

ニューラルネットワーク II

Python 中級
鈴木 敬彦

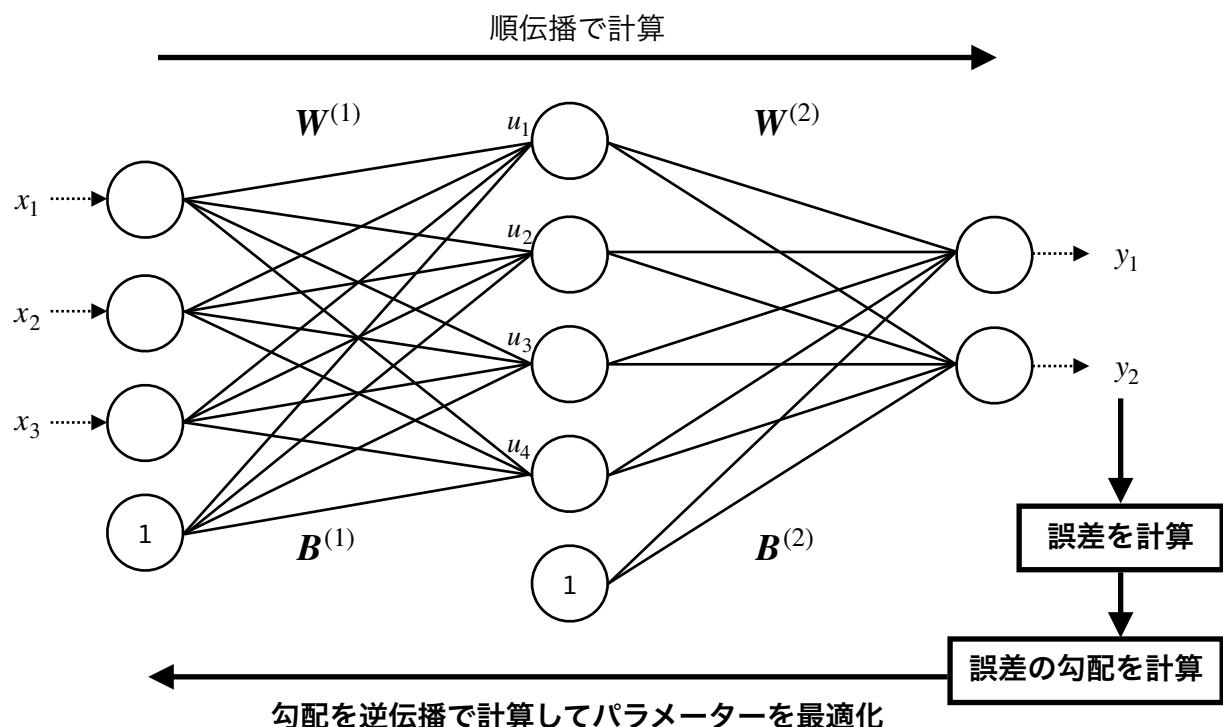
前回までに、MLP を例としてニューラルネットワークの構成と順伝播の計算を説明しました。今回は、ニューラルネットワークのウェイトとバイアスの最適化の手法（ニューラルネットワークに学習させる方法）について説明します。既に皆さんもご存知のように、ウェイトとバイアスが最適化されていないニューラルネットワークには何の意味もありません。ウェイトとバイアスが最適化されて初めて意味を持ちます。この最適化こそが最も重要で、最も難易度の高いパートです。

本稿では、MLP の時代から現代のディープニューラルネットワークに至るまでニューラルネットワークの最適化手法の定番として知られているバックプロパゲーション（Backpropagation、誤差逆伝播法）について説明します。かつては、勾配消失や勾配爆発の問題もあり、実用的とは言い難いものでしたが、その後の最適化手法を構成する諸要素の発展もあり、現在では誰でもDNNの最適化を行えるようになりました。

バックプロパゲーションを理解するにあたっては、微分についての理解が必須です。

バックプロパゲーション / Backpropagation

名前が表す通り、逆伝播させます。逆伝播させるものは誤差です。ニューラルネットワークに入力を与え順伝播の計算で得た出力と、正解であるデータ（理想とする出力。教師データ、教師信号とも呼びます。）との誤差を、ニューラルネットワークの文脈では損失（loss）と呼びます。ニューラルネットワークの出力の損失から、逆方向（出力層から入力層への方向）に沿って各層のウェイトとバイアスを最適化する方法をバックプロパゲーションと呼びます。

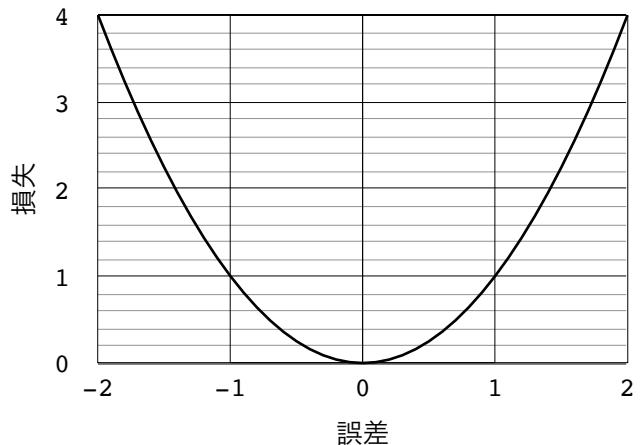


損失は、当たり前のことですが、正解から遠くなればなるほど大きくなります。ウェイトとバイアスの最適化とは、ウェイトとバイアスの誤差を可能な限り小さくすることです。

右の図は典型的な損失のグラフです。中央の誤差が 0 となる点はちょうど谷の底になり、その接線の傾きも 0 となります。中央から外れると接線は、誤差が正なら正の、負なら負の、傾きを持ちます。そして、その傾きは、誤差が大きくなればなるほど、大きくなっていきます。この傾きからネットワークパラメーターの最適化を行います。

ここでは簡便のため 2 次元で例示しました。この関数は 1 変数スカラー値関数で、この関数に

接する接線の傾きは、その関数の微分係数です。一方、ニューラルネットワークの計算では、入出力値の形状は最低でもベクトルで、損失は多くの場合スカラー値、つまり多変数スカラー値関数となります。この場合の「傾き」は勾配 (gradient) と呼ばれます。（例えば、3 次元の関数のグラフであれば、そのグラフは面になり、その関数に接する関数は接平面です。）入出力がこれより高階のテンソルであれば、ヤコビアン (Jacobian) と呼ばれます。勾配もヤコビアンも、傾きと同様に導関数から得ることができます。バックプロパゲーションでは、この損失の勾配を逆方向に伝播させ、ウェイトやバイアスの損失の勾配を計算してウェイトとバイアスを最適化していきます。



計算グラフ

ニューラルネットワークのバックプロパゲーションを説明する前に、連続する簡単な演算を例にバックプロパゲーションの基本的な概念を説明します。

$$y = f(x) = \sin(\ln((3x + 7)^3))$$

上の式を構成する演算はいずれも初等的な演算ばかりですが、いくつか組み合わさると複雑に見えます。これを演算毎に分解すると以下の様になります。中間出力は y_n としてあります。

$$y_1 = f_1(x) = 3x$$

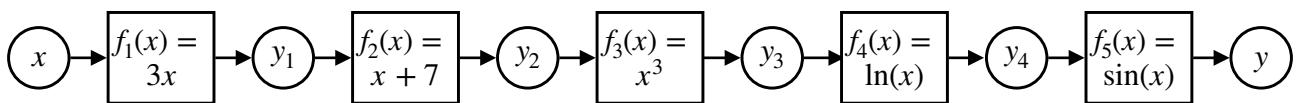
$$y_2 = f_2(y_1) = y_1 + 7$$

$$y_3 = f_3(y_2) = y_2^3$$

$$y_4 = f_4(y_3) = \ln(y_3)$$

$$y = f_5(y_4) = \sin(y_4)$$

上記の入出力と各関数をノード化してグラフで表すと以下の様に書くことができます。（各関数の入力は、便宜上、 x としてあります。）このように計算の流れをグラフとして表したもの計算グラフと呼びます。下記の例では、○は変数を、□は関数を表しています。



次に、先程の式 $y = f(x) = \sin(\ln((3x + 7)^3))$ の微分係数を得ることを考えます。もちろん、 $\frac{dy}{dx}$ を直接計算することができますが、その際に連鎖律を用いることは当然でしょうし、また、バックプロパレーションでは各層毎の勾配が必要となります。ここでは、連鎖律と計算グラフを用いて、各ノードの微分係数を得ます。

$\frac{dy}{dx}$ は、連鎖律から、以下の様に書けます。ここでは、逆伝播の入力を念頭に置いて $\frac{dy}{dy}$ も付け加えてあります。もちろん、常に $\frac{dy}{dy} = 1$ です。

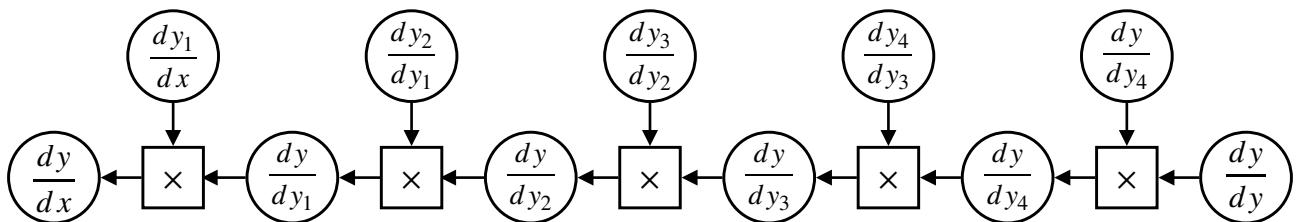
$$\frac{dy}{dx} = \frac{dy}{dy_4} \frac{dy_4}{dy_3} \frac{dy_3}{dy_2} \frac{dy_2}{dy_1} \frac{dy_1}{dx} = \frac{dy}{dy} \frac{dy}{dy_4} \frac{dy}{dy_3} \frac{dy}{dy_2} \frac{dy}{dy_1} \frac{dy}{dx}$$

上式を計算するにあたって、どこから計算を始めないといけないのかという規則はありませんが、バックプロパゲーションでは、後方（計算グラフの出力に近い側）から計算します。

$$\frac{dy}{dx} = \left(\left(\left(\left(\frac{dy}{dy_4} \frac{dy}{dy_3} \right) \frac{dy_4}{dy_3} \right) \frac{dy_3}{dy_2} \right) \frac{dy_2}{dy_1} \right) \frac{dy_1}{dx}$$

$\frac{dy}{dy_4} \frac{dy}{dy_3} = \frac{dy}{dy_4}$
 $\frac{dy}{dy_4} \frac{dy_4}{dy_3} = \frac{dy}{dy_3}$
 $\frac{dy}{dy_3} \frac{dy_3}{dy_2} = \frac{dy}{dy_2}$
 $\frac{dy}{dy_2} \frac{dy_2}{dy_1} = \frac{dy}{dy_1}$
 $\frac{dy}{dy_1} \frac{dy_1}{dx} = \frac{dy}{dx}$

これを計算グラフで表すと、以下の様になります。

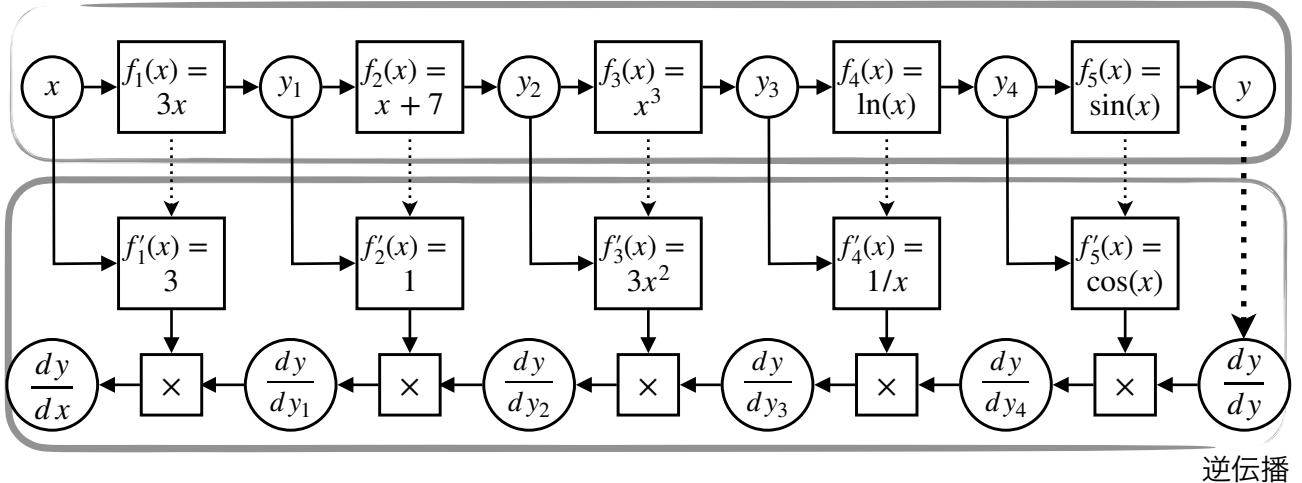


乗算 (\times) の入力は、 $\frac{dy}{dy}$ からの逆伝播と、順伝播での各関数の出力を入力で微分した $\frac{dy_n}{dy_{n-1}}$ です。

また、上の計算グラフはバックプロパゲーションの核心を示唆しています。各段の乗算の出力の分子は全て dy である事に注目してください。バックプロパゲーションでは、ネットワークの出力 y から損失 L を計算し、その勾配 $\frac{\partial L}{\partial y}$ を逆伝播させます。この時、上の計算グラフと同様に、各層に逆伝播していくものは損失の勾配です。

計算グラフの説明に戻ります。先程の $\frac{dy_n}{dy_{n-1}}$ は、 $y = f(x)$ を入力 x で微分したものなので、即ち $\frac{dy}{dx} = f'(x)$ です。これらから、順伝播と逆伝播の計算グラフをまとめると、以下の様になります。

