How Does Batch Normalization Help Optimization?

논문리뷰

이도연

목차

1. Introduction

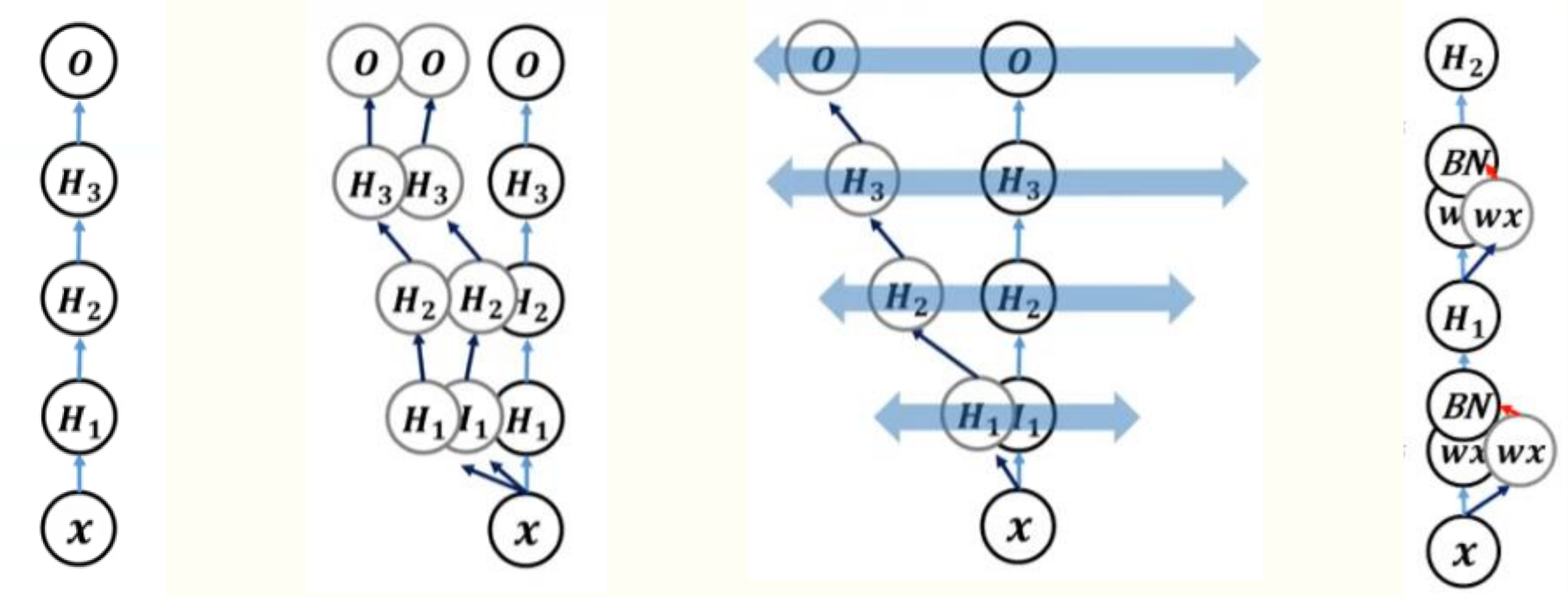
- Batch Normalization은 의심할 여지없이 유용
- DNN에서 더 빠르고 안정적인 학습을 가능하게 함
- 그렇지만 왜 잘 되는지에 대한 이해가 부족
- ICS(Internal Covariate Shift)를 감소시키는 효과가 있다고 알려져 있음

ICS를 줄이는 것은 성능과 관련이 없으며 Batch Normalization이 ICS를 감소시키지도 않는다

Batch Normalization의 진짜 효과는 Smoothing!!

