Algorithm 1 Minimax SGD

initialize all hyper-parameters, and initialize the primal variable θ and the dual variables $\mathbf{w}_1$ and $\mathbf{w}_2$.
for $t = 1 : T$ **do**
  sample ϵ to compute $\mathbf{v} = \mu + L\epsilon$.
  sample $\mathbf{w}$ to compute ARD kernel matrix.
  $g(\theta) = [P_\alpha(\mathbf{v}|\mathbf{w}), \log \frac{P(D|\mathbf{v})}{q(\mathbf{v}|\theta)}]$.
  $\mathbf{y} = \mathbf{e}_1 Sigmoid(\mathbf{e}_2\epsilon + \mathbf{b}_2) + \mathbf{b}_1$.
  update the primal variable:
  $\Delta\theta = \frac{\partial g_1(\theta)}{\partial \theta}\mathbf{y} - \frac{\partial g_2(\theta)}{\partial \theta}$.
  $\theta = \theta - \alpha\Delta\theta$.
  update the dual variable:
  $\mathbf{e}_1 = \mathbf{e}_1 + \beta\left(g_1(\theta) + 1/y\right)\frac{\partial \mathbf{y}}{\partial \mathbf{e}_1}$.
  $\mathbf{e}_2 = \mathbf{e}_2 + \beta\left(g_1(\theta) + 1/y\right)\frac{\partial \mathbf{y}}{\partial \mathbf{e}_2}$.
  $\mathbf{b}_1 = \mathbf{b}_1 + \beta\left(g_1(\theta) + 1/y\right)\frac{\partial \mathbf{y}}{\partial \mathbf{b}_1}$.
  $\mathbf{b}_2 = \mathbf{b}_2 + \beta\left(g(\theta) + 1/y\right)\frac{\partial \mathbf{y}}{\partial \mathbf{b}_2}$.

$$
\begin{aligned}
L(\theta) = & \int \log(E_{p_\alpha(\mathbf{w})}(p_\alpha(\mathbf{v}|\mathbf{w})))q(\mathbf{v}|\theta)\mathrm{d}\mathbf{v} + \int \log \frac{p_\alpha(\mathbf{D}|\mathbf{v})}{q(\mathbf{v}|\theta)} q(\mathbf{v}|\theta)\mathrm{d}\mathbf{v} \\
= & \int \log\left(\int p_\alpha(\mathbf{v}|\mathbf{w})p_\alpha(\mathbf{w})\mathrm{d}\mathbf{w}\right) q(\mathbf{v}|\theta)\mathrm{d}\mathbf{v} + \int \log \frac{p_\alpha(\mathbf{D}|\mathbf{v})}{q(\mathbf{v}|\theta)} q(\mathbf{v}|\theta)\mathrm{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)}\mathrm{d}\mathbf{w}\right) \frac{1}{|LL^T|^{1/2}} e^{-\frac{1}{2}\epsilon^T\epsilon}\mathrm{d}\mathbf{v} \\
+ & \int \log \frac{p_\alpha(\mathbf{D}|\mathbf{v})}{q(\mathbf{v}|\theta)} q(\mathbf{v}|\theta)\mathrm{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)}\mathrm{d}\mathbf{w}\right) \frac{1}{|LL^T|^{1/2}} e^{-\frac{1}{2}\epsilon^T\epsilon}\mathrm{d}\mathbf{v} \\
+ & \int (\log p_\alpha(\mathbf{D}|\mathbf{v}) - \log q(\mathbf{v}|\theta))q(\mathbf{v}|\theta)\mathrm{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)}\mathrm{d}\mathbf{w}\right) \frac{1}{|LL^T|} e^{-\epsilon^T\epsilon}\mathrm{d}\mathbf{v} \\
+ & \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}\mathrm{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)}\mathrm{d}\mathbf{w}\right) \frac{1}{|LL^T|} e^{-\epsilon^T\epsilon}\mathrm{d}(\mu+L\epsilon) \\
+ & \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}\mathrm{d}(\mu+L\epsilon)
\end{aligned}
$$

Objective:

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

Gradients w.r.t the primal variables:

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

Here,

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

where

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

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

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

Besides,

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

where

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

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

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

1 Experimental details

Initialize:

training data: 40

test data: 10

iterations: 1000

learning rate (primal): 10^{-3}

learning rate (dual): 10^{-3}

μ_0: median of pair-wise distance

u_0: 1

σ_0: 1

τ: 10^{-6}

ϵ: multivariate $N(0, 1)$

$\log w$: multivariate $N(0, 1)$

μ: zeros$(n, 1)$

L: Identity matrix: eye(n)

dual variable $\mathbf{y}_\phi = \mathbf{e}_1 \frac{1}{1+\exp(-\mathbf{e}_2\epsilon - \mathbf{b}_2)} + \mathbf{b}_1$

# of hidden layers: $m = n$

$\mathbf{e}_1 = -1 * ones(1, m)$

$\mathbf{e}_2 = ones(m, n)$

$\mathbf{b}_1 = 1$

$\mathbf{b}_2 = ones(m, 1)$

Output:

$\mathbf{y}_\phi$: 9.2×10^6

μ: $< 10^{-26}$ for elements corresponding to test data, > 0.02 for elements corresponding to training data

train likelyhold (log): -0.682443 its absolute value is decreasing

test likelyhold (log): -0.693147 its absolute value is a constant

References