

# Recursive Iteratively Re-weighted Least Squares for Multinomial Logistic Regression

November 28, 2025

## 1 Notes

These notes explore a possible extension of Recursive Least Squares (RLS) — an algorithm which incrementally updates linear regression parameter estimates as new data is observed — to the multinomial logistic regression setting using the Iteratively Reweighted Least Squares (IRLS) approach.

Consider pairs of observations that arrive sequentially over time. The pair  $(\mathbf{x}^{[t]}, \mathbf{y}^{[t]})$  at time  $t$  is defined

$$\begin{aligned} \mathbf{x}_{m \times 1}^{[t]} &= \begin{bmatrix} 1 \\ \mathbf{x}_*^{[t]} \end{bmatrix} = \begin{bmatrix} 1 \\ x_1^{[t]} \\ \vdots \\ x_{m-1}^{[t]} \end{bmatrix}, & \mathbf{y}_{k \times 1}^{[t]} &= \begin{bmatrix} y_1^{[t]} \\ \vdots \\ y_k^{[t]} \end{bmatrix} \end{aligned}$$

where  $\mathbf{x}_*^{[t]}$  is a  $(m-1)$ -vector of covariates and  $\mathbf{y}^{[t]}$  is a one-hot-encoded  $k$ -vector of outcomes (a.k.a category labels). That is,  $y_i^{[t]} = 1$  if the observation belongs to the  $i^{\text{th}}$  category and 0 otherwise. The observations are assumed to be i.i.d. (stationary) over time and the labels are conditionally distributed according to a Multinomial distribution:

$$\mathbf{y}^{[t]} | \mathbf{x}^{[t]} \sim \text{Multinomial}(1, \mathbf{p}^{[t]})$$

where  $\mathbf{p}^{[t]}$  is a  $k$ -vector of probabilities. Here,  $E(\mathbf{y}^{[t]} | \mathbf{x}^{[t]}) = \mathbf{p}^{[t]}$  with

$$p_i^{[t]} = \frac{\exp(\mathbf{x}^{[t]'} \boldsymbol{\theta}_i)}{1 + \sum_{j=1}^{k-1} \exp(\mathbf{x}^{[t]'} \boldsymbol{\theta}_j)} \quad \text{for } i = 1, \dots, k-1 \quad (1)$$

$$p_i^{[t]} = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(\mathbf{x}^{[t]'} \boldsymbol{\theta}_j)} \quad \text{for } i = k \quad (2)$$

where  $\boldsymbol{\Theta}$  is a  $(k-1) \times m$  parameter matrix.

Suppose that  $T$  samples have been observed and can thus be gathered in the matrices

$$\mathbf{X}_{T \times m} = \begin{bmatrix} \mathbf{x}^{[1]'} \\ \vdots \\ \mathbf{x}^{[T]'} \end{bmatrix}, \quad \mathbf{Y}_{T \times k} = \begin{bmatrix} \mathbf{y}^{[1]'} \\ \vdots \\ \mathbf{y}^{[T]'} \end{bmatrix}, \quad \mathbf{P}_{T \times k} = \begin{bmatrix} \mathbf{p}^{[1]'} \\ \vdots \\ \mathbf{p}^{[T]'} \end{bmatrix}.$$

The log-likelihood of  $\boldsymbol{\Theta}$  with respect to these observations is

$$\ell(\boldsymbol{\Theta}; \mathbf{X}, \mathbf{Y}) = \sum_{t=1}^T \sum_{i=1}^k y_i^{[t]} \ln(p_i^{[t]}). \quad (3)$$

