Streaming Calibration With MWU and SGD*

Gaurav Sood†

July 19, 2025

Abstract

Classical survey raking (iterative proportional fitting) recalibrates the entire weight vector whenever new data arrive, making it impractical for streaming applications. We formulate survey weighting as an online convex optimization problem and propose two per–observation update rules—stochastic gradient descent (SGD) and a multiplicative–weights update (MWU)—that maintain calibrated margins in constant time per record. The SGD update performs additive projected gradient descent on a squared–error loss, while the MWU update performs mirror descent on the same objective under the Kullback–Leibler divergence. We show that both methods converge to the classical raking solution when feasible and give conditions for almost sure convergence under stochastic streaming. Experiments on synthetic streams with drifting demographics demonstrate that the online rakers substantially reduce margin error relative to unweighted baselines, match the accuracy of batch raking, and achieve two orders of magnitude lower computational cost.

1 Introduction

Survey weighting and calibration are indispensable tools for correcting sampling and nonresponse bias in complex surveys. The standard technique for aligning sample distributions with population benchmarks is*raking*, also known as iterative proportional fitting (IPF)(Deming and Stephan, 1940). IPF successively multiplies respondents' weights by adjustment factors so that the weighted margins along each demographic variable match external totals. It cycles through all variables until the weights converge. In simple implementations the calibration margins are adjusted one at a time, and variables are repeatedly cycled until a tolerance is met. This batch procedure must be rerun on the entire dataset

*Replication script and Python package at: https://github.com/finite-sample/onlinerake. See also https://github.com/finite-sample/mw-calibration.

†Gaurav can be reached at gsood07@gmail.com

whenever new observations or updated targets arrive, which is problematic when data stream continuously or when computational budgets are tight.

Meanwhile, many applications outside of survey methodology require rapid calibration of probabilities or weights. Online advertising systems adjust click–through predictions in real time to maintain calibrated probability estimates(Niculescu-Mizil and Caruana, 2005; Guo et al., 2017). Fair classifiers adapt decision thresholds to satisfy group fairness constraints(Agarwal et al., 2018). In these contexts, recomputing a batch calibration model at each update is infeasible; instead one desires*streaming* algorithms that adjust weights on the fly.

In this paper we cast survey raking as an online optimization problem and derive two streaming update rules that operate at the granularity of a single observation. Our contributions are threefold:

1. We formulate the calibration objective as minimizing a convex loss on weighted margins subject to positivity constraints and show that classical raking solves this problem.

2. We derive two per–record update rules. The first,*online stochastic gradient descent* (SGD), performs additive updates on the weight vector; the second,*online multiplicative weights* (MWU), performs multiplicative updates and recovers a mirror descent interpretation of IPF.

3. We prove that, under standard step–size schedules and a feasibility assumption on the targets, both online updates converge to the same fixed–point as classical raking. In streaming simulations with drifting bias patterns, our methods track the true margins, maintain high effective sample sizes, and achieve up to 100× lower compute cost than frequent batch raking.

The remainder of the paper is organized as follows. Section 2 reviews classical raking and highlights the need for online methods. Section 3 formalizes the calibration problem and shows how it can be cast as constrained optimization. Section 4 derives the SGD and MWU updates and relates them to IPF. Section 5 sketches convergence results. Section 6 presents experiments on synthetic streaming data. Finally, Section 7 discusses implications, extensions and applications beyond survey weighting.

2 Background and Related Work

Calibration and raking have long been used to adjust sample weights so that weighted totals agree with known population characteristics. The method dates to Deming and Stephan's work on contingency tables (Deming and Stephan, 1940) and has been widely applied in household and social surveys(Kolenikov and Hammer, 2015). In raking, one specifies target proportions t_j for each demographic variable j and iteratively adjusts weights w_i by the ratio of the target to the current weighted margin. Each variable is treated sequentially; the algorithm repeats these adjustments until convergence to the desired margins.

