LETTER

Unsupervised Weight Parameter Estimation for Exponential Mixture Distribution Based on Symmetric Kullback-Leibler Divergence*

Masato UCHIDA[†a)], ***Member***

SUMMARY When there are multiple component predictors, it is promising to integrate them into one predictor for advanced reasoning. If each component predictor is given as a stochastic model in the form of probability distribution, an exponential mixture of the component probability distributions provides a good way to integrate them. However, weight parameters used in the exponential mixture model are difficult to estimate if there is no training samples for performance evaluation. As a suboptimal way to solve this problem, weight parameters may be estimated so that the exponential mixture model should be a balance point that is defined as an equilibrium point with respect to the distance from/to all component probability distributions. In this paper, we propose a weight parameter estimation method that represents this concept using a symmetric Kullback-Leibler divergence and generalize this method.

key words: *ensemble learning, parameter estimation, exponential mixture model, symmetric Kullback-Leibler divergence*

1. Introduction

A learning method that builds an advanced predictor by integrating multiple component predictors is generically referred to as ensemble learning. Representative ensemble learning methods include Bagging [2], Boosting [3], and their derivatives such as AdaBoost [4], LogitBoost [5], and MadaBoost [6]. The main focus of these methods is to build an efficient ensemble predictor by effectively using the given training samples for performance evaluation. For example, in the boosting method called AdaBoost [4], weak learners/predictors (slightly better than random guessing) are trained iteratively while the intensity is increased for the misclassified samples and decreased for correctly classified samples in the given training samples. The ensemble predictor is then built by integrating the trained component predictors in accordance with their performance with respect to the given training samples.

On the other hand, the simple ensemble learning method proposed in [7] builds an ensemble predictor using the weighted average of trained component predictors. The weights are determined in accordance with the performance of the ensemble predictor with respect to the given training samples. In our previous works [8], [9], the algorithmic structure of the simple ensemble learning method was investigated based on a general framework, which is characterized by the exponential mixture of probability distributions and the Kullback-Leibler (KL) divergence. Here, the exponential mixture of probability distributions corresponds to the ensemble predictor, and the KL divergence corresponds to the loss function of ensemble learning.

The limitation of these previous studies [2]–[7] on ensemble learning is that they cannot be applied when the training samples are not available for evaluating the performance of an ensemble predictor. However, even when the training samples are not given, the essential strategy of ensemble learning itself (i.e., integration of multiple component predictors) is potentially useful if multiple component predictors are available for building the ensemble predictor. In this paper, we propose an *unsupervised* method for estimating weight parameter of ensemble predictor without using the training samples.

The proposed *unsupervised* weight parameter estimation method is a variant of our previous method [10]. The main difference between these methods is that the proposed method uses a symmetric KL divergence while our previous method uses the ordinal (i.e., asymmetric) KL divergence. More specifically, the proposed method is designed on the basis of the policy that the ensemble predictor (i.e., the exponential mixture of probability distributions) should be a balance point that is defined as an equilibrium point with respect to the distance (i.e., the Kullback-Leibler divergence) from/to all component predictors (i.e., all component probability distributions).

This paper is organized as follows. Section 2 provides a brief review of supervised and unsupervised weight parameter estimation methods in ensemble learning that are based on the general framework provided in our previous works [8]–[10]. Section 3 describes the proposed unsupervised weight parameter estimation method in ensemble learning and generalizes it based on α-mixture model and α-divergnece. Section 4 concludes this paper.

2. Previous Works

2.1 Supervised Weight Parameter Estimation

Let $\mathcal{P}(\mathcal{Z})$ be the set of all probability density functions

Manuscript received June 5, 2015.
Manuscript revised July 7, 2015.

[†]The author is with the Department of Electrical, Electronics and Computer Engineering, Chiba Institute of Technology, Narashino-shi, 275-016 Japan.

