通信理論に特化した深層学習
第6回ゼミ資料
誤差逆伝播法

豊橋技術科学大学
電気・電子情報工学系
准教授　竹内啓悟

勾配の計算

損失関数を局所最小化するためには、順伝播型ネットワークに含まれるすべてのパラメータに関する損失関数の偏微分を効率的に計算する必要がある。

$$\Downarrow$$

「深層」合成関数のパラメータに関する偏微分を効率的に計算する問題

「深層」合成関数

$\boldsymbol{g}^t(\cdot,\boldsymbol{\theta}^t): \mathbb{R}^{N_{t-1}} \times \mathbb{R}^{N_t}$をパラメータ$\boldsymbol{\theta}^t$を持つ微分可能なベクトル値関数とし、

$$\boldsymbol{G}_{t_1}^{t_2}(\boldsymbol{z},\boldsymbol{\Theta}_{t_1}^{t_2}) = \boldsymbol{g}^{t_2}\circ \boldsymbol{g}^{t_2-1}\circ\cdots\circ\boldsymbol{g}^{t_1}(\boldsymbol{z},\boldsymbol{\theta}^{t_1}), \qquad \boldsymbol{\Theta}_{t_1}^{t_2}=\{\boldsymbol{\theta}^{t_1},\dots,\boldsymbol{\theta}^{t_2}\}.$$

注意

$N_T = 1$として、スカラー関数$g^T(\cdot,\boldsymbol{\theta}^T): \mathbb{R}^{N_{T-1}} \to \mathbb{R}$に損失関数を含める。

パラメータ$\boldsymbol{\Theta}_1^T$に関する$G_1^T(\boldsymbol{x},\boldsymbol{\Theta}_1^T)$のすべての偏微分を効率的に計算したい。

偏微分の計算

偏微分の連鎖測

t番目の関数のパラメータ$\boldsymbol{\theta}^t$に注目する。

$$\xrightarrow{\boldsymbol{x}} \boxed{\boldsymbol{G}_1^{t-1}(\cdot,\boldsymbol{\Theta}_1^{t-1})} \xrightarrow{\boldsymbol{z}^{t-1}} \boxed{\boldsymbol{g}^t(\cdot,\boldsymbol{\theta}^t)} \xrightarrow{\boldsymbol{z}^t} \boxed{G_{t+1}^T(\cdot,\boldsymbol{\Theta}_{t+1}^T)}$$

$$G_1^T(\boldsymbol{x},\boldsymbol{\Theta}_1^T) = G_{t+1}^T(\boldsymbol{g}^t(\boldsymbol{z}^{t-1},\boldsymbol{\theta}^t),\boldsymbol{\Theta}_{t+1}^T), \qquad \boldsymbol{z}^t = \boldsymbol{G}_1^t(\boldsymbol{x},\boldsymbol{\Theta}_1^t).$$

上記の表現の両辺を$\theta_i^t = [\boldsymbol{\theta}^t]_i$に関して偏微分すると、

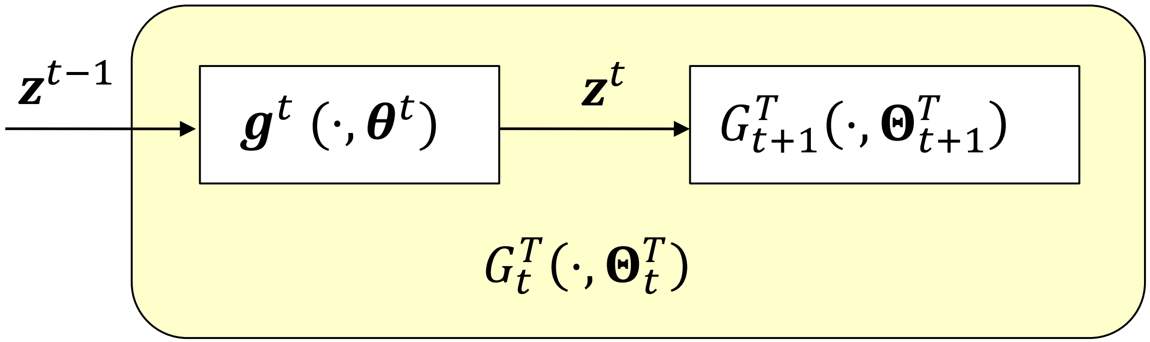
$$\frac{\partial}{\partial \theta_i^t} G_1^T(\boldsymbol{x},\boldsymbol{\Theta}_1^T) = \sum_{n=1}^{N_t} \partial_n G_{t+1}^T(\boldsymbol{z}^t,\boldsymbol{\Theta}_{t+1}^T)\,\frac{\partial g_n^t}{\partial \theta_i^t}(\boldsymbol{z}^{t-1},\boldsymbol{\theta}^t), \qquad g_n^t = [\boldsymbol{g}^t]_n.$$