Differentiating with respect to  $\theta_i$  gives the gradient

$$\mathbf{g}(\theta_i) := \frac{\partial \ell}{\partial \theta_i} \quad (4)$$

$$= \sum_{t=1}^T \mathbf{x}^{[t]} (y_i^{[t]} - p_i^{[t]}) \quad (5)$$

$$= \mathbf{X}' (\mathbf{Y}_{\cdot i} - \mathbf{P}_{\cdot i}) \quad (6)$$

where  $\mathbf{Y}_{\cdot i}$  and  $\mathbf{P}_{\cdot i}$  are the  $i^{\text{th}}$  columns of  $\mathbf{Y}$  and  $\mathbf{P}$ . Differentiating again gives the Hessian

$$\mathbf{H}(\theta_i) := \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_i'} \quad (7)$$

$$= - \sum_{t=1}^T \mathbf{x}^{[t]} \mathbf{x}^{[t]'} p_i^{[t]} (1 - p_i^{[t]}) \quad (8)$$

$$= -\mathbf{X}' \mathbf{W}_i \mathbf{X} \quad (9)$$

$$\mathbf{W}_i = \text{diag}(p_i^{[1]}(1-p_i^{[1]}), \dots, p_i^{[T]}(1-p_i^{[T]})) \quad (10)$$

The log-likelihood can be maximized to find the estimator  $\tilde{\Theta}$  using a partial (approximate) Newton-Raphson procedure.<sup>123</sup> The updating rule is

$$\tilde{\theta}_i^{\wedge} \leftarrow \tilde{\theta}_i^{\vee} - \mathbf{H}(\tilde{\theta}_i^{\vee})^{-1} \mathbf{g}(\tilde{\theta}_i^{\vee}) \quad (11)$$

$$= \tilde{\theta}_i^{\vee} + (\mathbf{X}' \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}' (\mathbf{Y}_{\cdot i} - \tilde{\mathbf{P}}_{\cdot i}) \quad (12)$$

where  $\tilde{\theta}_i^{\wedge}$  denotes the updated parameter and  $\tilde{\theta}_i^{\vee}$  denotes the current parameter. This rule is applied for each  $i = 1, \dots, k-1$  repeatedly until convergence is reached. In this procedure,  $\mathbf{W}$  and  $\tilde{\mathbf{P}}$  must be recomputed at each iteration based on the current  $\tilde{\Theta}^{\vee}$ . This method is referred to Iteratively Re-weighted Least-Squares (IRLS), as the update rule can be manipulated to look like the weighted least-squares solution.

This estimation procedure does not work in settings where the full dataset is not available all at once. Such a setting — referred to as “streaming data” — can arise if  $T$  is so large that  $\mathbf{H}$  cannot be computed in a computer’s RAM or if only single observations (or small batches of observations) are available at a time. In such settings, it is common to employ Recursive Least-Squares (RLS). The updating rule for RLS is<sup>4</sup>

$$\mathbf{M}^{[t]} \leftarrow \mathbf{M}^{[t-1]} - \frac{\mathbf{M}^{[t-1]} \mathbf{x}^{[t]} \mathbf{x}^{[t]'} \mathbf{M}^{[t-1]}}{1 + \mathbf{x}^{[t]} \mathbf{M}^{[t-1]} \mathbf{x}^{[t]'}} \quad (13)$$

$$\hat{\beta}^{[t]} \leftarrow \hat{\beta}^{[t-1]} + \left[ \mathbf{M}^{[t]} \mathbf{x}^{[t]} (\mathbf{y}^{[t]} - \hat{\beta}^{[t-1]'} \mathbf{x}^{[t]})' \right]' \quad (14)$$

where  $\hat{\beta}^{[t]}$  is a recursive estimate of the  $k \times m$  parameter matrix  $\beta$ . With good initialization, this procedure results in

$$\hat{\beta}^{[T]} \approx (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y} \quad (15)$$

which is the solution to the ordinary least squares problem.