[*]A part of this paper appeared in the Proceedings of SCIS & ISIS 2014 [1].

a) E-mail: masato.uchida@ieee.org

DOI: 10.1587/transfun.E98.A.2349

(pdfs) on $\mathcal{Z}$ ($\in \Re^m$). Let $\mathcal{P}_i(\mathcal{Z})$, $i = 1, \ldots, M$, be a subset of $\mathcal{P}(\mathcal{Z})$. If we have multiple component pdfs, $p_i(z)$ ($\in \mathcal{P}_i(\mathcal{Z})$), $i = 1, \ldots, M$, we can define a new pdf by using them as follows:

$$\bar{p}_{\boldsymbol{\beta}}(z) \stackrel{\text{def}}{=} \frac{\prod_{i=1}^{M} p_i(z)^{\beta_i}}{\int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(z)^{\beta_i} \mathrm{d}z}, \tag{1}$$

where we assume

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{M-1})^T \in \Re^{M-1}, \quad \sum_{i=1}^{M} \beta_i = 1,$$

$$\int_{\mathcal{Z}} \prod_{i=1}^{M} p_i(z)^{\beta_i} \mathrm{d}z < \infty.$$

We hereinafter refer to $\bar{p}_{\boldsymbol{\beta}}(z)$ as an *exponential mixture model*. Here, $\bar{p}_{\boldsymbol{\beta}}(z)$ and $p_i(z)$, $i = 1, \ldots, M$, correspond to ensemble and component predictors in ensemble learning, respectively.

In our previous works [8], [9], we have shown that the process of the simple ensemble learning method proposed in [7] can be reduced to a sequence of three operations:

$$\hat{p}_i(z) \stackrel{\text{def}}{=} \arg \min_{p(z) \in \mathcal{P}_i(\mathcal{Z})} D_{\mathrm{KL}}(p_* \| p), \tag{2}$$

$$\bar{\hat{p}}_{\boldsymbol{\beta}}(z) \stackrel{\text{def}}{=} \arg \min_{p(z) \in \mathcal{P}(\mathcal{Z})} \sum_{i=1}^{M} \beta_i D_{\mathrm{KL}}(p \| \hat{p}_i), \tag{3}$$

$$\hat{\boldsymbol{\beta}}_S \stackrel{\text{def}}{=} \arg \min_{\boldsymbol{\beta}} D_{\mathrm{KL}}(p_* \| \bar{\hat{p}}_{\boldsymbol{\beta}}), \tag{4}$$

where $p_*(z)$ ($\in \mathcal{P}(\mathcal{Z})$) denotes the target distribution from which training samples are drawn and $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes Kullback-Leibler divergence. The obtained exponential mixture model, $\bar{\hat{p}}_{\hat{\boldsymbol{\beta}}_S}(z)$, is used for problem solving and decision making. Note that the estimation in Eq. (4) is supervised in the sense that it is executed with the training samples from target distribution. The more detailed review about the above formulation can be found in [10].

2.2 Unsupervised Weight Parameter Estimation

The supervised weight parameter estimation that obtains $\hat{\boldsymbol{\beta}}_S$ through Eq. (4) is the recommended strategy. However, Eq. (4) can be executed only when the training samples (i.e., $p_*(z)$) are available. On the other hand, we previously proposed an *unsupervised* method for estimating the weight parameters without needing the training samples [10]. This *unsupervised* weight parameter estimation method is formulated on the basis of the general framework of ensemble learning provided in [8], [9] as follows:

$$\hat{\boldsymbol{\beta}}_{U_{\mathrm{KL}}} = \arg \max_{\boldsymbol{\beta}} \sum_{i=1}^{M} \beta_i D_{\mathrm{KL}}(\bar{p}_{\boldsymbol{\beta}} \| p_i). \tag{5}$$

Equation (5) is shown to be valid for weight parameter estimation in an unsupervised situation if the component predictors used to build the ensemble predictor perform with similar efficiency, i.e.,

$$D_{\mathrm{KL}}(p_* \| p_i) = D_{\mathrm{KL}}(p_* \| p_j) \tag{6}$$

holds ($\forall i, j = 1, 2, \ldots, M$). Although the condition given by Eq. (6) is arbitrary, it is acceptable when exact knowledge of the performances of the component predictors is unavailable.