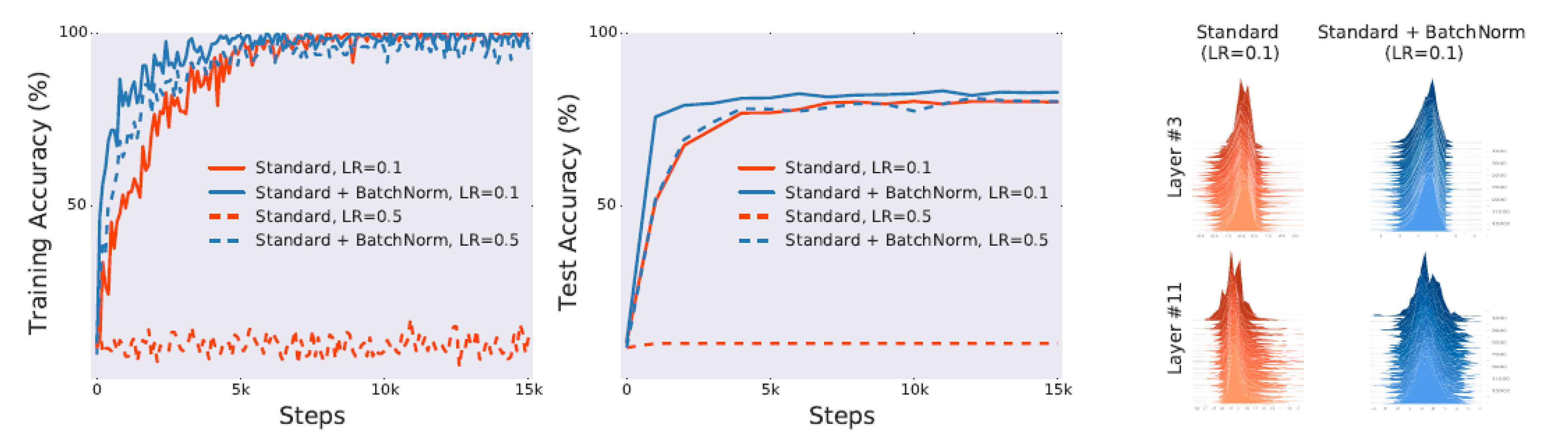
2. BatchNorm & ICS

Internal Covariate Shift



Input layer distribution의 변동

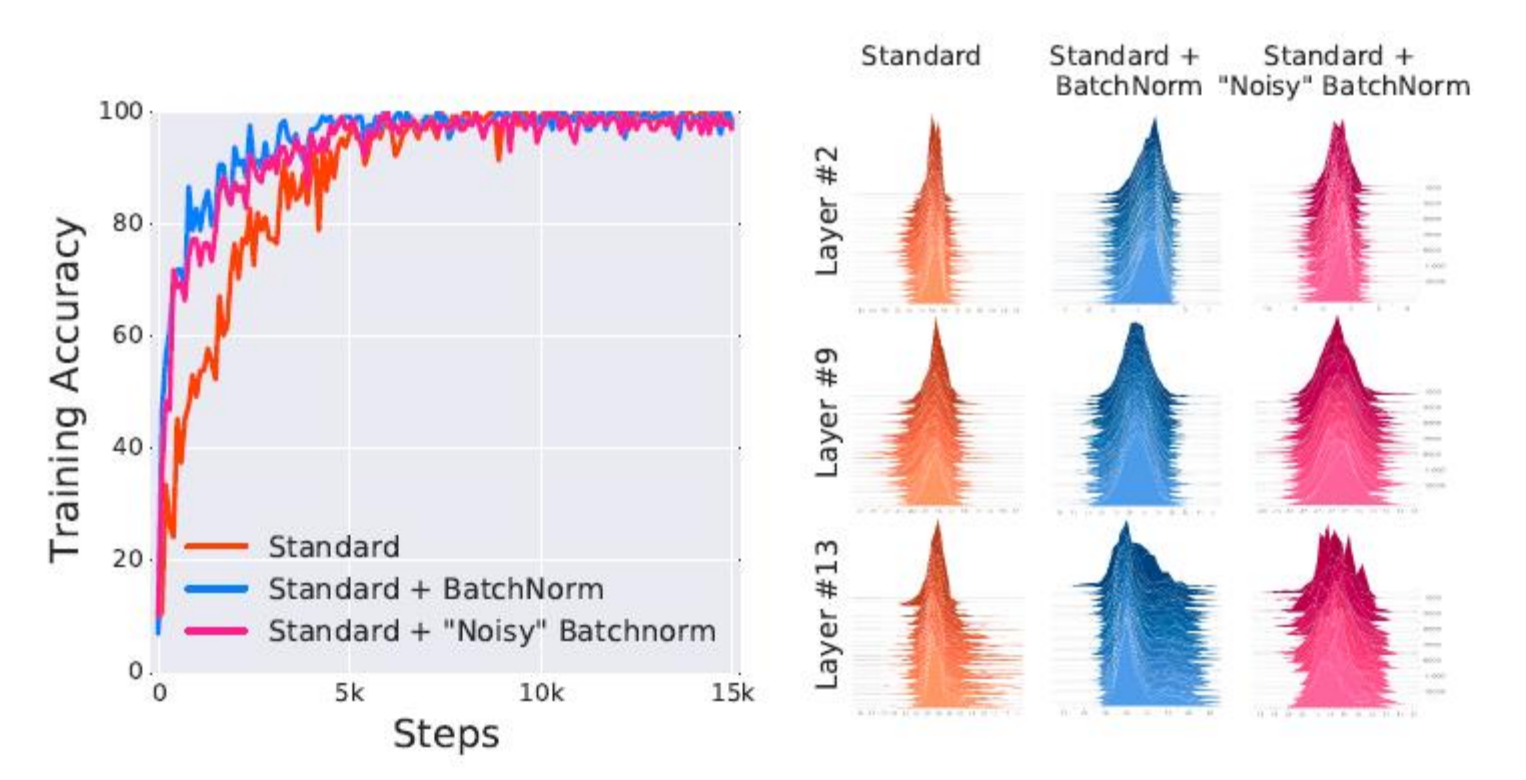
2. BatchNorm & ICS



Batch Normalization은 효과적이다

