End-to-End Learning for the Deep Multivariate Probit Model

Di Chen [1] **Yexiang Xue** [1] **Carla Gomes** [1]

7. Appendix

Theorem 1 *Let $\mu \in R^l$ and $\Sigma \in R^{l \times l}$ be the rescaled mean and the rescaled residual covariance matrix of the random variable $w^{(k)}$ in the equation (7) of the main text, then we have*

$$\Pr\left[\left|\frac{1}{M}\sum_{k=1}^{M}\prod_{j=1}^{l}\Phi(w_{i,j}^{(k)}) - \Pr(y_i|x_i)\right| \geq \epsilon \Pr(y_i|x_i)\right]$$

$$\leq \frac{\Phi\left(0;\begin{bmatrix}-\mu\\-\mu\end{bmatrix},\begin{bmatrix}\Sigma+I & \Sigma\\ \Sigma & \Sigma+I\end{bmatrix}\right) - \Phi^2(0;-\mu,\Sigma+I)}{M\Phi^2(0;-\mu,\Sigma+I)\epsilon^2} \tag{1}$$

$$\leq \frac{\left(\frac{\Phi(0;-\mu,2\Sigma+I)}{\Phi(0;-\mu,\Sigma+I)}\right)^2 |2\Sigma+I|^{1/2} - 1}{M\epsilon^2} \tag{2}$$

$$\leq \frac{\prod_{i=1}^{l} g(\mu_i)^2 |2\Sigma+I|^{1/2} - 1}{M\epsilon^2} \tag{3}$$

where $g(\mu_i) = \max_x \frac{\Phi(\sqrt{2}x+\mu_i)}{\Phi(x+\mu_i)}$. The function $g(\mu_i)$ does not have a closed form but it is a monotonous decreasing function, which converges to 1 as μ_i increases.

Proof. For the ease of expression, we omit the subscripts related to i-th data point in our proof. Without loss of generality, we can also assume the diagonal matrix V is an indentity matrix. Defining $\Pr(y|w) = \prod_{j=1}^{n}\Phi(w_j)$, $\Pr(y|x) = E_{w\sim N(\mu,\Sigma)}[\Pr(y|w)]$. We prove this convergence bound by analysing the first and second moment of random variable $\Pr(y|w)$.

$$\begin{aligned} E_w[\Pr(y|w)] &= \int_w \prod_{j=1}^{n}\Phi(w_j)Pr_w(w)\mathrm{d}w \\ &= \int_w Pr_z(z \preceq w|w)Pr_w(w)\mathrm{d}w \\ &= Pr_{z,w}(z \preceq w) \\ &= Pr_{z,w}(z - w \preceq 0) \end{aligned} \tag{4}$$

Here $z\ N(0, I)$ and $a \preceq b$ means $\forall a_i \leq b_i$

Since z is subject to multivariate gaussian distribution, $z - w$ is still a multivariate gaussian random variable, which is subject to $N(-\mu, \Sigma + I)$. Thus, $\Pr(y|x) = E_w[\Pr(y|w)] = \Phi(0; -\mu, \Sigma + I)$. ($\Phi(\cdot)$ denotes the cumulative function of multivariate gaussian distribution.)

Similarly, we can derive that

$$\begin{aligned} E[\Pr(y|w)^2] &= \Pr(z_1 \preceq w \wedge z_2 \preceq w) \\ &= \Pr\left(\begin{bmatrix}z_1\\z_2\end{bmatrix} \preceq \begin{bmatrix}r\\r\end{bmatrix}\right) \\ &= \Phi\left(0;\begin{bmatrix}-\mu\\-\mu\end{bmatrix},\begin{bmatrix}\Sigma+I & \Sigma\\ \Sigma & \Sigma+I\end{bmatrix}\right) \end{aligned}$$

Let $B = \begin{bmatrix}\Sigma+I & \Sigma\\ \Sigma & \Sigma+I\end{bmatrix}$, we have $|B| = \left|det\left(\begin{bmatrix}2\Sigma+I & \Sigma\\ 0 & I\end{bmatrix}\right)\right| = |2\Sigma + I|$. Since Σ is a positive definite matrix, we can decompose $\Sigma = UDU^T$, where U is an orthogonal matrix and D is a diagonal matrix. Similarly, we can decompose

$$B^{-1} = \begin{bmatrix}U & 0\\ 0 & U\end{bmatrix}\begin{bmatrix}(2D+I)^{-1}(D+I) & -(2D+I)^{-1}D\\ -(2D+I)^{-1}D & (2D+I)^{-1}(D+I)\end{bmatrix}\begin{bmatrix}U^T & 0\\ 0 & U^T\end{bmatrix}$$

