本節では摂動法 (permutation) による勾配推定について説明する．摂動法に含まれる手法は複数あるが，総じて次のような手法を指す．まず，あるモデル（ネットワーク）を用意し，その目的関数を $\mathcal{L}$ とする．次にモデルのパラメータや活動にランダムな微小変化（摂動）$\mathbf{v}$ を加え，摂動を受ける前後の目的関数の変化量 $\delta\mathcal{L}$ を取得する．この $\delta\mathcal{L}$ や $\mathbf{v}$ およびモデルの活動等を用いてパラメータを更新するのが摂動法である．

代表的なニューラルネットワークの摂動法はノード摂動法 (Node perturbation; NP) と重み摂動法 (weight perturbation; WP) である．ノード摂動法は各ノード（ニューロン）の活動に摂動を加える手法であり，重み摂動法は各パラメータ（シナプス結合等）に摂動を加える手法である．

まず，以下のように順伝播を行うネットワークを設定する $(\ell = 1, \ldots, L)$

$$\text{入力層：} \mathbf{z}_1 = \mathbf{x} \tag{1}$$
$$\text{隠れ層：} \mathbf{a}_\ell = \mathbf{W}_\ell \mathbf{z}_\ell + \mathbf{b}_\ell \tag{2}$$
$$\mathbf{z}_{\ell+1} = f_\ell(\mathbf{a}_\ell) \tag{3}$$
$$\text{出力層：} \hat{\mathbf{y}} = \mathbf{z}_{L+1} \tag{4}$$

損失は $\mathcal{L}(\mathbf{z}_{L+1}; \mathbf{x})$ とする．それぞれの手法において，以下のようにネットワークを摂動する．

$$\text{重み摂動法:} \quad \tilde{\mathbf{z}}_{\ell+1} = f_\ell\left((\mathbf{W}_\ell + \mathbf{V}_\ell)\tilde{\mathbf{z}}_\ell + \mathbf{b}_\ell + \mathbf{v}_\ell\right) \tag{5}$$
$$\text{ノード摂動法:} \quad \tilde{\mathbf{z}}_{\ell+1} = f_\ell\left(\mathbf{W}_\ell \tilde{\mathbf{z}}_\ell + \mathbf{b}_\ell + \mathbf{v}_\ell\right) \tag{6}$$

目的関数の変化量を

$$\delta\mathcal{L} = \mathcal{L}(\tilde{\mathbf{z}}_{L+1}; \mathbf{x}) - \mathcal{L}(\mathbf{z}_{L+1}; \mathbf{x}) \tag{7}$$

とする．SGD でパラメータを行う場合，

$$\text{重み摂動法:} \quad \Delta\mathbf{W}_\ell^{\mathrm{WP}} = -\eta\frac{\delta\mathcal{L}}{\sigma}\mathbf{V}_\ell, \qquad \Delta\mathbf{b}_\ell^{\mathrm{WP}} = -\eta\frac{\delta\mathcal{L}}{\sigma}\mathbf{v}_\ell \tag{8}$$
$$\text{ノード摂動法:} \quad \Delta\mathbf{W}_\ell^{\mathrm{NP}} = -\eta\frac{\delta\mathcal{L}}{\sigma}\mathbf{v}_\ell\mathbf{z}_\ell^\top, \qquad \Delta\mathbf{b}_\ell^{\mathrm{NP}} = -\eta\frac{\delta\mathcal{L}}{\sigma}\mathbf{v}_\ell \tag{9}$$

でパラメータを更新する．

1　不偏推定量であることの証明

各手法の更新則が勾配の不偏推定量 (unbiased estimator) であることを示す．まず方向微分 (Directional derivative) を導入する．関数 f について点 $\mathbf{u}$ における方向 $\mathbf{v}$ の方向微分は

$$\nabla_{\mathbf{v}} f(\mathbf{u}) := \lim_{h\to 0}\frac{f(\mathbf{u}+h\mathbf{v}) - f(\mathbf{u})}{h} \tag{10}$$

として定義される．また f が点 $\mathbf{u}$ において微分可能なら

$$\nabla_{\mathbf{v}} f(\mathbf{u}) = \nabla f(\mathbf{u}) \cdot \mathbf{v} \left(= \frac{\partial f(\mathbf{u})}{\partial \mathbf{u}} \cdot \mathbf{v} \right) \tag{11}$$

が成り立つ．ここで，$\nabla f(\mathbf{u}) \cdot \mathbf{v}$ を Jacobian-vector product (JVP) と呼び，$f(\mathbf{u}) \in \mathbb{R}$ の場合，$\nabla f(\mathbf{u}) \cdot \mathbf{v} \in \mathbb{R}$ となる．この JVP を有限差分 (finite difference) を用いて近似計算すると[*1]，

$$\nabla f(\mathbf{u}) \cdot \mathbf{v} \approx \frac{f(\mathbf{u} + \epsilon \mathbf{v}) - f(\mathbf{u})}{\epsilon} \tag{12}$$

となる $(0 < \epsilon \ll 1)$.