順伝播



順伝播が正しいことは直感的にわかると思いますが、逆伝播が本当に正しいのか実際に計算してみましょう。皆さんも上記の計算グラフの計算手順を理解し確認するためにも、実際に自分で計算してみてください。

入力

$$x = 3$$

順伝播と逆伝播

関数	導関数	微分係数	逆伝播
$y_1 = 3x = 9$	$dy_1/dx = 3$	$= 3$	$dy/dx = -0.25162653354$
$y_2 = y_1 + 7 = 16$	$dy_2/dy_1 = 1$	$= 1$	$dy/dy_1 = -0.08387551118$
$y_3 = y_2^3 = 4096$	$dy_3/dy_2 = 3y_2^2 = 768$	$= 768$	$dy/dy_2 = -0.08387551118$
$y_4 = \ln(y_3) = 8.31776616671$	$dy_4/dy_3 = 1/y_3 = 0.00024414063$	$= 0.00024414063$	$dy/dy_3 = -0.00010921290$
$y = \sin(y_4) = 0.89436594845$	$dy/dy_4 = \cos(y_4) = -0.44733605962$		$dy/dy_4 = -0.44733605962$
			$dy/dy = 1$

直接計算

入力

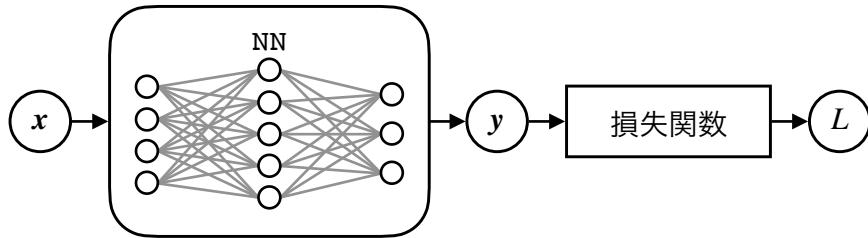
$x = x$	$= 3$	関数	$y = \sin(\ln((3x + 7)^3)) = 0.89436594845$
$y_1 = 3x$	$= 9$		$y = \sin(\ln((y_1 + 7)^3)) = 0.89436594845$
$y_2 = 3x + 7$	$= 16$		$y = \sin(\ln(y_2^3)) = 0.89436594845$
$y_3 = (3x + 7)^3$	$= 4096$		$y = \sin(\ln(y_3)) = 0.89436594845$
$y_4 = \ln((3x + 7)^3)$	$= 8.3177661667$		$y = \sin(y_4) = 0.89436594845$

導関数	微分係数
$dy/dx = 9 \cos(\ln((3x + 7)^3))/(3x + 7)$	$= -0.25162653354$
$dy/dy_1 = 3 \cos(\ln((y_1 + 7)^3))/(y_1 + 7)$	$= -0.08387551118$
$dy/dy_2 = 3 \cos(\ln(y_2^3))/y_2$	$= -0.08387551118$
$dy/dy_3 = \cos(\ln(y_3))/y_3$	$= -0.00010921290$
$dy/dy_4 = \cos(y_4)$	$= -0.44733605962$

ニューラルネットワークでのバックプロパゲーション

ニューラルネットワークでのバックプロパゲーションについて、もう少し細かく説明します。

先ず、ネットワークの出力 y から損失 L を計算（損失関数については後述）します。



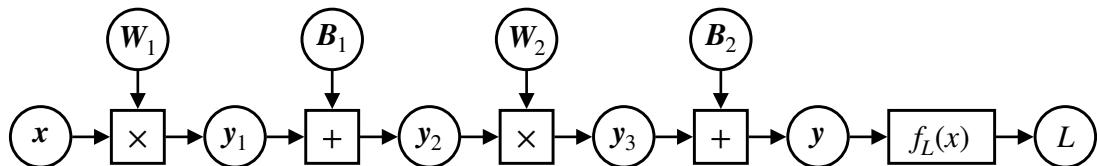
ニューラルネットワークの学習では、ここが折り返し点となります。ネットワークの逆伝播の入り口は出力層ですが、出力 y から損失 L を計算した過程も逆伝播させなくてはいけません。順伝播で出力層以降は、 y を入力とした損失関数から出力 L を得ています。この部分の逆伝播は、入力 y に対する L の勾配 $\frac{\partial L}{\partial y}$ で、これをネットワークの逆伝播の入力にします。

3 層の MLP を用いて、入力をベクトル x 、出力を y 、中間出力を y_n 、各層のウェイトとバイアスをそれぞれ W_n 、 B_n として、 W は入力の右から乗ずるものとすると、逆伝播は以下のようになります。損失関数は任意の微分可能な関数とします。

順伝播：

$$\begin{aligned}
 y_1 &= xW_1 & &= xW_1 \\
 y_2 &= y_1 + B_1 & &= y_1 + B_1 & &= xW_1 + B_1 \\
 y_3 &= y_2W_2 & = y_2W_2 & = (y_1 + B_1)W_2 & = (xW_1 + B_1)W_2 \\
 y &= y_3 + B_2 & = y_2W_2 + B_2 & = (y_1 + B_1)W_2 + B_2 & = (xW_1 + B_1)W_2 + B_2
 \end{aligned}$$

$$\begin{aligned}
 L &= f_L(y) \\
 &= f_L(y_3 + B_2) \\
 &= f_L(y_2W_2 + B_2) \\
 &= f_L((y_1 + B_1)W_2 + B_2) \\
 &= f_L((xW_1 + B_1)W_2 + B_2)
 \end{aligned}$$



逆伝播:

逆伝播では y_1, y_2, y_3, y に関する勾配だけではなく W_1, W_2 に関する勾配も計算します。ニューラルネットワークの順伝播の計算において W, B は共に定数ですが、逆伝播においてこれらは更新の対象であり変数です。逆伝播した損失の勾配から損失の W, B に関する勾配 $\frac{\partial L}{\partial W}, \frac{\partial L}{\partial B}$ を得て W, B の値を更新します。損失の連鎖律と各変数に関する勾配は以下の通りです。

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial y_3} \frac{\partial y_3}{\partial y_2} \frac{\partial y_2}{\partial y_1} \frac{\partial y_1}{\partial x}$$

$$\frac{\partial L}{\partial y} = f'_L(y)$$

$$\frac{\partial L}{\partial B_2} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial B_2} = \frac{\partial L}{\partial y} \frac{\partial}{\partial B_2} (y_3 + B_2) = \frac{\partial L}{\partial y}$$

$$\frac{\partial L}{\partial y_3} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial y_3} = \frac{\partial L}{\partial y} \frac{\partial}{\partial y_3} (y_3 + B_2) = \frac{\partial L}{\partial y}$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial y_3} \frac{\partial y_3}{\partial W_2} = \frac{\partial L}{\partial y_3} \frac{\partial y_3}{\partial W_2} = \frac{\partial L}{\partial y} \frac{\partial}{\partial W_2} y_2 W_2 = y_2^T \frac{\partial L}{\partial y}$$

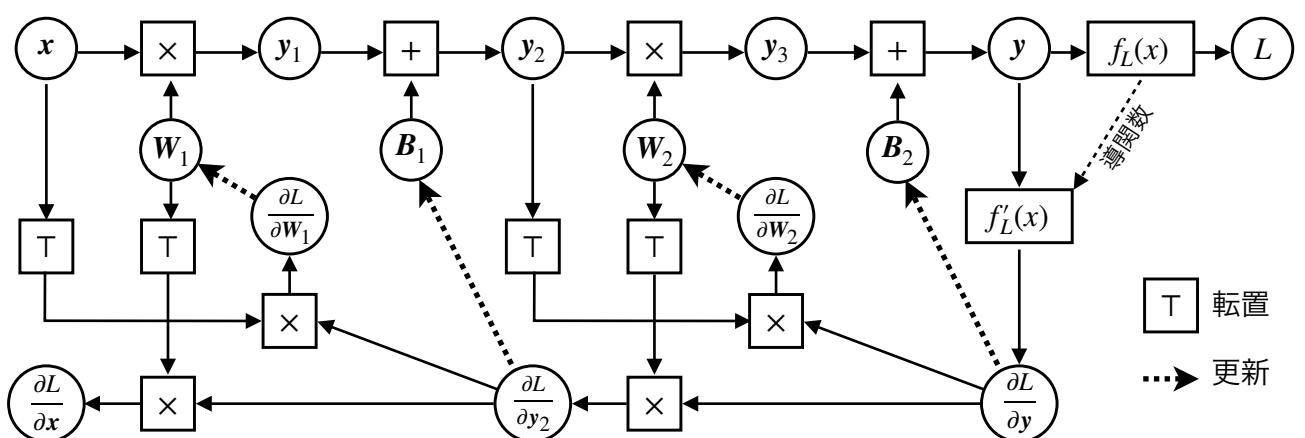
$$\frac{\partial L}{\partial y_2} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial y_3} \frac{\partial y_3}{\partial y_2} = \frac{\partial L}{\partial y_3} \frac{\partial y_3}{\partial y_2} = \frac{\partial L}{\partial y} \frac{\partial}{\partial y_2} y_2 W_2 = \frac{\partial L}{\partial y} W_2^T$$

$$\frac{\partial L}{\partial B_1} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial y_3} \frac{\partial y_3}{\partial y_2} \frac{\partial y_2}{\partial B_1} = \frac{\partial L}{\partial y_2} \frac{\partial y_2}{\partial B_1} = \frac{\partial L}{\partial y_2} \frac{\partial}{\partial B_1} (y_1 + B_1) = \frac{\partial L}{\partial y_2}$$

$$\frac{\partial L}{\partial y_1} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial y_3} \frac{\partial y_3}{\partial y_2} \frac{\partial y_2}{\partial y_1} = \frac{\partial L}{\partial y_2} \frac{\partial y_2}{\partial y_1} = \frac{\partial L}{\partial y_2} \frac{\partial}{\partial y_1} (y_1 + B_1) = \frac{\partial L}{\partial y_2}$$

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial y_3} \frac{\partial y_3}{\partial y_2} \frac{\partial y_2}{\partial y_1} \frac{\partial y_1}{\partial W_1} = \frac{\partial L}{\partial y_2} \frac{\partial y_1}{\partial W_1} = \frac{\partial L}{\partial y_2} \frac{\partial}{\partial W_1} x W_1 = x^T \frac{\partial L}{\partial y_2}$$

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial y_3} \frac{\partial y_3}{\partial y_2} \frac{\partial y_2}{\partial y_1} \frac{\partial y_1}{\partial x} = \frac{\partial L}{\partial y_2} \frac{\partial y_1}{\partial x} = \frac{\partial L}{\partial y_2} \frac{\partial}{\partial x} x W_1 = \frac{\partial L}{\partial y_2} W_1^T$$



バックプロパゲーションの勾配の計算とグラフは上記のようになります。

また、見ての通り、 $\frac{\partial L}{\partial B_2}, \frac{\partial L}{\partial y_3}, \frac{\partial L}{\partial B_1}, \frac{\partial L}{\partial y_1}, \frac{\partial L}{\partial x}$ の計算は不要です。

成分毎の計算や途中の計算を含めた例は冗長となるため、巻末 (p.14) に記載します。また、ここでの例に活性化関数は含まれていませんが、順伝播の計算に活性化関数が含まれていれば、当然、逆伝播の計算にも活性化関数の微分が含まれます。この計算は簡単なので、皆さんもぜひ挑戦してみてください。

損失関数 / Loss Function

誤差関数、コスト関数、目的関数とも呼ばれます。前述の通り、ニューラルネットワークの順伝播の出力と教師信号の誤差を損失と呼び、損失を算出する関数を損失関数と呼びます。様々な損失関数が提案されていますが、ここでは定番の損失関数ともいべき平均二乗誤差と、フーバー損失を紹介します。損失がスカラ一値であることに注意してください。

平均二乗誤差 / Mean Squared Error

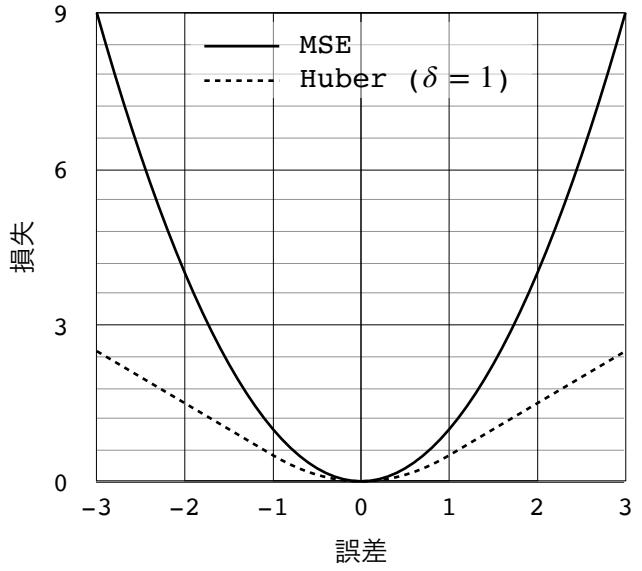
$$L = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

n はデータ数、 y_i は教師信号 y の i 番目の成分、 \hat{y}_i はニューラルネットワークの出力 \hat{y} の i 番目の成分を意味します。

フーバー損失 / Huber Loss

$$L_\delta(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq \delta, \\ \delta(|x| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases}$$

$$L = \frac{1}{n} \sum_{i=1}^n L_\delta(y_i - \hat{y}_i)$$



フーバー損失は平均二乗誤差と比較して、誤差が大きくても発散しづらいといわれ、深層学習でも多く用いられています。

ウェイトとバイアスの最適化

ウェイトやバイアスの更新には、前述の通り、損失のウェイトやバイアスに関する勾配を用います。この方法は最急降下法 (Gradient Descent) と呼ばれていて、以下の式でウェイトやバイアスを更新します。

$$\theta \leftarrow \theta - \alpha \nabla_\theta L(\theta)$$

ここで θ はウェイトやバイアス等のニューラルネットワークのパラメーターを \leftarrow は更新を、 $\nabla_\theta L(\theta)$ は損失の θ に関する勾配を表します。 α は学習率 (learning rate) もしくはステップサイズ (step size) と呼ばれるハイパーパラメーターで、その範囲は一般に 0.0~1.0 です。ハ

