SUPPLEMENTARY MATERIALS – PAC-BAYES GENERALIZATION ERROR BOUND

Anonymous authors
Paper under double-blind review

We briefly introduce the basic settings for PAC-Bayes generalization error. The expected risk is defined as $\mathbb{E}_{x\sim\mathcal{P}(x)}\ell(w,x)$. Suppose the parameter follows a distribution with density $p(w)$, the expected risk in terms of $p(w)$ is defined as $\mathbb{E}_{w\sim p(w),x\sim\mathcal{P}(x)}\ell(w,x)$. The empirical risk in terms of $p(w)$ is defined as $\mathbb{E}_{w\sim p(w)}L(w) = \mathbb{E}_{w\sim p(w)}\frac{1}{n}\sum_{i=1}^{n}\ell(w,x_i)$. Suppose the prior distribution over the parameter space is $p'(w)$ and $p(w)$ is the distribution on the parameter space expressing the learned hypothesis function. For $\text{SGF}_{\Sigma_2^{w^*}(w)}$, $p(w)$ is its stationary distribution and we choose $p'(w)$ to be Gaussian distribution with center w^* and covariance matrix I. Then we can get the following theorem.

Theorem 1 *Suppose that $w\in\mathbb{R}^d$ and $\kappa > \frac{d}{2}$. For $\delta > 0$, with probability at least $1-\delta$, the stationary distribution of $\text{SGF}_{\Sigma_2^{w^*}(w)}$ has the following generalization error bound,*

$$\mathbb{E}_{w\sim p(w),x\sim\mathcal{P}(x)}\ell(w,x) \le \mathbb{E}_{w\sim p(w)}L(w) + \sqrt{\frac{KL(p||p') + \log\frac{1}{\delta} + \log n + 2}{n-1}}, \tag{1}$$

where $KL(p||p') \le \frac{1}{2}\log\frac{\det(H_{w^*})}{\det(\Sigma_{g_{w^*}})} + \frac{Tr(\frac{\eta}{m}\Sigma_{g_{w*}}H_{w^*}^{-1})-2d}{4\left(1-\frac{1}{\kappa_{w^*}}\left(\frac{d}{2}-1\right)\right)} + \frac{d}{2}\log\frac{2m}{\eta}$, $p(w)$ is the stationary distribution of d-dimensional $\text{SGF}_{\Sigma_2^{w^*}(w)}$, $p'(w)$ is a prior distribution which is selected to be standard Gaussian distribution, and $\mathcal{P}(x)$ is the underlying distribution of data x, $\det(\cdot)$ and $Tr(\cdot)$ are the determinant and trace of a matrix, respectively.

Proof: [1] Eq.(1) directly follows the results in (McAllester, 1999). Here we calculate the Kullback–Leibler (KL) divergence between prior distribution and the stationary distribution of $\text{SGF}_{\Sigma_2^{w^*}(w)}$. The prior distribution is selected to be standard Gaussion distribution with distribution density $p'(w) = \frac{1}{\sqrt{(2\pi)^d\det(I)}}\exp\{-\frac{1}{2}(w-w^*)^TI(w-w^*)\}$. The posterior distribution density is the stationary distribution for $\text{SGF}_{\Sigma_2^{w^*}(w)}$, i.e., $p(w) = \frac{1}{Z}\cdot(1+\frac{1}{\tilde{\eta}\kappa}\cdot(w-w^*)^TH\Sigma_g^{-1}(w-w^*))^{-\kappa}$ according to Lemma 5 in appendix.

Suppose $H\Sigma_g^{-1}$ are symmetric matrix. Then there exist orthogonal matrix Q and diagonal matrix $\Lambda = diag(\lambda_1,\cdots,\lambda_d)$ that satisfy $H\Sigma_g^{-1} = Q^T\Lambda Q$. We also denote $v = Q(w-w^*)$.