まず，重み摂動法について考える．モデルのパラメータを $\boldsymbol{\theta} \in \mathbb{R}^P$ とする．これは $\mathbf{W}_\ell$ および $\mathbf{b}_\ell$ をまとめたベクトルであり，P はパラメータ空間の次元である．$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_P)$ $(\in \mathbb{R}^P)$ とすると，$\sigma \to 0$ の場合，

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \cdot \mathbf{v} = \frac{\mathcal{L}(\boldsymbol{\theta} + \sigma \mathbf{v}) - \mathcal{L}(\boldsymbol{\theta})}{\sigma} = \frac{\delta \mathcal{L}}{\sigma} \tag{13}$$

となるので，

$$\mathbb{E}\left[\frac{\delta \mathcal{L}}{\sigma} \mathbf{v} \right] = \cdot \mathbb{E}\left[\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \cdot \mathbf{v} \right) \mathbf{v} \right] \tag{14}$$

$$= \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \mathbb{E}[\mathbf{v}\mathbf{v}^\top] = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \tag{15}$$

が成立する．SGD でパラメータ更新する場合は

$$\mathbb{E}[\Delta \mathbf{W}_\ell] = -\eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}_\ell}, \quad \mathbb{E}[\Delta \mathbf{b}_\ell] = -\eta \frac{\partial \mathcal{L}}{\partial \mathbf{b}_\ell} \tag{16}$$

であればいいので，$(\boldsymbol{\theta}, \mathbf{v}) \to (\mathbf{W}_\ell, \mathbf{V}_\ell), (\mathbf{b}_\ell, \mathbf{v}_\ell)$ と置き換えて

$$\Delta \mathbf{W}_\ell^{\mathrm{WP}} := -\eta \frac{\delta \mathcal{L}}{\sigma} \mathbf{V}_\ell, \quad \Delta \mathbf{b}_\ell^{\mathrm{WP}} := -\eta \frac{\delta \mathcal{L}}{\sigma} \mathbf{v}_\ell \tag{17}$$

となる．ノード摂動法は重み摂動法におけるバイアス項のみを摂動すると解釈できるため，$\Delta \mathbf{b}_\ell^{\mathrm{NP}} := \Delta \mathbf{b}_\ell^{\mathrm{WP}}$ である．ここで

[*1] JVP は順方向自動微分 (Forward-mode Automatic Differentiation; Forward AD) により計算でき，有限差分法よりも数値的に安定する (順方向自動微分は Python ライブラリの JAX 等に実装されている)．Forward Gradient は順方向自動微分を採用して重み摂動法をより安定させた手法である．

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_\ell} = \frac{\partial \mathcal{L}}{\partial \mathbf{z}_\ell}\frac{\partial \mathbf{z}_\ell}{\partial \mathbf{a}_\ell}\frac{\partial \mathbf{a}_\ell}{\partial \mathbf{W}_\ell} \tag{18}$$

$$= \left(\frac{\partial \mathcal{L}}{\partial \mathbf{z}_\ell}\frac{\partial \mathbf{z}_\ell}{\partial \mathbf{a}_\ell}\frac{\partial \mathbf{a}_\ell}{\partial \mathbf{b}_\ell}\right)\mathbf{z}_\ell^\top \quad \left(\because \frac{\partial \mathbf{a}_\ell}{\partial \mathbf{b}_\ell} = \mathbf{1}\right) \tag{19}$$

$$= \frac{\partial \mathcal{L}}{\partial \mathbf{b}_\ell}\mathbf{z}_\ell^\top \tag{20}$$

が成り立つので，ノード摂動法の更新則は

$$\Delta \mathbf{W}_\ell^{\mathrm{NP}} := \Delta \mathbf{b}_\ell^{\mathrm{NP}}\mathbf{z}_\ell^\top = -\eta\frac{\delta \mathcal{L}}{\sigma}\mathbf{v}_\ell \mathbf{z}_\ell^\top, \quad \Delta \mathbf{b}_\ell^{\mathrm{NP}} := -\eta\frac{\delta \mathcal{L}}{\sigma}\mathbf{v}_\ell \tag{21}$$

と設定できる．