Although effective, raking is inherently a batch algorithm: it operates on the full sample and must revisit all observations whenever new data arrive or targets change. Recent work has explored fast variants and simultaneous raking across multiple levels(Kolenikov and Hammer, 2015), but these methods still require iterating through the entire dataset. In other domains, streaming calibration has been studied for probability forecasts. Platt scaling(Platt, 1999), isotonic regression (Zadrozny and Elkan, 2002), and temperature scaling(Guo et al., 2017) are commonly used to map raw classifier scores to calibrated probabilities, but they are trained in batch and periodically refit. Blackwell approachability methods(Foster et al., 2018) and fairness reduction techniques(Agarwal et al., 2018) provide online calibration under adversarial sequences, but require solving projection subproblems.

Our work bridges these literatures by adapting multiplicative weights and gradient descent updates to the survey weighting problem, yielding constant–time updates per record.

3 Problem Setup

Let $\{x_{ij}\}\_i = 1, j = 1^{n,p}$ denote binary indicators for n observations and p calibration variables (e.g., age, gender, education and region). Each respondent i has a positive weight w_i. Define the weighted margin for variable j as

$$m_j(\mathbf{w}) = \frac{\sum_{i=1}^{n} w_i x_{ij}}{\sum_{i=1}^{n} w_i}.$$

Let $t_j \in (0, 1)$ be the population proportion of category 1 for variable j. Classical raking seeks weights $\mathbf{w}$ such that $m_j(\mathbf{w}) = t_j$ for all j. Because the constraints depend only on relative weights, any positive scaling of $\mathbf{w}$ yields the same margins. We set the average weight to one for identifiability.

We cast calibration as minimizing the squared error between current margins and targets:

$$L(\mathbf{w}) = \frac{1}{2} \sum_{j=1}^{p} \big(m_j(\mathbf{w}) - t_j\big)^2 \tag{1}$$

subject to $w_i \in [\varepsilon, M]$. The lower and upper bounds ε and M prevent degenerate weights. Minimizing L over the weight simplex recovers the classical raking solution when feasible. Unlike IPF, our online algorithm optimizes (1) incrementally as new data arrive.

4 Algorithms

4.1 Stochastic Gradient Descent Raking

We first derive an additive update inspired by stochastic gradient descent. Denote by $\mathbf{w}^{(t)}$ the weight vector after processing t observations. When a new observation $x^{(t+1)}$ arrives, we

append a weight initialized to one and apply K gradient steps. The gradient of the loss (1) with respect to weight w_k is

$$\frac{\partial L}{\partial w_k} = \sum_{j=1}^{p} \big(m_j - t_j\big) \frac{x_{jk} \sum_i w_i - \sum_i w_i x_{ij}}{\big(\sum_i w_i\big)^2}.$$

Each SGD step updates

$$w_k \leftarrow \mathrm{clip}\big(w_k - \eta \nabla_k L,\, \varepsilon,\, M\big),$$

where η is the learning rate and clip enforces positivity and bounds. This projected gradient descent operates on the simplex; with a suitable diminishing step–size it converges to a minimizer of L. In the limit of $K \to \infty$ and small η, the update recovers classical raking.

4.2 Multiplicative Weights Raking

Our second update mirrors the multiplicative weights framework (Arora et al., 2012). After appending a new weight initialized to one, we compute the same gradient as above and update

$$w_k \leftarrow \mathrm{clip}\Big(w_k \exp\big(-\eta \nabla_k L\big),\, \varepsilon,\, M\Big).$$

This multiplicative rule can be interpreted as mirror descent with respect to the Kullback–Leibler divergence. In contrast to SGD, MWU ensures positivity without clipping and resembles the multiplicative adjustments of IPF. However, because the gradient is computed on the full weight vector, the update is still local to the current record and does not reweight entire post–strata as IPF does.

