Supplementary material for the paper: Stopping criterion for active learning based on deterministic generalization bounds

Hideaki Ishibashi
Kyushu Institute of Technology

Hideitsu Hino
The Institute of Statistical Mathematics/RIKEN AIP

1 Proof of Theorem 1 and Corollary 1

We demonstrate the following three lemmas to prove Theorem 1 and Corollary 1.

Lemma 1. *(Donsker and Varadhan, 1975; McAllester, 2003) Let $\phi : \mathcal{F} \to \mathbb{R}$ be any measurable function. Then, the following inequality holds:*

$$\mathop{\mathbb{E}}_{p(f)}[\phi(f)] \leq D_{KL}[p(f)||p'(f)] + \log \mathop{\mathbb{E}}_{p'(f)}\left[e^{\phi(f)}\right]. \quad (1)$$

Here, p and p' are the probability distributions on $\mathcal{F}$.

Lemma 2. *(Simic, 2008) Let $h : X \to \mathbb{R}$ be a concave function, where $X \in [a, b]$. p is a probability distribution with respect to X. We denote the difference of Jensen's inequality by $J(p, X)$, that is,*

$$J(p, X) = h(\mathop{\mathbb{E}}_{p}[X]) - \mathop{\mathbb{E}}_{p}[h(X)]. \quad (2)$$

Then, the following inequality holds:

$$J(p, X) \leq 2h(\frac{a+b}{2}) - h(a) - h(b). \quad (3)$$

Lemma 3. *Let $q(f|S)$ and $q(f|S')$ be the posteriors with respect to f given $S = (X, Y)$ and $S' = (X', Y')$, respectively. We assume that the prior of $q(f|S)$ is the same as that of $q(f|S')$. Then, the following inequality holds:*

$$D_{KL}[q(f|S)||q(f|S')] = D_{KL}\left[q(\mathbf{f}_{X_+}|S))||q(\mathbf{f}_{X_+}|S'))\right], \quad (4)$$

where $X_+ := X \cup X'$.

Proof. Let X_Ω be a universal set of input data. We denote X_Ω / X_+ by X_*. Then, from the chain rule of KL divergence (Gray, 2011), the following equation hold:

$$\begin{aligned} & D_{KL}[q(f|S)||q(f|S')] & (5) \\ =& D_{KL}\left[q(\mathbf{f}_{X_+}|S)||q(\mathbf{f}_{X_+}|S')\right] \\ & + \mathop{\mathbb{E}}_{q(\mathbf{f}_{X_+}|S)}\left[D_{KL}\left[q(\mathbf{f}_{X_*}|\mathbf{f}_{X_+}, S)||q(\mathbf{f}_{X_*}|\mathbf{f}_{X_+}, S')\right]\right]. & (6) \end{aligned}$$

We denote the prior of $q(\mathbf{f}_{X_*}, \mathbf{f}_{X_+}|S)$ and $q(\mathbf{f}_{X_*}, \mathbf{f}_{X_+}|S')$ by $p(\mathbf{f}_{X_*}, \mathbf{f}_{X_+})$. Then, from the Bayesian theorem, the following equation holds:

$$\begin{aligned} q(\mathbf{f}_{X_*}|\mathbf{f}_{X_+}, S)) =& \frac{p(\mathbf{f}_{X_*}, \mathbf{f}_{X_+}|S)}{p(\mathbf{f}_{X_+}|S)} & (7) \\ =& \frac{p(Y|\mathbf{f}_{X_+}, X)p(\mathbf{f}_{X_*}|\mathbf{f}_{X_+})p(\mathbf{f}_{X_+})}{p(Y|X)} \\ & \times \frac{p(Y|X)}{p(Y|\mathbf{f}_{X_+}, X)p(\mathbf{f}_{X_+})} & (8) \\ =& \frac{p(\mathbf{f}_{X_*}, \mathbf{f}_{X_+})}{p(\mathbf{f}_{X_+})} = p(\mathbf{f}_{X_*}|\mathbf{f}_{X_+}). & (9) \end{aligned}$$