We have

$$\begin{aligned}&\log\left(\frac{p(w)}{p'(w)}\right)\\ &= -\kappa\log(1+\frac{1}{\tilde{\eta}\kappa}\cdot(w-w^*)^TH\Sigma_g^{-1}(w-w^*)) - \log Z + \frac{1}{2}(w-w^*)^TI(w-w^*) + \frac{d}{2}\log 2\pi\end{aligned}$$

The KL-divergence is defined as $KL(p(w)||p'(w)) = \int_w p(w)\log\left(\frac{p(w)}{p'(w)}\right)dw$. Putting $v = Q(w-w^*)$ in the integral, we have

$$\begin{aligned}&KL(p(w)||p'(w))\\ =&\frac{d}{2}\log 2\pi - \log Z + \frac{1}{2Z}\int_v v^Tv\left(1+\frac{1}{\tilde{\eta}\kappa}\cdot v^T\Lambda v\right)^{-\kappa}dv - \frac{1}{Z\tilde{\eta}}\int_v v^T\Lambda v\cdot(1+\frac{1}{\tilde{\eta}\kappa}\cdot v^T\Lambda v)^{-\kappa}dv,\end{aligned} \tag{2}$$

[1]We omit the notation w^* in the following context for simplicity since our result is applied for a fixed w^*. Moreover, we let $\tilde{\eta} = \frac{\eta}{m}$ for simplicity.

where we use the approximation that $\log(1+x) \approx x$. We define a sequence as $T_k = 1 + \frac{1}{\tilde{\eta}\kappa} \cdot \sum_{j=k}^{d} \lambda_j v_j^2$ for $k = 1, \cdots, d$. We first calculate the normalization constant Z.

$$
\begin{aligned}
Z = &\int (1 + \frac{1}{\tilde{\eta}\kappa} \cdot v^T \Lambda v)^{-\kappa} dw = \int (1 + \frac{1}{\tilde{\eta}\kappa} \cdot \sum_{j=1}^{d} \lambda_j v_j^2)^{-\kappa} dv \\
=&((\tilde{\eta}\kappa)^{-1}\lambda_1)^{-\frac{1}{2}} \int T_2^{-\kappa+\frac{1}{2}} B(\frac{1}{2}, \kappa - \frac{1}{2}) dv = \prod_{j=1}^{d} ((\tilde{\eta}\kappa)^{-1}\lambda_j)^{-\frac{1}{2}} B(\frac{1}{2}, \kappa - \frac{j}{2}) \\
=&\prod_{j=1}^{d} ((\tilde{\eta}\kappa)^{-1}\lambda_j)^{-\frac{1}{2}} \cdot \frac{\sqrt{\pi^d}\Gamma(\kappa - \frac{d}{2})}{\Gamma(\kappa)}
\end{aligned}
$$

We define $Z_j = ((\tilde{\eta}\kappa)^{-1}\lambda_j)^{-\frac{1}{2}} B\left(\frac{1}{2}, \kappa - \frac{j}{2}\right)$. For the third term in Eq.(2), we have

$$
\begin{aligned}
&2Z \cdot III \\
=&\int_v v^T v (1 + \frac{1}{\tilde{\eta}\kappa} v^T \Lambda v)^{-\kappa} dv \\
=&\int_{v_2,\cdots v_d} \int_{v_1} v_1^2 \left(1 + \frac{1}{\tilde{\eta}\kappa} \cdot v^T \Lambda v\right)^{-\kappa} dv_1 + Z_1 \left(\sum_{j=2}^{d} v_j^2\right) \left(1 + \frac{1}{\tilde{\eta}\kappa} \cdot \sum_{j=2}^{d} \lambda_j v_j^2\right)^{-\kappa+\frac{1}{2}} d_{v_2\cdots,v_d} \\
=&\int_{v_2,\cdots v_d} T_2^{-\kappa} \int_{v_1} v_1^2 \left(1 + \frac{(\tilde{\eta}\kappa)^{-1}\lambda_1 v_1^2}{T_2}\right)^{-\kappa} dv_1 + Z_1 \left(\sum_{j=2}^{d} v_j^2\right) \left(1 + \frac{1}{\tilde{\eta}\kappa} \cdot \sum_{j=2}^{d} \lambda_j v_j^2\right)^{-\kappa+\frac{1}{2}} d_{v_2\cdots,v_d} \\
=&\int_{v_2,\cdots,v_d} T_2^{-\kappa} \int \left(\frac{T_2}{(\tilde{\eta}\kappa)^{-1}\lambda_1}\right)^{\frac{3}{2}} y^{\frac{1}{2}} (1+y)^{-\kappa} dy + Z_1 \left(\sum_{j=2}^{d} v_j^2\right) \left(1 + \frac{1}{\tilde{\eta}\kappa} \cdot \sum_{j=2}^{d} \lambda_j v_j^2\right)^{-\kappa+\frac{1}{2}} d_{v_2\cdots,v_d} \\
=&\int_{v_2,\cdots,v_d} ((\tilde{\eta}\kappa)^{-1}\lambda_1)^{-\frac{3}{2}} T_2^{-\kappa+\frac{3}{2}} B\left(\frac{3}{2}, \kappa - \frac{3}{2}\right) + Z_1 \left(\sum_{j=2}^{d} v_j^2\right) \left(1 + \frac{1}{\tilde{\eta}\kappa} \cdot \sum_{j=2}^{d} \lambda_j v_j^2\right)^{-\kappa+\frac{1}{2}} d_{v_2\cdots,v_d} \\
=&(\frac{\lambda_1}{\tilde{\eta}\kappa})^{-\frac{3}{2}} B\left(\frac{3}{2}, \kappa - \frac{3}{2}\right) \int_{v_2,\cdots,v_d} T_2^{-\kappa+\frac{3}{2}} d_{v_2\cdots,v_d} + \int_{v_2,\cdots,v_d} Z_1 \left(\sum_{j=2}^{d} v_j^2\right) \left(1 + \frac{1}{\tilde{\eta}\kappa} \cdot \sum_{j=2}^{d} \lambda_j v_j^2\right)^{-\kappa+\frac{1}{2}} d_{v_2\cdots,v_d}
\end{aligned}
$$