Unfortunately, the RLS procedure cannot be used to estimate  $\Theta$  as it does not maximize the multinomial log-likelihood function. To estimate  $\Theta$  on streaming data, we propose to combine the RLS and IRLS

<sup>1</sup>[https://people.stat.sc.edu/gregorkb/Tutorials/MultLogReg\\_Algs.pdf](https://people.stat.sc.edu/gregorkb/Tutorials/MultLogReg_Algs.pdf)

<sup>2</sup><https://hastie.su.domains/Papers/glmnet.pdf>

<sup>3</sup>[https://waxworksmath.com/Authors/G\\_M/Hastie/WriteUp/Weatherwax\\_Epstein\\_Hastie\\_Solution\\_Manual.pdf](https://waxworksmath.com/Authors/G_M/Hastie/WriteUp/Weatherwax_Epstein_Hastie_Solution_Manual.pdf)

<sup>4</sup><https://www.jld-stats.com/2020/03/27/updating-a-linear-regression-with-new-data>

procedures (RIRLS). For this, the updating rules are

$$\mathbf{M}_i^{[t]} \leftarrow \mathbf{M}_i^{[t-1]} - \frac{w_i^{[t]} \mathbf{M}_i^{[t-1]} \mathbf{x}^{[t]} \mathbf{x}^{[t]'} \mathbf{M}_i^{[t-1]}}{1 + w_i^{[t]} \mathbf{x}^{[t]'} \mathbf{M}_i^{[t-1]} \mathbf{x}^{[t]}} \quad (16)$$

$$\hat{\boldsymbol{\theta}}_i^{[t]} \leftarrow \hat{\boldsymbol{\theta}}_i^{[t-1]} + \mathbf{M}_i^{[t]} \mathbf{x}^{[t]} (y_i^{[t]} - \hat{p}_i^{[t]}) \quad (17)$$

$$w_i^{[t]} = \hat{p}_i^{[t]} (1 - \hat{p}_i^{[t]}) \quad (18)$$

$$\hat{p}_i^{[t]} = \frac{\exp(\mathbf{x}^{[t]'} \hat{\boldsymbol{\theta}}_i^{[t-1]})}{1 + \sum_{j=1}^{k-1} \exp(\mathbf{x}^{[t]'} \hat{\boldsymbol{\theta}}_j^{[t-1]})} \quad (19)$$

for  $i = 1, \dots, k-1$ . Here,  $\mathbf{M}_i^{[t]}$  approximates the inverse of the Hessian of  $\hat{\boldsymbol{\theta}}_i^{[t]}$  computed with observations  $1, \dots, t$ . Hopefully, given “good” initialization and a sufficient number of observations,  $\hat{\boldsymbol{\Theta}}^{[T]}$  is approximately equal to the estimate  $\tilde{\boldsymbol{\Theta}}$  obtained using IRLS.

A further approximation — which is more memory efficient — calls for using an aggregated inverse Hessian matrix  $\bar{\mathbf{M}}$  instead of a set of  $k-1$  inverse Hessian matrices (RIRLS-agg). That is,

$$\bar{\mathbf{M}}^{[t]} \leftarrow \bar{\mathbf{M}}^{[t-1]} - \frac{\bar{w}^{[t]} \bar{\mathbf{M}}^{[t-1]} \mathbf{x}^{[t]} \mathbf{x}^{[t]'} \bar{\mathbf{M}}^{[t-1]}}{1 + \bar{w}^{[t]} \mathbf{x}^{[t]'} \bar{\mathbf{M}}^{[t-1]} \mathbf{x}^{[t]}} \quad (20)$$

$$\hat{\boldsymbol{\Theta}}^{[t]} \leftarrow \hat{\boldsymbol{\Theta}}^{[t-1]} + [\bar{\mathbf{M}}^{[t]} \mathbf{x}^{[t]} (\mathbf{y}_*^{[t]} - \hat{\mathbf{p}}_*^{[t]})']' \quad (21)$$

$$\bar{w}^{[t]} = \hat{\mathbf{p}}_*^{[t]'} (\mathbf{1}_{k-1} - \hat{\mathbf{p}}_*^{[t]}) / (k-1) \quad (22)$$

where

$$\mathbf{y}_*^{[t]} = \begin{bmatrix} y_1^{[t]} \\ \vdots \\ y_{k-1}^{[t]} \end{bmatrix}', \quad \hat{\mathbf{p}}_*^{[t]} = \begin{bmatrix} \hat{p}_1^{[t]} \\ \vdots \\ \hat{p}_{k-1}^{[t]} \end{bmatrix}.$$

## 2 Convergence

*proof in progress...*

## 3 Simulation results

See figures 1, 2, and 3.