イパーコンピューターとは、ニューラルネットワークの層の数や層毎のノードの数を始めとした人間が決定するニューラルネットワークのパラメーターのことです。

先程の式を、今までと同じ記法で行列と成分毎の表記で表すと、以下のようになります。

$$w_{j,k}^{(i)} \leftarrow w_{j,k}^{(i)} - \alpha \left(\frac{\partial L}{\partial w^{(i)}} \right)_{j,k}, \quad w^{(i)} \leftarrow w^{(i)} - \alpha \frac{\partial L}{\partial w^{(i)}}, \quad w^{(i)} \leftarrow w^{(i)} - \alpha \nabla_{w^{(i)}} L(w^{(i)})$$

i 番目の層のウェイト $W^{(i)}$ の j, k 成分 $w_{j,k}^{(i)}$ を、現在の値 $w_{j,k}^{(i)}$ から 損失 L の $W^{(i)}$ に関する勾配に学習率 α を乗じた値を減じた値で更新します。

様々な入力（訓練データ）でこの更新を繰り返し、ニューラルネットワークのパラメーターを最適化します。全てのニューラルネットワークのパラメーターの損失が最小値に収束した状態が理想的な状態です。

適切な訓練データを準備することは当然としても、ハイパーコンピューターの選択が適切でなければ、ニューラルネットワーク最適化に成功する可能性は低くなります。例えば、学習率一つをとっても、ニューラルネットワーク最適化の成否に大きく影響します。仮に、現在の誤差と損失が右図の点の位置であるとします。勾配は負の値なので、ウェイトを更新すれば谷方向に更新されます。しかし、もしその時の学習率が大き過ぎたり小さ過ぎたら、どのような事が起きるでしょうか。

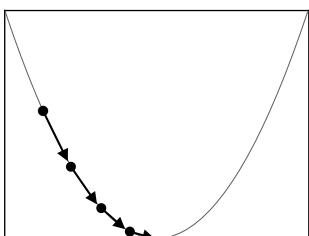
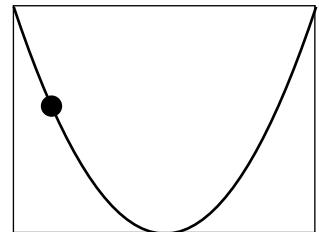


図 1: 理想

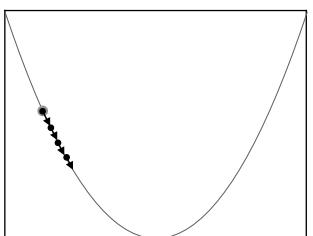


図 2: 小さい

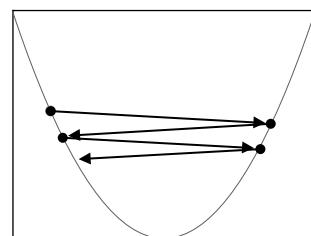


図 3: 大きい

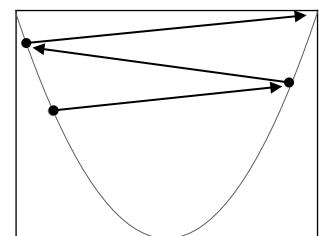


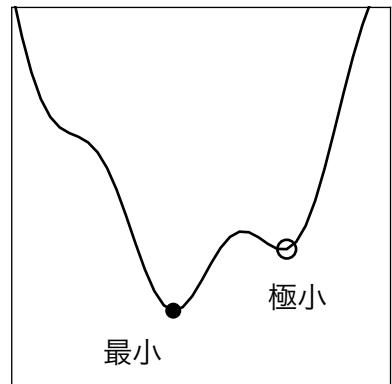
図 4: 大き過ぎ

徐々に損失が 0 となる地点に近づく（図 1）のが理想的な学習の進行です。しかし、学習率が小さければ、学習が中々進行しない状況（図 2）となります。この様な状況は、一般に「学習が遅い」と呼ばれます。逆に学習率が大きくても同様の状況（図 3）に陥ってしまいます。更に、学習率が大きすぎると損失が最小となる地点から離れていく（図 4）しまいます。特に、最後の（図 4）のような状況は、当該パラメーターが大きく（或いは小さく）なり続け、最終的にはオーバーフローしてしまいます。この問題は「勾配爆発問題 (exploding gradient problem)¹」と呼ばれています。

また、ニューラルネットワークの最急降下法を用いた学習時の問題はこれだけではなく、他にも広く知られている問題として極小 (local minimum) の問題が挙げられます。これまでの誤差-損失曲線

¹ 似た様な名称で「勾配消失問題 (vanishing gradient problem)」という問題もあります。これは、入力層に近づくにつれ勾配が小さくなり過ぎてしまい、学習を制御できない状況に陥る問題です。（主な原因是学習率ではなく、活性化関数にあります。）

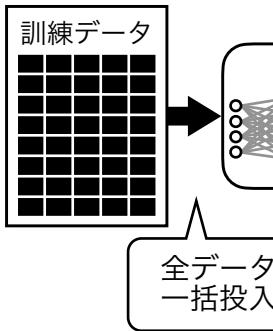
失のグラフの例では、簡便のため、単純な二次曲線を用いてきましたが、実際の誤差-損失の曲線（曲面）がどのような形状なのかはわかりません。最適化を進めて我々が到達したい点は、誤差が最小 (global minimum) となる点です。しかし、局所的な最小値、つまり極小値 (= 勾配が 0)、に留まってしまってうまく学習を進められない状況も起こり得ます。（ニューラルネットワークの文脈では、最小ではない極小値を局所解、最小値を最適解と呼ぶこともあります。）



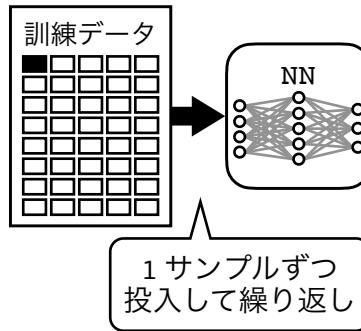
確率的勾配降下法 / Stochastic Gradient Descent

確率的勾配降下法のウェイトの更新式は、基本的に最急降下法 (Gradient Descent) の更新式と同じです。しかし、更新を行うタイミングや更新に用いるサンプル数等が異なってきます。最急降下法はバッチ勾配降下法と呼ばれることもあります。この場合の「バッチ (batch)」とは、1回のニューラルネットワークのパラメーター更新に全ての訓練データを用いることを意味し、下記のミニバッチに対してフルバッチと呼ばれることもあります。対する確率的勾配降下法では1回のパラメーター更新に訓練データ1サンプルを使用して、これを繰り返します。また、中間的な位置付けでミニバッチ (minibatch) 勾配降下法と呼ばれる手法もあります。ミニバッチは、全訓練データの一部を更新に用いることを意味していて、確率的勾配降下法と直接的な関係はありません。ミニバッチ勾配降下法は、1回のパラメーター更新にミニバッチを用いて勾配降下法を適用します。

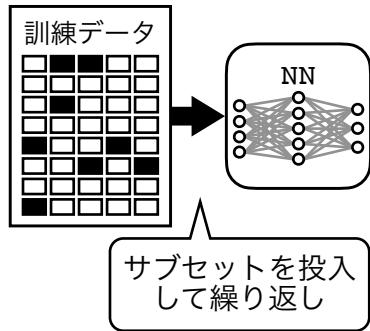
最急降下法



確率的勾配降下法



ミニバッチ勾配降下法



訓練データの全サンプル数を N 、1回の更新に使用する訓練データのサンプル数（バッチサイズ）を m 、1エポック当たりの更新回数（イテレーション数、ステップ、バッチ数）は以下の様になります。1エポックは、通算で N 件の訓練データを処理する期間を表します。

	最急降下法	確率的勾配降下法	ミニバッチ
バッチサイズ	N	1	m
イテレーション/エポック	1	N	N/m

確率的勾配降下法やミニバッチ勾配降下法では、一般的に、数エポック以上の更新を行います。また、最急降下法のように予測と最適化が分離されている学習法を「オンライン学習」と呼び（バッチ学習、フルバッチと呼ばれることも。）、確率的勾配降下法のように予測と最適化が交互に行われるタイプの学習法を「オンライン学習」と呼びます。

バッチ、シングルバッチ、ミニバッチには、それぞれ以下の様なメリットとデメリットがあります。

バッチ：

- 全ての訓練データを使って更新するため、ノイズや振動が少なく学習の進行が安定している。
- 並列化によって効率的に処理できる。
- ✗ ノイズや振動が少ないため極小から抜け出せない。
- ✗ 訓練データが多いため、大量の記憶領域を必要とする。
- ✗ 訓練データが多いため、計算に時間が掛かる。

シングルバッチ：

- 大規模な訓練データでは、頻繁な更新で、より早い収束を期待できる
- ノイズや振動が多いため、極小から抜け出せる。
- 訓練データが 1 つなので、少ない記憶領域で済む。
- 訓練データが 1 つなので、計算に時間が掛からない。
- ✗ ノイズや振動が多いため、収束するまでに時間が掛かる。
- ✗ ノイズや振動が多いため、異なる方向へ向かう可能性がある。
- ✗ 並列化できない。

ミニバッチ：

- 平均化により、安定した誤差勾配と収束が得られる。
- 極小に陥っても、ノイズの多いステップで抜け出せる。
- 比較的少ない記憶領域で済む。
- 並列化によって効率的に処理できる。

このようにミニバッチは、バッチとシングルバッチの中間でバランスのとれた最適化の方法だと言えます。現に、現在では、ミニバッチによる勾配降下法が最適化手法の主流となっています。

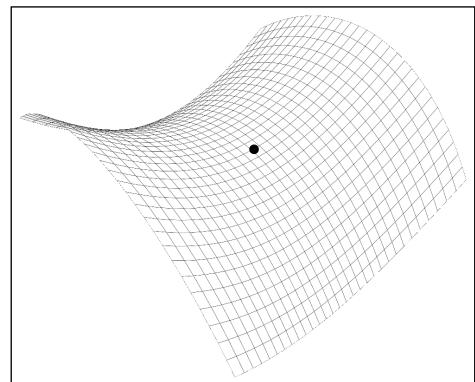
ミニバッチのバッチサイズは、一般に 32 から 512 程度です。バッチサイズの選択によって、それぞれ上記の様なバッチとシングルバッチの特性が現れてきます。また、バッチサイズは他のハイパーパラメーターにも影響を及ぼします。例えば、バッチサイズが小さくなればイテレーション数は大きくなるため、学習率も小さくすべきです。

最適化における問題点は他にも過学習と言った基本的な問題から *Pathological Curvature* と言った問題まで様々ですが、それらに対する最適化の技術も発展し、様々な最適化手法や工夫が提案されています。

Pathological Curvature を敢えて訳すなら「病的な曲率」となるでしょう。連續性を欠いた曲線等を意味します。誤差-損失平面に幅の狭い急峻な谷がある場合に最小値へ向かう途中で振動

してしまう問題です。この問題はニュートン法を用いると早く且つ正確に解決することは知られていますが、二階微分を必要し、その計算量は最急降下法、確率的勾配降下法の計算量の二乗となるため、現時点では未だ実用化されていません。

次に挙げる最適化手法はいずれも Pathological Curvature を抑制するために確率的勾配降下法を基として開発された手法です。



鞍点の例 (グラフは $z = x^2 - y^2$)

モーメンタム / Momentum²

確率的勾配降下法と並んで非常によく用いられる手法の一つです。momentum は、日本語で運動量や勢いを意味します。モーメンタムでは現在の勾配だけではなく過去のステップの勾配も含めて進むべき方向を決定し、振動を抑制します。パラメーターの更新式は以下の通りです。

$$\begin{cases} \Delta w \leftarrow \eta \Delta w - \alpha \frac{\partial L}{\partial w} \\ w \leftarrow w + \Delta w \end{cases}$$

ここで、 Δw の Δ は単に w の差分と言う意味でしかありません。 η はモーメンタム係数と呼ばれ、0.0~1.0 の範囲の値を取ります。 α は以前と同じく学習率です。各パラメーターの一般的な値は $\alpha = 0.1$ 、 $\eta = 1 - \alpha = 0.9$ です。

まず、1 行目の式の意味について考えてみましょう。 η が 0 であれば 2 行目の式は以前と変わりありませんが、 η がそれ以外の値を取った場合にどうなるかを見てみましょう。ここでは簡便のため n 回目の更新の Δw と勾配をそれぞれ Δw_n 、 G_n と書きます。

$$\begin{aligned} n = 0: \quad & \Delta w_0 = 0 \\ n = 1: \quad & \Delta w_1 = \eta \Delta w_0 - \alpha G_1 = -\alpha G_1 \\ n = 2: \quad & \Delta w_2 = \eta \Delta w_1 - \alpha G_2 = -\eta \alpha G_1 - \alpha G_2 = -\alpha(\eta G_1 + G_2) \\ n = 3: \quad & \Delta w_3 = \eta \Delta w_2 - \alpha G_3 = \eta[-\alpha(\eta G_1 + G_2)] - \alpha G_3 = -\alpha(\eta^2 G_1 + \eta G_2 + G_3) \\ \therefore \quad & \Delta w_n = -\alpha(\eta^{n-1} G_1 + \eta^{n-2} G_2 + \cdots + \eta^1 G_{n-1} + \eta^0 G_n) = -\alpha \sum_{i=1}^n \eta^{n-i} G_i \end{aligned}$$

この様に Δw は過去の勾配に影響を受け続けます。古い勾配ほど影響は小さく、新しい勾配ほど影響は大きくなります。一般に指数平均と呼ばれています。これまでの式をまとめると以下のようになります。（ n 回目の更新とします。）

$$w_{n+1} = w_n - \alpha \left(\frac{\partial L}{\partial w_n} + \sum_{i=1}^{n-1} \eta^{n-i} \frac{\partial L}{\partial w_i} \right)$$

² Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". Nature. 323 (6088): 533–536.

