mt math

Tianyi Zhou

February 2022

1 Question

$$\begin{array}{ll} \min_A & L\left(D_{val}; V^*\left(W^*(A)\right)\right) \\ \text{s.t} & V^*\left(W^*(A)\right) = \operatorname{argmin}_V [\sum_{i=1}^{M} l\left(u_i, g\left(u_i; W^*(A)\right); V\right) + \lambda \sum_{j=1}^{K} l\left(p_j, q_j; V\right)] \\ & W^*(A) = \operatorname{argmin}_W \sum_{i=1}^{N} a_i l\left(x_i, y_i; W\right) \end{array}$$

2 Optimization

One-step gradient decent of $\widetilde{W}(A)$:

$$\widetilde{W}(A) \doteq \bar{W}(A) = W - \xi_W \nabla_W \sum_{i=1}^{N} a_i l\left(x_i, y_i, W\right)$$

Approximation of $\widetilde{V}(\widetilde{W}(A))$:

$$\widetilde{V}(\widetilde{W}(A)) \doteq \bar{V}(\bar{W}(A)) = V - \xi_v \nabla_v [\sum_{i=1}^{M} l\left(u_i, g\left(u_i; W^*(A)\right); V\right) + \lambda \sum_{j=1}^{K} l\left(p_j, q_j; V\right)]$$

For architecture:

$$\min_A L\left(D_{val}, \bar{V}(\bar{W}(A))\right)$$

$$A = A - \xi_A \nabla_A \bar{V}(\bar{W}(A)) \cdot \nabla_{\bar{V}} L\left(D_{val}, \bar{V}\right) \tag{1}$$

Left part in equation 1

$$\begin{aligned} \nabla_A \bar{V}(\bar{W}(A)) &= \nabla_A \left(V - \xi_v \nabla_v [\sum_{i=1}^{M} l\left(u_i, g\left(u_i; W^*(A)\right); V\right) + \lambda \sum_{j=1}^{K} l\left(p_j, q_j; V\right)]\right) \\ &= \xi_V \nabla_V \nabla_A \left(\sum_{i=1}^{M} l\left(u_i, g\left(u_i; W^*(A)\right); V\right)\right) \\ &= \xi_V \cdot \nabla_V \left[\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V\right) \cdot \nabla_A \bar{W}(A)\right] \end{aligned}$$

Now equation 1 becomes:

$$A = A - \xi_A \xi_V \cdot \nabla_V \nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V\right) \cdot \nabla_{\bar{V}} L\left(D_{val}, \bar{V}\right) \cdot \nabla_A \bar{W}(A) \tag{2}$$

Using Finite difference method to approximate the first part in equation 2

$$\nabla_V \nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V\right) \cdot \nabla_{\bar{V}} L\left(D_{val}, \bar{V}\right)$$

$$= \frac{1}{2\alpha_W} \left(\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^+\right) - \nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^-\right)\right)$$

where:

$$V^{\pm} = V \pm \alpha_W \nabla_{\bar{V}} L\left(D_{val}, \bar{V}\right)$$
$$\alpha_W = \frac{0.01}{\left\|\nabla_{\bar{V}} l\left(D_{val}, \bar{V}\right)\right\|_2}$$

For the second term in equation 2

$$\nabla_A \bar{W}(A) = \nabla_A \left(W - \xi \nabla_W \sum_{i=1}^{N} \left(\alpha_i l\left(x_i, y_; W\right)\right)\right.$$
$$= \xi_W \nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W\right)$$

Now equation 2 becomes:

$$A = A - \xi_A \xi_V \left[\frac{1}{2\alpha_W}\left(\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{+}\right) - \nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{-}\right)\right)\right] \cdot \left[\xi_W \nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W\right)\right]$$
$$A = A - \xi_W \xi_A \xi_V \frac{1}{2\alpha_W}\left[\left(\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{+}\right) - \nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{-}\right)\right)\right] \cdot \left[\nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W\right)\right]$$
$$A = A - \xi_W \xi_A \xi_V \frac{1}{2\alpha_W} \cdot R \tag{3}$$

where

$$R = \left(\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{+}\right)\right) \cdot \nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W\right) - \left(\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{-}\right)\right) \cdot \nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W\right)$$

For the first term in R:
Using Finite difference method to approximate

$$\left[\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{+}\right)\right] \cdot \left[\nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W\right)\right]$$
$$= \frac{1}{2\alpha_A}\left(\nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W^{+}\right) - \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W^{-}\right)\right)$$

where:

$$W^{\pm} = W \pm \alpha_A \left(\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{+}\right)\right)$$
$$\alpha_A = \frac{0.01}{\left\|\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{+}\right)\right\|_2}$$

For the second term in R:
Using Finite difference method to approximate

$$\left[\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{-}\right)\right] \cdot \left[\nabla_W \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W\right)\right]$$
$$= \frac{1}{2\alpha_A}\left(\nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W^{+}\right) - \nabla_A \sum_{i=1}^{n} \alpha_i l\left(x_i, y_i; W^{-}\right)\right)$$

where:

$$W^{\pm} = W \pm \alpha_A \left(\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{-}\right)\right)$$
$$\alpha_A = \frac{0.01}{\left\|\nabla_{\bar{W}} \sum_{i=1}^{M} l\left(u_i, g\left(u_i, \bar{W}\right); V^{-}\right)\right\|_2}$$