Figure 1 supports the claim that the RIRLS estimator converges to the IRLS estimator as the number of observations increases.

## 4 Reading

Iteratively Reweighted Least Squares (multinomial regression):

- [https://people.stat.sc.edu/gregorkb/Tutorials/MultLogReg\\_Algs.pdf](https://people.stat.sc.edu/gregorkb/Tutorials/MultLogReg_Algs.pdf)
- <https://hastie.su.domains/Papers/glmnet.pdf>
- <https://arxiv.org/pdf/1404.3177> (page 8)
- [https://waxworksmath.com/Authors/G\\_M/Hastie/WriteUp/Weatherwax\\_Epstein\\_Hastie\\_Solution\\_Manual.pdf](https://waxworksmath.com/Authors/G_M/Hastie/WriteUp/Weatherwax_Epstein_Hastie_Solution_Manual.pdf) (page 79)

Recursive Least Squares:

- <https://www.jld-stats.com/2020/03/27/updating-a-linear-regression-with-new-data>

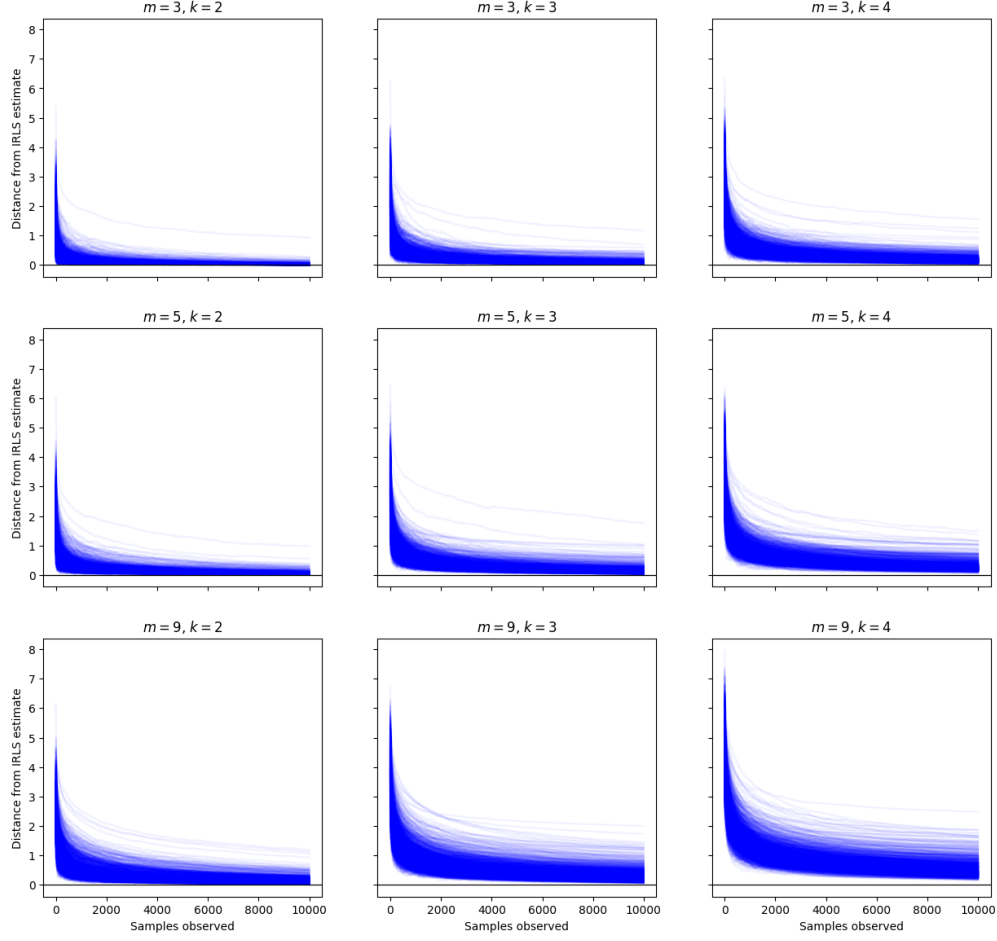


Figure 1: Convergence of iterative RIRLS estimates toward IRLS estimates for different numbers of covariates ( $m - 1$ ) and categories ( $k$ ). The plots show the distance of the RIRLS estimate from the converged IRLS estimate over repeated trials with simulated data. Distance is measured as  $\|\hat{\Theta}^{[t]} - \tilde{\Theta}\|_F$ .