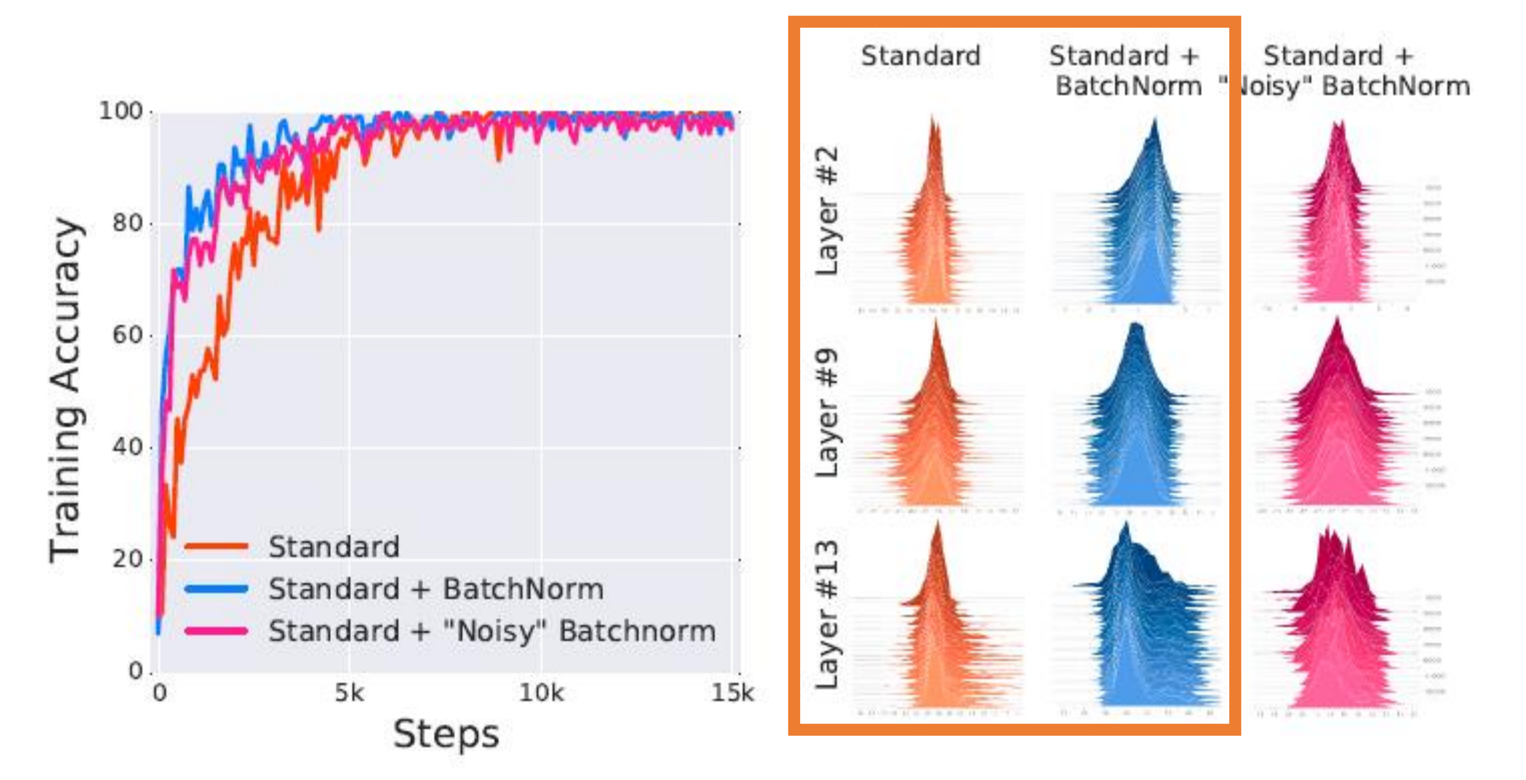
Input layer distribution의 결과는 명확하지 않다

2. BatchNorm & ICS



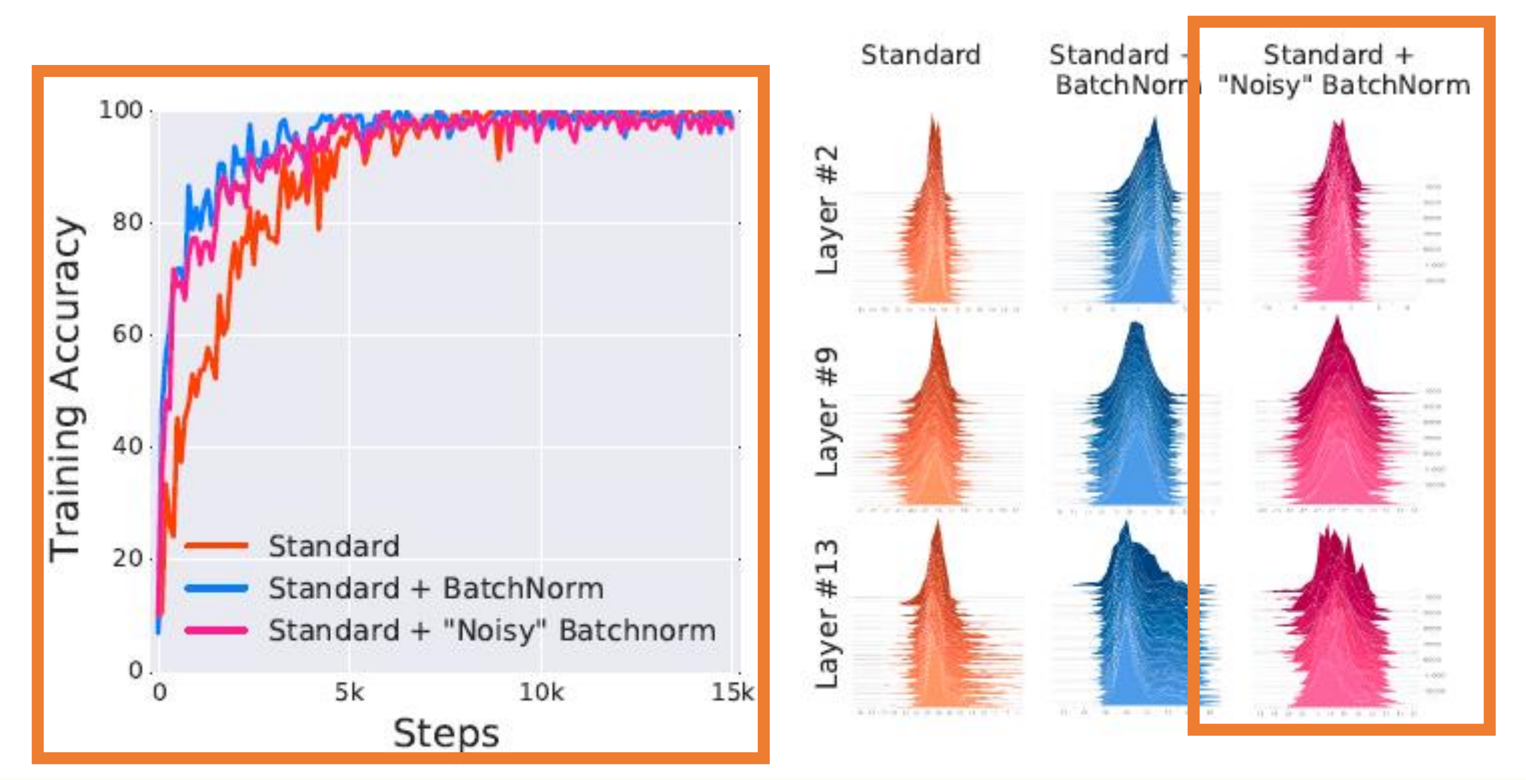
Noise를 삽입해 강제로 ICS를 발생시켜 보았다

2. BatchNorm & ICS



BatchNorm이 ICS를 감소시키나? 아닌 거 같은데?

