Generalization error for Tweedie models: decomposition and error reduction with bagging

Michel Denuit
Institute of Statistics, Biostatistics and Actuarial Science
UCLouvain
Louvain-la-Neuve, Belgium

Julien Trufin
Department of Mathematics
Université Libre de Bruxelles (ULB)
Brussels, Belgium

January 18, 2021

Abstract

Wüthrich and Buser (2020) studied the generalization error for Poisson regression models. This short note aims to extend their results to the Tweedie family of distributions, to which the Poisson law belongs. In case of bagging, a new condition emerges that becomes increasingly binding with the power parameter involved in the Tweedie variance function.

Keywords: Generalization error, Supervised learning, Exponential dispersion family, Tweedie, Bagging.

1 Introduction and motivation

In many applications, the analyst targets the conditional expectation $\mu(\boldsymbol{X}) = \mathrm{E}[Y|\boldsymbol{X}]$ of the response Y given the available information summarized in the vector $\boldsymbol{X}$. The function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is generally unknown and is approximated by an estimator $\boldsymbol{x} \mapsto \widehat{\mu}(\boldsymbol{x})$. The goal is to produce the most accurate approximation to the true $\widehat{\mu}(\boldsymbol{x})$. Lack of accuracy for $\widehat{\mu}(\boldsymbol{x})$ is defined by the generalization error

$$Err(\widehat{\mu}) = \mathrm{E}\left[L(Y, \widehat{\mu}(\boldsymbol{X}))\right], \tag{1.1}$$

where $L(.,.)$ is a function measuring the discrepancy between its two arguments, called loss function, and the expected value is over the joint distribution of $(Y, \boldsymbol{X})$. We refer the readers to Hastie et al. (2009) for more details about the generalization error. We aim to find a function $\widehat{\mu}(\boldsymbol{x})$ of the features minimizing the generalization error (1.1).

In practice, loss functions $L(.,.)$ often correspond to negative log-likelihood functions associated with distributions that belong to the Tweedie family. The Tweedie class regroups the members of the Exponential Dispersion family having power variance functions $V(\mu) = \mu^{\xi}$ for some ξ. We refer the readers to Denuit et al. (2019) for an extensive treatment of the Exponential Dispersion family and the Tweedie class in the context of insurance. Specifically, the Tweedie class contains continuous distributions such as the Normal, Gamma and Inverse Gaussian distributions. It also includes the Poisson and compound Poisson-Gamma distributions. Compound Poisson-Gamma distributions can be used for modeling data having a positive probability mass at zero and a continuous distribution on the positive real numbers such as yearly insurance losses or rainfall meteorological data.

The following table gives a list of all Tweedie distributions:

	Type	Name
$\xi < 0$	Continuous	-
$\xi = 0$	Continuous	Normal
$0 < \xi < 1$	Non existing	-
$\xi = 1$	Discrete	Poisson
$1 < \xi < 2$	Mixed, non-negative	Compound Poisson-Gamma
$\xi = 2$	Continuous, positive	Gamma
$2 < \xi < 3$	Continuous, positive	-
$\xi = 3$	Continuous, positive	Inverse Gaussian
$\xi > 3$	Continuous, positive	-

Negative values of ξ gives continuous distributions on the whole real axis. There is no probability distribution in the Tweedie class corresponding to power parameters $0 < \xi < 1$. In this paper, we consider non-negative data and we restrict our analysis to $\xi \geq 1$.

