Доказать утверждение из [1] :

Пусть L – дифференцируемая функция, такая что все стационарные точки L являются локальными минимумами. Пусть также гессиан $\mathbf{H}^{-1}$ функции потерь L является обратимым в каждой стационарной точке
Тогда

$$\nabla_{\mathbf{h}} Q(T(\boldsymbol{\theta}_0, \mathbf{h}), \mathbf{h}) = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}^\eta, \mathbf{h}) - \nabla_{\mathbf{h}} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^\eta, \mathbf{h})^T \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^\eta, \mathbf{h}).$$

Доказательство
В стационарных точках:

$$\nabla_{\boldsymbol{\theta}} L(T(\boldsymbol{\theta}_0, \mathbf{h})) = 0$$

Продифференциируем по $\mathbf{h}$ и воспользуемся chain rule [1]

$$\nabla_{\mathbf{h}} \nabla_\theta L(T(\boldsymbol{\theta}_0, \mathbf{h})) = \nabla_{\boldsymbol{\theta}, \mathbf{h}} L(\boldsymbol{\theta}^\eta, \mathbf{h}) + \nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^\eta, \mathbf{h}) \frac{\partial \boldsymbol{\theta}}{\partial \mathbf{h}} = 0$$

$$\nabla_{\boldsymbol{\theta}, \mathbf{h}} L(\boldsymbol{\theta}^\eta, \mathbf{h}) = -\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^\eta, \mathbf{h}) \frac{\partial \boldsymbol{\theta}}{\partial \mathbf{h}} = -\mathbf{H} \frac{\partial \boldsymbol{\theta}}{\partial \mathbf{h}}$$

Поскольку $\mathbf{H}^{-1}$ обратим во всех стационарных точках, домножим на $-\mathbf{H}^{-1}$

$$\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{h}} = -\mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}, \mathbf{h}} L(\boldsymbol{\theta}^\eta, \mathbf{h}).$$

Возьмём градиент Q по $\mathbf{h}$

$$\nabla_{\mathbf{h}} Q(T(\theta_0, \mathbf{h})) = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}^\eta, \mathbf{h}) + \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^\eta, \mathbf{h})^T \frac{\partial \boldsymbol{\theta}}{\partial \mathbf{h}}$$

Подставим $\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{h}}$ из предыдущего выражения

$$\nabla_{\mathbf{h}} Q(T(\boldsymbol{\theta}_0, \mathbf{h})) = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}^\eta, \mathbf{h}) - \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^\eta, \mathbf{h})^T \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}, \mathbf{h}} L(\boldsymbol{\theta}^\eta, \mathbf{h}) =$$

$$= \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}^\eta, \mathbf{h}) - \nabla_{\boldsymbol{\theta}, \mathbf{h}} L(\boldsymbol{\theta}^\eta, \mathbf{h})^T \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^\eta, \mathbf{h})$$

Список литературы

[1] *F. Pedregosa* Hyperparameter optimization with approximate gradient // arxiv.org, 2016