2. BatchNorm & ICS



Noise를 삽입해 강제로 ICS를 발생해도 성능 개선은 명확

2. BatchNorm & ICS

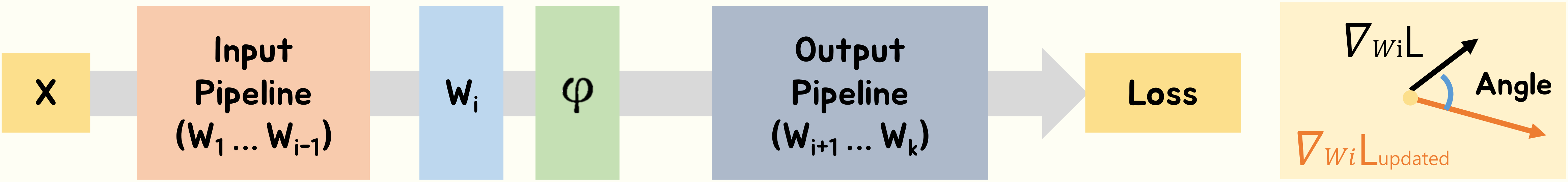
ICS를 계산하는 방법 제시

Definition 2.1. *Let $\mathcal{L}$ be the loss, $W_1^{(t)}, \ldots, W_k^{(t)}$ be the parameters of each of the k layers and $(x^{(t)}, y^{(t)})$ be the batch of input-label pairs used to train the network at time t. We define* internal covariate shift (ICS) *of activation i at time t to be the difference $||G_{t,i} - G'_{t,i}||_2$, where*

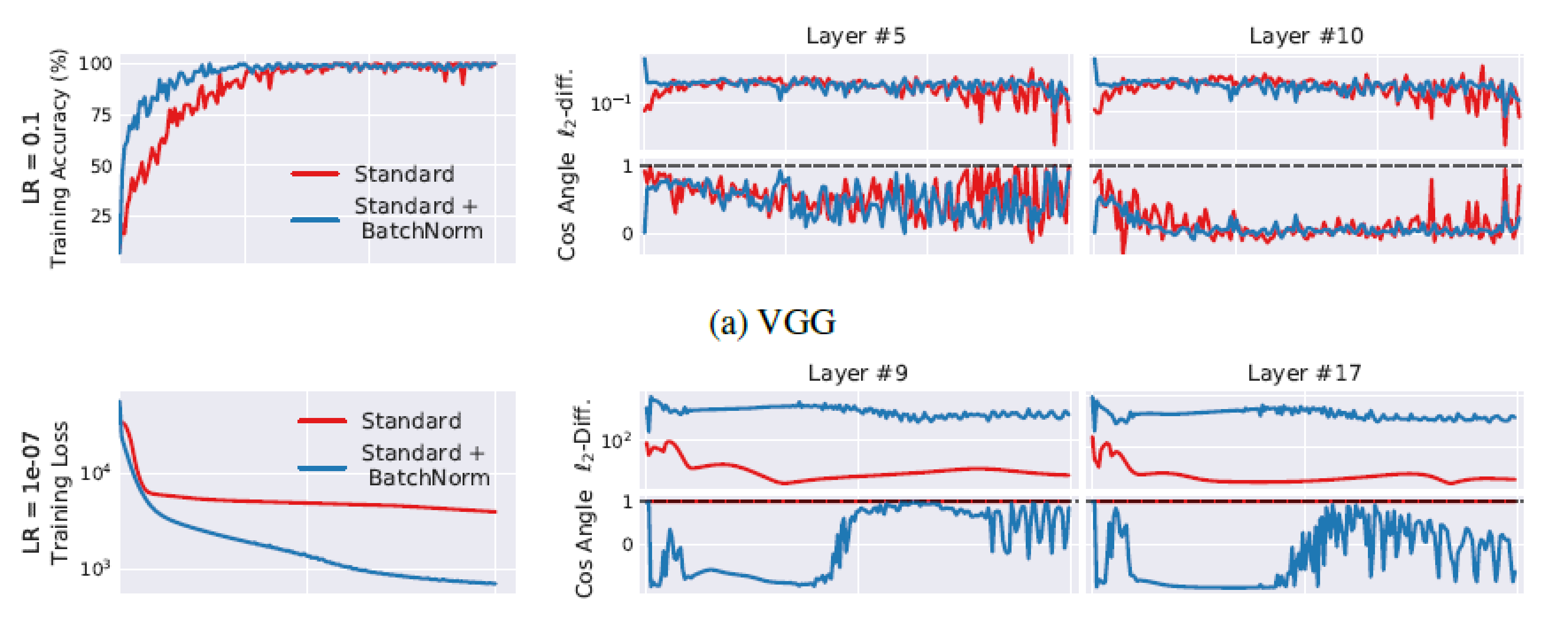
$$G_{t,i} = \nabla_{W_i^{(t)}} \mathcal{L}(W_1^{(t)}, \ldots, W_k^{(t)}; x^{(t)}, y^{(t)})$$

$$G'_{t,i} = \nabla_{W_i^{(t)}} \mathcal{L}(W_1^{(t+1)}, \ldots, W_{i-1}^{(t+1)}, W_i^{(t)}, W_{i+1}^{(t)}, \ldots, W_k^{(t)}; x^{(t)}, y^{(t)}).$$