We denote by

$$\mathcal{D} = \{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\} \tag{1.2}$$

the set of observations used to fit the model $\widehat{\mu}$, called training set. An estimate $\widehat{\mu}(\boldsymbol{x})$ to $\mu(\boldsymbol{x})$ is obtained by minimizing the total loss on the training set $\mathcal{D}$, that is, $\sum_{i=1}^{n} L(y_i, \widehat{\mu}(\boldsymbol{x}_i))$ that can be seen as the empirical version of (1.1). The loss function corresponding to the Tweedie

deviance is

$$L(y,\widehat{\mu}(\boldsymbol{x})) = \begin{cases} 2\left(y\ln\frac{y}{\widehat{\mu}(\boldsymbol{x})} - (y-\widehat{\mu}(\boldsymbol{x}))\right) & \text{for } \xi = 1 \\ 2\left(-\ln\frac{y}{\widehat{\mu}(\boldsymbol{x})} + \frac{y}{\widehat{\mu}(\boldsymbol{x})} - 1\right) & \text{for } \xi = 2 \\ 2\left(\frac{y^{2-\xi}}{(1-\xi)(2-\xi)} - \frac{y\widehat{\mu}(\boldsymbol{x})^{1-\xi}}{1-\xi} + \frac{\widehat{\mu}(\boldsymbol{x})^{2-\xi}}{2-\xi}\right) & \text{for } \xi \in]1,+\infty[\setminus\{2\}. \end{cases} \tag{1.3}$$

Notice that the Tweedie loss functions (1.3) are particular cases of Bregman loss functions so that the expected loss is minimum for the mean response, whatever the value of the power parameter $\xi \geq 1$, as shown in Savage (1971).

This paper aims to decompose the generalization error (1.1) for loss functions (1.3) derived from Tweedie deviance into a sum of the error for the true model and an estimation error. This extends the classical decomposition that is known to hold for the squared error loss, corresponding to the Normal distribution ($\xi = 0$) and for the loss function derived from the Poisson deviance ($\xi = 1$) as established by Wüthrich and Buser (2020) to the whole Tweedie class with $\xi \geq 1$. The condition under which bagging reduces the error is then obtained, depending on the power parameter ξ. Interestingly, this condition becomes increasingly binding when ξ increases.

2 Generalization error

The generalization error $Err(\widehat{\mu})$ given in (1.1) can also be defined for a fixed value $\boldsymbol{X} = \boldsymbol{x}$ as

$$Err(\widehat{\mu}(\boldsymbol{x})) = \mathrm{E}\left[L(Y,\widehat{\mu}(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right]. \tag{2.1}$$

Notice that averaging the local errors $Err(\widehat{\mu}(\boldsymbol{x}))$ enables to recover the generalization error $Err(\widehat{\mu})$, that is,

$$Err(\widehat{\mu}) = \mathrm{E}\left[Err(\widehat{\mu}(\boldsymbol{X}))\right]. \tag{2.2}$$

The generalization error of $\widehat{\mu}$ at $\boldsymbol{X} = \boldsymbol{x}$ can be expressed as follows.

Proposition 2.1. *We have*

$$Err(\widehat{\mu}(\boldsymbol{x})) = Err(\mu(\boldsymbol{x})) + \mathcal{E}_\xi(\widehat{\mu}(\boldsymbol{x})) \tag{2.3}$$

with

$$\mathcal{E}_\xi(\widehat{\mu}(\boldsymbol{x})) = \begin{cases} 2\mu(\boldsymbol{x})\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})} - 1 - \ln\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right) & \textit{for } \xi = 1 \\ 2\left(\frac{\mu(\boldsymbol{x})}{\widehat{\mu}(\boldsymbol{x})} - 1 - \ln\left(\frac{\mu(\boldsymbol{x})}{\widehat{\mu}(\boldsymbol{x})}\right)\right) & \textit{for } \xi = 2 \\ \frac{2}{(2-\xi)(1-\xi)}\left(\widehat{\mu}(\boldsymbol{x})^{2-\xi}(1-\xi) + \mu(\boldsymbol{x})^{2-\xi} + (\xi-2)\mu(\boldsymbol{x})\widehat{\mu}(\boldsymbol{x})^{1-\xi}\right) & \textit{for } \xi \in]1,+\infty[\setminus\{2\}. \end{cases} \tag{2.4}$$

Proof. For $\xi = 1$, this result can be found in Wüthrich and Buser (2020) (see Equation

