1 Loss Function and its Derivatives

In this Section, we write down the math. We will use the symbol $\bigoplus$ to denote the log-space summation operator: $\bigoplus_{i=1}^{N} \log(a_i) = \log\left(\sum_{i=1}^{N} a_i\right)$. Writing expressions using this notation will allow us to obtain numerically stable computational expressions.

Let $m \in \{1, \ldots, M\}$ index component predictive models and $i \in \{1, \ldots, N\}$ index prediction cases in the training data. For example, in a seasonal time series prediction context i may index times at which we make predictions for seasonal quantities, or combination of a time at which we make a prediction and the prediction horizon for predictions at individual weeks. Let $f_m(y_i|\mathbf{x}_i)$ denote the predictive density from model m for the value of the random variable Y_i. The $\mathbf{x}_i$ is a vector observed covariates which may be used by any of the component models as conditioning variables in forming the predictive distribution, and is also used in calculating the component model weights. Note that the component models and computation of the component model weights may use a proper subset of the variables in $\mathbf{x}_i$.

The combined predictive density for case i is

$$f(y_i|\mathbf{x}_i) = \sum_{m=1}^{M} \pi_m(\mathbf{x}_i) f_m(y_i|\mathbf{x}_i), \text{ where} \tag{1}$$

$$\pi_m(\mathbf{x}_i) = \frac{\exp\{\rho_m(\mathbf{x}_i)\}}{\sum_{m'=1}^{M} \exp\{\rho_{m'}(\mathbf{x}_i)\}} \tag{2}$$

In Equation (1) the $\pi_m(\mathbf{x}_i)$ are the model weights, which we regard as functions of $\mathbf{x}_i$. These weights must be non-negative and sum to 1 across m. We ensure that these constraints are met by parameterizing the $\pi_m(\mathbf{x}_i)$ in terms of the softmax transformation of real-valued functions $\rho_m(\mathbf{x}_i)$ in Equation (2). For notational brevity, we will suppress the expression of these quantities as functions of $\mathbf{x}_i$ and write $\rho_m(\mathbf{x}_i) = \rho_{mi}$ with $\boldsymbol{\rho} = (\rho_{11}, \ldots, \rho_{MN})$, and $\pi_m(\mathbf{x}_i) = \pi_{mi}$ with $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{MN})$.

Our goal is to estimate the functions ρ_{mi}. To do this, we require as inputs cross-validated estimates of $\log\{f_m(y_i|\mathbf{x}_i)\}$ for each component model m and case i in the training data; we denote these values by $\log\{f_m^{cv}(y_i|\mathbf{x}_i)\}$. We will focus on optimization of the log-score of the combined predictive distribution for now; we may consider other loss functions in the future. Considered as a function of the vector of values ρ_{mi} for each combination of m and i, this loss function is given by

$$L(\boldsymbol{\rho}) = \sum_{i=1}^{N} \log\{f(y_i|\mathbf{x}_i)\} \tag{3}$$

$$= \sum_{i=1}^{N} \log\left\{\sum_{m=1}^{M} \pi_{mi} f_m(y_i|\mathbf{x}_i)\right\} \tag{4}$$

$$= \sum_{i=1}^{N} \log\left\{\sum_{m=1}^{M} \frac{\exp(\rho_{mi})}{\sum_{m'=1}^{M} \exp(\rho_{m'i})} f_m(y_i|\mathbf{x}_i)\right\} \tag{5}$$

We must find the first and second order partial derivatives of L with respect to each $\rho_{m^*i^*}$.

