
Algorithm 1 Minimax SGD

```
1: initialize all hyper-parameters, and initialize the primal variable  $\theta$  and the dual variables  $\mathbf{w}_1$  and  $\mathbf{w}_2$ .
2: for  $t = 1 : T$  do
3:   sample  $\epsilon$  to compute  $\mathbf{v} = \mu + L\epsilon$ .
4:   sample  $\mathbf{w}$  to compute ARD kernel matrix.
5:    $g(\theta) = [P_\alpha(\mathbf{v}|\mathbf{w}), \log \frac{P(\mathbf{D}|\mathbf{v})}{q(\mathbf{v}|\theta)}]$ .
6:    $\mathbf{y} = \mathbf{e}_1 \text{Sigmoid}(\mathbf{e}_2\epsilon + \mathbf{b}_2) + \mathbf{b}_1$ .
7:   update the primal variable:
8:    $\Delta\theta = \frac{\partial g_1(\theta)}{\partial \theta} \mathbf{y} - \frac{\partial g_2(\theta)}{\partial \theta}$ .
9:    $\theta = \theta - \alpha \Delta\theta$ .
10:  update the dual variable:
11:   $\mathbf{e}_1 = \mathbf{e}_1 + \beta (g_1(\theta) + 1/y) \frac{\partial \mathbf{y}}{\partial \mathbf{e}_1}$ .
12:   $\mathbf{e}_2 = \mathbf{e}_2 + \beta (g_1(\theta) + 1/y) \frac{\partial \mathbf{y}}{\partial \mathbf{e}_2}$ .
13:   $\mathbf{b}_1 = \mathbf{b}_1 + \beta (g_1(\theta) + 1/y) \frac{\partial \mathbf{y}}{\partial \mathbf{b}_1}$ .
14:   $\mathbf{b}_2 = \mathbf{b}_2 + \beta (g_1(\theta) + 1/y) \frac{\partial \mathbf{y}}{\partial \mathbf{b}_2}$ .
```

$$\begin{aligned} L(\theta) &= \int \log(E_{p_\alpha(\mathbf{w})}(p_\alpha(\mathbf{v}|\mathbf{w})))q(\mathbf{v}|\theta)d\mathbf{v} + \int \log \frac{p_\alpha(\mathbf{D}|\mathbf{v})}{q(\mathbf{v}|\theta)}q(\mathbf{v}|\theta)d\mathbf{v} \\ &= \int \log \left(\int p_\alpha(\mathbf{v}|\mathbf{w})p_\alpha(\mathbf{w})d\mathbf{w} \right) q(\mathbf{v}|\theta)d\mathbf{v} + \int \log \frac{p_\alpha(\mathbf{D}|\mathbf{v})}{q(\mathbf{v}|\theta)}q(\mathbf{v}|\theta)d\mathbf{v} \\ &= \int \log \left(\int \frac{1}{|K_{nn}|^{1/2}} e^{-\frac{1}{2}(\mu+L\epsilon)K_{nn}^{-1}(\mu+L\epsilon)} e^{-\frac{1}{2}(\log \mathbf{w}-\mu_0)(\mathbf{e}_0^2)^{-1}(\log \mathbf{w}-\mu_0)} d\mathbf{w} \right) \frac{1}{|LL^T|^{1/2}} e^{-\frac{1}{2}\epsilon^T \epsilon} d\mathbf{v} \\ &\quad + \int \log \frac{p_\alpha(\mathbf{D}|\mathbf{v})}{q(\mathbf{v}|\theta)}q(\mathbf{v}|\theta)d\mathbf{v} \\ &= \int \log \left(\int \frac{1}{|K_{nn}|^{1/2}} e^{-\frac{1}{2}(\mu+L\epsilon)K_{nn}^{-1}(\mu+L\epsilon)} e^{-\frac{1}{2}(\log \mathbf{w}-\mu_0)(\mathbf{e}_0^2)^{-1}(\log \mathbf{w}-\mu_0)} d\mathbf{w} \right) \frac{1}{|LL^T|^{1/2}} e^{-\frac{1}{2}\epsilon^T \epsilon} d\mathbf{v} \\ &\quad + \int (\log p_\alpha(\mathbf{D}|\mathbf{v}) - \log q(\mathbf{v}|\theta))q(\mathbf{v}|\theta)d\mathbf{v} \\ &= \int \log \left(\int \frac{1}{|K_{nn}|} e^{-(\mu+L\epsilon)K_{nn}^{-1}(\mu+L\epsilon)} e^{-(\log \mathbf{w}-\mu_0)(\mathbf{e}_0^2)^{-1}(\log \mathbf{w}-\mu_0)} d\mathbf{w} \right) \frac{1}{|LL^T|} e^{-\epsilon^T \epsilon} d\mathbf{v} \\ &\quad + \int \left(\sum_{i=1}^{n-n_{test}} (-\log(1 + e^{-r_i(\mu_i+L_i\epsilon)}) + \log |LL^T| + \epsilon^T \epsilon) \right) \frac{1}{|LL^T|} e^{-\epsilon^T \epsilon} d\mathbf{v} \\ &= \int \log \left(\int \frac{1}{|K_{nn}|} e^{-(\mu+L\epsilon)K_{nn}^{-1}(\mu+L\epsilon)} e^{-(\log \mathbf{w}-\mu_0)(\mathbf{e}_0^2)^{-1}(\log \mathbf{w}-\mu_0)} d\mathbf{w} \right) \frac{1}{|LL^T|} e^{-\epsilon^T \epsilon} d(\mu + L\epsilon) \\ &\quad + \int \left(\sum_{i=1}^{n-n_{test}} (-\log(1 + e^{-r_i(\mu_i+L_i\epsilon)}) + \log |LL^T| + \epsilon^T \epsilon) \right) \frac{1}{|LL^T|} e^{-\epsilon^T \epsilon} d(\mu + L\epsilon) \end{aligned}$$