後者の因子は順伝播時に計算可能

逆伝播

表記法

関数fのn番目の変数に関する偏微分を$\partial_n f$と書く。



$$G_{t}^{T}(\boldsymbol{z},\boldsymbol{\Theta}_{t}^{T}) = G_{t+1}^{T}(\boldsymbol{g}^{t}(\boldsymbol{z},\boldsymbol{\theta}^{t}),\boldsymbol{\Theta}_{t+1}^{T})$$

上記の表現の両辺を$z_n=[\boldsymbol{z}]_n$に関して点$\boldsymbol{z}=\boldsymbol{z}^{t-1}$で偏微分すると、

$$\partial_n G_{t}^{T}(\boldsymbol{z}^{t-1},\boldsymbol{\Theta}_{t}^{T}) = \sum_{n'=1}^{N_t} \partial_n g_{n'}^{t}(\boldsymbol{z}^{t-1},\boldsymbol{\theta}^{t})\,\partial_{n'} G_{t+1}^{T}(\boldsymbol{z}^{t},\boldsymbol{\Theta}_{t+1}^{T})$$

青色部分は逆伝播時に逐次的に計算可能

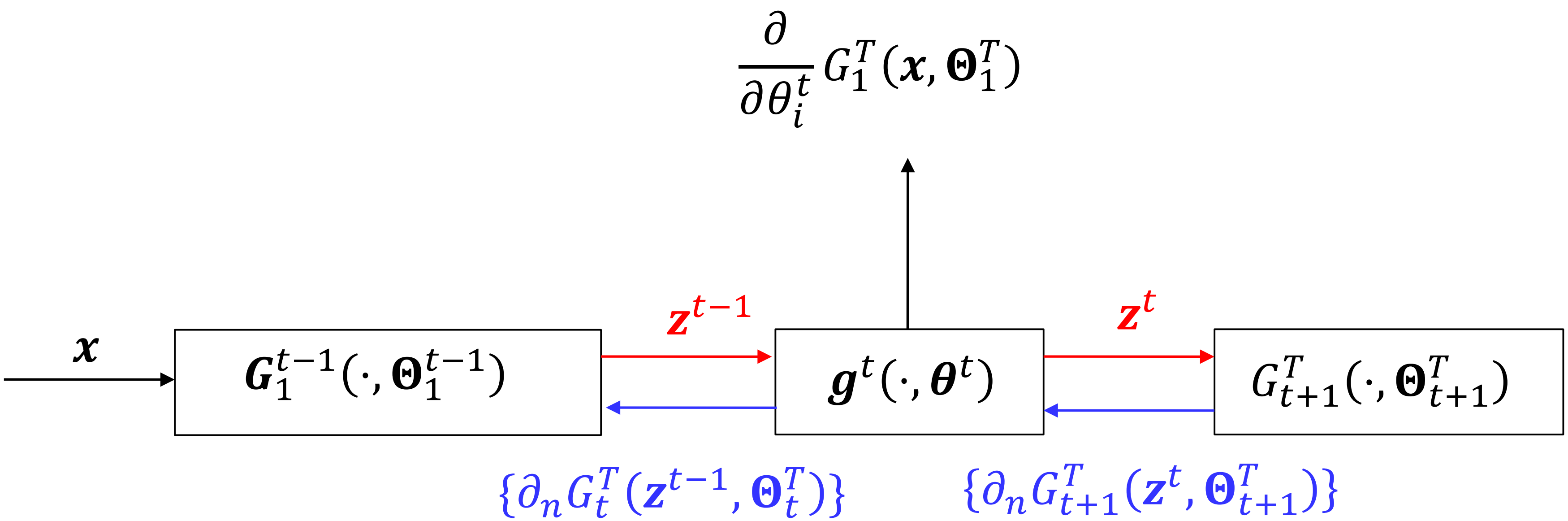
誤差逆伝播法（Back propagation）

手順1

順伝播により、$\{\boldsymbol{z}^t, \partial \boldsymbol{g}^t(\boldsymbol{z}^{t-1}, \boldsymbol{\theta}^t)/\partial \theta_i^t, \partial_n \boldsymbol{g}^t(\boldsymbol{z}^{t-1}, \boldsymbol{\theta}^t)\}$を計算する。