元の勾配降下の式との違いは、勾配に \sum の項が追加されているだけです。 \sum の項は指数平均で、前述の通り、過去の勾配にステップ毎に重みを掛け総和を取ったもので、即ち、過去の勾配のモーメンタムです。モーメンタムはこれまでの勾配が指示してきた進むべき方向を表すもので、Pathological Curvature に出会ったとしても、振動を抑制し最小値へと向かいます。ただし、オーバーシュート (overshoot、最小値を通り越してしまう事) してしまう事もあるため、学習率の調整等が必要です。（具体的にはアニーリングと呼ばれる手法を使います。）

因みに $\eta = 1 - \alpha$ となる場合は、一般的に走行平均、移動平均 (running average、moving average) と呼ばれる平均値となります。これは、音響処理や映像処理において信号の平滑化に用いられる手法と同じものです。音響においては所謂ローパスフィルター (low-pass filter) であり、映像であれば動体検出に用いる背景映像となります。

RMSProp³

前出の Momentum とは異なる方法で振動を抑制する手法です。この手法は特に論文等では発表されておらず、Geoffrey Hinton が彼の講義で提案した手法で、近年の深層学習ブームの切掛ともなった論文⁴でも用いられている手法です。RMS は Root Mean Square (二乗平均平方根) の、Prop は propagation の略です。パラメーターの更新式は以下の通りです。

$$\begin{cases} \nu \leftarrow \rho\nu + (1 - \rho)\left(\frac{\partial L}{\partial w}\right)^2 \\ w \leftarrow w - \frac{\alpha}{\sqrt{\nu + \epsilon}} \frac{\partial L}{\partial w} \end{cases}$$

ρ は0.0~1.0 の範囲の値を取る忘却率 (forgetting factor) と呼ばれる係数で、所謂、減衰率です。 ϵ はゼロ除算を回避するための措置で、勾配に大きく影響しない程度の小さな値を設定します。各パラメーターの一般的な値は $\alpha = 0.01$ 、 $\rho = 0.9$ 、 $\epsilon = 1e-7 \sim 1e-10$ 程度です。

1 行目の式は、勾配が二乗であることを除けば、モーメンタムの 1 行目の式と同じです。つまり、勾配の二乗の移動平均です。

$$\begin{aligned} n = 0: \quad & \nu_0 = 0 \\ n = 1: \quad & \nu_1 = \rho\nu_0 + (1 - \rho)G_1^2 = (1 - \rho)G_1^2 \\ n = 2: \quad & \nu_2 = \rho\nu_1 + (1 - \rho)G_2^2 = (1 - \rho)(\rho G_1^2 + G_2^2) \\ n = 3: \quad & \nu_3 = \rho\nu_2 + (1 - \rho)G_3^2 = (1 - \rho)(\rho^2 G_1^2 + \rho G_2^2 + G_3^2) \\ \therefore \quad & \nu_n = (1 - \rho)(\rho^{n-1} G_1^2 + \rho^{n-2} G_2^2 + \dots + \rho^1 G_{n-1}^2 + \rho^0 G_n^2) = (1 - \rho) \sum_{i=1}^n \rho^{n-i} G_i^2 \end{aligned}$$

³ Hinton, G., Srivastava, N., & Swersky, K., 2012, "Neural Networks for Machine Learning"
http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

⁴ Volodymyr Mnih, et.al, 2013, "Playing Atari with Deep Reinforcement Learning", [arXiv:1312.5602 \[cs.LG\]](https://arxiv.org/abs/1312.5602)
 Volodymyr Mnih, et.al, 2015, "Human-level control through deep reinforcement learning", [Nature](https://www.nature.com/articles/nature14236)

2 行目の式は、モーメンタムの式とは大きく異なります。RMSProp では、勾配の二乗の移動平均の平方根（二乗平均平方根）で学習率を制御し振動の抑制を行います。勾配の絶対値が大きい時は学習率が小さくなり、勾配の絶対値が小さい時は学習率が大きくなります。そのため、モーメンタムよりもオーバーシュートが少なくなります。

Adam⁵

Adaptive Moment Estimation の略で、モーメンタムと RMSProp を合わせた方法です。パラメーターの更新式は以下の通りです。

$$\left\{ \begin{array}{l} \mu \leftarrow \beta_1 \mu + (1 - \beta_1) \frac{\partial L}{\partial w} \\ \nu \leftarrow \beta_2 \nu + (1 - \beta_2) \left(\frac{\partial L}{\partial w} \right)^2 \\ \hat{\mu} = \frac{\mu}{1 - \beta_1} \\ \hat{\nu} = \frac{\nu}{1 - \beta_2} \\ w \leftarrow w - \alpha \frac{\hat{\mu}}{\sqrt{\hat{\nu}} + \epsilon} \end{array} \right.$$

μ は勾配の、 ν は勾配の二乗の移動平均です。 β_1 、 β_2 はそれぞれの忘却率です。各パラメーターの一般的な値は $\alpha = 0.001$ 、 $\beta_1 = 0.9$ 、 $\beta_2 = 0.999$ です。

その他

他にも数多くの最適化手法があります。ぜひ、色々と調べて試してみてください。

AdaGrad、AdaDelta、アニーリング等々…

⁵ Diederik P. Kingma, Jimmy Ba, 2015, "Adam: A Method for Stochastic Optimization", [arXiv:1412.6980 \[cs.LG\]](https://arxiv.org/abs/1412.6980)

バックプロパゲーション 計算例

3 層の MLP で、入力 \mathbf{x} を $1 \times p$ の行ベクトル、出力 \mathbf{y} を $1 \times r$ の行ベクトル、中間出力を $\mathbf{y}^{(n)}$ として、各層のウェイトとバイアスをそれぞれ $\mathbf{W}^{(n)}$ 、 $\mathbf{B}^{(n)}$ として、 \mathbf{W} は入力の右から乗ずるものとします。活性化関数は省略します。

順伝播

$$\begin{aligned}\mathbf{y}^{(1)} &= \mathbf{x} \mathbf{W}^{(1)} \\ \mathbf{y}^{(2)} &= \mathbf{y}^{(1)} + \mathbf{B}^{(1)} = \mathbf{x} \mathbf{W}^{(1)} + \mathbf{B}^{(1)} \\ \mathbf{y}^{(3)} &= \mathbf{y}^{(2)} \mathbf{W}^{(2)} = (\mathbf{y}^{(1)} + \mathbf{B}^{(1)}) \mathbf{W}^{(2)} = (\mathbf{x} \mathbf{W}^{(1)} + \mathbf{B}^{(1)}) \mathbf{W}^{(2)} \\ \mathbf{y} &= \mathbf{y}^{(3)} + \mathbf{B}^{(2)} = \mathbf{y}^{(2)} \mathbf{W}^{(2)} + \mathbf{B}^{(2)} = (\mathbf{y}^{(1)} + \mathbf{B}^{(1)}) \mathbf{W}^{(2)} + \mathbf{B}^{(2)} = (\mathbf{x} \mathbf{W}^{(1)} + \mathbf{B}^{(1)}) \mathbf{W}^{(2)} + \mathbf{B}^{(2)}\end{aligned}$$

順伝播 成分表記

$$\begin{aligned}\mathbf{y}^{(1)} &= \begin{pmatrix} y_1^{(1)} & y_2^{(1)} & \cdots & y_q^{(1)} \end{pmatrix} = \mathbf{x} \mathbf{W}^{(1)} \\ &= (x_1 \ x_2 \ \cdots \ x_p) \begin{pmatrix} W_{1,1}^{(1)} & W_{1,2}^{(1)} & \cdots & W_{1,q}^{(1)} \\ W_{2,1}^{(1)} & W_{2,2}^{(1)} & \cdots & W_{2,q}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{p,1}^{(1)} & W_{p,2}^{(1)} & \cdots & W_{p,q}^{(1)} \end{pmatrix} = \begin{pmatrix} x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} \\ x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} \\ \vdots \\ x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} \end{pmatrix}^T \\ \mathbf{y}^{(2)} &= \begin{pmatrix} y_1^{(2)} & y_2^{(2)} & \cdots & y_q^{(2)} \end{pmatrix} = \mathbf{y}^{(1)} + \mathbf{B}^{(1)} = \mathbf{x} \mathbf{W}^{(1)} + \mathbf{B}^{(1)} \\ &= \begin{pmatrix} y_1^{(1)} & y_2^{(1)} & \cdots & y_q^{(1)} \end{pmatrix} + \begin{pmatrix} B_1^{(1)} & B_2^{(1)} & \cdots & B_q^{(1)} \end{pmatrix} = \begin{pmatrix} y_1^{(1)} + B_1^{(1)} & y_2^{(1)} + B_2^{(1)} & \cdots & y_q^{(1)} + B_q^{(1)} \end{pmatrix} \\ &= \begin{pmatrix} x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)} \\ x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)} \\ \vdots \\ x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)} \end{pmatrix} \\ \mathbf{y}^{(3)} &= \begin{pmatrix} y_1^{(3)} & y_2^{(3)} & \cdots & y_r^{(3)} \end{pmatrix} = \mathbf{y}^{(2)} \mathbf{W}^{(2)} = (\mathbf{y}^{(1)} + \mathbf{B}^{(1)}) \mathbf{W}^{(2)} = (\mathbf{x} \mathbf{W}^{(1)} + \mathbf{B}^{(1)}) \mathbf{W}^{(2)} \\ &= \begin{pmatrix} y_1^{(2)} & y_2^{(2)} & \cdots & y_q^{(2)} \end{pmatrix} \begin{pmatrix} W_{1,1}^{(2)} & W_{1,2}^{(2)} & \cdots & W_{1,r}^{(2)} \\ W_{2,1}^{(2)} & W_{2,2}^{(2)} & \cdots & W_{2,r}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{q,1}^{(2)} & W_{q,2}^{(2)} & \cdots & W_{q,r}^{(2)} \end{pmatrix} = \begin{pmatrix} y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \cdots + y_q^{(2)} W_{q,1}^{(2)} \\ y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \cdots + y_q^{(2)} W_{q,2}^{(2)} \\ \vdots \\ y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \cdots + y_q^{(2)} W_{q,r}^{(2)} \end{pmatrix}^T\end{aligned}$$

$$\begin{aligned}
&= \begin{pmatrix} (y_1^{(1)} + B_1^{(1)})W_{1,1}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,1}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)})W_{q,1}^{(2)} \\ (y_1^{(1)} + B_1^{(1)})W_{1,2}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,2}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)})W_{q,2}^{(2)} \\ \vdots \\ (y_1^{(1)} + B_1^{(1)})W_{1,r}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,r}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)})W_{q,r}^{(2)} \end{pmatrix}^\top \\
&= \begin{pmatrix} (x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)})W_{1,1}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)})W_{2,1}^{(2)} \\ + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)})W_{q,1}^{(2)} \\ (x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)})W_{1,2}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)})W_{2,2}^{(2)} \\ + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)})W_{q,2}^{(2)} \\ \vdots \\ (x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)})W_{1,r}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)})W_{2,r}^{(2)} \\ + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)})W_{q,r}^{(2)} \end{pmatrix}^\top
\end{aligned}$$

$$\begin{aligned}
\mathbf{y} = (y_1 \ y_2 \ \cdots \ y_r) &= \mathbf{y}^{(3)} + \mathbf{B}^{(2)} = \mathbf{y}^{(2)} \mathbf{W}^{(2)} + \mathbf{B}^{(2)} \\
&= (\mathbf{y}^{(1)} + \mathbf{B}^{(1)}) \mathbf{W}^{(2)} + \mathbf{B}^{(2)} = (\mathbf{x} \mathbf{W}^{(1)} + \mathbf{B}^{(1)}) \mathbf{W}^{(2)} + \mathbf{B}^{(2)} \\
&= (y_1^{(3)} \ y_2^{(3)} \ \cdots \ y_r^{(3)}) + (B_1^{(2)} \ B_2^{(2)} \ \cdots \ B_r^{(2)}) = (y_1^{(3)} + B_1^{(2)} \ y_2^{(3)} + B_2^{(2)} \ \cdots \ y_r^{(3)} + B_r^{(2)})
\end{aligned}$$

$$\begin{aligned}
&= \begin{pmatrix} y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \cdots + y_q^{(2)} W_{q,1}^{(2)} \\ y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \cdots + y_q^{(2)} W_{q,2}^{(2)} \\ \vdots \\ y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \cdots + y_q^{(2)} W_{q,r}^{(2)} \end{pmatrix}^\top \begin{pmatrix} B_1^{(2)} \\ B_2^{(2)} \\ \vdots \\ B_r^{(2)} \end{pmatrix}^\top = \begin{pmatrix} y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \cdots + y_q^{(2)} W_{q,1}^{(2)} + B_1^{(2)} \\ y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \cdots + y_q^{(2)} W_{q,2}^{(2)} + B_2^{(2)} \\ \vdots \\ y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \cdots + y_q^{(2)} W_{q,r}^{(2)} + B_r^{(2)} \end{pmatrix}^\top
\end{aligned}$$

$$\begin{aligned}
&= \begin{pmatrix} (y_1^{(1)} + B_1^{(1)})W_{1,1}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,1}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)})W_{q,1}^{(2)} + B_1^{(2)} \\ (y_1^{(1)} + B_1^{(1)})W_{1,2}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,2}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)})W_{q,2}^{(2)} + B_2^{(2)} \\ \vdots \\ (y_1^{(1)} + B_1^{(1)})W_{1,r}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,r}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)})W_{q,r}^{(2)} + B_r^{(2)} \end{pmatrix}^\top \\
&= \begin{pmatrix} (x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)})W_{1,1}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)})W_{2,1}^{(2)} \\ + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)})W_{q,1}^{(2)} + B_1^{(2)} \\ (x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)})W_{1,2}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)})W_{2,2}^{(2)} \\ + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)})W_{q,2}^{(2)} + B_2^{(2)} \\ \vdots \\ (x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)})W_{1,r}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)})W_{2,r}^{(2)} \\ + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)})W_{q,r}^{(2)} + B_r^{(2)} \end{pmatrix}^\top
\end{aligned}$$