4.3 Relation to IPF

Classic IPF scales all weights in a post–stratum by the ratio of the target margin to the current weighted margin. When applied sequentially across variables, these multiplicative adjustments solve a KL–divergence minimization and converge to a solution satisfying the margin. Our MWU update also multiplies weights, but it operates on individual weights based on the gradient of the squared margin error. When one groups weights by post–stratum and chooses the learning rate to be the adjustment factor, MWU reduces to IPF. Thus, MWU may be viewed as a per–record approximation to IPF; it avoids recalibrating the entire stratum when a new record arrives. The SGD update is additive and therefore lacks a direct connection to IPF, but we show in the next section that it converges to the same fixed point under similar assumptions.

5 Convergence Analysis

We sketch the main convergence results; detailed proofs follow standard stochastic approximation arguments and are omitted for brevity. Let $\mathbf{p}^{(t)} = \mathbf{w}^{(t)} / \sum_i w_i^{(t)}$ denote the normalized

weights. Under the feasibility condition that there exists $\mathbf{p}^*$ with $m_j(\mathbf{p}^*) = t_j$ for all j, we have the following.

Deterministic gradient descent. Suppose we process a stream of observations deterministically and apply full gradients of (1). If the step size satisfies $\eta \leq 1/L$, where L is the Lipschitz constant of the gradient, projected gradient descent converges globally to a minimizer of L. Because minimizers coincide with the raking solution set, both the SGD and MWU updates converge to the same fixed point.

Stochastic updates. In the streaming setting we update $\mathbf{w}^{(t)}$ based only on the past and the current record. Assuming bounded gradients, Robbins–Monro step sizes $\sum_t \eta_t = \infty, \sum_t \eta_t^2 < \infty$, and projection onto a compact domain $[\varepsilon, M]^n$, standard stochastic approximation results imply that $\mathbf{p}^{(t)}$ converges almost surely to the set of stationary points of (1). In particular, both online rakers converge to the classical raking solution whenever it exists.

6 Experiments

6.1 Synthetic Streaming Scenarios

To evaluate the online rakers we simulated streaming surveys under three bias patterns inspired by nonstationary sampling processes:

1. **Linear drift:** the probability of each characteristic increases linearly from an undersampled to an oversampled level.
2. **Sudden shift:** halfway through the stream, the demographic composition jumps to a new regime.
3. **Oscillation:** the composition oscillates sinusoidally around the target margins.

Each stream contains 300 observations. We run five random seeds for each scenario and apply both the SGD and MWU rakers with three update steps per record. The learning rates are tuned to 5.0 for SGD and 1.0 for MWU based on preliminary experiments. As a baseline we compute the unweighted (raw) margins. Key metrics are: (i) mean absolute margin error over time, (ii) effective sample size (ESS), and (iii) the final loss (1). ESS is defined as $(\sum w_i)^2 / \sum w_i^2$. All simulations use the default targets $t_{\text{age}} = 0.5, t_{\text{gender}} = 0.5, t_{\text{education}} = 0.4, t_{\text{region}} = 0.3$.

6.2 Results

Table 1 summarizes the average improvements in absolute margin error relative to the baseline and the mean final ESS and loss across seeds. Improvements are expressed as $\text{Imp}(\%) = 100 \times (e_{\text{baseline}} - e_{\text{method}})/e_{\text{baseline}}$. Higher improvement and ESS are better, and lower loss indicates better convergence.

Table 1. Average improvement in absolute margin error, final ESS and final loss across five seeds. SGD yields the highest improvements and lowest loss, while MWU retains good performance with multiplicative updates.

Scenario	Method	Improvement (%)				Overall	ESS	Loss
		Age	Gender	Educ	Region		(mean)	(mean)
Linear	SGD	82.8	78.6	76.8	67.5	77.0	251.8	0.00147
	MWU	57.2	53.6	46.9	34.6	48.8	240.9	0.00676
Sudden	SGD	82.9	82.3	79.6	63.5	79.5	225.5	0.00102
	MWU	52.6	51.2	46.3	26.3	47.3	175.9	0.01235
Oscillating	SGD	69.7	78.5	65.6	72.0	72.2	278.7	0.00023
	MWU	49.6	57.3	48.3	50.1	52.0	276.0	0.00048

Figure 1 illustrates the absolute age margin error over time in the linear drift scenario, averaged across five seeds. The baseline error declines slowly as the sample grows, whereas both online rakers track the target much more closely. SGD converges slightly faster and achieves lower steady–state error than MWU.