앞 레이어만 업데이트한 상황에서 동일한 입력에 대해 기울기를 다시 계산



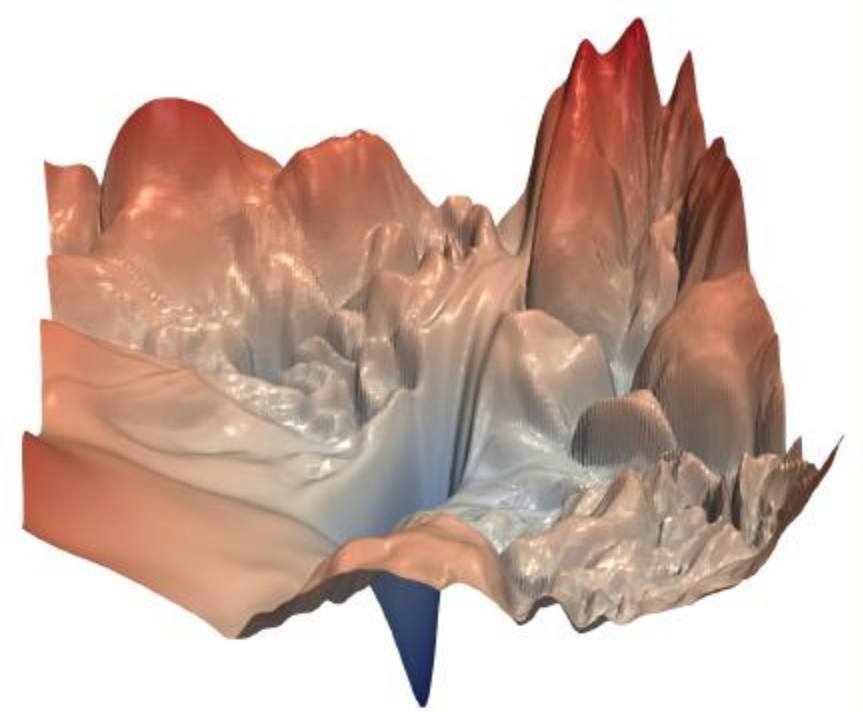
2. BatchNorm & ICS

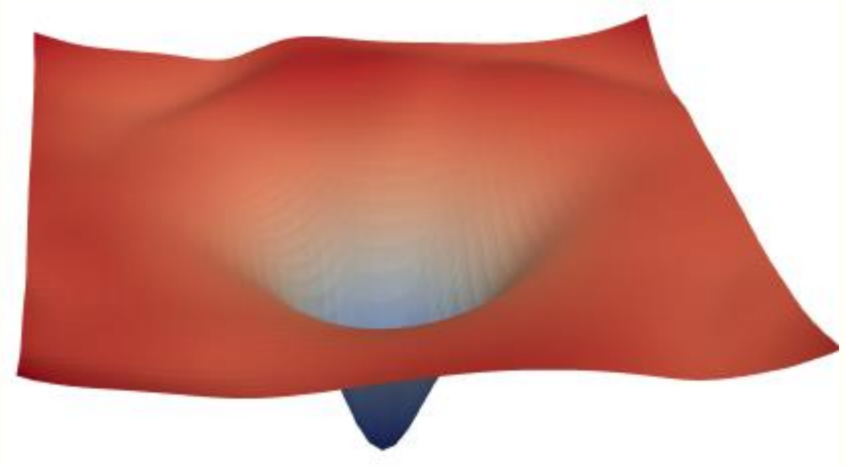


BatchNorm을 사용했을 때 별 차이가 없거나 오히려 ICS가 증가하기도 했다

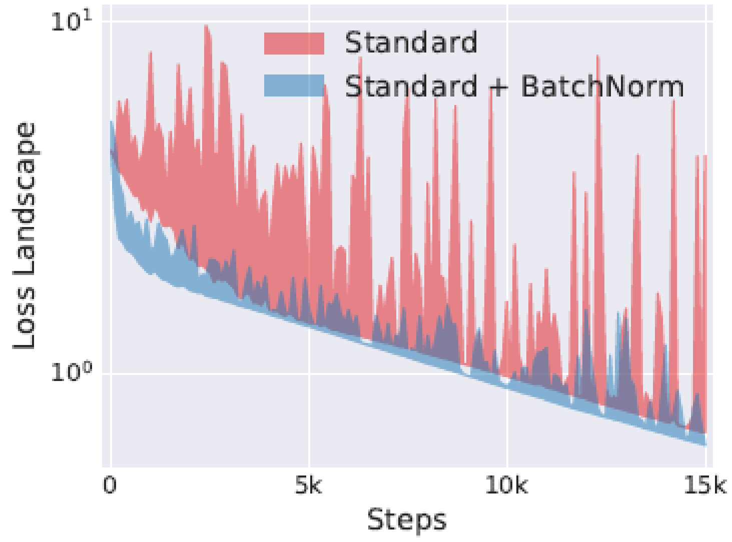
2. BatchNorm & ICS

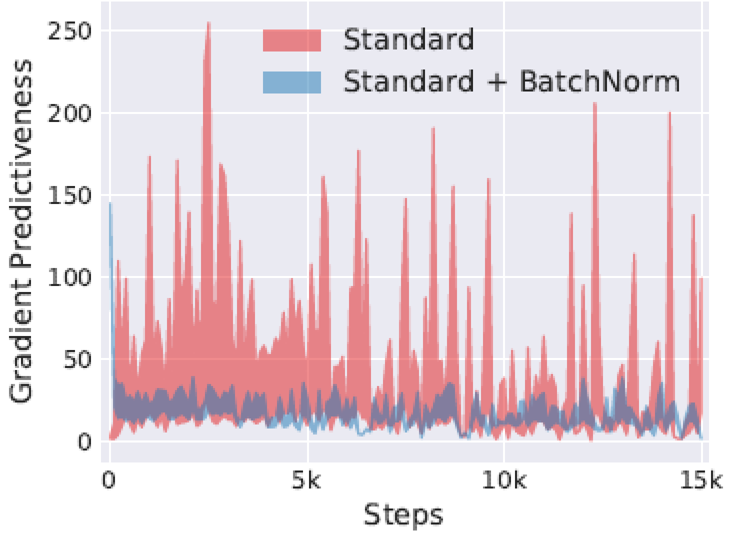
그렇다면 왜 잘되지?
바로, Optimization Landscape를 부드럽게 만드는 Smoothing 효과

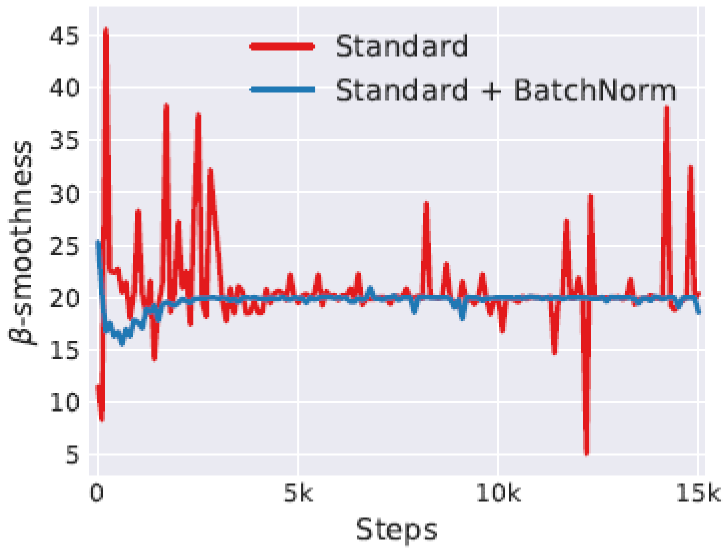




3. Why dose BatchNorm work?







$$\mathcal{L}(x + \eta \nabla \mathcal{L}(x))$$

Loss 값의 변화

$$||\nabla \mathcal{L}(x) - \nabla \mathcal{L}(x + \eta \nabla \mathcal{L}(x))||$$