Similarly, $q(\mathbf{f}_{X_*}|\mathbf{f}_{X_+}, S') = p(\mathbf{f}_{X_*}|\mathbf{f}_{X_+})$ also holds. Therefore, if the prior of $q(f|S)$ is the same as that of $q(f|S')$, the second term of Eq. (6) is zero. □

Proof of Theorem 1

Proof. By using Lemmas 1 and 2, the upper bound for $\mathcal{R}(q(f|S), q(f|S'))$ is obtained as follows:

$$\begin{aligned} & \mathcal{R}(q(f|S), q(f|S')) & (10) \\ \leq & D_{KL}[q(f|S)||q(f|S')] \\ & + \log \mathop{\mathbb{E}}_{q(f|S')}\left[e^{\mathcal{L}_{\mathcal{D}}(f)}\right] - \mathop{\mathbb{E}}_{q(f|S')}\left[\log e^{\mathcal{L}_{\mathcal{D}}(f)}\right] & (11) \\ \leq & D_{KL}[q(f|S)||q(f|S')] + 2\log\frac{e^a + e^b}{2} - a - b. & (12) \end{aligned}$$

By applying Lemma 1 to Eq. (1), we obtain Eq. (11). Because the sum of the second and third terms of Eq. (11) is the difference of Jensen's inequality, Lemma 2 can be applied to it. Moreover, from $l \in [a, b]$, $\mathcal{L}_{\mathcal{D}}(f) \in [a, b]$ holds. Therefore, we obtain Eq. (12). □

Proof of Corollary 1

Proof. The proof is evident from Theorem 1 and Lemma 3. □

2 Calculation of KL divergence between GPs

Lemma 4. *Let $q(f|S_t)$ and $q(f|S_{t+1})$ be the GP posteriors given $S_t = \{(x_i, y_i)\}_{i=1}^{t}$ and $S_{t+1} = \{(x_i, y_i)\}_{i=1}^{t+1}$, respectively. We assume that the prior of $q(f|S_t)$ is the same as that of $q(f|S_{t+1})$. Let μ_t and σ_t and β be the mean and covariance functions of $q(f|S_t)$ and accuracy of Gaussian noise, respectively.Then the following equation holds:*

$$\begin{aligned}&D_{KL}\left[q(f|S_t)||q(f|S_{t+1})\right]\\ =&\frac{1}{2}\beta\sigma_t(x_{t+1},x_{t+1}) - \frac{1}{2}\log\left(1+\beta\sigma_t(x_{t+1},x_{t+1})\right)\\ &+\frac{1}{2}\frac{\beta\sigma_t(x_{t+1},x_{t+1})}{\sigma_t(x_{t+1},x_{t+1})+\beta^{-1}}(y_{t+1}-\mu_t(x_{t+1}))^2. \quad (13)\end{aligned}$$

Proof. From Lemma 3, the following equation holds:

$$D_{KL}\left[q(f|S_t)||q(f|S_{t+1})\right] = D_{KL}\left[q(\mathbf{f}|S_t)||q(\mathbf{f}|S_{t+1})\right], \quad (14)$$

where $\mathbf{f} := (f(x_1), f(x_2), \cdots, f(x_{t+1}))$. When $S_{t+1} = (X_{t+1}, Y_{t+1})$ is observed, $q(\mathbf{f}|S_{t+1})$ can be described as follows:

$$\begin{aligned}q(\mathbf{f}|S_{t+1}) &= \frac{p(Y_{t+1}|\mathbf{f}, X_{t+1})p(\mathbf{f})}{p(Y_{t+1}|X_{t+1})}\\ &= \frac{p(y_{t+1}|\mathbf{f}, x_{t+1})p(Y_t|\mathbf{f}, X_t)p(\mathbf{f})}{\int p(y_{t+1}|\mathbf{f}', x_{t+1})p(Y_t|\mathbf{f}', X_t)p(\mathbf{f}')d\mathbf{f}'}\\ &= \frac{p(y_{t+1}|\mathbf{f}, x_{t+1})p(Y_t|X_t)q(\mathbf{f}|S_t)}{\int p(y_{t+1}|\mathbf{f}', x_{t+1})p(Y_t|X_t)q(\mathbf{f}'|S_t)d\mathbf{f}'}\\ &= \frac{p(y_{t+1}|\mathbf{f}, x_{t+1})q(\mathbf{f}|S_t)}{p(y_{t+1}|x_{t+1})}. \quad (15)\end{aligned}$$