For term $\int_{v_2,\cdots,v_d} T_2^{-\frac{1}{\kappa}+\frac{3}{2}} d_{v_2\cdots,v_d}$ in above equation, we have

$$
\begin{aligned}
&\int_{v_2,\cdots,v_d} T_2^{-\kappa+\frac{3}{2}} d_{v_2\cdots,v_d} \\
=&\int_{v_3,\cdots,v_d} T_3^{-\kappa+2} ((\tilde{\eta}\kappa)^{-1}\lambda_2)^{-\frac{1}{2}} B\left(\frac{1}{2}, \kappa - 2\right) d_{v_3,\cdots,v_d} \\
=&\int_{v_4,\cdots,v_d} T_4^{-\kappa+\frac{5}{2}} ((\tilde{\eta}\kappa)^{-1}\lambda_2)^{-\frac{1}{2}} ((\tilde{\eta}\kappa)^{-1}\lambda_3)^{-\frac{1}{2}} B\left(\frac{1}{2}, \kappa - \frac{5}{2}\right) B\left(\frac{1}{2}, \kappa - 2\right) d_{v_4,\cdots,v_d} \\
=&\int_{v_d} T_d^{-\kappa+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}\times d} \prod_{j=2}^{d-1} ((\tilde{\eta}\kappa)^{-1}\lambda_j)^{-\frac{1}{2}} \prod_{j=2}^{d-1} B\left(\frac{1}{2}, \kappa - (\frac{j}{2}+1)\right) d_{v_d} \\
=&\prod_{j=2}^{d} ((\tilde{\eta}\kappa)^{-1}\lambda_j)^{-\frac{1}{2}} \prod_{j=2}^{d} B\left(\frac{1}{2}, \kappa - (\frac{j}{2}+1)\right)
\end{aligned}
$$

Let $A_j = ((\tilde{\eta}\kappa)^{-1}\lambda_j)^{-\frac{3}{2}} B\left(\frac{3}{2}, \kappa - (\frac{j}{2}+1)\right)$. According to the above two equations, we can get the recursion