$$\frac{\partial}{\partial \rho_{m^*i^*}} L(\boldsymbol{\rho}) = \frac{\partial}{\partial \rho_{m^*i^*}} \sum_{i=1}^{N} \log\left\{\sum_{m=1}^{M} \frac{\exp(\rho_{mi})}{\sum_{m'=1}^{M} \exp(\rho_{m'i})} f_m(y_i|\mathbf{x}_i)\right\} \tag{6}$$

$$= \left\{\frac{1}{\sum_{m=1}^{M} \frac{\exp(\rho_{mi^*})}{\sum_{m'=1}^{M} \exp(\rho_{m'i^*})} f_m(y_{i^*}|\mathbf{x}_{i^*})}\right\} \times \frac{\partial}{\partial \rho_{m^*i^*}} \sum_{m=1}^{M} \frac{\exp(\rho_{mi^*})}{\sum_{m'=1}^{M} \exp(\rho_{m'i^*})} f_m(y_{i^*}|\mathbf{x}_{i^*}) \tag{7}$$

Now note that for $m^* = m$,

$$\begin{aligned}
\frac{\partial}{\partial \rho_{m^* i^*}} \pi_{m i^*} &= \frac{\partial}{\partial \rho_{m^* i^*}} \frac{\exp(\rho_{m i^*})}{\sum_{m'=1}^{M} \exp(\rho_{m' i^*})} \\
&= \frac{\exp(\rho_{m i^*}) \left\{ \sum_{m'=1}^{M} \exp(\rho_{m' i^*}) \right\} - \exp(\rho_{m i^*})^2}{\left\{ \sum_{m'=1}^{M} \exp(\rho_{m' i^*}) \right\}^2} \\
&= \pi_{m i^*} - \pi_{m i^*}^2 .
\end{aligned}$$

For $m^* \neq m$,

$$\begin{aligned}
\frac{\partial}{\partial \rho_{m^* i^*}} \pi_{m i^*} &= \frac{\partial}{\partial \rho_{m^* i^*}} \frac{\exp(\rho_{m i^*})}{\sum_{m'=1}^{M} \exp(\rho_{m' i^*})} \\
&= \frac{-\exp(\rho_{m i^*}) \exp(\rho_{m^* i^*})}{\left\{ \sum_{m'=1}^{M} \exp(\rho_{m' i^*}) \right\}^2} \\
&= -\pi_{m i^*} \pi_{m^* i^*} .
\end{aligned}$$

Substituting these results into Equation (7), we obtain

$$\frac{\partial}{\partial \rho_{m^* i^*}} L(\boldsymbol{\rho}) = \left\{ \frac{1}{\sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})} \right\} \left\{ \pi_{m^* i^*} \left(f_{m^*}(y_{i^*} | \mathbf{x}_{i^*}) - \sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*}) \right) \right\} \tag{8}$$

$$= \frac{\pi_{m^* i^*} f_{m^*}(y_{i^*} | \mathbf{x}_{i^*})}{\sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})} - \pi_{m^* i^*} \tag{9}$$

Now, we calculate the second order derivative as

$$\frac{\partial^2}{\partial \rho_{m^* i^*}^2} L(\boldsymbol{\rho}) = \frac{\partial}{\partial \rho_{m^* i^*}} \left[\frac{\partial}{\partial \rho_{m^* i^*}} L(\boldsymbol{\rho}) \right] \tag{10}$$

$$= \frac{\partial}{\partial \rho_{m^* i^*}} \left[\frac{\pi_{m^* i^*} f_{m^*}(y_{i^*} | \mathbf{x}_{i^*})}{\sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})} - \pi_{m^* i^*} \right] \tag{11}$$

$$= \frac{\left(\pi_{m^* i^*} - \pi_{m^* i^*}^2 \right) f_{m^*}(y_{i^*} | \mathbf{x}_{i^*}) \sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})}{\left\{ \sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*}) \right\}^2} \tag{12}$$

$$- \frac{\pi_{m^* i^*} f_{m^*}(y_{i^*} | \mathbf{x}_{i^*}) \left\{ \pi_{m^* i^*} \left(f_{m^*}(y_{i^*} | \mathbf{x}_{i^*}) - \sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*}) \right) \right\}}{\left\{ \sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*}) \right\}^2} \tag{13}$$