(7.5)). Turning to the case $\xi = 2$, we get

$$\begin{aligned}
Err\left(\widehat{\mu}(\boldsymbol{x})\right) &= 2\mathrm{E}\left[-\ln\left(\frac{Y}{\widehat{\mu}(\boldsymbol{x})}\right)+\frac{Y}{\widehat{\mu}(\boldsymbol{x})}-1\Big|\boldsymbol{X}=\boldsymbol{x}\right]\\
&= 2\mathrm{E}\left[-\ln\left(\frac{Y}{\mu(\boldsymbol{x})}\right)+\frac{Y}{\mu(\boldsymbol{x})}-1\Big|\boldsymbol{X}=\boldsymbol{x}\right]\\
&\quad +2\mathrm{E}\left[-\ln\left(\frac{Y}{\widehat{\mu}(\boldsymbol{x})}\right)+\ln\left(\frac{Y}{\mu(\boldsymbol{x})}\right)+\frac{Y}{\widehat{\mu}(\boldsymbol{x})}-\frac{Y}{\mu(\boldsymbol{x})}\Big|\boldsymbol{X}=\boldsymbol{x}\right]\\
&= Err\left(\mu(\boldsymbol{x})\right)+2\mathrm{E}\left[\ln\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)+Y\left(\frac{\mu(\boldsymbol{x})-\widehat{\mu}(\boldsymbol{x})}{\widehat{\mu}(\boldsymbol{x})\mu(\boldsymbol{x})}\right)\Big|\boldsymbol{X}=\boldsymbol{x}\right]\\
&= Err\left(\mu(\boldsymbol{x})\right)+2\left(\ln\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)+\mathrm{E}\left[Y|\boldsymbol{X}=\boldsymbol{x}\right]\left(\frac{\mu(\boldsymbol{x})-\widehat{\mu}(\boldsymbol{x})}{\widehat{\mu}(\boldsymbol{x})\mu(\boldsymbol{x})}\right)\right)\\
&= Err\left(\mu(\boldsymbol{x})\right)+2\left(\frac{\mu(\boldsymbol{x})}{\widehat{\mu}(\boldsymbol{x})}-1-\ln\left(\frac{\mu(\boldsymbol{x})}{\widehat{\mu}(\boldsymbol{x})}\right)\right).
\end{aligned}$$

Finally, for the remaining cases, we have

$$\begin{aligned}
Err\left(\widehat{\mu}(\boldsymbol{x})\right) &= 2\mathrm{E}\left[\frac{Y^{2-\xi}}{(1-\xi)(2-\xi)}-\frac{Y\widehat{\mu}(\boldsymbol{x})^{1-\xi}}{1-\xi}+\frac{\widehat{\mu}(\boldsymbol{x})^{2-\xi}}{2-\xi}\Big|\boldsymbol{X}=\boldsymbol{x}\right]\\
&= 2\mathrm{E}\left[\frac{Y^{2-\xi}}{(1-\xi)(2-\xi)}-\frac{Y\mu(\boldsymbol{x})^{1-\xi}}{1-\xi}+\frac{\mu(\boldsymbol{x})^{2-\xi}}{2-\xi}\Big|\boldsymbol{X}=\boldsymbol{x}\right]\\
&\quad -2\mathrm{E}\left[\frac{Y\left(\widehat{\mu}(\boldsymbol{x})^{1-\xi}-\mu(\boldsymbol{x})^{1-\xi}\right)}{1-\xi}-\frac{\widehat{\mu}(\boldsymbol{x})^{2-\xi}-\mu(\boldsymbol{x})^{2-\xi}}{2-\xi}\Big|\boldsymbol{X}=\boldsymbol{x}\right]\\
&= Err\left(\mu(\boldsymbol{x})\right)+\frac{2\widehat{\mu}(\boldsymbol{x})^{2-\xi}-2\mu(\boldsymbol{x})^{2-\xi}}{2-\xi}-\frac{2\mu(\boldsymbol{x})\widehat{\mu}(\boldsymbol{x})^{1-\xi}}{1-\xi}+\frac{2\mu(\boldsymbol{x})^{2-\xi}}{1-\xi}.
\end{aligned}$$

This ends the proof. □

The smallest generalization error coincides with the one associated to the true model.