勾配の計算

$$\frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial y^{(3)}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{y}^{(1)}} \frac{\partial \mathbf{y}^{(1)}}{\partial \mathbf{x}}$$

$$\frac{\partial L}{\partial \mathbf{y}}$$

損失関数は任意の損失関数として、 L の \mathbf{y} に関する勾配を下記の通りとします。

$$\frac{\partial L}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial L}{\partial y_1} & \frac{\partial L}{\partial y_2} & \dots & \frac{\partial L}{\partial y_r} \end{pmatrix}$$

$$\frac{\partial L}{\partial \mathbf{B}^{(2)}}$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{B}^{(2)}} &= \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{B}^{(2)}} = \begin{pmatrix} \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial B_1^{(2)}} & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial B_2^{(2)}} & \dots & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial B_r^{(2)}} \end{pmatrix} \\ \frac{\partial \mathbf{y}}{\partial B_1^{(2)}} &= \frac{\partial}{\partial B_1^{(2)}} (\mathbf{y}^{(3)} + \mathbf{B}^{(2)}) = \left(\frac{\partial}{\partial B_1^{(2)}} (y_1^{(3)} + B_1^{(2)}) \quad \frac{\partial}{\partial B_1^{(2)}} (y_2^{(3)} + B_2^{(2)}) \quad \dots \quad \frac{\partial}{\partial B_1^{(2)}} (y_r^{(3)} + B_r^{(2)}) \right)^T \\ &= (1 \quad 0 \quad \dots \quad 0)^T \end{aligned}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{B}^{(2)}} = \left(\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \dots \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \mathbf{I}$$

∴

$$\frac{\partial L}{\partial \mathbf{B}^{(2)}} = \left(\frac{\partial L}{\partial y} [1 \quad 0 \quad \dots \quad 0]^T \quad \frac{\partial L}{\partial y} [0 \quad 1 \quad \dots \quad 0]^T \quad \dots \quad \frac{\partial L}{\partial y} [0 \quad 0 \quad \dots \quad 1]^T \right) = \frac{\partial L}{\partial \mathbf{y}}$$

$$\frac{\partial L}{\partial \mathbf{y}^{(3)}}$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{y}^{(3)}} &= \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{y}^{(3)}} = \begin{pmatrix} \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial y_1^{(3)}} & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial y_2^{(3)}} & \dots & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial y_r^{(3)}} \end{pmatrix} \\ \frac{\partial \mathbf{y}}{\partial y_1^{(3)}} &= \frac{\partial}{\partial y_1^{(3)}} (\mathbf{y}^{(3)} + \mathbf{B}^{(2)}) = \left(\frac{\partial}{\partial y_1^{(3)}} (y_1^{(3)} + B_1^{(2)}) \quad \frac{\partial}{\partial y_1^{(3)}} (y_2^{(3)} + B_2^{(2)}) \quad \dots \quad \frac{\partial}{\partial y_1^{(3)}} (y_r^{(3)} + B_r^{(2)}) \right)^T \\ &= (1 \quad 0 \quad \dots \quad 0)^T \end{aligned}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{y}^{(3)}} = \left(\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \dots \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \mathbf{I}$$

∴

$$\frac{\partial L}{\partial \mathbf{y}^{(3)}} = \left(\frac{\partial L}{\partial y} [1 \quad 0 \quad \dots \quad 0]^T \quad \frac{\partial L}{\partial y} [0 \quad 1 \quad \dots \quad 0]^T \quad \dots \quad \frac{\partial L}{\partial y} [0 \quad 0 \quad \dots \quad 1]^T \right) = \frac{\partial L}{\partial \mathbf{y}}$$

$$\partial L / \partial \mathbf{W}^{(2)}$$

$$\begin{aligned}
& \frac{\partial L}{\partial \mathbf{W}^{(2)}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{y}^{(3)}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{W}^{(2)}} = \left(\begin{array}{cccc} \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{1,1}^{(2)}} & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{1,2}^{(2)}} & \cdots & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{1,r}^{(2)}} \\ \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{2,1}^{(2)}} & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{2,2}^{(2)}} & \cdots & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{2,r}^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{q,1}^{(2)}} & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{q,2}^{(2)}} & \cdots & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial W_{q,r}^{(2)}} \end{array} \right) \\
& \frac{\partial \mathbf{y}^{(3)}}{\partial W_{1,1}^{(2)}} = \frac{\partial}{\partial W_{1,1}^{(2)}} \mathbf{y}^{(2)} \mathbf{W}^{(2)} = \left(\begin{array}{c} \frac{\partial}{\partial W_{1,1}^{(2)}} (y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \cdots + y_q^{(2)} W_{q,1}^{(2)}) \\ \frac{\partial}{\partial W_{1,1}^{(2)}} (y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \cdots + y_q^{(2)} W_{q,2}^{(2)}) \\ \vdots \\ \frac{\partial}{\partial W_{1,1}^{(2)}} (y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \cdots + y_q^{(2)} W_{q,r}^{(2)}) \end{array} \right) = \begin{pmatrix} y_1^{(2)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
& \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{W}^{(2)}} = \left(\begin{array}{ccccc} [y_1^{(2)} \ 0 \ \cdots \ 0]^\top & [0 \ y_1^{(2)} \ \cdots \ 0]^\top & \cdots & [0 \ 0 \ \cdots \ y_1^{(2)}]^\top \\ [y_2^{(2)} \ 0 \ \cdots \ 0]^\top & [0 \ y_2^{(2)} \ \cdots \ 0]^\top & \cdots & [0 \ 0 \ \cdots \ y_2^{(2)}]^\top \\ \vdots & \vdots & \ddots & \vdots \\ [y_q^{(2)} \ 0 \ \cdots \ 0]^\top & [0 \ y_q^{(2)} \ \cdots \ 0]^\top & \cdots & [0 \ 0 \ \cdots \ y_q^{(2)}]^\top \end{array} \right) \\
& \vdots \\
& \frac{\partial L}{\partial \mathbf{W}^{(2)}} = \left(\begin{array}{ccccc} \frac{\partial L}{\partial \mathbf{y}} [y_1^{(2)} \ 0 \ \cdots \ 0]^\top & \frac{\partial L}{\partial \mathbf{y}} [0 \ y_1^{(2)} \ \cdots \ 0]^\top & \cdots & \frac{\partial L}{\partial \mathbf{y}} [0 \ 0 \ \cdots \ y_1^{(2)}]^\top \\ \frac{\partial L}{\partial \mathbf{y}} [y_2^{(2)} \ 0 \ \cdots \ 0]^\top & \frac{\partial L}{\partial \mathbf{y}} [0 \ y_2^{(2)} \ \cdots \ 0]^\top & \cdots & \frac{\partial L}{\partial \mathbf{y}} [0 \ 0 \ \cdots \ y_2^{(2)}]^\top \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial \mathbf{y}} [y_q^{(2)} \ 0 \ \cdots \ 0]^\top & \frac{\partial L}{\partial \mathbf{y}} [0 \ y_q^{(2)} \ \cdots \ 0]^\top & \cdots & \frac{\partial L}{\partial \mathbf{y}} [0 \ 0 \ \cdots \ y_q^{(2)}]^\top \end{array} \right) \\
& = \begin{pmatrix} \frac{\partial L}{\partial y_1} y_1^{(2)} & \frac{\partial L}{\partial y_2} y_1^{(2)} & \cdots & \frac{\partial L}{\partial y_r} y_1^{(2)} \\ \frac{\partial L}{\partial y_1} y_2^{(2)} & \frac{\partial L}{\partial y_2} y_2^{(2)} & \cdots & \frac{\partial L}{\partial y_r} y_2^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial y_1} y_q^{(2)} & \frac{\partial L}{\partial y_2} y_q^{(2)} & \cdots & \frac{\partial L}{\partial y_r} y_q^{(2)} \end{pmatrix} = \begin{pmatrix} y_1^{(2)} \\ y_2^{(2)} \\ \vdots \\ y_q^{(2)} \end{pmatrix} \begin{pmatrix} \frac{\partial L}{\partial y_1} & \frac{\partial L}{\partial y_2} & \cdots & \frac{\partial L}{\partial y_r} \end{pmatrix} = \mathbf{y}^{(2)\top} \frac{\partial L}{\partial \mathbf{y}}
\end{aligned}$$

$$\partial L / \partial \mathbf{y}^{(2)}$$

$$\frac{\partial L}{\partial \mathbf{y}^{(2)}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{y}^{(3)}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{y}^{(2)}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{y}^{(2)}} = \begin{pmatrix} \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial y_1^{(2)}} & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial y_2^{(2)}} & \dots & \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}^{(3)}}{\partial y_q^{(2)}} \end{pmatrix}$$

$$\frac{\partial \mathbf{y}^{(3)}}{\partial y_1^{(2)}} = \frac{\partial}{\partial y_1^{(2)}} \mathbf{y}^{(2)} \mathbf{W}^{(2)} = \begin{pmatrix} \frac{\partial}{\partial y_1^{(2)}} (y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \dots + y_q^{(2)} W_{q,1}^{(2)}) \\ \frac{\partial}{\partial y_1^{(2)}} (y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \dots + y_q^{(2)} W_{q,2}^{(2)}) \\ \vdots \\ \frac{\partial}{\partial y_1^{(2)}} (y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \dots + y_q^{(2)} W_{q,r}^{(2)}) \end{pmatrix} = \begin{pmatrix} W_{1,1}^{(2)} \\ W_{1,2}^{(2)} \\ \vdots \\ W_{1,r}^{(2)} \end{pmatrix}$$

$$\frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{y}^{(2)}} = \begin{pmatrix} \begin{bmatrix} W_{1,1}^{(2)} \\ W_{1,2}^{(2)} \\ \vdots \\ W_{1,r}^{(2)} \end{bmatrix} & \begin{bmatrix} W_{2,1}^{(2)} \\ W_{2,2}^{(2)} \\ \vdots \\ W_{2,r}^{(2)} \end{bmatrix} & \dots & \begin{bmatrix} W_{q,1}^{(2)} \\ W_{q,2}^{(2)} \\ \vdots \\ W_{q,r}^{(2)} \end{bmatrix} \end{pmatrix} = \begin{pmatrix} W_{1,1}^{(2)} & W_{2,1}^{(2)} & \dots & W_{q,1}^{(2)} \\ W_{1,2}^{(2)} & W_{2,2}^{(2)} & \dots & W_{q,2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,r}^{(2)} & W_{2,r}^{(2)} & \dots & W_{q,r}^{(2)} \end{pmatrix} = \mathbf{W}^{(2)\top}$$

$$\therefore$$

$$\frac{\partial L}{\partial \mathbf{y}^{(2)}} = \begin{pmatrix} \frac{\partial L}{\partial y_1} \begin{bmatrix} W_{1,1}^{(2)} \\ W_{1,2}^{(2)} \\ \vdots \\ W_{1,r}^{(2)} \end{bmatrix} & \frac{\partial L}{\partial y_2} \begin{bmatrix} W_{2,1}^{(2)} \\ W_{2,2}^{(2)} \\ \vdots \\ W_{2,r}^{(2)} \end{bmatrix} & \dots & \frac{\partial L}{\partial y_r} \begin{bmatrix} W_{q,1}^{(2)} \\ W_{q,2}^{(2)} \\ \vdots \\ W_{q,r}^{(2)} \end{bmatrix} \end{pmatrix} = \begin{pmatrix} \frac{\partial L}{\partial y_1} W_{1,1}^{(2)} + \frac{\partial L}{\partial y_2} W_{1,2}^{(2)} + \dots + \frac{\partial L}{\partial y_r} W_{1,r}^{(2)} \\ \frac{\partial L}{\partial y_1} W_{2,1}^{(2)} + \frac{\partial L}{\partial y_2} W_{2,2}^{(2)} + \dots + \frac{\partial L}{\partial y_r} W_{2,r}^{(2)} \\ \vdots \\ \frac{\partial L}{\partial y_1} W_{q,1}^{(2)} + \frac{\partial L}{\partial y_2} W_{q,2}^{(2)} + \dots + \frac{\partial L}{\partial y_r} W_{q,r}^{(2)} \end{pmatrix}^\top$$

$$= \begin{pmatrix} \frac{\partial L}{\partial y_1} & \frac{\partial L}{\partial y_2} & \dots & \frac{\partial L}{\partial y_r} \end{pmatrix} \begin{pmatrix} W_{1,1}^{(2)} & W_{2,1}^{(2)} & \dots & W_{q,1}^{(2)} \\ W_{1,2}^{(2)} & W_{2,2}^{(2)} & \dots & W_{q,2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,r}^{(2)} & W_{2,r}^{(2)} & \dots & W_{q,r}^{(2)} \end{pmatrix} = \frac{\partial L}{\partial \mathbf{y}} \mathbf{W}^{(2)\top}$$

$$\partial L / \partial \mathbf{B}^{(1)}$$

$$\frac{\partial L}{\partial \mathbf{B}^{(1)}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{y}^{(3)}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{B}^{(1)}} = \frac{\partial L}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{B}^{(1)}} = \begin{pmatrix} \frac{\partial L}{\partial y_1^{(2)}} \frac{\partial y_1^{(2)}}{\partial B_1^{(1)}} & \frac{\partial L}{\partial y_2^{(2)}} \frac{\partial y_2^{(2)}}{\partial B_2^{(1)}} & \dots & \frac{\partial L}{\partial y_r^{(2)}} \frac{\partial y_r^{(2)}}{\partial B_q^{(1)}} \end{pmatrix}$$