$$- \left(\pi_{m^* i^*} - \pi_{m^* i^*}^2 \right) \tag{14}$$

$$= (1 - \pi_{m^* i^*}) \frac{\pi_{m^* i^*} f_{m^*}(y_{i^*} | \mathbf{x}_{i^*})}{\sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})} \tag{15}$$

$$- \left[\frac{\pi_{m^* i^*} f_{m^*}(y_{i^*} | \mathbf{x}_{i^*})}{\sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})} \right]^2 + \pi_{m^* i^*} \frac{\pi_{m^* i^*} f_{m^*}(y_{i^*} | \mathbf{x}_{i^*})}{\sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})} \tag{16}$$

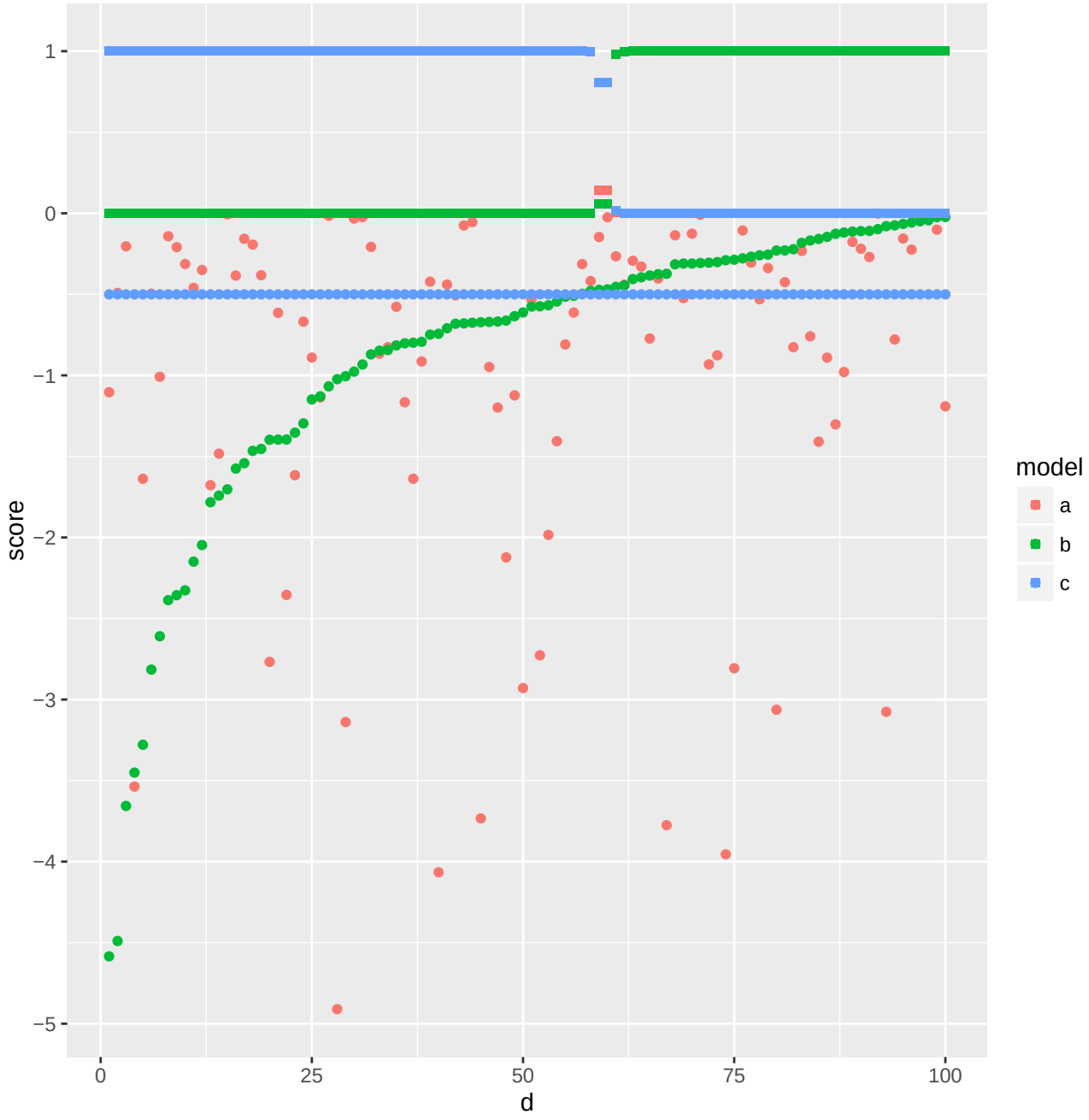
$$- \left(\pi_{m^* i^*} - \pi_{m^* i^*}^2 \right) \tag{17}$$