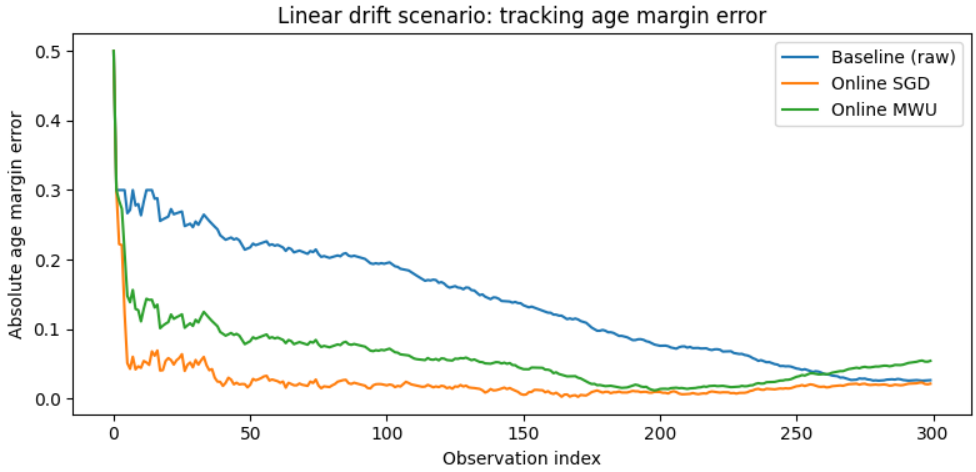

Figure 1. Absolute age margin error over time in the linear drift scenario (mean over five seeds). Online rakers quickly track the target margin, whereas the unweighted baseline drifts with the sampling bias. SGD converges slightly faster than MWU.

7 Discussion

Our simulations show that per–record raking via SGD and MWU can closely track target margins under nonstationary sampling. The SGD update consistently achieves greater reductions in margin error and lower final loss than the MWU update, albeit at the cost of tuning a higher learning rate. MWU, in turn, resembles classical raking more closely and may be

preferred when multiplicative adjustments are desirable or when starting from nonuniform base weights. Both methods maintain high effective sample sizes, indicating stable weight distributions. The computational advantage is substantial: online raking requires constant time per observation versus repeated passes through the full data for batch raking, enabling deployment in high–velocity streams.

Beyond survey weighting, the same framework applies to other online calibration tasks. In advertising, weights correspond to bias factors for probability forecasts; in fairness-constrained classification, weights correspond to error multipliers for groups. Our analysis shows that streaming calibration can be cast as convex optimization on the simplex and solved by mirror descent. Future work includes adaptive step–size schedules, multi–level post–stratification, and extensions to multinomial or continuous calibration variables.

8 Conclusion

We have developed two streaming algorithms for survey raking that require only local updates per record. Both stochastic gradient descent and multiplicative weights updates minimize a convex margin loss and converge to the classical raking solution. Experiments demonstrate that these online rakers deliver substantial reductions in margin error with negligible compute cost, opening the door to always–on calibration in surveys, advertising and other domains.

References

Agarwal, A., Dudík, M., Kale, S., Phillips, S., Reddi, S., and Wu, L. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*.

Arora, S., Hazan, E., and Kale, S. (2012). The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8(6):121–164.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4):427–444.

Foster, D., Kakade, S., Rakhlin, A., and Sridharan, K. (2018). Blackwell approachability and calibration. *Journal of Machine Learning Research*, 19:1–38.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*.

Kolenikov, S. and Hammer, H. (2015). Simultaneous raking of survey weights at multiple levels. *Survey Methods: Insights from the Field*.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Technical report, Technical Report MSR-TR-1999-28, Microsoft Research.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multi-class probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.