3. Proposed Method

3.1 Formulation

In this paper, we propose a new *unsupervised* weight parameter estimation method that is a variant of our previous method given by Eq. (5). The proposed method is formulated as

$$\hat{\boldsymbol{\beta}}_{U_{\text{S-KL}}} = \arg \max_{\boldsymbol{\beta}} \sum_{i=1}^{M} \beta_i D_{\text{S-KL}}(\bar{p}_{\boldsymbol{\beta}} \| p_i), \tag{7}$$

where $D_{\text{S-KL}}(\cdot\|\cdot)$ is a symmetric KL divergence defined as

$$D_{\text{S-KL}}(p\|q) = \frac{1}{2}\left\{D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p)\right\}, \quad \forall p(z),\ q(z) \in \mathcal{P}(\mathcal{Z}). \tag{8}$$

Note that the proposed method can be executed without using the training samples (i.e., $p_*(z)$) as well as our previous method given by Eq. (5). However, the objective function used in the proposed method is symmetric with respect to $\bar{p}_{\boldsymbol{\beta}}(z)$ and $p_i(z)$, while that used in our previous method is asymmetric. That is, the proposed method represents a policy that the ensemble predictor (i.e., the exponential mixture of probability distributions, $\bar{p}_{\boldsymbol{\beta}}(z)$) should be a balance point that is the equilibrium point with respect to the distance (i.e., the Kullback-Leibler divergence) from/to all component predictors (i.e., all component probability distributions, $p_i(z)$, $i = 1, 2, \ldots, M$).

Here, it is important to note that

$$\sum_{i=1}^{M} \beta_i \{D_{\mathrm{KL}}(\bar{p}_{\boldsymbol{\beta}} \| p_i) + D_{\mathrm{KL}}(p_i \| \bar{p}_{\boldsymbol{\beta}})\} = \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_i \beta_j D_{\mathrm{KL}}(p_i \| p_j) \tag{9}$$

holds. The derivation of Eq. (9) is given in Appendix A. By using Eq. (9), Eq. (7) can be transformed into simple quadratic optimization problem as

$$\hat{\boldsymbol{\beta}}_{U_{\text{S-KL}}} = \arg \max_{\boldsymbol{\beta}} \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_i \beta_j D_{\text{S-KL}}(p_i \| p_j). \tag{10}$$

3.2 Weight Parameter Estimation

Let us define the objective function in Eq. (10) as

$$c(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \sum_{i=1}^{M}\sum_{j=1}^{M}\beta_i\beta_j D_{\text{S-KL}}(p_i\|p_j).$$

Using $\beta_M = 1 - \sum_{i=1}^{M-1}\beta_i$, we obtain

$$\nabla_{\boldsymbol{\beta}} c(\boldsymbol{\beta}) = \boldsymbol{\delta} - A\boldsymbol{\beta}, \tag{11}$$

where $A = (a_{i,j})_{(M-1)\times(M-1)}$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_{M-1})^T$ are defined as

$$a_{i,j} = D_{\text{S-KL}}(p_i\|p_M) + D_{\text{S-KL}}(p_j\|p_M) - D_{\text{S-KL}}(p_i\|p_j),$$
$$\delta_i = D_{\text{S-KL}}(p_i\|p_M).$$

If A is non-singular, $\nabla_{\boldsymbol{\beta}} c(\boldsymbol{\beta}) = 0$ yields

$$\hat{\boldsymbol{\beta}}_{U_{\text{S-KL}}} = A^{-1}\boldsymbol{\delta}.$$

For an interesting example, let us consider a case that satisfies

$$D_{\text{S-KL}}(p_i\|p_j) = \varepsilon, \quad (\forall i, j = 1, 2, \ldots, M,\ i \neq j) \tag{12}$$

where ε is a positive constant. In this case, we obtain

$$A^{-1} = \frac{1}{\varepsilon}\begin{pmatrix} \frac{M-1}{M} & -\frac{1}{M} & -\frac{1}{M} & \cdots & -\frac{1}{M} \\ -\frac{1}{M} & \frac{M-1}{M} & -\frac{1}{M} & \cdots & -\frac{1}{M} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -\frac{1}{M} & \cdots & -\frac{1}{M} & \frac{M-1}{M} & -\frac{1}{M} \\ -\frac{1}{M} & \cdots & -\frac{1}{M} & -\frac{1}{M} & \frac{M-1}{M} \end{pmatrix},$$
$$\boldsymbol{\delta} = (\varepsilon, \varepsilon, \ldots, \varepsilon)^T.$$

Let $\hat{\boldsymbol{\beta}}_A$ be the value of $\boldsymbol{\beta}$ when Eq. (12) holds. We then obtain

$$\hat{\boldsymbol{\beta}}_A = \left(\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right)^T.$$

The above property characterizes the meaning of a simple average in the context of unsupervised weight parameter estimation.