$$= \frac{\pi_{m^* i^*} f_{m^*}(y_{i^*} | \mathbf{x}_{i^*})}{\sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})} - \left[\frac{\pi_{m^* i^*} f_{m^*}(y_{i^*} | \mathbf{x}_{i^*})}{\sum_{m=1}^{M} \pi_{m i^*} f_m(y_{i^*} | \mathbf{x}_{i^*})} \right]^2 - \left(\pi_{m^* i^*} - \pi_{m^* i^*}^2 \right) \tag{18}$$

2 Simulated Application

```
library(tidyr)
library(dplyr)

##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##    filter, lag
## The following objects are masked from 'package:base':
##
##    intersect, setdiff, setequal, union

library(xgboost)

##
## Attaching package:  'xgboost'
## The following object is masked from 'package:dplyr':
##
##    slice

library(xgbstack)
library(ggplot2)

### For now, let's just make up some data for the purposes of method development
### this will need to go into test code too
set.seed(9873)
loso_pred_res <- data.frame(
  model = paste0("log_score_", rep(letters[1:3], each = 100)),
  d = rep(1:100, times = 3),
  loso_log_score = c(
    log(runif(100, 0, 1)), # model a's performance not related to d
    sort(log(runif(100, 0, 1))), # model b's performance increasing in d
    rep(-0.5, 100))  # model c's performance constant
) %>%
  spread(model, loso_log_score)

## Obtain stacking model fit
fit <- xgbstack(log_score_a + log_score_b + log_score_c ~ d,
  data = loso_pred_res)

## Let's see what the model weights look like as a function of d

## Get predictions -- need to refactor this into a user-friendly method
## Create some test data
dtest <- xgb.DMatrix(
  data = as.matrix(1:100) %>%
      `storage.mode<-`("double")
  )

predictions <- predict(fit, newdata = dtest)

preds <- preds_to_matrix(preds = predictions, num_models = 3)
```

```r
component_model_scores_df <- as.data.frame(as.matrix(
      loso_pred_res[, paste0("log_score_", letters[1:3]), drop = FALSE]
    ) %>%
    `storage.mode<-`("double")) %>%
  `colnames<-`(letters[1:3]) %>%
  gather_("model", "score", letters[1:3]) %>%
  mutate(d = rep(1:100, 3))

## Convert predictions to model weights -- refactor
log_denom <- logspace_sum_matrix_rows(preds)
component_model_weights_df <- as.data.frame(exp(sweep(preds, 1, log_denom, `-`))) %>%
  `colnames<-`(letters[1:3])

component_model_weights_df <- component_model_weights_df %>%
  gather_("model", "weight", letters[1:3]) %>%
  mutate(d = rep(1:100, 3))

ggplot() +
  geom_point(aes(x = d, y = score, colour = model), data = component_model_scores_df) +
  geom_point(aes(x = d, y = weight, colour = model), shape = 15, data = component_model_weights_df)
```


1
0
−1
−2
−3
−4
−5
score
0
25
50
75
100
d
model
a
b
c