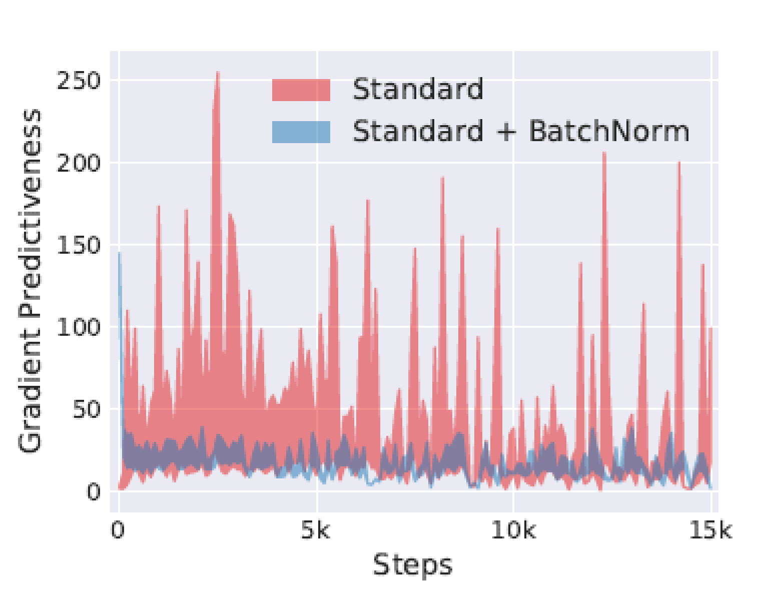
기울기 예측성

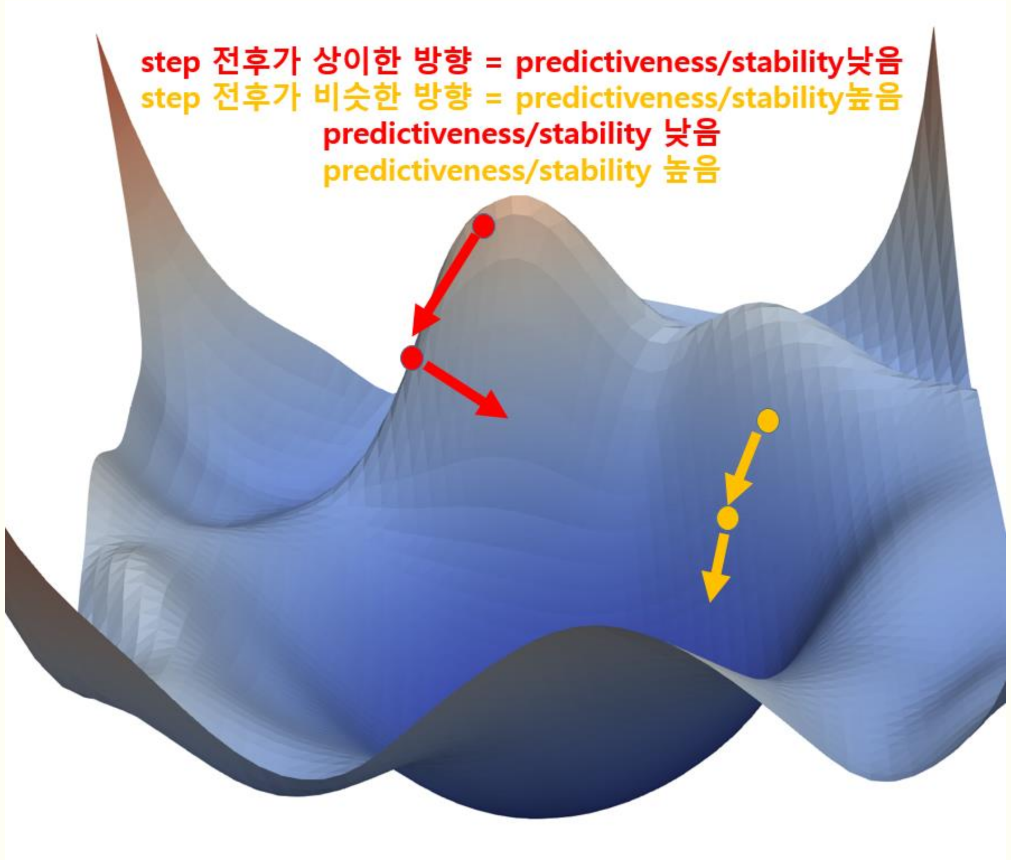
Loss Gradient의 변화

기울기 값 변화에 대한

Lipschitzness

3. Why dose BatchNorm work?

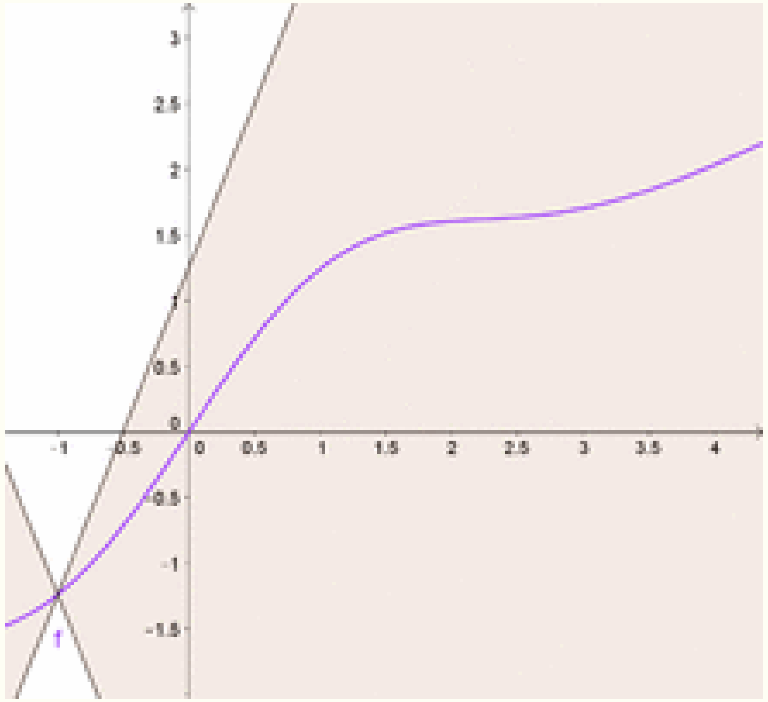




기울기 예측성
Loss Gradient의 변화

3. Why dose BatchNorm work?

립시츠 연속 함수 (Lipschitz-continuous function)