$$\frac{\partial \mathbf{y}^{(2)}}{\partial B_1^{(1)}} = \frac{\partial}{\partial B_1^{(1)}} (\mathbf{y}^{(1)} + \mathbf{B}^{(1)}) = \begin{pmatrix} \frac{\partial}{\partial B_1^{(1)}} (y_1^{(1)} + B_1^{(1)}) & \frac{\partial}{\partial B_1^{(1)}} (y_2^{(1)} + B_2^{(1)}) & \dots & \frac{\partial}{\partial B_1^{(1)}} (y_q^{(1)} + B_q^{(1)}) \end{pmatrix}^\top$$

$$= (1 \ 0 \ \dots \ 0)^\top$$

$$\frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{B}^{(1)}} = \begin{pmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} & \dots & \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \mathbf{I}$$

$$\therefore \frac{\partial L}{\partial \mathbf{B}^{(1)}} = \left(\frac{\partial L}{\partial y_1^{(2)}} [1 \ 0 \ \dots \ 0]^\top \ \frac{\partial L}{\partial y_2^{(2)}} [0 \ 1 \ \dots \ 0]^\top \ \dots \ \frac{\partial L}{\partial y_r^{(2)}} [0 \ 0 \ \dots \ 1]^\top \right) = \frac{\partial L}{\partial \mathbf{y}^{(2)}}$$

$$\partial L / \partial \mathbf{y}^{(1)}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{y}^{(1)}} &= \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{y}^{(3)}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{y}^{(1)}} = \frac{\partial L}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{y}^{(1)}} = \left(\frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial y_1^{(1)}} \quad \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial y_2^{(1)}} \quad \dots \quad \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(2)}}{\partial y_q^{(1)}} \right) \\ \frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{y}^{(1)}} &= \frac{\partial}{\partial y_1^{(1)}} (\mathbf{y}^{(1)} + \mathbf{B}^{(1)}) = \left(\frac{\partial}{\partial y_1^{(1)}} (y_1^{(1)} + B_1^{(1)}) \quad \frac{\partial}{\partial y_1^{(1)}} (y_2^{(1)} + B_2^{(1)}) \quad \dots \quad \frac{\partial}{\partial y_1^{(1)}} (y_q^{(1)} + B_q^{(1)}) \right)^T \\ &= (1 \quad 0 \quad \dots \quad 0)^T\end{aligned}$$

$$\frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{y}^{(1)}} = \left(\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \dots \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \mathbf{I}$$

$$\therefore \frac{\partial L}{\partial \mathbf{y}^{(1)}} = \left(\frac{\partial L}{\partial y^{(2)}} [1 \quad 0 \quad \dots \quad 0]^T \quad \frac{\partial L}{\partial y^{(2)}} [0 \quad 1 \quad \dots \quad 0]^T \quad \dots \quad \frac{\partial L}{\partial y^{(2)}} [0 \quad 0 \quad \dots \quad 1]^T \right) = \frac{\partial L}{\partial \mathbf{y}^{(2)}}$$

$$\partial L / \partial \mathbf{W}^{(1)}$$

$$\frac{\partial L}{\partial \mathbf{W}^{(1)}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{y}^{(3)}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{y}^{(1)}} \frac{\partial \mathbf{y}^{(1)}}{\partial \mathbf{W}^{(1)}} = \frac{\partial L}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(1)}}{\partial \mathbf{W}^{(1)}} = \begin{pmatrix} \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{1,1}^{(1)}} & \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{1,2}^{(1)}} & \dots & \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{1,q}^{(1)}} \\ \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{2,1}^{(1)}} & \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{2,2}^{(1)}} & \dots & \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{2,q}^{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{p,1}^{(1)}} & \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{p,2}^{(1)}} & \dots & \frac{\partial L}{\partial y^{(2)}} \frac{\partial y^{(1)}}{\partial W_{p,q}^{(1)}} \end{pmatrix}$$

$$\frac{\partial \mathbf{y}^{(1)}}{\partial W_{1,1}^{(1)}} = \frac{\partial}{\partial W_{1,1}^{(1)}} \mathbf{x} \mathbf{W}^{(1)} = \begin{pmatrix} \frac{\partial}{\partial W_{1,1}^{(1)}} (x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \dots + x_p W_{p,1}^{(1)}) \\ \frac{\partial}{\partial W_{1,1}^{(1)}} (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \dots + x_p W_{p,2}^{(1)}) \\ \vdots \\ \frac{\partial}{\partial W_{1,1}^{(1)}} (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \dots + x_p W_{p,q}^{(1)}) \end{pmatrix} = \begin{pmatrix} x_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\frac{\partial \mathbf{y}^{(1)}}{\partial \mathbf{W}^{(1)}} = \begin{pmatrix} [x_1 \quad 0 \quad \dots \quad 0]^T \quad [0 \quad x_1 \quad \dots \quad 0]^T \quad \dots \quad [0 \quad 0 \quad \dots \quad x_1]^T \\ [x_2 \quad 0 \quad \dots \quad 0]^T \quad [0 \quad x_2 \quad \dots \quad 0]^T \quad \dots \quad [0 \quad 0 \quad \dots \quad x_2]^T \\ \vdots & \vdots & \ddots & \vdots \\ [x_p \quad 0 \quad \dots \quad 0]^T \quad [0 \quad x_p \quad \dots \quad 0]^T \quad \dots \quad [0 \quad 0 \quad \dots \quad x_p]^T \end{pmatrix}$$

\therefore

$$\frac{\partial L}{\partial \mathbf{W}^{(1)}} = \begin{pmatrix} \frac{\partial L}{\partial y^{(2)}} [x_1 \quad 0 \quad \dots \quad 0]^T \quad \frac{\partial L}{\partial y^{(2)}} [0 \quad x_1 \quad \dots \quad 0]^T \quad \dots \quad \frac{\partial L}{\partial y^{(2)}} [0 \quad 0 \quad \dots \quad x_1]^T \\ \frac{\partial L}{\partial y^{(2)}} [x_2 \quad 0 \quad \dots \quad 0]^T \quad \frac{\partial L}{\partial y^{(2)}} [0 \quad x_2 \quad \dots \quad 0]^T \quad \dots \quad \frac{\partial L}{\partial y^{(2)}} [0 \quad 0 \quad \dots \quad x_2]^T \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial y^{(2)}} [x_p \quad 0 \quad \dots \quad 0]^T \quad \frac{\partial L}{\partial y^{(2)}} [0 \quad x_p \quad \dots \quad 0]^T \quad \dots \quad \frac{\partial L}{\partial y^{(2)}} [0 \quad 0 \quad \dots \quad x_p]^T \end{pmatrix}$$

$$\begin{aligned}
 &= \begin{pmatrix} \frac{\partial L}{\partial y_1^{(2)}} x_1 & \frac{\partial L}{\partial y_2^{(2)}} x_1 & \cdots & \frac{\partial L}{\partial y_q^{(2)}} x_1 \\ \frac{\partial L}{\partial y_1^{(2)}} x_2 & \frac{\partial L}{\partial y_2^{(2)}} x_2 & \cdots & \frac{\partial L}{\partial y_q^{(2)}} x_2 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial y_1^{(2)}} x_p & \frac{\partial L}{\partial y_2^{(2)}} x_p & \cdots & \frac{\partial L}{\partial y_q^{(2)}} x_p \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \begin{pmatrix} \frac{\partial L}{\partial y_1^{(2)}} & \frac{\partial L}{\partial y_2^{(2)}} & \cdots & \frac{\partial L}{\partial y_q^{(2)}} \end{pmatrix} = \mathbf{x}^\top \frac{\partial L}{\partial \mathbf{y}^{(2)}}
 \end{aligned}$$

 $\partial L / \partial \mathbf{x}$

$$\frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{y}^{(3)}} \frac{\partial \mathbf{y}^{(3)}}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(2)}}{\partial \mathbf{y}^{(1)}} \frac{\partial \mathbf{y}^{(1)}}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{y}^{(2)}} \frac{\partial \mathbf{y}^{(1)}}{\partial \mathbf{x}} = \left(\frac{\partial L}{\partial y_1^{(2)}} \frac{\partial y^{(1)}}{\partial x_1} \quad \frac{\partial L}{\partial y_2^{(2)}} \frac{\partial y^{(1)}}{\partial x_2} \quad \cdots \quad \frac{\partial L}{\partial y_q^{(2)}} \frac{\partial y^{(1)}}{\partial x_p} \right)$$

$$\frac{\partial \mathbf{y}^{(1)}}{\partial x_1} = \frac{\partial}{\partial x_1} \mathbf{x} \mathbf{W}^{(1)} = \begin{pmatrix} \frac{\partial}{\partial x_1} \left(x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} \right) \\ \frac{\partial}{\partial x_1} \left(x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} \right) \\ \vdots \\ \frac{\partial}{\partial x_1} \left(x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} \right) \end{pmatrix} = \begin{pmatrix} W_{1,1}^{(1)} \\ W_{1,2}^{(1)} \\ \vdots \\ W_{1,q}^{(1)} \end{pmatrix}$$

$$\frac{\partial \mathbf{y}^{(1)}}{\partial \mathbf{x}} = \begin{pmatrix} \begin{bmatrix} W_{1,1}^{(1)} \\ W_{1,2}^{(1)} \\ \vdots \\ W_{1,q}^{(1)} \end{bmatrix} & \begin{bmatrix} W_{2,1}^{(1)} \\ W_{2,2}^{(1)} \\ \vdots \\ W_{2,q}^{(1)} \end{bmatrix} & \cdots & \begin{bmatrix} W_{p,1}^{(1)} \\ W_{p,2}^{(1)} \\ \vdots \\ W_{p,q}^{(1)} \end{bmatrix} \end{pmatrix} = \begin{pmatrix} W_{1,1}^{(1)} & W_{2,1}^{(1)} & \cdots & W_{p,1}^{(1)} \\ W_{1,2}^{(1)} & W_{2,2}^{(1)} & \cdots & W_{p,2}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,q}^{(1)} & W_{2,q}^{(1)} & \cdots & W_{p,q}^{(1)} \end{pmatrix} = \mathbf{W}^{(1)\top}$$

 \therefore

$$\frac{\partial L}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial L}{\partial y^{(2)}} \begin{bmatrix} W_{1,1}^{(1)} \\ W_{1,2}^{(1)} \\ \vdots \\ W_{1,q}^{(1)} \end{bmatrix} & \frac{\partial L}{\partial y^{(2)}} \begin{bmatrix} W_{2,1}^{(1)} \\ W_{2,2}^{(1)} \\ \vdots \\ W_{2,q}^{(1)} \end{bmatrix} & \cdots & \frac{\partial L}{\partial y^{(2)}} \begin{bmatrix} W_{p,1}^{(1)} \\ W_{p,2}^{(1)} \\ \vdots \\ W_{p,q}^{(1)} \end{bmatrix} \end{pmatrix} = \begin{pmatrix} \frac{\partial L}{\partial y_1^{(2)}} W_{1,1}^{(1)} + \frac{\partial L}{\partial y_2^{(2)}} W_{1,2}^{(1)} + \cdots + \frac{\partial L}{\partial y_q^{(2)}} W_{1,q}^{(1)} \\ \frac{\partial L}{\partial y_1^{(2)}} W_{2,1}^{(1)} + \frac{\partial L}{\partial y_2^{(2)}} W_{2,2}^{(1)} + \cdots + \frac{\partial L}{\partial y_q^{(2)}} W_{2,q}^{(1)} \\ \vdots \\ \frac{\partial L}{\partial y_1^{(2)}} W_{p,1}^{(1)} + \frac{\partial L}{\partial y_2^{(2)}} W_{p,2}^{(1)} + \cdots + \frac{\partial L}{\partial y_q^{(2)}} W_{p,q}^{(1)} \end{pmatrix}^\top$$

$$= \left(\frac{\partial L}{\partial y_1^{(2)}} \quad \frac{\partial L}{\partial y_2^{(2)}} \quad \cdots \quad \frac{\partial L}{\partial y_q^{(2)}} \right) \begin{pmatrix} W_{1,1}^{(1)} & W_{2,1}^{(1)} & \cdots & W_{p,1}^{(1)} \\ W_{1,2}^{(1)} & W_{2,2}^{(1)} & \cdots & W_{p,2}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,q}^{(1)} & W_{2,q}^{(1)} & \cdots & W_{p,q}^{(1)} \end{pmatrix} = \frac{\partial L}{\partial \mathbf{y}^{(2)}} \mathbf{W}^{(1)\top}$$