Here, we consider the special case in which

$$p_i(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(z-\mu_i)}{\sigma^2}\right\}, \quad (i = 1,\ 2,\ \ldots,\ M). \tag{13}$$

In this case, Eqs. (5) and (7) provide the equivalent results in weight parameter estimation because the following equation holds.

$$D_{\text{KL}}(\bar{p}_{\boldsymbol{\beta}}\|p_i) = D_{\text{KL}}(p_i\|\bar{p}_{\boldsymbol{\beta}}).$$

3.3 Generalization

We can generalize the proposed method given by Eqs. (7) and (10) by replacing KL divergence with α-divergence, where α-divergence between p and q is an extension of KL divergence defined as

$$D_\alpha(p\|q) = \frac{1}{\alpha(1-\alpha)}\int_{\mathcal{Z}} p(z)\left\{1 - \left(\frac{q(z)}{p(z)}\right)^\alpha\right\}\mathrm{d}z.$$

Note that

$$\lim_{\alpha\to 0} D_\alpha(p\|q) = D_{\text{KL}}(p\|q), \quad \lim_{\alpha\to 1} D_\alpha(p\|q) = D_{\text{KL}}(q\|p)$$

hold. Here, we define symmetric α-divergence between $p(z)$ and $q(z)$ as

$$D_{\text{S-}\alpha}(p\|q) = \frac{1}{2}\left\{D_\alpha(p\|p) + D_\alpha(q\|p)\right\}. \tag{14}$$

and the α-mixture of probability distributions as

$$\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}(z) = \frac{\{\sum_{i=1}^{M}\beta_i p_i(z)^\alpha\}^{\frac{1}{\alpha}}}{\int_{\mathcal{Z}}\{\sum_{i=1}^{M}\beta_i p_i^\alpha(z)\}^{\frac{1}{\alpha}}\mathrm{d}z}. \tag{15}$$

Note that

$$\lim_{\alpha\to 0} D_{\text{S-}\alpha}(p\|q) = D_{\text{S-KL}}(p\|q),$$
$$\lim_{\alpha\to 1} D_{\text{S-}\alpha}(p\|q) = D_{\text{S-KL}}(q\|p)$$

and

$$\lim_{\alpha\to 0}\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}(z) = \bar{p}_{\boldsymbol{\beta}}(z),$$
$$\lim_{\alpha\to 1}\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}(z) \stackrel{\text{def}}{=} \tilde{p}_{\boldsymbol{\beta}}(z) = \sum_{i=1}^{M}\beta_i p_i(z)$$

hold. Therefore, the α-mixture model includes the exponential and linear mixture models as special cases.

Then, the generalization of Eqs. (7) and (10) are given as

$$\hat{\boldsymbol{\beta}}^{(1)}_{U_{\text{S-}\alpha}} = \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^{M}\beta_i D_{\text{S-}\alpha}(\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}\|p_i) \tag{16}$$

and

$$\hat{\boldsymbol{\beta}}^{(2)}_{U_{\text{S-}\alpha}} = \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_i\beta_j D_{\text{S-}\alpha}(p_i\|p_j), \tag{17}$$

respectively. Note that the proposed methods given by Eqs. (7) and (10) are equivalent to each other because Eq. (9) holds. However, two generalizations given by Eqs. (16) and (17) are not equivalent because

$$\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_i\beta_j D_\alpha(p_i\|p_j)$$
$$= \sum_{i=1}^{M}\beta_i\left\{D_\alpha(\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}\|p_i) + D_\alpha(p_i\|\bar{p}_{\boldsymbol{\beta}}^{(\alpha)})\right\}$$
$$- \alpha(1-\alpha)\left(\sum_{i=1}^{M}\beta_i D_\alpha(\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}\|p_i)\right)\left(\sum_{i=1}^{M}\beta_i D_\alpha(p_i\|\bar{p}_{\boldsymbol{\beta}}^{(\alpha)})\right) \tag{18}$$

holds. The derivation of Eq. (18) is given in Appendix B.