- <https://dsbaero.engin.umich.edu/wp-content/uploads/sites/441/2019/08/RLSCSM.pdf>

Incremental updates:

- <https://www.tandfonline.com/doi/full/10.1080/10618600.2022.2035231#d1e213>
- <https://pmc.ncbi.nlm.nih.gov/articles/PMC9006691/>
- <https://academic.oup.com/biometrics/article/68/1/23/7390679> (interesting)
- <https://www.sciencedirect.com/science/article/abs/pii/S0895435607002132>

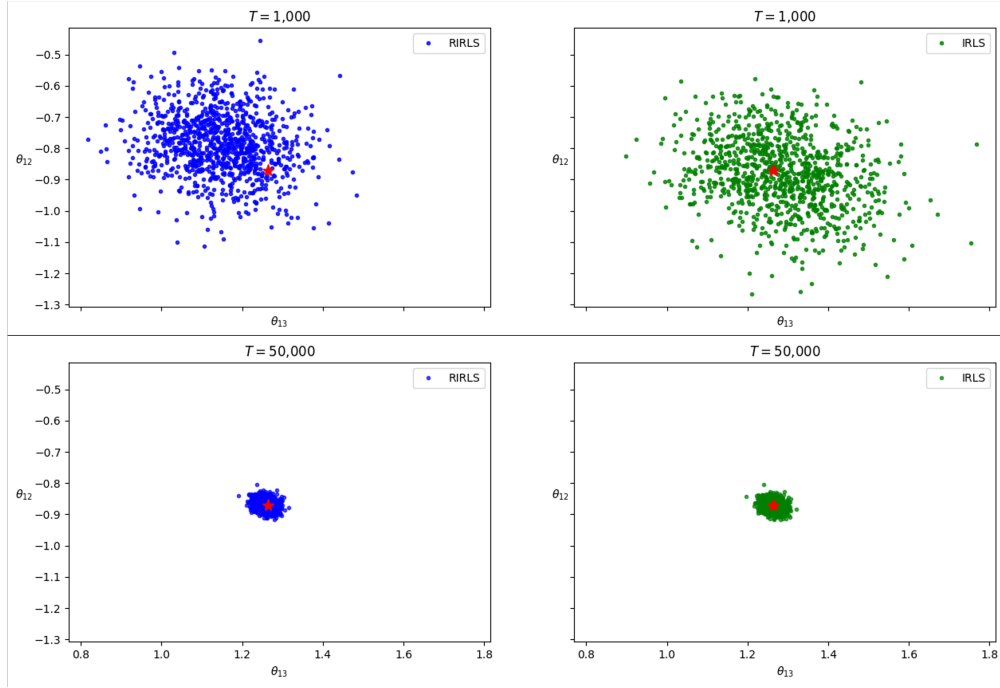


Figure 2: Simulated data is generated according to a fixed parameter vector  $\Theta$ . There are two normally distributed covariates (corresponding to  $\theta_{12}$  and  $\theta_{13}$ ) and a column of ones (corresponding to  $\theta_{11}$ , the intercept). In the top panel, 1,000 observations are generated for each trial; in the bottom panel, 50,000 observations are generated for each trial. The location of the true parameters  $(\theta_{12}, \theta_{13})$  is fixed for all trials and marked by a red star. The location of parameter estimates are marked by circles. RIRLS estimates are shown on the left and IRLS estimates are shown on the right.