Let $x_1, x_2 \in R^l$, $y_1 = U^T(x_1 + \mu)$, $y_2 = U^T(x_1 + \mu)$ and $D = diag(d_1, ..., d_l)$, then we have,

$$\begin{aligned} E[\Pr(y|r)^2] &= \Phi\left(0;\begin{bmatrix}-\mu\\-\mu\end{bmatrix},\begin{bmatrix}\Sigma+I & \Sigma\\ \Sigma & \Sigma+I\end{bmatrix}\right) \\ &= \frac{1}{(2\pi)^l|B|^{1/2}}\int_{(-\infty,0]^l} e^{-\frac{1}{2}(\sum_{i=1}^{l}(y_{1,i}^2+y_{2,i}^2)\frac{d_i+1}{2d_i+1} - 2\sum_{i=1}^{l} y_{1,i}y_{2,i}\frac{d_i}{2d_i+1})}\mathrm{d}x_1\mathrm{d}x_2 \\ &\leq \frac{1}{(2\pi)^l|B|^{1/2}}\int_{(-\infty,0]^l} e^{-\frac{1}{2}(\sum_{i=1}^{l}(y_{1,i}^2+y_{2,i}^2)\frac{1}{2d_i+1})}\mathrm{d}x_1\mathrm{d}x_2 \\ &= |2\Sigma+I|^{1/2}\Phi\left(0;\begin{bmatrix}-\mu\\-\mu\end{bmatrix},\begin{bmatrix}2\Sigma+I & 0\\ 0 & 2\Sigma+I\end{bmatrix}\right) \end{aligned}$$

Thus,

$$E[\Pr(y|r)^2]^{1/2} \leq |2\Sigma + I|^{1/4}\Phi(0; -\mu, 2\Sigma + I)$$

Using the inverse transformation in equation (4), we have

$$\begin{aligned} &\Phi(0; -\mu, 2\Sigma + I) \\ &= \frac{1}{(2\pi)^{l/2}|2\Sigma|^{1/2}}\int\prod\Phi(x)e^{\frac{1}{4}(x-\mu)^T\Sigma^{-1}(x-\mu)}\mathrm{d}x \\ &= \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}}\int\prod\Phi(\sqrt{2}y + \mu_i)e^{\frac{1}{2}y^T\Sigma^{-1}y}\mathrm{d}y \end{aligned} \tag{5}$$

[1]Computer Science Department, Cornell University, Ithaca, NY, US 14850. Correspondence to: Di Chen <di@cs.cornell.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Let $g(\mu_i) = \max_x \frac{\Phi(\sqrt{2}x+\mu_i)}{\Phi(x+\mu_i)}$, then we have

$$
\begin{aligned}
&\Phi(0; -\mu, 2\Sigma + I) \\
&= \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \int \prod \Phi(\sqrt{2}y + \mu) e^{\frac{1}{2}y^T\Sigma^{-1}y} \mathrm{d}y \\
&\leq \frac{\prod_{i=1}^{l} g(\mu_i)}{(2\pi)^{l/2}|\Sigma|^{1/2}} \int \prod \Phi(y + \mu) e^{\frac{1}{2}y^T\Sigma^{-1}y} \mathrm{d}y \\
&= \prod_{i=1}^{l} g(\mu_i) \Phi(\mu|\Sigma + I) \\
&= \prod_{i=1}^{l} g(\mu_i) \Pr(y|x)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
E[\Pr(y|w)^2]^{1/2} &\leq |2\Sigma + I|^{1/4} \Phi(0; -\mu, 2\Sigma + I) \\
&\leq |2\Sigma + I|^{1/4} \prod_{i=1}^{l} g(\mu_i) \Phi(0; -\mu, \Sigma + I)
\end{aligned}
$$