Note that, by taking $\alpha \to 0$ and $\alpha \to 1$ as limits, Eq. (18) reduces to

$$\sum_{i=1}^{M}\sum_{j=1}^{M}\beta_i\beta_j D_{\mathrm{KL}}(p_i\|p_j)$$
$$= \sum_{i=1}^{M}\beta_i\left\{D_{\mathrm{KL}}(\bar{p}_{\boldsymbol{\beta}}\|p_i) + D_{\mathrm{KL}}(p_i\|\bar{p}_{\boldsymbol{\beta}})\right\} \quad (19)$$
$$= \sum_{i=1}^{M}\beta_i\left\{D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\beta}}\|p_i) + D_{\mathrm{KL}}(p_i\|\tilde{p}_{\boldsymbol{\beta}})\right\}. \quad (20)$$

Therefore, the weight parameters of the exponential mixture model, $\bar{p}_{\boldsymbol{\beta}}(z)$, estimated by the proposed method (i.e., the maximization of Eq. (19), see Eqs. (7) and (10)) are equivalent to the weight parameters of the liner mixture model, $\tilde{p}_{\boldsymbol{\beta}}(z)$, estimated by the maximization of Eq. (20) with respect to $\boldsymbol{\beta}$. This means that, in the context of the proposed method, the weight parameters of the exponential mixture model, $\bar{p}_{\boldsymbol{\beta}}(z)$, play the same role in the linear mixture model, $\tilde{p}_{\boldsymbol{\beta}}(z)$.

3.4 Alternative Perspective of the Proposed Method

Note that

$$\frac{1}{2}\left\{D_{\mathrm{KL}}(p_*\|\bar{p}_{\boldsymbol{\beta}}) + D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\beta}}\|p_*) - D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\beta}}\|\bar{p}_{\boldsymbol{\beta}})\right\}$$
$$= \sum_{i=1}^{M}\beta_i D_{\text{S-KL}}(p_*\|p_i) - \sum_{i=1}^{M}\sum_{j=1}^{M}\beta_i\beta_j D_{\text{S-KL}}(p_i\|p_j) \quad (21)$$

holds. The derivation of Eq. (21) is given in Appendix C. Here, along the same lines as the discussion in Sect. 2.2, let us assume the component predictors used to build the ensemble predictor perform with similar efficiency as is the case in Eq. (6) with respect to the symmetric KL divergence, i.e.,

$$D_{\text{S-KL}}(p_*\|p_i) = D_{\text{S-KL}}(p_*\|p_j) \quad (22)$$

holds ($\forall i, j = 1, 2, \ldots, M$). Under this assumption, Eqs. (7) and (10) are equivalent to the minimization of

$$D_{\mathrm{KL}}(p_*\|\bar{p}_{\boldsymbol{\beta}}) + D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\beta}}\|p_*) - D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\beta}}\|\bar{p}_{\boldsymbol{\beta}}) \quad (23)$$

with respect to $\boldsymbol{\beta}$. This provides a viewpoint from which the formulation of the proposed method defined as Eqs. (7) and (10) can be derived. That is, the proposed method is intended to solve the problem to find weight parameters that can be commonly used for both the exponential mixture model, $\bar{p}_{\boldsymbol{\beta}}(z)$, and the liner mixture model, $\tilde{p}_{\boldsymbol{\beta}}(z)$ (see the first and second terms of Eq. (23)), while $\bar{p}_{\boldsymbol{\beta}}(z)$ and $\tilde{p}_{\boldsymbol{\beta}}(z)$ should differ from each other (see the third term of Eq. (23)).