手順2

逆伝播により、出力側から勾配を順に計算する。

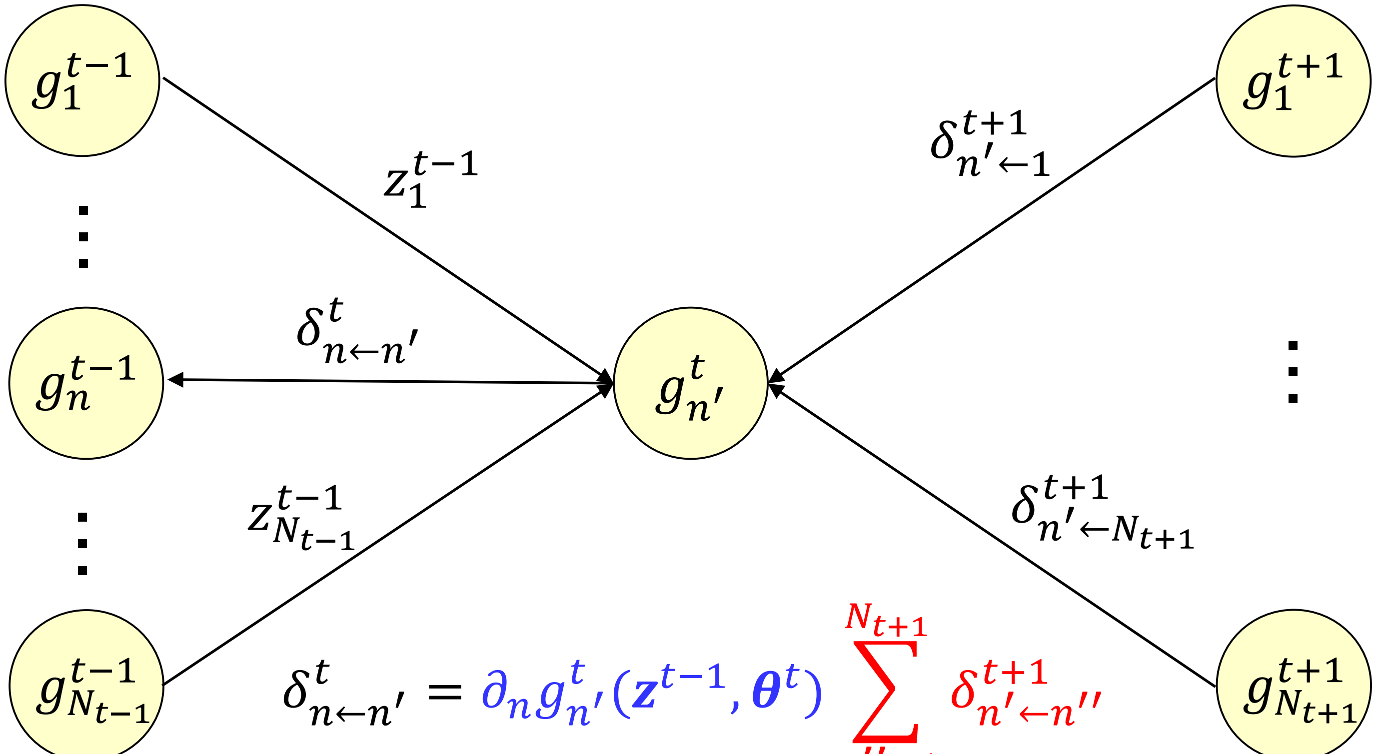


計算グラフ

計算結果が誤差逆伝播法による計算結果と等価になるように、
ノード間でやり取りされるメッセージを定義する。

関数g_n^tをt層n番目のノードに対応付ける。

$\delta_{n\leftarrow n'}^t$：$t$層$n'$番目のノードから$t-1$層$n$番目のノードに送られるメッセージ



$$\delta_{n\leftarrow n'}^t = \partial_n g_{n'}^t(\boldsymbol{z}^{t-1}, \boldsymbol{\theta}^t) \sum_{n''=1}^{N_{t+1}} \delta_{n'\leftarrow n''}^{t+1}$$

入力メッセージの和を取り、宛先ノードに関する偏微分を乗算する。

計算結果の等価性

誤差逆伝播法

$$\partial_n G_t^T(\mathbf{z}^{t-1}, \mathbf{\Theta}_t^T) = \sum_{n'=1}^{N_t} \partial_n g_{n'}^t(\mathbf{z}^{t-1}, \boldsymbol{\theta}^t)\partial_{n'} G_{t+1}^T(\mathbf{z}^t, \mathbf{\Theta}_{t+1}^T).$$

計算グラフ

$$\delta_{n\leftarrow n'}^t = \partial_n g_{n'}^t(\mathbf{z}^{t-1}, \boldsymbol{\theta}^t)\partial_{n'} G_{t+1}^T(\mathbf{z}^t, \mathbf{\Theta}_{t+1}^T).$$