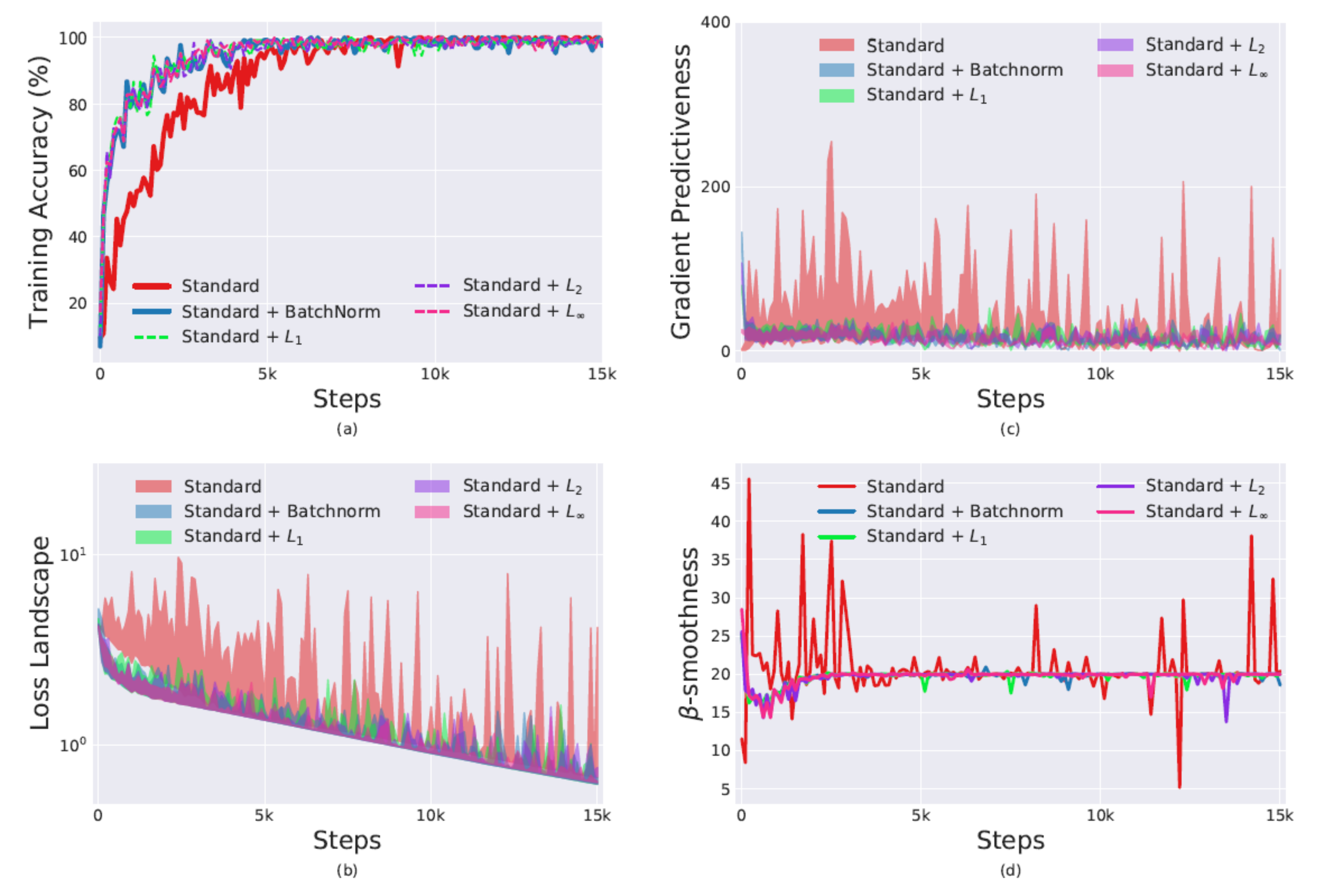


연속적이고, 미분 가능하며,
어떠한 두 점을 잡아도 기울기가 K보다 작은 함수

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|, \text{ for all } x_1 \text{ and } x_2$$

BatchNorm을 사용하면
Loss의 Lipschitzness를 향상시킨다.
즉, 안정적인 학습을 할 수 있다!

3. Why dose BatchNorm work?



$$L_1 = (\sum_i^n |x_i|) = |x_1| + |x_2| + |x_3| + \ldots + |x_n|$$

$$L_2 = \sqrt{\sum_i^n x_i^2} = \sqrt{x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2}$$

$$L_\infty = max(|x_1|, |x_2|, |x_3|, \ldots, |x_n|)$$

다른 Normalization 기법도
성능을 향상시킬 수 있다

+ Theoretical Analysis

Theorem 4.1 (The effect of BatchNorm on the Lipschitzness of the loss). *For a BatchNorm network with loss $\widehat{\mathcal{L}}$ and an identical non-BN network with (identical) loss $\mathcal{L}$,*

$$\left|\left|\nabla_{\boldsymbol{y_j}}\widehat{\mathcal{L}}\right|\right|^2 \leq \frac{\gamma^2}{\sigma_j^2}\left(\left|\left|\nabla_{\boldsymbol{y_j}}\mathcal{L}\right|\right|^2 - \frac{1}{m}\left\langle \boldsymbol{1}, \nabla_{\boldsymbol{y_j}}\mathcal{L}\right\rangle^2 - \frac{1}{m}\left\langle \nabla_{\boldsymbol{y_j}}\mathcal{L}, \hat{\boldsymbol{y}}_j\right\rangle^2\right).$$