4. Conclusion

This paper proposed an *unsupervised* weight parameter estimation method for an exponential mixture distribution. The proposed method is formulated by using a symmetric Kullback-Leibler divergence. In addition, we generalized the proposed method based on α-mixture model and α-divergence. We will numerically evaluate these methods in a future study.

Acknowledgment

This work was supported in part by the Japan Society for the Promotion of Science through Grants-in-Aid for Scientific Research (C) (26330112).

References

[1] M. Uchida, "Unsupervised weight parameter estimation for exponential mixture distribution based on symmetric kullback-leibler divergence," Proc. Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems (SCIS & ISIS 2014), Kitakyushu, Japan, pp.1126–1129, Dec. 2014.
[2] L. Breiman, "Bagging predictors," Machine Learning, vol.24, no.2, pp.123–140, Aug. 1996.
[3] R.E. Schapire, "The strength of weak learnability," Machine Learning, vol.5, no.2, pp.197–227, June 1990.
[4] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., vol.55, no.1, pp.119–139, Aug. 1997.
[5] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," The Annals of Statistics, vol.28, no.2, pp.337–407, April 2000.
[6] C. Domingo and O. Watanabe, "Madaboost: A modification of adaboost," Proc. 13th Annual Conference on Computational Learning Theory (COLT'00), pp.180–189, Stanford, CA, USA, June–July 2000.
[7] N. Ueda and R. Nakano, "Generalization error of ensemble estimators," Proc. International Conference on Neural Networks 1996 (ICNN'96), vol.1, pp.90–95, Washington, D.C., WA, USA, June 1996.
[8] M. Uchida, H. Shioya, and T. Da-te, "Analysis and extension of ensemble learning," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J84-D-II, no.7, pp.1537–1542, July 2001.
[9] M. Uchida and H. Shioya, "A study on assignment of weight parameters in ensemble learning model," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J86-D-II, no.7, pp.1131–1134, July 2003.
[10] M. Uchida, H. Shioya, and Y. Maehara, "Unsupervised weight parameter estimation method for ensemble learning," J. Mathematical Modelling and Algorithms, vol.10, no.4, pp.307–322, Dec. 2011.

Appendix A: Derivation of Eq. (9)

Let us note that

$$D_{\mathrm{KL}}(p\|\bar{p}_{\boldsymbol{\beta}}) = \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p\|p_i) - \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(\bar{p}_{\boldsymbol{\beta}}\|p_i),$$
$$(\forall p(z) \in \mathcal{P}(\mathcal{Z})). \quad (\mathrm{A}\cdot 1)$$

holds [10]. Therefore, by substituting $p_j(z)$ into $p(z)$ in (A· 1), we have

$$\sum_{j=1}^{M}\beta_j D_{\mathrm{KL}}(p_j\|\bar{p}_{\boldsymbol{\beta}})$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{M}\beta_i\beta_j D_{\mathrm{KL}}(p_j\|p_i) - \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(\bar{p}_{\boldsymbol{\beta}}\|p_i).$$

This indicates that Eq. (9) holds.

Appendix B: Derivation of Eq. (18)

Let us note that

$$\begin{aligned}
&C(\alpha)\\
&\stackrel{\mathrm{def}}{=} \left[\int_{\mathcal{Z}}\left\{\sum_{i=1}^{M}\beta_i p_i(z)^{\alpha}\right\}^{\frac{1}{\alpha}}\mathrm{d}z\right]^{\alpha}\\
&= \int_{\mathcal{Z}}\frac{\int_{\mathcal{Z}}\left\{\sum_{i=1}^{M}\beta_i p_i(z)^{\alpha}\right\}^{\frac{1-\alpha}{\alpha}}}{\left[\int_{\mathcal{Z}}\left\{\sum_{i=1}^{M}\beta_i p_i(z')^{\alpha}\right\}^{\frac{1}{\alpha}}\mathrm{d}z'\right]^{1-\alpha}}\left\{\sum_{i=1}^{M}\beta_i p_i(z)^{\alpha}\right\}\mathrm{d}z\\
&= \sum_{i=1}^{M}\beta_i\int_{\mathcal{Z}}\left\{\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}(z)\right\}^{1-\alpha}p_i(z)^{\alpha}\mathrm{d}z\\
&= 1-\alpha(1-\alpha)\sum_{i=1}^{M}\beta_i D_{\alpha}(\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}\|p_i)
\end{aligned}$$