Corollary 2.2.

$$Err(\widehat{\mu}(\boldsymbol{x}))\geq Err(\mu(\boldsymbol{x})). \tag{2.5}$$

Proof. For $\xi = 1$ and $\xi = 2$, it suffices to notice that $\mathcal{E}_\xi\left(\widehat{\mu}(\boldsymbol{x})\right)$ is always positive since $y \to y - 1 - \ln y$ is positive on $\mathbb{R}^+$. For the remaining cases, define the function f on $\mathbb{R}^+$ as

$$f(y)=\frac{2}{(2-\xi)(1-\xi)}\left(y^{2-\xi}(1-\xi)+\mu(\boldsymbol{x})^{2-\xi}+(\xi-2)\mu(\boldsymbol{x})y^{1-\xi}\right). \tag{2.6}$$

We have

$$f'(y)=2\left(y^{1-\xi}-\mu(\boldsymbol{x})y^{-\xi}\right)\ \text{ and }\ f''(y)=2y^{-\xi}\left(1-\xi+\xi\frac{\mu(\boldsymbol{x})}{y}\right).$$

Hence, for $y > 0$, $f'(y) = 0$ if, and only if, $y = \mu(\boldsymbol{x})$ and

$$f''(\mu(\boldsymbol{x}))=2\mu(\boldsymbol{x})^{-\xi}>0,$$

so that $f(y) \geq f(\mu(\boldsymbol{x})) = 0$ for all $y > 0$, which completes the proof. □

The generalization error of $\widehat{\mu}$ can be expressed as the sum of the generalization error of the true model μ and an estimation error that is positive. The generalization error of the true model is called the residual error and is thus irreducible. Corollary 2.2 can be found in Gneiting (2011) (see Theorem 7), where the loss functions derived from Tweedie models correspond to the Patton's family of homogeneous scoring functions on the positive half line.

3 Bagging models and expected generalization error

Bagging is one of the first ensemble methods proposed in the literature by Breiman (1996), who showed that aggregating multiple versions of an estimator into an ensemble improves the model accuracy. Consider a model fitted to our training set $\mathcal{D}$, obtaining the prediction $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$ at point $\boldsymbol{x}$. Bootstrap aggregation or bagging averages this prediction over a set of bootstrap samples in order to reduce its variability with respect to the data used to build it.

The probability distribution of the random vector $(Y, \boldsymbol{X})$ is usually not known. It can be approximated by its empirical version which puts an equal probability $\frac{1}{n}$ on each of the observations $\{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\}$ of the training set $\mathcal{D}$. Hence, instead of simulating B training sets $\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^B$ from the probability distribution of $(Y, \boldsymbol{X})$, which is not possible in practice, the idea of bagging is rather to simulate B bootstrap samples $\mathcal{D}^{*1}, \mathcal{D}^{*2}, \ldots, \mathcal{D}^{*B}$ of the training set $\mathcal{D}$ from its empirical counterpart. Specifically, a bootstrap sample is thus a random sample of $\mathcal{D}$ taken with replacement which has the same size as $\mathcal{D}$.

Let $\mathcal{D}^{*1}, \mathcal{D}^{*2}, \ldots, \mathcal{D}^{*B}$ be B bootstrap samples of the training set $\mathcal{D}$. For each $\mathcal{D}^{*b}$, $b = 1, \ldots, B$, we fit a model, giving prediction $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}) = \widehat{\mu}_{\mathcal{D}^{*b}}(\boldsymbol{x})$. The bagging prediction is then defined by

$$\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}), \tag{3.1}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_B)$. Random vectors $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_B$ fully capture the randomness of the training procedure. For bagging, $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_B$ are independent and identically distributed so that $\boldsymbol{\Theta}_b$ is a vector of n integers randomly and uniformly drawn in $\{1, 2, \ldots, n\}$. Each component of $\boldsymbol{\Theta}_b$ indexes one observation of the training set selected in $\mathcal{D}^{*b}$.