Theorem 4.2 (The effect of BN to smoothness). *Let $\hat{\boldsymbol{g}}_{\boldsymbol{j}} = \nabla_{\boldsymbol{y_j}}\mathcal{L}$ and $\boldsymbol{H}_{jj} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{y_j}\partial \boldsymbol{y_j}}$ be the gradient and Hessian of the loss with respect to the layer outputs respectively. Then*

$$\left(\nabla_{\boldsymbol{y_j}}\widehat{\mathcal{L}}\right)^\top \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j}\left(\nabla_{\boldsymbol{y_j}}\widehat{\mathcal{L}}\right) \leq \frac{\gamma^2}{\sigma^2}\left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j}\right)^\top \boldsymbol{H}_{jj}\left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j}\right) - \frac{\gamma}{m\sigma^2}\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\rangle \left|\left|\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j}\right|\right|^2$$

If we also have that the $\boldsymbol{H}_{jj}$ preserves the relative norms of $\hat{\boldsymbol{g}}_{\boldsymbol{j}}$ and $\nabla_{\boldsymbol{y_j}}\widehat{\mathcal{L}}$,

$$\left(\nabla_{\boldsymbol{y_j}}\widehat{\mathcal{L}}\right)^\top \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j}\left(\nabla_{\boldsymbol{y_j}}\widehat{\mathcal{L}}\right) \leq \frac{\gamma^2}{\sigma^2}\left(\hat{\boldsymbol{g}}_j^\top \boldsymbol{H}_{jj}\hat{\boldsymbol{g}}_j - \frac{1}{m\gamma}\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\rangle \left|\left|\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j}\right|\right|^2\right)$$

Theorem 4.4 (Minimax bound on weight-space Lipschitzness). *For a BatchNorm network with loss $\widehat{\mathcal{L}}$ and an identical non-BN network (with identical loss $\mathcal{L}$), if*

$$g_j = \max_{||X|| \leq \lambda} ||\nabla_W \mathcal{L}||^2, \qquad \hat{g}_j = \max_{||X|| \leq \lambda}\left|\left|\nabla_W \widehat{\mathcal{L}}\right|\right|^2 \implies \hat{g}_j \leq \frac{\gamma^2}{\sigma_j^2}\left(g_j^2 - m\mu_{g_j}^2 - \lambda^2\left\langle \nabla_{y_j}\mathcal{L}, \hat{\boldsymbol{y}}_j\right\rangle^2\right).$$

Lemma 4.5 (BatchNorm leads to a favourable initialization). *Let $\boldsymbol{W}^*$ and $\widehat{\boldsymbol{W}}^*$ be the set of local optima for the weights in the normal and BN networks, respectively. For any initialization W_0*

$$\left|\left|W_0 - \widehat{W}^*\right|\right|^2 \leq ||W_0 - W^*||^2 - \frac{1}{||W^*||^2}\left(||W^*||^2 - \langle W^*, W_0\rangle\right)^2,$$

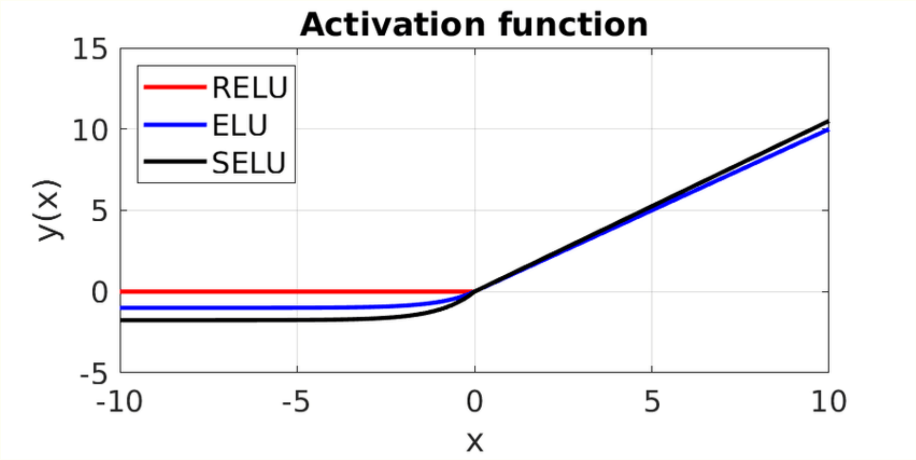
if $\langle W_0, W^*\rangle > 0$, where $\widehat{W}^*$ and W^* are closest optima for BN and standard network, respectively.

+ Related Work

- Batch Normalization의 대안으로 Layer 정규화, Batch Subset, 이미지 차원 등

- Weight Normalization은 Activation 대신 Weight를 정규화하는 보완 방식

- ELU, SELU를 Batch Normalization의 대안으로 사용할 수 있음



이외에도 몇몇 이야기가 더 있음.. 논문 참고..

4. Conclusion

- DNN에서 BatchNorm의 효과를 연구
- ICS(분포 안정성 관점)는 성능 향상에 대한 좋은 설명이 아니었다.
- BatchNorm & ICS 크게 관계 없다.
- BatchNorm은 Loss 관점에서 stable 하고 smooth 하게 optimization 한다.
- 이를 통해 예측 가능하고, 빠르고 효과적인 최적화가 진행된다.
- 몇몇 다른 normalization 방법들이 유사한 효과를 냈다.
- 추가적으로 이 논문은 Training 에서의 BN 효과에 집중했지만, BatchNrom이 Generalization을 향상시키는 경향이 있는 것 같다.

(특히 BN의 Smoothing Effect가 Training 과정에서 more flat minima에 수렴)

Q&A

참고

https://www.youtube.com/watch?v=TDx8iZHwFtM&t=619s PR-021: Batch Normalization

https://www.youtube.com/watch?v=58fuWVu5DVU&t=3289s 나동빈님의 배치 정규화

https://ml-dnn.tistory.com/6 How Does Batch Normalization Help Optimization? 논문 정리