Using the Chebyshev's inequality, we have

$$
\begin{aligned}
&\Pr[|\frac{1}{M}\sum_{k=1}^{M}\prod_{j=1}^{l} \Phi(w_{i,j}^{(k)}) - \Pr(y_i|x_i)| \geq \epsilon \Pr(y_i|x_i)] \\
&= \Pr[|\frac{1}{M}\sum_{k=1}^{M} \Pr(y_i|w_i^{(k)}) - \Pr(y_i|x_i)| \geq \epsilon \Pr(y_i|x_i)] \\
&= \Pr[|\frac{1}{M}\sum_{k=1}^{M} \Pr(y_i|w_i^{(k)}) - \Pr(y_i|x_i)|^2 \geq \epsilon^2 \Pr(y_i|x_i)^2] \\
&\leq \frac{E[(\frac{1}{M}\sum_{k=1}^{M} \Pr(y_i|w_i^{(k)}) - \Pr(y_i|x_i))^2]}{\epsilon^2 \Pr(y_i|x_i)^2} \\
&= \frac{\prod_{i=1}^{l} g^2(\mu_i)|2\Sigma + I|^{1/2} - 1}{M\epsilon^2} \qquad \blacksquare
\end{aligned}
$$

The function $g(\mu_i)$ does not have a closed form but it is a monotonous decreasing function, which converges to 1 as μ_i increases. The figure (1) is the visualization of function

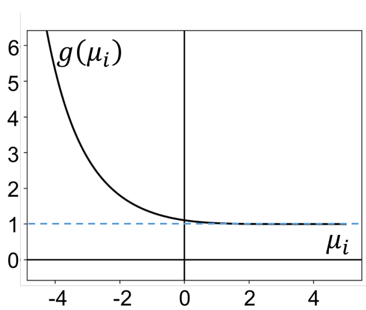

Figure 1. The visualization of function $g(\mu_i)$.

$g(\mu_i)$. As you see, the function $g(\mu_i)$ is very close to 1 when μ_i is positive. The following lemma provides a more analytical upper bound for function $g(\mu_i)$.

Lemma 1 *For any y,* $\Phi(\sqrt{2}y+\mu) \leq g(\mu)\Phi(y+\mu)$, *where*

$$
g(\mu) \leq \begin{cases} \sqrt{2}e^{\frac{3-2\sqrt{2}}{2}\mu^2} & \textit{if} \quad \mu < 0 \\ 1.182 & \textit{if} \quad \mu \geq 0 \end{cases}
$$

Proof. $\frac{\Phi(\sqrt{2}y+\mu)}{\Phi(y+\mu)}$ achieves the maximum when its derivative is equal to zero, i.e.,

$$
\begin{aligned}
&\left(\frac{\Phi(\sqrt{2}y + \mu)}{\Phi(y + \mu)}\right)' = 0 \Longrightarrow \\
&\frac{\frac{1}{\sqrt{2\pi}}(\sqrt{2}e^{-\frac{1}{2}(\sqrt{2}y+\mu)^2}\Phi(y+\mu) - e^{-\frac{1}{2}(y+\mu)^2}\Phi(\sqrt{2}y+\mu)}{\Phi^2(y+\mu)} = 0 \\
&\Longrightarrow \frac{\Phi(\sqrt{2}y + \mu)}{\Phi(y + \mu)} = \sqrt{2}e^{-\frac{1}{2}(y^2+2(\sqrt{2}-1)\mu y)}
\end{aligned}
$$

Since $\Phi(x)$ is a monotonic increasing function, $max_y\sqrt{2}e^{-\frac{1}{2}(y^2+2(\sqrt{2}-1)\mu y)} = \sqrt{2}e^{\frac{3-2\sqrt{2}}{2}\mu^2}$ when $\mu < 0$. Similarly, when $\mu \geq 0$, we know $y^* = argmax_y \frac{\Phi(\sqrt{2}y+\mu)}{\Phi(y+\mu)} \geq 0$. Thus, $\Phi(y^* + \mu) \geq \frac{1}{2}$. By analysing the maximal value of $\Phi(\sqrt{2}y + \mu) - \Phi(y + \mu)$ as well as the fact that $\Phi(\sqrt{2}y + \mu) - \Phi(y + \mu) \leq (\sqrt{2} - 1)y * \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y+\mu)^2}$, we could know that $\Phi(\sqrt{2}y + \mu) - \Phi(y + \mu) \leq 0.091$. That is,

$$
g(\mu) \leq \begin{cases} \sqrt{2}e^{\frac{3-2\sqrt{2}}{2}\mu^2} & \text{if} \quad \mu < 0 \\ 1.182 & \text{if} \quad \mu \geq 0 \end{cases}
$$

.