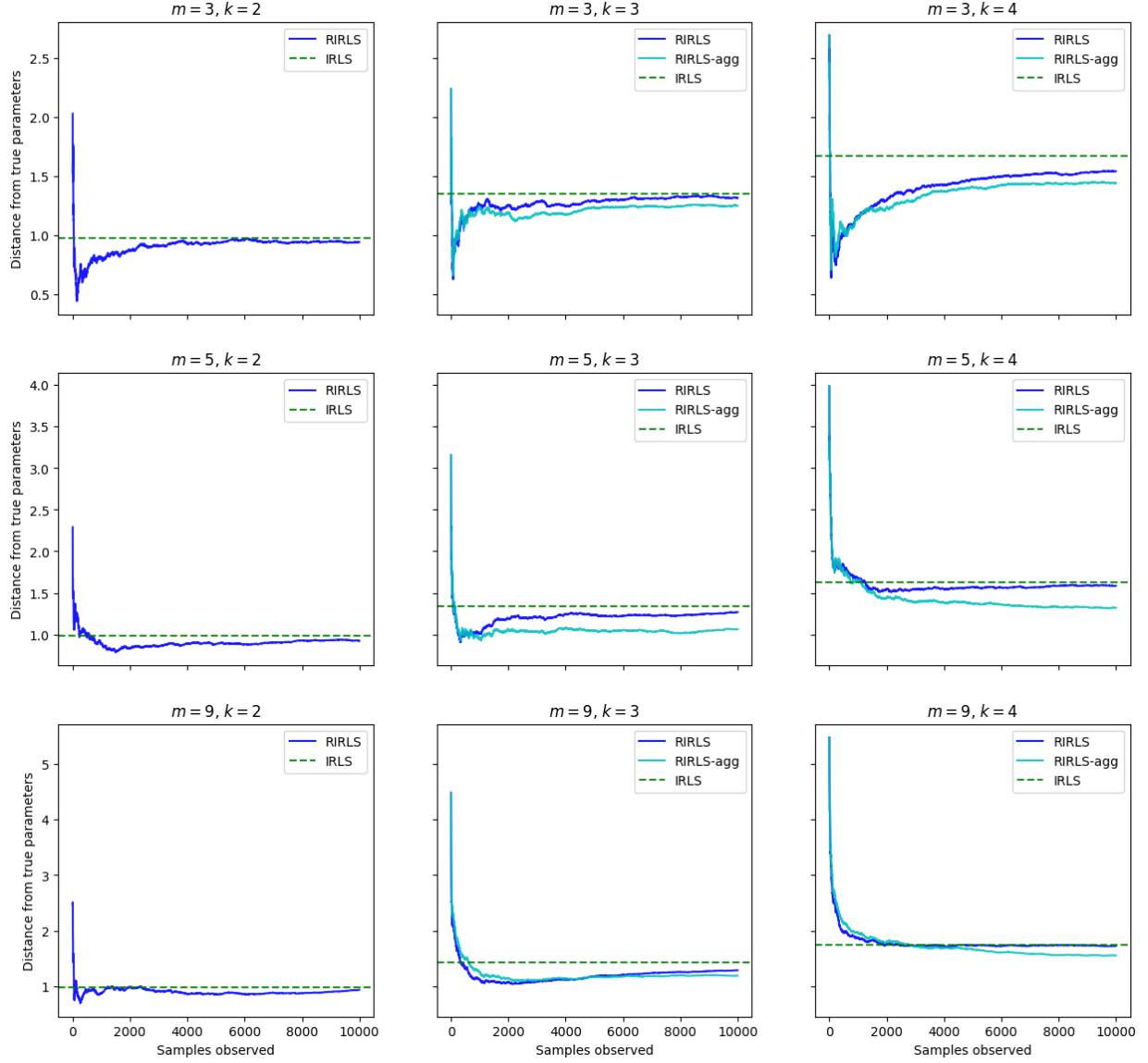


Figure 3: Convergence of parameter estimates is measured for different numbers of covariates ( $m - 1$ ) and categories ( $k$ ). The plots show the distance of the estimates from the true parameter as a function of the number of samples observed ( $t$ ). The RIRLS (standard and aggregated) estimates are shown as solid blue curves. The dashed green line shows the distance for the converged IRLS estimate.