1 Objective:

$$\min_{g_\theta} \mathcal{L}(\theta) = \min_{\mathbf{y}_v} \max_{g_\theta} g_1(\theta) \mathbf{y}_v + 1 + \log(-\mathbf{y}_v) - g_2(\theta)$$

2 Gradients w.r.t the primal variables:

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \left(\frac{\partial \mathcal{L}(\theta)}{\partial \mu}, \frac{\partial \mathcal{L}(\theta)}{\partial L} \right)^T.$$

3 Here,

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \mu} &= \mathbf{y}_v \frac{\partial g_1(\theta)}{\partial \mu} - \frac{\partial g_2(\theta)}{\partial \mu} \\ &= \mathbf{y}_v \frac{\partial g_1(\theta)}{\partial \mu} - \left(\frac{\partial \log P(\mathbf{D}|\mathbf{v})}{\partial \mu} - \frac{\partial \log q(\mathbf{v}|\theta)}{\partial \mu} \right)\end{aligned}$$

4 where

$$\frac{\partial g_1(\theta)}{\partial \mu} = \frac{\partial P_\alpha(\mathbf{v}|\mathbf{w})}{\partial \mu} = \exp \left(-\frac{1}{2}(\mu + L\epsilon)^T K^{-1}(\mu + L\epsilon) \right) K^{-1}(\mu + L\epsilon),$$

$$\frac{\partial \log P(\mathbf{D}|\mathbf{v})}{\partial \mu} = \left(\mathbf{0}_{1 \times n_{test}}, \left[\frac{y_i}{1 + \exp(y_i(\mu_i + L_i\epsilon))} \right]_{i:n_{test}+1 \rightarrow n} \right)^T$$

$$\frac{\partial \log q(\mathbf{v}|\theta)}{\partial \mu} = \mathbf{0}$$

5 Besides,

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial L} &= \mathbf{y}_v \frac{\partial g_1(\theta)}{\partial L} - \frac{\partial g_2(\theta)}{\partial L} \\ &= \mathbf{y}_v \frac{\partial g_1(\theta)}{\partial L} - \left(\frac{\partial \log P(\mathbf{D}|\mathbf{v})}{\partial L} - \frac{\partial \log q(\mathbf{v}|\theta)}{\partial L} \right)\end{aligned}$$

6 where

$$\frac{\partial g_1(\theta)}{\partial L} = \frac{P_\alpha(\mathbf{v}|\mathbf{w})}{\partial L} = \exp \left(-\frac{1}{2}(\mu + L\epsilon)^T K^{-1}(\mu + L\epsilon) \right) \frac{\partial(\mu + L\epsilon)^T K^{-1}(\mu + L\epsilon)}{\partial L},$$

$$\frac{\partial \log P(\mathbf{D}|\mathbf{v})}{\partial L} = \left(\mathbf{0}_{n_{test} \times n}, \left[\frac{y_i}{1 + \exp(y_i(\mu_i + L_i\epsilon))} \epsilon^T \right]_{i:n_{test}+1 \rightarrow n} \right)^T$$

$$\frac{\partial \log q(\mathbf{v}|\theta)}{\partial L} = (L^T)^{-1}$$

7 **References**