勾配の計算（成分毎）

この計算例では、損失関数に教師信号を t とした平均二乗誤差を用います。

L

$$\begin{aligned}
 L &= \frac{1}{r} \sum_{i=1}^r (t_i - y_i)^2 = \frac{1}{r} \left\{ [t_1 - y_1]^2 + [t_2 - y_2]^2 + \cdots + [t_r - y_r]^2 \right\} \\
 &= \frac{1}{r} \left\{ [t_1 - (y_1^{(3)} + B_1^{(2)})]^2 + [t_2 - (y_2^{(3)} + B_2^{(2)})]^2 + \cdots + [t_r - (y_r^{(3)} + B_r^{(2)})]^2 \right\} \\
 &= \frac{1}{r} \left\{ [t_1 - (y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \cdots + y_q^{(2)} W_{q,1}^{(2)} + B_1^{(2)})]^2 + \right. \\
 &\quad [t_2 - (y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \cdots + y_q^{(2)} W_{q,2}^{(2)} + B_2^{(2)})]^2 + \cdots \\
 &\quad \left. [t_r - (y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \cdots + y_q^{(2)} W_{q,r}^{(2)} + B_r^{(2)})]^2 \right\} \\
 &= \frac{1}{r} \left\{ [t_1 - ((y_1^{(1)} + B_1^{(1)}) W_{1,1}^{(2)} + (y_2^{(1)} + B_2^{(1)}) W_{2,1}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)}) W_{q,1}^{(2)} + B_1^{(2)})]^2 + \right. \\
 &\quad [t_2 - ((y_1^{(1)} + B_1^{(1)}) W_{1,2}^{(2)} + (y_2^{(1)} + B_2^{(1)}) W_{2,2}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)}) W_{q,2}^{(2)} + B_2^{(2)})]^2 + \cdots \\
 &\quad \left. [t_r - ((y_1^{(1)} + B_1^{(1)}) W_{1,r}^{(2)} + (y_2^{(1)} + B_2^{(1)}) W_{2,r}^{(2)} + \cdots + (y_q^{(1)} + B_q^{(1)}) W_{q,r}^{(2)} + B_r^{(2)})]^2 \right\} \\
 &= \frac{1}{r} \left\{ \left[t_1 - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,1}^{(2)} + \right. \right. \right. \\
 &\quad (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,1}^{(2)} + \cdots + \\
 &\quad \left. \left. \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,1}^{(2)} + B_1^{(2)} \right) \right]^2 + \right. \\
 &\quad \left[t_2 - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,2}^{(2)} + \right. \right. \\
 &\quad (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,2}^{(2)} + \cdots + \\
 &\quad \left. \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,2}^{(2)} + B_2^{(2)} \right) \right]^2 + \cdots + \right. \\
 &\quad \left[t_r - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,r}^{(2)} + \right. \right. \\
 &\quad (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,r}^{(2)} + \cdots + \\
 &\quad \left. \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,r}^{(2)} + B_r^{(2)} \right) \right]^2 \right\}
 \end{aligned}$$

$$\partial L / \partial \mathbf{y}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{y}} &= \begin{pmatrix} \frac{\partial L}{\partial y_1} & \frac{\partial L}{\partial y_2} & \cdots & \frac{\partial L}{\partial y_r} \end{pmatrix} \\ \frac{\partial L}{\partial y_1} &= \frac{\partial}{\partial y_1} \frac{1}{r} ((t_1 - y_1)^2 + (t_2 - y_2)^2 + \cdots + (t_r - y_r)^2) = \frac{1}{r} \frac{\partial}{\partial y_1} (t_1 - y_1)^2 = \frac{2}{r} (y_1 - t_1) \\ \therefore \frac{\partial L}{\partial \mathbf{y}} &= \frac{2}{r} (y_1 - t_1 \quad y_2 - t_2 \quad \cdots \quad y_r - t_r)\end{aligned}$$

$$\partial L / \partial \mathbf{B}^{(2)}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{B}^{(2)}} &= \begin{pmatrix} \frac{\partial L}{\partial B_1^{(2)}} & \frac{\partial L}{\partial B_2^{(2)}} & \cdots & \frac{\partial L}{\partial B_r^{(2)}} \end{pmatrix} \\ \frac{\partial L}{\partial B_1^{(2)}} &= \frac{\partial}{\partial B_1^{(2)}} \frac{1}{r} \left\{ [t_1 - (y_1^{(3)} + B_1^{(2)})]^2 + [t_2 - (y_2^{(3)} + B_2^{(2)})]^2 + \cdots + [t_r - (y_r^{(3)} + B_r^{(2)})]^2 \right\} \\ &= \frac{1}{r} \frac{\partial}{\partial B_1^{(2)}} (t_1 - y_1^{(3)} - B_1^{(2)})^2 = \frac{2}{r} (y_1^{(3)} + B_1^{(2)} - t_1) = \frac{2}{r} (y_1 - t_1) \\ \therefore \frac{\partial L}{\partial \mathbf{B}^{(2)}} &= \frac{2}{r} (y_1 - t_1 \quad y_2 - t_2 \quad \cdots \quad y_r - t_r) = \frac{\partial L}{\partial \mathbf{y}}\end{aligned}$$

$$\partial L / \partial \mathbf{y}^{(3)}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{y}^{(3)}} &= \begin{pmatrix} \frac{\partial L}{\partial y_1^{(3)}} & \frac{\partial L}{\partial y_2^{(3)}} & \cdots & \frac{\partial L}{\partial y_r^{(3)}} \end{pmatrix} \\ \frac{\partial L}{\partial y_1^{(3)}} &= \frac{\partial}{\partial y_1^{(3)}} \frac{1}{r} \left\{ [t_1 - (y_1^{(3)} + B_1^{(2)})]^2 + [t_2 - (y_2^{(3)} + B_2^{(2)})]^2 + \cdots + [t_r - (y_r^{(3)} + B_r^{(2)})]^2 \right\} \\ &= \frac{1}{r} \frac{\partial}{\partial y_1^{(3)}} (t_1 - y_1^{(3)} - B_1^{(2)})^2 = \frac{2}{r} (y_1^{(3)} + B_1^{(2)} - t_1) = \frac{2}{r} (y_1 - t_1) \\ \therefore \frac{\partial L}{\partial \mathbf{y}^{(3)}} &= \frac{2}{r} (y_1 - t_1 \quad y_2 - t_2 \quad \cdots \quad y_r - t_r) = \frac{\partial L}{\partial \mathbf{y}}\end{aligned}$$

$$\partial L / \partial \mathbf{W}^{(2)}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{W}^{(2)}} &= \begin{pmatrix} \frac{\partial L}{\partial W_{1,1}^{(2)}} & \frac{\partial L}{\partial W_{1,2}^{(2)}} & \cdots & \frac{\partial L}{\partial W_{1,r}^{(2)}} \\ \frac{\partial L}{\partial W_{2,1}^{(2)}} & \frac{\partial L}{\partial W_{2,2}^{(2)}} & \cdots & \frac{\partial L}{\partial W_{2,r}^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial W_{q,1}^{(2)}} & \frac{\partial L}{\partial W_{q,2}^{(2)}} & \cdots & \frac{\partial L}{\partial W_{q,r}^{(2)}} \end{pmatrix} \\ \frac{\partial L}{\partial W_{1,1}^{(2)}} &= \frac{\partial}{\partial W_{1,1}^{(2)}} \frac{1}{r} \left\{ [t_1 - (y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \cdots + y_q^{(2)} W_{q,1}^{(2)} + B_1^{(2)})]^2 \right. \\ &\quad \left. [t_2 - (y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \cdots + y_q^{(2)} W_{q,2}^{(2)} + B_2^{(2)})]^2 + \cdots \right\}\end{aligned}$$

$$\begin{aligned}
& \left[t_r - (y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \dots + y_q^{(2)} W_{q,r}^{(2)} + B_r^{(2)}) \right]^2 \Big\} \\
&= \frac{2}{r} \left(y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \dots + y_q^{(2)} W_{q,1}^{(2)} + B_1^{(2)} - t_1 \right) y_1^{(2)} = \frac{2}{r} (y_1 - t_1) y_1^{(2)} \\
\therefore \frac{\partial L}{\partial \mathbf{W}^{(2)}} &= \frac{2}{r} \begin{pmatrix} (y_1 - t_1) y_1^{(2)} & (y_2 - t_2) y_1^{(2)} & \dots & (y_r - t_r) y_1^{(2)} \\ (y_1 - t_1) y_2^{(2)} & (y_2 - t_2) y_2^{(2)} & \dots & (y_r - t_r) y_2^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ (y_1 - t_1) y_q^{(2)} & (y_2 - t_2) y_q^{(2)} & \dots & (y_r - t_r) y_q^{(2)} \end{pmatrix} \\
&= \frac{2}{r} \begin{pmatrix} y_1^{(2)} \\ y_2^{(2)} \\ \vdots \\ y_q^{(2)} \end{pmatrix} (y_1 - t_1 \ y_2 - t_2 \ \dots \ y_r - t_r) = \mathbf{y}^{(2)\top} \frac{\partial L}{\partial \mathbf{y}}
\end{aligned}$$

$$\partial L / \partial \mathbf{y}^{(2)}$$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{y}^{(2)}} &= \left(\frac{\partial L}{\partial y_1^{(2)}} \quad \frac{\partial L}{\partial y_2^{(2)}} \quad \dots \quad \frac{\partial L}{\partial y_q^{(2)}} \right) \\
\frac{\partial L}{\partial y_1^{(2)}} &= \frac{\partial}{\partial y_1^{(2)}} \frac{1}{r} \left\{ \left[t_1 - (y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \dots + y_q^{(2)} W_{q,1}^{(2)} + B_1^{(2)}) \right]^2 \right. \\
&\quad \left. [t_2 - (y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \dots + y_q^{(2)} W_{q,2}^{(2)} + B_2^{(2)})]^2 + \dots \right. \\
&\quad \left. [t_r - (y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \dots + y_q^{(2)} W_{q,r}^{(2)} + B_r^{(2)})]^2 \right\} \\
&= \frac{2}{r} \left\{ \left(y_1^{(2)} W_{1,1}^{(2)} + y_2^{(2)} W_{2,1}^{(2)} + \dots + y_q^{(2)} W_{q,1}^{(2)} + B_1^{(2)} - t_1 \right) W_{1,1}^{(2)} + \right. \\
&\quad \left(y_1^{(2)} W_{1,2}^{(2)} + y_2^{(2)} W_{2,2}^{(2)} + \dots + y_q^{(2)} W_{q,2}^{(2)} + B_2^{(2)} - t_2 \right) W_{1,2}^{(2)} + \dots \\
&\quad \left. \left(y_1^{(2)} W_{1,r}^{(2)} + y_2^{(2)} W_{2,r}^{(2)} + \dots + y_q^{(2)} W_{q,r}^{(2)} + B_r^{(2)} - t_r \right) W_{1,r}^{(2)} \right\} \\
&= \frac{2}{r} \left\{ (y_1 - t_1) W_{1,1}^{(2)} + (y_2 - t_2) W_{1,2}^{(2)} + \dots + (y_r - t_r) W_{1,r}^{(2)} \right\} \\
\therefore \frac{\partial L}{\partial \mathbf{y}^{(2)}} &= \frac{2}{r} (y_1 - t_1 \ y_2 - t_2 \ \dots \ y_r - t_r) \begin{pmatrix} W_{1,1}^{(2)} & W_{2,1}^{(2)} & \dots & W_{q,1}^{(2)} \\ W_{1,2}^{(2)} & W_{2,2}^{(2)} & \dots & W_{q,2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,r}^{(2)} & W_{2,r}^{(2)} & \dots & W_{q,r}^{(2)} \end{pmatrix} = \frac{\partial L}{\partial \mathbf{y}} \mathbf{W}^{(2)\top}
\end{aligned}$$

$$\partial L / \partial \mathbf{B}^{(1)}$$

$$\frac{\partial L}{\partial \mathbf{B}^{(1)}} = \left(\frac{\partial L}{\partial B_1^{(1)}} \quad \frac{\partial L}{\partial B_2^{(1)}} \quad \dots \quad \frac{\partial L}{\partial B_q^{(1)}} \right)$$

$$\begin{aligned}
\frac{\partial L}{\partial B_1^{(1)}} &= \frac{\partial}{\partial B_1^{(1)}} \frac{1}{r} \left\{ [t_1 - ((y_1^{(1)} + B_1^{(1)})W_{1,1}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,1}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,1}^{(2)} + B_1^{(2)})]^2 \right. \\
&\quad + [t_2 - ((y_1^{(1)} + B_1^{(1)})W_{1,2}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,2}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,2}^{(2)} + B_2^{(2)})]^2 \\
&\quad + \dots + [t_r - ((y_1^{(1)} + B_1^{(1)})W_{1,r}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,r}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,r}^{(2)} + B_r^{(2)})]^2 \left. \right\} \\
&= \frac{2}{r} \left\{ [(y_1^{(1)} + B_1^{(1)})W_{1,1}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,1}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,1}^{(2)} + B_1^{(2)} - t_1]W_{1,1}^{(2)} \right. \\
&\quad + [(y_1^{(1)} + B_1^{(1)})W_{1,2}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,2}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,2}^{(2)} + B_2^{(2)} - t_2]W_{1,2}^{(2)} + \dots \\
&\quad + [(y_1^{(1)} + B_1^{(1)})W_{1,r}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,r}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,r}^{(2)} + B_r^{(2)} - t_r]W_{1,r}^{(2)} \left. \right\} \\
&= \frac{2}{r} \left\{ (y_1 - t_1) W_{1,1}^{(2)} + (y_2 - t_2) W_{1,2}^{(2)} + \dots + (y_r - t_r) W_{1,r}^{(2)} \right\} \\
\therefore \frac{\partial L}{\partial B_1^{(1)}} &= \frac{2}{r} (y_1 - t_1 \ y_2 - t_2 \ \dots \ y_r - t_r) \begin{pmatrix} W_{1,1}^{(2)} & W_{2,1}^{(2)} & \dots & W_{q,1}^{(2)} \\ W_{1,2}^{(2)} & W_{2,2}^{(2)} & \dots & W_{q,2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,r}^{(2)} & W_{2,r}^{(2)} & \dots & W_{q,r}^{(2)} \end{pmatrix} = \frac{\partial L}{\partial \mathbf{y}} \mathbf{W}^{(2)\top} = \frac{\partial L}{\partial \mathbf{y}^{(2)}}
\end{aligned}$$