holds. In addition,

$$\begin{aligned}
&C(\alpha)D_{\alpha}(p\|\bar{p}_{\boldsymbol{\beta}}^{(\alpha)})\\
&= \frac{1}{1-\alpha}\left[1-\int_{\mathcal{Z}}p(z)^{1-\alpha}\left\{\sum_{i=1}^{M}\beta_i p_i(z)^{\alpha}\right\}\mathrm{d}z\right]\\
&\quad -\frac{1}{1-\alpha}\{1-C(\alpha)\}\\
&= \sum_{i=1}^{M}\beta_i D_{\alpha}(p\|p_i) - \sum_{i=1}^{M}\beta_i D_{\alpha}(\bar{p}_{\boldsymbol{\beta}}^{(\alpha)}\|p_i) \qquad (\mathrm{A}\cdot 2)
\end{aligned}$$

holds. Therefore, by substituting $p_j(z)$ into $p(z)$ in (A· 2), we have Eq. (18).

Appendix C: Derivation of Eq. (21)

In Eq. (A· 2), by replacing $p(z)$ with $p_*(z)$ and taking $\alpha \to 0$ and $\alpha \to 1$ as limits, we obtain

$$D_{\mathrm{KL}}(p_*\|\bar{p}_{\boldsymbol{\beta}}) = \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_*\|p_i) - \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(\bar{p}_{\boldsymbol{\beta}}\|p_i),$$

$$D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\beta}}\|p_*) = \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|p_*) - \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|\tilde{p}_{\boldsymbol{\beta}}),$$

respectively (the first equation can be also given by replacing $p(z)$ with $p_*(z)$ in Eq. (A· 1)). Therefore, we have

$$\begin{aligned}
&D_{\mathrm{KL}}(p_*\|\bar{p}_{\boldsymbol{\beta}}) + D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\beta}}\|p_*)\\
&= \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_*\|p_i) + \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|p_*)\\
&\quad -\sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(\bar{p}_{\boldsymbol{\beta}}\|p_i) - \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|\bar{p}_{\boldsymbol{\beta}})\\
&\quad +\sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|\bar{p}_{\boldsymbol{\beta}}) - \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|\tilde{p}_{\boldsymbol{\beta}})\\
&= 2\sum_{i=1}^{M}\beta_i D_{\text{S-KL}}(p_*\|p_i) - 2\sum_{i=1}^{M}\beta_i D_{\text{S-KL}}(\bar{p}_{\boldsymbol{\beta}}\|p_i)\\
&\quad +\sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|\bar{p}_{\boldsymbol{\beta}}) - \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|\tilde{p}_{\boldsymbol{\beta}}). \qquad (\mathrm{A}\cdot 3)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|\bar{p}_{\boldsymbol{\beta}}) - \sum_{i=1}^{M}\beta_i D_{\mathrm{KL}}(p_i\|\tilde{p}_{\boldsymbol{\beta}})\\
&= -\sum_{i=1}^{M}\beta_i\int_{\mathcal{Z}}p_i(z)\ln\bar{p}_{\boldsymbol{\beta}}(z)\mathrm{d}z\\
&\quad +\sum_{i=1}^{M}\beta_i\int_{\mathcal{Z}}p_i(z)\ln\tilde{p}_{\boldsymbol{\beta}}(z)\mathrm{d}z\\
&= -\int_{\mathcal{Z}}\tilde{p}_{\boldsymbol{\beta}}(z)\ln\bar{p}_{\boldsymbol{\beta}}(z)\mathrm{d}z + \int_{\mathcal{Z}}\tilde{p}_{\boldsymbol{\beta}}(z)\ln\tilde{p}_{\boldsymbol{\beta}}(z)\mathrm{d}z\\
&= D_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\beta}}\|\bar{p}_{\boldsymbol{\beta}}) \qquad (\mathrm{A}\cdot 4)
\end{aligned}$$

holds. Therefore, by substituting Eq. (A· 4) into Eq. (A· 3), we obtain Eq. (21).