From this equation, $D_{KL}\left[q(\mathbf{f}|S_{t-1})||q(\mathbf{f}|S_t)\right]$ can be rewritten as follows:

$$\begin{aligned}&D_{KL}\left[q(\mathbf{f}|S_t)||q(\mathbf{f}|S_{t+1})\right]\\ =&\mathop{\mathbb{E}}_{q(\mathbf{f}|S_t)}\left[\log\frac{q(\mathbf{f}|S_t)p(y_{t+1}|x_{t+1})}{p(y_{t+1}|\mathbf{f}, x_{t+1})q(\mathbf{f}|S_t)}\right]\\ =&\log p(y_{t+1}|x_{t+1}) - \mathop{\mathbb{E}}_{q(\mathbf{f}|S_t)}\left[\log p(y_{t+1}|\mathbf{f}, x_{t+1})\right]\\ =&\log\int p(y_{t+1}|f_{t+1})q(f_{t+1}|S_t)df_{t+1}\\ &-\int q(f_{t+1}|S_t)\log p(y_{t+1}|f_{t+1})df_{t+1}, \quad (16)\end{aligned}$$

where $f_{t+1} := f(x_{t+1})$. The first term of Eq. (16) becomes logarithm of a normal distribution since $p(y_{t+1}|f_{t+1})$ and $q(f_{t+1}|S_t)$ are normal distributions. Specifically, from $p(y_{t+1}|f_{t+1}) = \mathcal{N}(y_{t+1}|f_{t+1}, \beta^{-1})$ and $p(f_{t+1}|S_t) = \mathcal{N}(f_{t+1}|\mu_t(x_{t+1}), \sigma_t(x_{t+1}, x_{t+1}))$, the following equation holds:

$$\begin{aligned}&\log\int p(y_{t+1}|f_{t+1})q(f_{t+1}|S_t)df_{t+1}\\ &=\log\mathcal{N}(y_{t+1}|\mu_t(x_{t+1}), \sigma_t(x_{t+1},x_{t+1})+\beta^{-1}) \quad (17)\end{aligned}$$

The second term can be rewritten as follows:

$$\begin{aligned}&-\int q(f_{t+1}|S_t)\log p(y_{t+1}|f_{t+1})df_{t+1}\\ =&\mathop{\mathbb{E}}_{q(f_{t+1}|S_t)}\left[\frac{\beta}{2}(y_{t+1}-f_{t+1})^2\right]+\frac{1}{2}\log 2\pi\beta^{-1}\\ =&\frac{\beta}{2}\left(y_{t+1}^2 - 2y_{t+1}\mathbb{E}[f_{t+1}] + \mathbb{E}[f_{t+1}^2]\right)+\frac{1}{2}\log 2\pi\beta^{-1}\\ =&\frac{\beta}{2}(y_{t+1}-\mu_t(x_{t+1}))^2+\frac{\beta}{2}\sigma_t(x_{t+1},x_{t+1})+\frac{1}{2}\log 2\pi\beta^{-1}\\ &\quad (18)\end{aligned}$$