帰納法による証明のスケッチ

$$\delta_{n\leftarrow n'}^{t-1} = \partial_n g_{n'}^{t-1}(\mathbf{z}^{t-2}, \boldsymbol{\theta}^{t-1}) \sum_{n''=1}^{N_t} \delta_{n'\leftarrow n''}^t$$

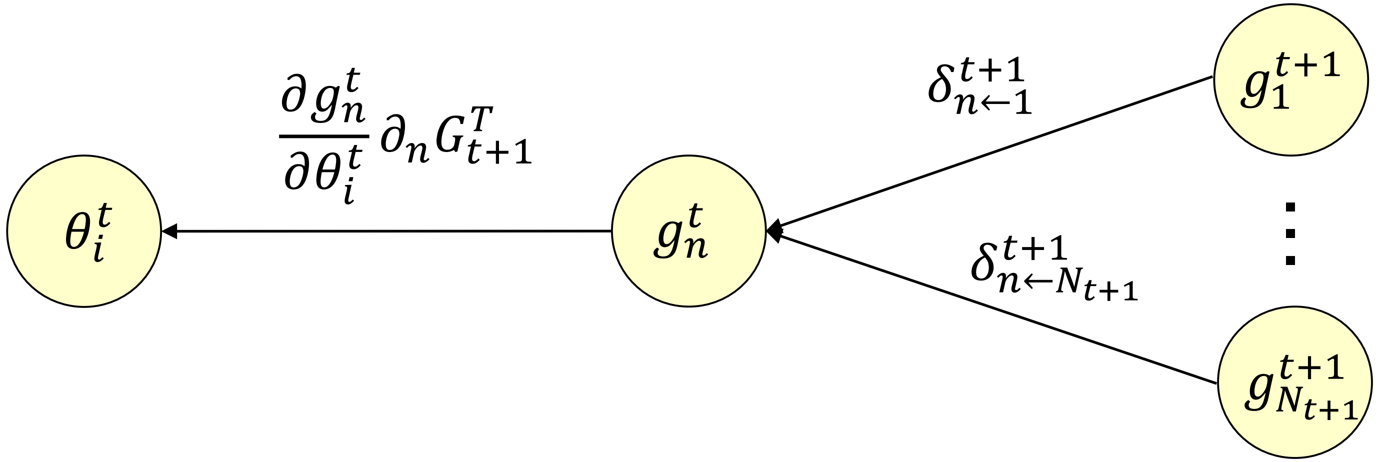
$$= \partial_n g_{n'}^{t-1}(\mathbf{z}^{t-2}, \boldsymbol{\theta}^{t-1})\partial_{n'} G_t^T(\mathbf{z}^{t-1}, \mathbf{\Theta}_t^T).$$

最後の等号は、帰納法の仮定と逆伝播の定義式から従う。 ■

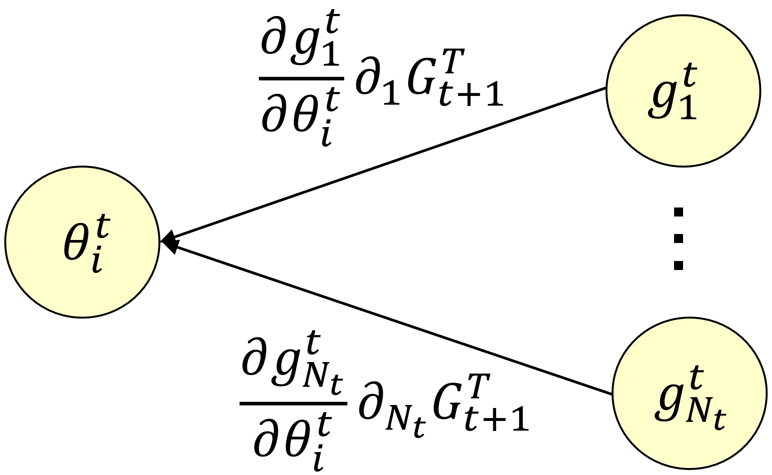
三つの添え字を持つメッセージ$\delta_{n\leftarrow n'}^t$（3階のテンソル）が計算グラフ上を流れる。

TensorFlowの名前の由来

計算グラフによる勾配計算



入力メッセージの和を取り、宛先ノードに関する偏微分をかける。



入力メッセージの和を取る。

$$\frac{\partial}{\partial \theta_i^t} G_1^T(\boldsymbol{x}, \boldsymbol{\Theta}_1^T) = \sum_{n=1}^{N_t} \partial_n G_{t+1}^T(\boldsymbol{z}^t, \boldsymbol{\Theta}_{t+1}^T) \frac{\partial g_n^t}{\partial \theta_i^t}(\boldsymbol{z}^{t-1}, \boldsymbol{\theta}^t).$$