Problem

Notations

The objective loss function is:

$$L(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|K_{nn}| + \frac{1}{2}(\mu + L\epsilon)^T K_{nn}^{-1}(\mu + L\epsilon) + (-\sum_{i=1}^{n-n_{test}} \log(1 + \exp^{-label(i)(\mu_i + L_i\epsilon)}) - (-\frac{n}{2}(\log(2\pi)) - \frac{1}{2}\log|LL^T| - \frac{1}{2}\epsilon^T \epsilon$$

.

Loss function: $L(\theta) = \log g_1 + g_2$ where $g = [g_1, g_2] = [P_\alpha(v|w), \log\frac{P(D|v)}{q(v|\theta)}]$. (Section 4.1, Eq. (9))

$P_\alpha(v|w) = \frac{1}{(2\pi)^{n/2}|K_{nn}|^{1/2}}\exp(\frac{1}{2}(\mu + L\epsilon)^T K_{nn}^{-1}(\mu + L\epsilon))$. (Section 4.1, Eq. (9))

$P(D|v) = \Pi_{i=1}^{n}\frac{1}{1+\exp^{-label(i)(\mu_i + L_i\epsilon)}}$. (Section 4.1, Eq. (9))

$q(v|\theta) = \frac{1}{(2\pi)^{n/2}|LL^T|^{1/2}}\exp(-1/2\epsilon^T\epsilon)$. (Section 4.1, Eq. (9))

$\theta = [\mu, vec(L)]$

Update of primal variables

$\theta = \theta - \alpha\langle\nabla g(\theta), y\rangle$, and $\nabla g(\theta) = [\frac{\partial P_\alpha(v|w)}{\partial\theta}, \frac{\partial \log\frac{P(D|v)}{q(v|\theta)}}{\partial\theta}]$

$\frac{\partial P_\alpha(v|w)}{\partial\theta} = \frac{1}{(2\pi)^{n/2}|K_{nn}|^{1/2}}\exp(-\frac{1}{2}(\mu + L\epsilon)^T K_{nn}^{-1}(\mu + L\epsilon))K_{nn}^{-1}(\mu + L\epsilon)\frac{\partial\mu + L\epsilon}{\partial\theta}$ Here, $\exp(-\frac{1}{2}(\mu + L\epsilon)^T K_{nn}^{-1}(\mu + L\epsilon))$ is very small. The reason is that $\frac{1}{2}(\mu + L\epsilon)^T K_{nn}^{-1}(\mu + L\epsilon)$ is large (> 10000). Therefore, when I begin to compute the gradient of $P_\alpha(v|w)$ with respect to $\theta = [\mu, vec(L)]$, I find that the gradient is very small (see the figure).

stoc_nabla_mu_L_1

2652x1 double

	1	2
1	2.5986e-09	
2	6.8391e-09	
3	-1.3308e-09	
4	-3.4339e-09	
5	4.9560e-09	
6	6.2637e-10	
7	7.8143e-09	
8	-5.5226e-09	
9	-5.0651e-10	
10	-2.6263e-09	
11	-2.2707e-09	
12	5.1343e-10	
13	-2.6969e-09	
14	-6.2604e-08	
15	-7.8483e-09	
16	1.2442e-09	
17	5.1988e-11	
18	2.4678e-08	
19	1.9726e-09	
20	-5.9031e-08	
21	6.5674e-08	
22	6.0161e-09	
23	4.2297e-09	
24	1.6973e-09	

The second item of g consist of $P(D|v)$ and $q(v|\theta)$. The gradient of $P(D|v)$ is computed as following codes:

```
%the second item of g
    stoc_nabla_mu_L_temp_2 = zeros(n+n*n,1);
    for j=1:n
        if j<=n_test
            continue;% During training, the test data is discarded due to lack of labels.
        end
        stoc_nabla_mu_L_temp_2 = stoc_nabla_mu_L_temp_2 +
(label(j)*transpose(Q(j,:)))/(1+exp(label(j)*Q(j,:)*theta));
    end
```

Its gradeint with respect to $\mu_{testdata}$ is 0 because the labels of test data is not used during the training of parameters.

The gradient of $q(v|\theta)$ with respect to μ is 0. Because it is a function with respect to L.

Therefore, during training iterations, the μ corresponding to the test data (dimensions from 1 to 10) do not have any changes:

theta_sequence × | theta_avg × | train_loss × | test_loss × | mu_temp ×

2652x100 double

	1	2	3	4	5	6	7	8	9
1	0.0070	0.0070	0.0070	0.0070	0.0070	0.0070	0.0070	0.0070	0.0070
2	0.0057	0.0057	0.0057	0.0057	0.0057	0.0057	0.0057	0.0057	0.0057
3	0.0063	0.0063	0.0063	0.0063	0.0063	0.0063	0.0063	0.0063	0.0063
4	0.0088	0.0088	0.0088	0.0088	0.0088	0.0088	0.0088	0.0088	0.0088
5	0.0066	0.0066	0.0066	0.0066	0.0066	0.0066	0.0066	0.0066	0.0066
6	0.0088	0.0088	0.0088	0.0088	0.0088	0.0088	0.0088	0.0088	0.0088
7	0.0047	0.0047	0.0047	0.0047	0.0047	0.0047	0.0047	0.0047	0.0047
8	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014
9	6.8134e-...	6.8134e-...	6.8134e-...	6.8134e-04	6.8134e-...	6.8134e-...	6.8134e-...	6.8134e-...	6.8134e-...
10	0.0071	0.0071	0.0071	0.0071	0.0071	0.0071	0.0071	0.0071	0.0071
11	-0.0021	-0.0021	-0.0021	0.0075	0.0075	0.0075	0.0077	0.0077	0.0239
12	0.0117	0.0117	0.0117	0.0021	0.0021	0.0021	0.0021	0.0021	-0.0142
13	0.0013	-0.0029	0.0052	0.0052	0.0192	0.0333	0.0498	0.0653	0.0653
14	0.0101	0.0101	0.0101	5.3257e-04	5.3257e-...	5.3257e-...	4.4034e-...	4.4034e-...	-0.0159
15	0.0119	0.0119	0.0119	0.0023	-0.0028	-0.0028	-0.0028	-0.0028	-0.0191
16	0.0099	0.0099	0.0099	3.0357e-04	3.0357e-...	3.0357e-...	-0.0017	-0.0017	-0.0180
17	-2.1036e...	-2.1036e...	-2.1036e...	0.0094	0.0094	0.0094	0.0096	0.0096	0.0259
18	0.0099	0.0142	0.0060	0.0060	-0.0080	-0.0221	-0.0348	-0.0504	-0.0504
19	0.0043	0.0043	0.0043	0.0139	0.0139	0.0139	0.0299	0.0299	0.0461
20	0.0088	0.0131	0.0049	0.0049	-0.0091	-0.0232	-0.0397	-0.0553	-0.0553
21	-0.0040	-0.0040	-0.0040	0.0057	0.0057	0.0057	0.0154	0.0154	0.0317
22	-0.0028	-0.0028	-0.0028	0.0068	0.0068	0.0068	0.0068	0.0068	0.0231
23	0.0117	0.0159	0.0078	0.0078	-0.0062	-0.0203	-0.0369	-0.0524	-0.0524

Update of dual variables

$y = y + \beta(g(\theta) - \nabla f^*(y))$