The generalization error $Err(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x}))$ is evaluated conditional on the bootstrap samples $\mathcal{D}^{*1}, \mathcal{D}^{*2}, \ldots, \mathcal{D}^{*B}$ of D, so that it gives an idea of the general accuracy of the bagging training procedure for the particular bootstrap samples of $\mathcal{D}$. In order to assess the general performance of the bagging training procedure, we use the expected generalization error, which averages the generalization error $Err(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x}))$ over $\mathcal{D}$ and $\boldsymbol{\Theta}$, that is, $\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right]$. From Proposition 2.1, we get

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] = Err\left(\mu(\boldsymbol{x})\right) + \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\mathcal{E}_\xi\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right]. \tag{3.2}$$

For $\xi = 1$, Wüthrich and Buser (2020) showed that

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x}))\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}))\right],$$

meaning that the bagging prediction $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ outperforms the individual sample estimate $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$ in the Poisson case. The next proposition extends this result for $\xi \geq 1$ subject to the additional condition (3.3). Condition (3.3) implies that if the individual sample estimates $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$ do not overestimate too much the true prediction $\mu(\boldsymbol{x})$, then it is beneficial to aggregate them in the sense that the aggregation reduces the local generalization error.

Proposition 3.1. *If the individual sample estimates satisfy*

$$\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})} \leq \frac{\xi}{\xi - 1}, \tag{3.3}$$

then we have

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err(\widehat{\mu}^{bag}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x}))\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}))\right]. \tag{3.4}$$

Proof. The case $\xi = 1$ is due to Wüthrich and Buser (2020) (see Proposition 7.2). Notice that the upper bound in (3.3) is not binding when $\xi = 1$. For $\xi = 2$, by Proposition 2.1, one sees that inequality (3.4) is fulfilled if, and only if,

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\phi\left(\frac{\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\phi\left(\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right], \tag{3.5}$$

where $\phi : y > 0 \to \frac{1}{y} + \ln y$. The latter inequality holds true when the individual sample estimates satisfy

$$\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})} \leq 2, \tag{3.6}$$

which, in turn, guarantees that $\frac{\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})} \leq 2$. Indeed, the function $\phi(y)$ is convex for $y \leq 2$, so that Jensen's inequality implies

$$\begin{aligned}
\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\phi\left(\frac{\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right] &= \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\phi\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right] \\
&\leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\frac{1}{B}\sum_{b=1}^{B}\phi\left(\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right] \\
&= \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\phi\left(\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right]
\end{aligned} \tag{3.7}$$

provided that condition (3.6) holds.

In the remaining cases for ξ, by Proposition 2.1, inequality (3.4) is satisfied if, and only if, $\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[f\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[f\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right]$, where the function $f(\cdot)$ is defined in (2.6). Now, we have $f''(y) \geq 0$ if, and only if, $\frac{y}{\mu(\boldsymbol{x})} \leq \frac{\xi}{\xi-1}$. Therefore, when $\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})} \leq \frac{\xi}{\xi-1}$ and hence $\frac{\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})} \leq \frac{\xi}{\xi-1}$, we have $f''(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})) \geq 0$ and $f''(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})) \geq 0$, so that Jensen's inequality leads to

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[f\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[f\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right].$$

This ends the proof. □

References

- Breiman, L. (1996). Bagging predictors. Machine Learning 24, 123-140.

- Denuit, M., Hainaut, D., Trufin, J. (2019). Effective Statistical Learning Methods for Actuaries I: GLM and Extensions. Springer Actuarial Lecture Notes Series.

- Gneiting, T. (2011). Making and evaluating point forecasts. Journal of the American Statistical Association 106, 746-762.

- Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition. Springer Series in Statistics.

- Savage, L.J. (1971). Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66, 783-810.

- Wüthrich, M. V., Buser, C. (2020). Data Analytics for Non-Life Insurance Pricing. Lecture notes. Available at SSRN, http://dx.doi.org/10.2139/ssrn.2870308.