$$\boxed{\frac{\partial L}{\partial \mathbf{y}^{(1)}}}$$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{y}^{(1)}} &= \left(\frac{\partial L}{\partial y_1^{(1)}} \quad \frac{\partial L}{\partial y_2^{(1)}} \quad \dots \quad \frac{\partial L}{\partial y_q^{(1)}} \right) \\
\frac{\partial L}{\partial y_1^{(1)}} &= \frac{\partial}{\partial y_1^{(1)}} \frac{1}{r} \left\{ [t_1 - ((y_1^{(1)} + B_1^{(1)})W_{1,1}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,1}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,1}^{(2)} + B_1^{(2)})]^2 \right. \\
&\quad + [t_2 - ((y_1^{(1)} + B_1^{(1)})W_{1,2}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,2}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,2}^{(2)} + B_2^{(2)})]^2 \\
&\quad + \dots + [t_r - ((y_1^{(1)} + B_1^{(1)})W_{1,r}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,r}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,r}^{(2)} + B_r^{(2)})]^2 \left. \right\} \\
&= \frac{2}{r} \left\{ [(y_1^{(1)} + B_1^{(1)})W_{1,1}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,1}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,1}^{(2)} + B_1^{(2)} - t_1]W_{1,1}^{(2)} \right. \\
&\quad + [(y_1^{(1)} + B_1^{(1)})W_{1,2}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,2}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,2}^{(2)} + B_2^{(2)} - t_2]W_{1,2}^{(2)} + \dots \\
&\quad + [(y_1^{(1)} + B_1^{(1)})W_{1,r}^{(2)} + (y_2^{(1)} + B_2^{(1)})W_{2,r}^{(2)} + \dots + (y_q^{(1)} + B_q^{(1)})W_{q,r}^{(2)} + B_r^{(2)} - t_r]W_{1,r}^{(2)} \left. \right\} \\
&= \frac{2}{r} \left\{ (y_1 - t_1) W_{1,1}^{(2)} + (y_2 - t_2) W_{1,2}^{(2)} + \dots + (y_r - t_r) W_{1,r}^{(2)} \right\} \\
\therefore \frac{\partial L}{\partial \mathbf{y}^{(1)}} &= \frac{2}{r} (y_1 - t_1 \ y_2 - t_2 \ \dots \ y_r - t_r) \begin{pmatrix} W_{1,1}^{(2)} & W_{2,1}^{(2)} & \dots & W_{q,1}^{(2)} \\ W_{1,2}^{(2)} & W_{2,2}^{(2)} & \dots & W_{q,2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,r}^{(2)} & W_{2,r}^{(2)} & \dots & W_{q,r}^{(2)} \end{pmatrix} = \frac{\partial L}{\partial \mathbf{y}} \mathbf{W}^{(2)\top} = \frac{\partial L}{\partial \mathbf{y}^{(2)}}
\end{aligned}$$

$$\partial L / \partial \mathbf{W}^{(1)}$$

$$\frac{\partial L}{\partial \mathbf{W}^{(1)}} = \begin{pmatrix} \frac{\partial L}{\partial W_{1,1}^{(1)}} & \frac{\partial L}{\partial W_{1,2}^{(1)}} & \cdots & \frac{\partial L}{\partial W_{1,q}^{(1)}} \\ \frac{\partial L}{\partial W_{2,1}^{(1)}} & \frac{\partial L}{\partial W_{2,2}^{(1)}} & \cdots & \frac{\partial L}{\partial W_{2,q}^{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial W_{p,1}^{(1)}} & \frac{\partial L}{\partial W_{p,2}^{(1)}} & \cdots & \frac{\partial L}{\partial W_{p,q}^{(1)}} \end{pmatrix}$$

$$\frac{\partial L}{\partial W_{1,1}^{(1)}} = \frac{\partial}{\partial W_{1,1}^{(1)}} \frac{1}{r} \left\{ \left[t_1 - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,1}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,1}^{(2)} + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,1}^{(2)} + B_1^{(2)} \right)^2 + \left[t_2 - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,2}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,2}^{(2)} + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,2}^{(2)} + B_2^{(2)} \right)^2 + \left[t_r - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,r}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,r}^{(2)} + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,r}^{(2)} + B_r^{(2)} \right)^2 + \right] \right] \right\}^2 + \frac{2}{r} \left\{ \left[(x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,1}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,1}^{(2)} + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,1}^{(2)} + B_1^{(2)} - t_1 \right] x_1 W_{1,1}^{(2)} + \left[(x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,2}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,2}^{(2)} + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,2}^{(2)} + B_2^{(2)} - t_2 \right] x_1 W_{1,2}^{(2)} + \cdots + \left[(x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,r}^{(2)} + (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,r}^{(2)} + \cdots + (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,r}^{(2)} + B_r^{(2)} - t_r \right] x_1 W_{1,r}^{(2)} + \cdots + \right\}$$

$$\begin{aligned}
& \left(x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)} \right) W_{2,r}^{(2)} + \cdots + \\
& \left(x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)} \right) W_{q,r}^{(2)} + B_r^{(2)} - t_r \Big] x_1 W_{1,r}^{(2)} \Big\} \\
& = \frac{2}{r} \left\{ (y_1 - t_1) W_{1,1}^{(2)} + (y_2 - t_2) W_{1,2}^{(2)} + \cdots + (y_r - t_r) W_{1,r}^{(2)} \right\} x_1 \\
& \vdots \\
& \left(\begin{array}{l} \left\{ (y_1 - t_1) W_{1,1}^{(2)} + (y_2 - t_2) W_{1,2}^{(2)} + \cdots + (y_r - t_r) W_{1,r}^{(2)} \right\} x_1 \\ \left\{ (y_1 - t_1) W_{2,1}^{(2)} + (y_2 - t_2) W_{2,2}^{(2)} + \cdots + (y_r - t_r) W_{2,r}^{(2)} \right\} x_1 \\ \cdots \\ \left\{ (y_1 - t_1) W_{q,1}^{(2)} + (y_2 - t_2) W_{q,2}^{(2)} + \cdots + (y_r - t_r) W_{q,r}^{(2)} \right\} x_1 \\ \left\{ (y_1 - t_1) W_{1,1}^{(2)} + (y_2 - t_2) W_{1,2}^{(2)} + \cdots + (y_r - t_r) W_{1,r}^{(2)} \right\} x_2 \\ \left\{ (y_1 - t_1) W_{2,1}^{(2)} + (y_2 - t_2) W_{2,2}^{(2)} + \cdots + (y_r - t_r) W_{2,r}^{(2)} \right\} x_2 \\ \cdots \\ \left\{ (y_1 - t_1) W_{q,1}^{(2)} + (y_2 - t_2) W_{q,2}^{(2)} + \cdots + (y_r - t_r) W_{q,r}^{(2)} \right\} x_2 \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \left\{ (y_1 - t_1) W_{1,1}^{(2)} + (y_2 - t_2) W_{1,2}^{(2)} + \cdots + (y_r - t_r) W_{1,r}^{(2)} \right\} x_p \\ \left\{ (y_1 - t_1) W_{2,1}^{(2)} + (y_2 - t_2) W_{2,2}^{(2)} + \cdots + (y_r - t_r) W_{2,r}^{(2)} \right\} x_p \\ \cdots \\ \left\{ (y_1 - t_1) W_{q,1}^{(2)} + (y_2 - t_2) W_{q,2}^{(2)} + \cdots + (y_r - t_r) W_{q,r}^{(2)} \right\} x_p \end{array} \right) \\
& \frac{\partial L}{\partial \mathbf{W}^{(1)}} = \frac{2}{r} \left(\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_p \end{array} \right) (y_1 - t_1 \quad y_2 - t_2 \quad \cdots \quad y_r - t_r) \left(\begin{array}{cccc} W_{1,1}^{(2)} & W_{2,1}^{(2)} & \cdots & W_{q,1}^{(2)} \\ W_{1,2}^{(2)} & W_{2,2}^{(2)} & \cdots & W_{q,2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,r}^{(2)} & W_{2,r}^{(2)} & \cdots & W_{q,r}^{(2)} \end{array} \right) \\
& = \mathbf{x}^\top \frac{\partial L}{\partial \mathbf{y}} \mathbf{W}^{(2)\top} = \mathbf{x}^\top \frac{\partial L}{\partial \mathbf{y}^{(2)}}
\end{aligned}$$

$$\partial L / \partial \mathbf{x}$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{x}} &= \left(\frac{\partial L}{\partial x_1} \quad \frac{\partial L}{\partial x_2} \quad \cdots \quad \frac{\partial L}{\partial x_p} \right) \\ \frac{\partial L}{\partial x_1} &= \frac{\partial}{\partial x_1} \frac{1}{r} \left\{ \left[t_1 - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,1}^{(2)} + \right. \right. \right. \\ &\quad \left. \left. \left. (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,1}^{(2)} + \cdots + \right. \right. \right. \\ &\quad \left. \left. \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,1}^{(2)} + B_1^{(2)} \right) \right]^2 + \right. \\ &\quad \left[t_2 - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,2}^{(2)} + \right. \right. \\ &\quad \left. \left. (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,2}^{(2)} + \cdots + \right. \right. \\ &\quad \left. \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,2}^{(2)} + B_2^{(2)} \right) \right]^2 + \cdots + \right. \\ &\quad \left[t_r - \left((x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,r}^{(2)} + \right. \right. \\ &\quad \left. \left. (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,r}^{(2)} + \cdots + \right. \right. \\ &\quad \left. \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,r}^{(2)} + B_r^{(2)} \right) \right]^2 \right\} \\ &= \frac{2}{r} \left\{ \left[(x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,1}^{(2)} + \right. \right. \\ &\quad \left. \left. (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,1}^{(2)} + \cdots + \right. \right. \\ &\quad \left. \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,1}^{(2)} + B_1^{(2)} - t_1 \right] \right. \\ &\quad \times \left(W_{1,1}^{(1)} W_{1,1}^{(2)} + W_{1,2}^{(1)} W_{2,1}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,1}^{(2)} \right) + \\ &\quad \left[(x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,2}^{(2)} + \right. \\ &\quad \left. (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,2}^{(2)} + \cdots + \right. \\ &\quad \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,2}^{(2)} + B_2^{(2)} - t_2 \right] \\ &\quad \times \left(W_{1,1}^{(1)} W_{1,2}^{(2)} + W_{1,2}^{(1)} W_{2,2}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,2}^{(2)} \right) + \cdots + \\ &\quad \left[(x_1 W_{1,1}^{(1)} + x_2 W_{2,1}^{(1)} + \cdots + x_p W_{p,1}^{(1)} + B_1^{(1)}) W_{1,r}^{(2)} + \right. \\ &\quad \left. (x_1 W_{1,2}^{(1)} + x_2 W_{2,2}^{(1)} + \cdots + x_p W_{p,2}^{(1)} + B_2^{(1)}) W_{2,r}^{(2)} + \cdots + \right. \\ &\quad \left. (x_1 W_{1,q}^{(1)} + x_2 W_{2,q}^{(1)} + \cdots + x_p W_{p,q}^{(1)} + B_q^{(1)}) W_{q,r}^{(2)} + B_r^{(2)} - t_r \right] \\ &\quad \times \left(W_{1,1}^{(1)} W_{1,r}^{(2)} + W_{1,2}^{(1)} W_{2,r}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,r}^{(2)} \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{2}{r} \left\{ \begin{aligned}
&(y_1 - t_1) \left(W_{1,1}^{(1)} W_{1,1}^{(2)} + W_{1,2}^{(1)} W_{2,1}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,1}^{(2)} \right) + \\
&(y_2 - t_2) \left(W_{1,1}^{(1)} W_{1,2}^{(2)} + W_{1,2}^{(1)} W_{2,2}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,2}^{(2)} \right) + \cdots + \\
&(y_r - t_r) \left(W_{1,1}^{(1)} W_{1,r}^{(2)} + W_{1,2}^{(1)} W_{2,r}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,r}^{(2)} \right)
\end{aligned} \right\}
\end{aligned}$$

⋮

$$\begin{aligned}
&\frac{\partial L}{\partial \mathbf{x}} = \frac{2}{r} \begin{pmatrix} \begin{aligned} &(y_1 - t_1) \left(W_{1,1}^{(1)} W_{1,1}^{(2)} + W_{1,2}^{(1)} W_{2,1}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,1}^{(2)} \right) \\ &+ (y_2 - t_2) \left(W_{1,1}^{(1)} W_{1,2}^{(2)} + W_{1,2}^{(1)} W_{2,2}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,2}^{(2)} \right) \\ &+ \cdots + (y_r - t_r) \left(W_{1,1}^{(1)} W_{1,r}^{(2)} + W_{1,2}^{(1)} W_{2,r}^{(2)} + \cdots + W_{1,q}^{(1)} W_{q,r}^{(2)} \right) \end{aligned} \\ &\vdots \\ &\begin{aligned} &(y_1 - t_1) \left(W_{p,1}^{(1)} W_{1,1}^{(2)} + W_{p,2}^{(1)} W_{2,1}^{(2)} + \cdots + W_{p,q}^{(1)} W_{q,1}^{(2)} \right) \\ &+ (y_2 - t_2) \left(W_{p,1}^{(1)} W_{1,2}^{(2)} + W_{p,2}^{(1)} W_{2,2}^{(2)} + \cdots + W_{p,q}^{(1)} W_{q,2}^{(2)} \right) \\ &+ \cdots + (y_r - t_r) \left(W_{p,1}^{(1)} W_{1,r}^{(2)} + W_{p,2}^{(1)} W_{2,r}^{(2)} + \cdots + W_{p,q}^{(1)} W_{q,r}^{(2)} \right) \end{aligned} \end{pmatrix}^\top \\ &= \frac{2}{r} (y_1 - t_1 \ y_2 - t_2 \ \cdots \ y_r - t_r) \begin{pmatrix} W_{1,1}^{(2)} & W_{2,1}^{(2)} & \cdots & W_{q,1}^{(2)} \\ W_{1,2}^{(2)} & W_{2,2}^{(2)} & \cdots & W_{q,2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,r}^{(2)} & W_{2,r}^{(2)} & \cdots & W_{q,r}^{(2)} \end{pmatrix} \begin{pmatrix} W_{1,1}^{(1)} & W_{2,1}^{(1)} & \cdots & W_{p,1}^{(1)} \\ W_{1,2}^{(1)} & W_{2,2}^{(1)} & \cdots & W_{p,2}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,q}^{(1)} & W_{2,q}^{(1)} & \cdots & W_{p,q}^{(1)} \end{pmatrix} \\
&= \frac{\partial L}{\partial \mathbf{y}} \mathbf{W}^{(2)\top} \mathbf{W}^{(1)\top} = \frac{\partial L}{\partial \mathbf{y}^{(2)}} \mathbf{W}^{(1)\top}
\end{aligned}$$