$$
\begin{aligned}
& 2Z \int v^T v T_1^{-\kappa} dv \\
=& A_1 \cdot \int T_2^{-\kappa+\frac{3}{2}} + Z_1 \int_{v_2,\cdots,v_d} \left(\sum_{j=2}^{d} v_j^2\right) T_2^{-\kappa+\frac{1}{2}} d_{v_2\cdots,v_d} \\
=& A_1 \cdot \int T_2^{-\kappa+\frac{3-1}{2}} d_{v_2\cdots v_d} + Z_1 \cdot A_2 \int T_3^{-\kappa+\frac{4}{2}} d_{v_3\cdots,v_d} + Z_1 Z_2 \int \left(\sum_{j=3}^{d} v_j^2\right) T_3^{-\kappa+\frac{1}{2}} d_{v_3\cdots,v_d} \\
=& \sum_{j=1}^{d-1} A_j \prod_{k=1}^{j-1} Z_k \int T_{j+1}^{-\kappa+\frac{j+1+1}{2}} d_{v_{j+1},\cdots,v_d} + \prod_{k=1}^{d-1} Z_k \int v_d^2 T_d^{-\kappa+\frac{d-1}{2}} dv_d \\
=& \sum_{j=1}^{d-1} (\frac{\lambda_j}{\tilde{\eta}\kappa})^{-\frac{3}{2}} B\left(\frac{3}{2}, \kappa - (\frac{j}{2}+1)\right) \prod_{k=1}^{j-1} (\frac{\lambda_k}{\tilde{\eta}\kappa})^{-\frac{1}{2}} B\left(\frac{1}{2}, \kappa - \frac{k}{2}\right) \prod_{s=j+1}^{d} ((\frac{\lambda_s}{\tilde{\eta}\kappa})^{-\frac{1}{2}} \prod_{s=j+1}^{d} B\left(\frac{1}{2}, \kappa - (\frac{s}{2}+1)\right) \\
&+ \prod_{j=1}^{d-1} (\frac{\lambda_j}{\tilde{\eta}\kappa})^{-\frac{1}{2}} B(\frac{1}{2}, \kappa - \frac{j}{2} - 1) \cdot (\frac{\lambda_d}{\tilde{\eta}\kappa})^{-\frac{3}{2}} B(\frac{3}{2}, \kappa - (\frac{d}{2}+1)) \\
=& \frac{\sqrt{\pi^d}\Gamma(\kappa - \frac{d}{2} - 1) Tr(H^{-1}\Sigma_g)}{2\Gamma(\kappa)\sqrt{(\tilde{\eta}\kappa)^{-(d+2)} \det(H^{-1}\Sigma_g)}}
\end{aligned}
$$

We have

$$
\begin{aligned}
III =& \frac{\sqrt{\pi^d}\Gamma(\kappa - \frac{d}{2} - 1) Tr(H^{-1}\Sigma_g)}{4\Gamma(\kappa)\sqrt{(\tilde{\eta}\kappa)^{-(d+2)} \det(H^{-1}\Sigma_g)}} \cdot \prod_{j=1}^{d} ((\tilde{\eta}\kappa)^{-1}\lambda_j)^{\frac{1}{2}} \cdot \frac{\Gamma(\kappa)}{\sqrt{\pi^d}\Gamma(\kappa - \frac{d}{2})} \\
=& \frac{\tilde{\eta}\kappa Tr(H^{-1}\Sigma_g)}{4(\kappa - \frac{d}{2} - 1)}
\end{aligned}
$$

Similarly, for the fourth term in Eq.(2), we have $IV = \frac{\kappa d}{2(\kappa - \frac{d}{2} - 1)}$. Combining all the results together, we can get $KL(p||p') = \frac{1}{2}\log\frac{\det(H)}{(\tilde{\eta}\kappa)^d \det(\Sigma_g)} + \log\frac{\Gamma(\kappa)}{\Gamma(\kappa - \frac{d}{2})} + \frac{Tr(\tilde{\eta}\Sigma_g H^{-1}) - 2d}{4\left(1 - \frac{1}{\kappa}\left(\frac{d}{2} - 1\right)\right)} + \frac{d}{2}\log 2$. Using the fact that $\log\frac{\Gamma(\kappa)}{\Gamma(\kappa - \frac{d}{2})} \leq \frac{d}{2}\log\kappa$, we have $KL(p||p') \leq \frac{1}{2}\log\frac{\det(H)}{\det(\Sigma_g)} + \frac{Tr(\tilde{\eta}\Sigma_g H^{-1}) - 2d}{4\left(1 - \frac{1}{\kappa}\left(\frac{d}{2} - 1\right)\right)} + \frac{d}{2}\log\frac{2}{\tilde{\eta}}$.

REFERENCES

McAllester, David A. 1999. PAC-Bayesian model averaging. *Pages 164–170 of: Proceedings of the twelfth annual conference on Computational learning theory.*