From the above, the lemma is derived as follows:

$$\begin{aligned}&D_{KL}\left[q(f|S_t)||q(f|S_{t+1})\right]\\ =&-\frac{(y_{t+1}-\mu_t(x_{t+1}))^2}{2(\sigma_t(x_{t+1},x_{t+1})+\beta^{-1})}-\frac{1}{2}\log 2\pi(\sigma_t(x_{t+1},x_{t+1})+\beta^{-1})\\ &+\frac{\beta}{2}(y_{t+1}-\mu_t(x_{t+1}))^2+\frac{\beta}{2}\sigma_t(x_{t+1},x_{t+1})+\frac{1}{2}\log 2\pi\beta^{-1}\\ =&\frac{1}{2}\beta\sigma_t(x_{t+1},x_{t+1})-\frac{1}{2}\log\left(1+\beta\sigma_t(x_{t+1},x_{t+1})\right)\\ &+\frac{1}{2}\frac{\beta\sigma_t(x_{t+1},x_{t+1})}{\sigma_t(x_{t+1},x_{t+1})+\beta^{-1}}(y_{t+1}-\mu_t(x_{t+1}))^2. \quad (19)\end{aligned}$$

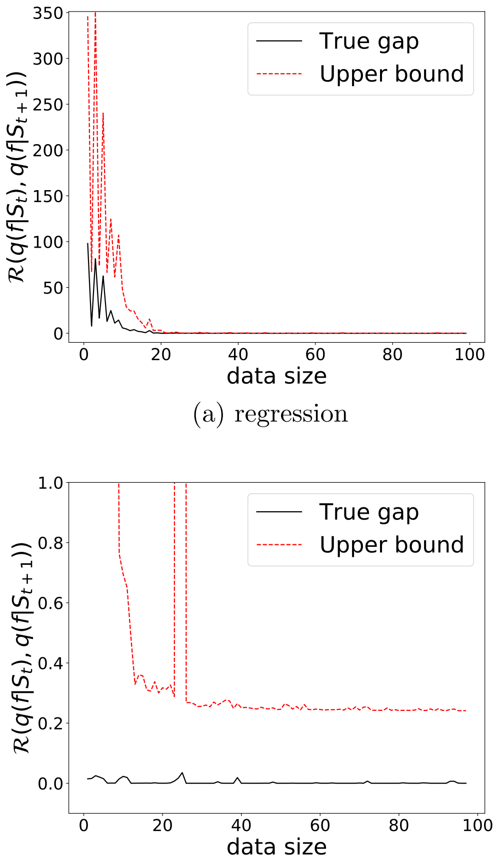
□

3 Tightness of the proposed upper bound

We experimentally evaluated the tightness of the proposed upper bound of a gap between expected generalization errors before and after adding a new sample by comparing to the true gap approximated by using a large amount of test data.

For the regression task, we used the artificial data used in the experiment of Section 5, while, for the classification task, we used the generated data $y_i = \mathrm{sgn}(\sin(2\pi x_i))$, where $\mathrm{sgn}(\cdot)$ is the sign function. The kernel function and its hyperparameter are determined in the same manner explained in Section 5.

Figures 1 (a) and 1 (b) show the gaps between the expected generalization errors and their upper bounds for (a) regression and (b) classification tasks. We see that (1) increasing the data size leads to a tight upper bound in both cases of regression and classification, and the KL-divergence term converges to zero. Moreover, (2) the bound could be trivial when the KL divergence takes a large value (> 1), particularly in classification setting. Also, as the KL-divergence is always



(a) regression

(b) classification

Figure 1: Gap between the expected generalization errors, and its upper bound for (a) regression and (b) classification settings.

non-negative, when the gap of the expected generalization error is negative, the bound is meaningless. As can be seen from Fig. 1 (a), the bound works well in a regression setting and the offers reasonable tightness.

References

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975. doi: 10.1002/cpa.3160280102. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160280102`.

David A. McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, Apr 2003. ISSN 1573-0565. doi: 10.1023/A:1021840411064. URL `https://doi.org/10.1023/A:1021840411064`.

Slavko Simic. On a global upper bound for jensen's inequality. *Journal of Mathematical Analysis and Applications*, 343(1):414 – 419, 2008. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2008.01.060. URL `http://www.sciencedirect.com/science/article/pii/S0022247X08000814`.

Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag New York, Inc., 2011. doi: 10.1007/978-1-4419-7970-4.