \lambda \log \lambda \right\} \end{aligned}$$

To evaluate the $f^*(t)$ or the supremum of $tu - f(u)$, take first derivative of $tu - f(u)$ to find where the supremum is.

$$\frac{d}{du}tu - f(u) = t - f'(u)$$
$$t - \log u + \log(u + \lambda) = 0$$
$$\therefore u = \frac{\lambda}{e^{-t} - 1}$$

Thus,

$$\begin{aligned} f^*(t) &= \frac{t\lambda}{e^{-t} - 1} - \frac{\lambda}{e^{-t} - 1}\log e^t - \lambda \log(1 - e^t) - (1 + \lambda)\log(1 + \lambda) + \lambda \log \lambda \\ &= -\lambda \log(1 - e^t) - (1 + \lambda)\log(1 + \lambda) + \lambda \log \lambda \end{aligned}$$

Recall

$$D_f(p_{true}||p_\theta) = \sup_{T \in \mathcal{T}} \left\{ \mathbb{E}_{X \sim p_{true}}[T(X)] - \mathbb{E}_{\tilde{X} \sim p_\theta}[f^*(T(\tilde{X}))] \right\}$$

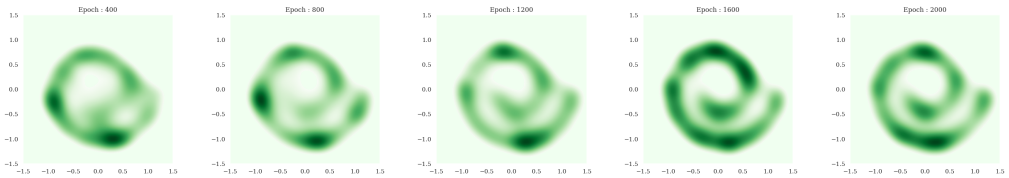
Although the bias term in $f^*(t)$ do affect to the maximum and minimum, it cannot affect to the gradient of the loss function and the optimum without taking the bias term into the consideration is same to that with bias term. This give us the background of deleting the bias term of $f^*(t)$. Therefore, the f-divergence is given as follow.

$$D_f(p_{true}||q) = \sup_{T \in \mathcal{T}} \left\{ \mathbb{E}_{X \sim p_{true}}[T(X)] + \lambda \mathbb{E}_{\tilde{X} \sim p_\theta} \left[\log(1 - e^{T(\tilde{X})}) \right] \right\}$$

By replacing $T(x) = \log D_\phi(x)$, one can observe that the given minmax problem is identical to GAN with non-uniform weights.

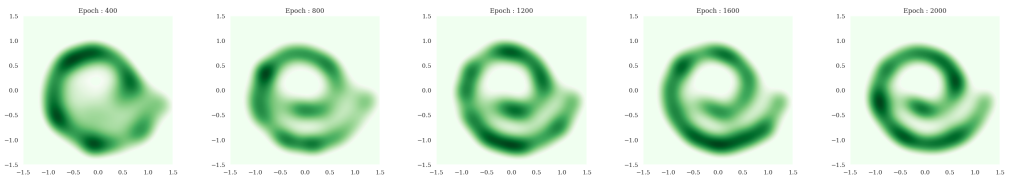
3,4

To implement the GAN and VAE, I refered to the example code in lecture. As Figure 2,3 depicts, the reconstruction of Swiss roll data is success, noting that the final data distribution is resemble to the that of original one.



(a) (b) (c) (d) (e)

Figure 2: (a) The reconstruction of Swiss roll data via VAE at 400 epochs.(b) The reconstruction of Swiss roll data via VAE at 800 epochs. (c) The reconstruction of Swiss roll data via VAE at 1200 epochs. (d) The reconstruction of Swiss roll data via VAE at 1600 epochs. (e) The reconstruction of Swiss roll data via VAE at 2000 epochs.



(a) (b) (c) (d) (e)

Figure 3: (a) The reconstruction of Swiss roll data via GAN at 400 epochs.(b) The reconstruction of Swiss roll data via GAN at 800 epochs. (c) The reconstruction of Swiss roll data via GAN at 1200 epochs. (d) The reconstruction of Swiss roll data via GAN at 1600 epochs. (e) The reconstruction of Swiss roll data via GAN at 2000 epochs.