Theorem 2 *Let* $\mu \in R^l$ *and* $\Sigma \in R^{l\times l}$ *be the rescaled mean and rescaled residual covariance matrix of the random variable* $w^{(k)}$ *in equation (7) of the main text, we have*

$$
\begin{aligned}
&\Pr\left[\left|\frac{\partial \frac{1}{M}\sum_{k=1}^{M}\prod_{j=1}^{l}\Phi(w_{i,j}^{k})}{\partial\mu_i} - \frac{\partial \Pr(y_i|x_i)}{\partial\mu_i}\right| \geq \epsilon\frac{\partial \Pr(y_i|x_i)}{\partial\mu_i}\right] \\
&\leq \frac{e^{\frac{\mu_i^2}{2(\Sigma_{i,i}+1)}}(\Sigma_{i,i}+1)\lambda_{max}\prod_{j\neq i}^{l} g(\mu_j')^2|2\Sigma+I|^{1/2}-1}{M\epsilon^2}
\end{aligned} \tag{6}
$$

Here λ_{max} *denotes the largest eigenvalue of* Σ *and* $\mu' = \mu - \frac{\mu_i}{v+1}\Sigma^{1/2}b_i$. *(*$b_i$ *denotes the i-th row of* $\Sigma_{1/2}$.*)*

Proof. For the ease of symbolism, we omit all the subscript

i related to the index of i-th data point. For any $1 \leq i \leq l$,

$$\frac{\partial \Pr(y|x)}{\partial \mu_i} = E_{w\sim N(\mu,\Sigma)}\left[\frac{\partial \prod_{j=1}^{l}\Phi(w_j)}{\partial \mu_i}\right]$$
$$= \int \prod_{j\neq i}^{l} \Phi(w_j) * \phi(w_i)\phi(w|\mu,\Sigma)\mathrm{d}w$$
$$= \int \prod_{j\neq i}^{l} \Phi(\Sigma_j^{1/2}x+\mu_j) * \phi(\Sigma_i^{1/2}x+\mu_i)\phi(x|0,I)\mathrm{d}x$$

Let $B = \Sigma^{1/2}$ and let b_j denote the j-th row of B.

$$= \int \prod_{j\neq i}^{l} \Phi(b_j^T x+\mu_j) * \phi(b_i^T x+\mu_i)\phi(x|0,I)\mathrm{d}x$$

let $v = b_i^T b_i = \Sigma_{i,i}$ and $C = I - \frac{b_i b_i^T}{v+1}$ ($C^{-1} = I + b_i b_i^T$).

$= \phi(\frac{\mu_i}{v+1}) * |C|^{1/2} \int \prod_{j\neq i}^{l} \Phi(b_j^T x+\mu_j) * \phi(x| -\frac{\mu_i}{v+1}b_i, C)\mathrm{d}x$

$= \phi(\frac{\mu_i}{v+1}) * |C|^{1/2} * \Pr(\forall j \neq i, z_j \leq b_j^T x + \mu_j)$

(where $x \sim N(-\frac{\mu_i}{v+1}b_i, C)$ and $z \sim N(0, I)$.)

$= \phi(\frac{\mu_i}{v+1}) * |C|^{1/2} * \Pr(z \preceq w)$

(where $w \sim N(\mu_{-i} - \frac{\mu_i}{v+1}B_{-i}b_i, B_{-i}CB_{-i}^T)$,
$\mu_{-i} \in R^{l-1}$ denotes the vector derived from μ by eliminating the i-th entry. $B_{-i} \in R^{l-1\times l}$ denotes the matrix derived from B by eliminating the i-th row.)

Thus, using the transformation above, we can transform the derivative in terms of μ_i into the form similar to theorem (1). Because $B_{-i}CB_{-i}^T = B_{-i}B_{-i}^T - \frac{(B_{-i}b_i)(B_{-i}b_i)^T}{v+1}$, where $B_{-i}B_{-i}^T$ is a principal submatrix of Σ, whose eigenvalues are interlaced with the eigenvalues of Σ, and $\frac{(B_{-i}b_i)(B_{-i}b_i)^T}{v+1}$ is a rank-1 matrix, we have $|2B_{-i}CB_{-i}^T + I| \leq |2\Sigma + I| * \lambda_{max}$.

