TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Replacing the Loss Gradient

with the Margin Gradient

GANs

The generator tries to fool the discriminator.

$$\Phi^* = \operatorname*{argmax}_{\Phi} \; \min_{\Psi} \; E_{\langle i, y \rangle \sim \tilde{p}_{\Phi}} \; - \ln P_{\Psi}(i|y)$$

Assuming universality of both the generator p_{Φ} and the discriminator P_{Ψ} we have $p_{\Phi^*} = \mathrm{pop}$.

The Discrimator Tends to Win

The log loss for the binary discrimination classifier is quickly driven to very near zero.

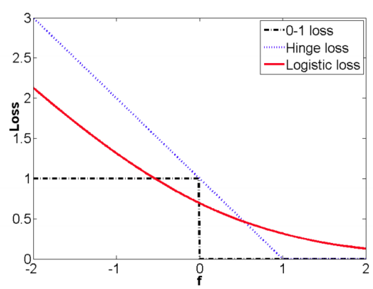
This causes the learning gradient to also become essentially zero and the learning stops.

Review of Binary Classification

In the case of binary classification cross-entropy loss becomes the log loss of the margin

$$
\begin{aligned}
\Psi^* &= \underset{\Psi}{\mathrm{argmin}}\;\; E_{(i,y)\sim\tilde{p}_\Phi}\;\; -\ln P_\Psi(i|y) \\
&= \underset{\Psi}{\mathrm{argmin}}\, E_{(i,y)\sim\tilde{p}_\Phi}\quad \ln(1+e^{-m})
\end{aligned}
$$

$$
m = 2is_\Psi(i|y) \quad \text{for} \quad \begin{aligned} s_\Psi(-1|y) &= -s_\Psi(1|y) \\ P_\Psi(i|y) &= \mathrm{softmax}_i\;\, s_\Psi(i|y) \end{aligned}
$$



Vanishing Gradients

For $i = 1$ and $y \sim \mathrm{pop}$:

$$\Psi \;\texttt{+=}\; \eta \, \frac{e^{-m}}{1+e^{-m}} \, \nabla_\Psi \, m \quad \approx 0 \text{ for } m >> 1$$

For $i = -1$ and $y \sim p_\Phi$:

$$\Psi \;\texttt{+=}\; \eta \, \frac{e^{-m}}{1+e^{-m}} \, \nabla_\Psi \, m \quad \approx 0 \text{ for } m >> 1$$

$$\Phi \;\texttt{-=}\; \eta \, \frac{e^{-m}}{1+e^{-m}} \, \nabla_\Phi \, m \quad \approx 0 \text{ for } m >> 1$$

The gradients vanish when the discriminator achieves large margins.

A Heuristic Patch

Replace

$$\Phi \mathrel{\texttt{-=}} \eta \, \frac{e^{-m}}{1+e^{-m}} \, \nabla_\Phi \, m \quad \approx 0 \text{ for } m >> 1$$

with

$$\Phi \mathrel{\texttt{-=}} \eta \, \nabla_\Phi \, m$$

This allows the generator to recover.

END