Утверждение 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

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

Доказательство

Рассматриваем $\mathcal{Q}(T(\boldsymbol{\theta}_0,\mathbf{h})) \Longrightarrow L(T(\boldsymbol{\theta}_0,\mathbf{h}))$.

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

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

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

$$\frac{\partial\boldsymbol{\theta}}{\partial\mathbf{h}} = -(\nabla^2_{\boldsymbol{\theta}}L(\boldsymbol{\theta}^{\eta},\mathbf{h}))^{-1}\nabla_{\boldsymbol{\theta}}\nabla_{\mathbf{h}}L(\boldsymbol{\theta}^{\eta},\mathbf{h})$$

Также известно, что $T(\boldsymbol{\theta},\mathbf{h}) = \boldsymbol{\theta} - \beta\nabla L(\boldsymbol{\theta},\mathbf{h})$

$$\Longrightarrow \nabla_{\mathbf{h}}\mathcal{Q}(T(\boldsymbol{\theta}_0,\mathbf{h})) = \nabla_{\mathbf{h}}\mathcal{Q}(\boldsymbol{\theta}^{\eta},\mathbf{h}) + \nabla_{\boldsymbol{\theta}}\mathcal{Q}(\boldsymbol{\theta}^{\eta},\mathbf{h})^T\frac{\partial\boldsymbol{\theta}}{\partial\mathbf{h}}$$

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