In terms of the second moment of the derivative of μ_i, we have,

$$E_{w\sim N(\mu,\Sigma)}\left[\left(\frac{\partial \prod_{j=1}^{l}\Phi(w_j)}{\partial \mu_i}\right)^2\right]$$
$$= \int \prod_{j\neq i}^{l} \Phi^2(\Sigma_j^{1/2}x+\mu_j) * \phi^2(\Sigma_i^{1/2}x+\mu_i)\phi(x|0,I)\mathrm{d}x$$
$$\leq \int \prod_{j\neq i}^{l} \Phi^2(\Sigma_j^{1/2}x+\mu_j) * \phi(\Sigma_i^{1/2}x+\mu_i)\phi(x|0,I)\mathrm{d}x$$

$= \phi(\frac{\mu_i}{v+1}) * |C|^{1/2} \int \prod_{j\neq i}^{l} \Phi^2(b_j^T x+\mu_j) * \phi(x| -\frac{\mu_i}{v+1}b_i, C)\mathrm{d}x$

$= \phi(\frac{\mu_i}{v+1}) * |C|^{1/2} * \Pr(z^1 \preceq w \wedge z^2 \preceq w)$

Here we use the same notation as the proof above.

Using the similar trick as theorem (1), we have

$$\Pr\left[\left|\frac{\partial \frac{1}{M}\sum_{k=1}^{M}\prod_{j=1}^{l}\Phi(w_{i,j}^k)}{\partial \mu_i} - \frac{\partial \Pr(y_i|x_i)}{\partial \mu_i}\right| \geq \epsilon \frac{\partial \Pr(y_i|x_i)}{\partial \mu_i}\right]$$
$$\leq \frac{e^{\frac{\mu_i^2}{2(v+1)}}|C^{-1}|\lambda_{max}\prod_{j\neq i}^{l} g(\mu_j')^2 |2\Sigma+I|^{1/2}-1}{M\epsilon^2}$$
$$\leq \frac{e^{\frac{\mu_i^2}{2(\Sigma_{i,i}+1)}}(\Sigma_{i,i}+1)\lambda_{max}\prod_{j\neq i}^{l} g(\mu_j')^2 |2\Sigma+I|^{1/2}-1}{M\epsilon^2}$$

Here $\mu' = \mu - \frac{\mu_i}{v+1}\Sigma^{1/2}b_i$.

In this way, we bound the convergence of the derivatives in terms of μ, so that the derivatives in term of the parameters in feature network can be derived by chain rule. However, because the derivatives of $\Sigma^{1/2}$ could be negative or zero, we can not apply the Chebyshev's inequality to have a similar multiplicative error bound. Nevertheless, because all the data points share a global residual covariance matrix, empirical experiments show that $\Sigma^{1/2}$ converges well on all the datasets.

Here we show that the variance of our sampling process is strictly lower than the rejection sampling.

Theorem 3 *Here we follow the notation of equation(7) in the main paper. Let θ_1 be the reject sampling estimator of $\Phi(0; -\mu, \Sigma)$, where $E[\theta_1] = E_{r\sim N(0,\Sigma)}[I\{r \preccurlyeq \mu\}]$. Let θ_2 be the estimator of DMVP's sampling process, where $E[\theta_2] = E_{w\sim N(0,\Sigma_r)}[\Pr(z \preccurlyeq (w+\mu)|w)]$ and $z \sim N(0, V)$. We have $Var[\theta_2] < Var[\theta_1]$.*
Proof.

$$
\begin{aligned}
&Var[\theta_2] = E[(\theta_2 - E[\theta_2])^2] \\
&= E_{w\sim N(0,\Sigma_r)}[(\Pr(z \preccurlyeq (w+\mu)|w) - E[\theta_2])^2] \\
&= E_{w\sim N(0,\Sigma_r)}[(E_{z\sim N(0,V)}[I\{z \preccurlyeq (w+\mu)\} - E[\theta_2]|w])^2] \\
&< E_{w\sim N(0,\Sigma_r)}[E_{z\sim N(0,V)}[(I\{z \preccurlyeq (w+\mu)\} - E[\theta_2])^2|w]] \\
&= E_{r\sim N(0,\Sigma)}[(I\{r \preccurlyeq \mu\} - E[\theta_1])^2] \\
&\qquad \textit{(Here } r = z - w \textit{ and } E[\theta_1] = E[\theta_2]) \\
&= E[(\theta_1 - E[\theta_1])^2] = Var[\theta_1]
\end{aligned}
$$

The inequality follows the fact that $E[x^2] > E[x]^2$ given $Var[x] \neq 0$.