On the Convergence of Multicalibration Gradient Boosting

Daniel Haimovich[1] **Fridolin Linder**[1] **Lorenzo Perini**[1] **Niek Tax**[1] **Milan Vojnović**[2]
[1]Meta, London, United Kingdom
[2]Department of Statistics, LSE, London, United Kingdom
{danielha,flinder,lorenzoperini,niek}@meta.com, m.vojnovic@lse.ac.uk

Abstract

Multicalibration gradient boosting has recently emerged as a scalable method that empirically produces approximately multicalibrated predictors and has been deployed at web scale. Despite this empirical success, its convergence properties are not well understood. In this paper, we provide computational guarantees for multicalibration gradient boosting algorithms. We show that the magnitude of successive prediction updates decays at $O(1/\sqrt{T})$ (in the minimum over T rounds), which implies the same convergence rate bound for the empirical multicalibration error over rounds. For the squared-error loss, this rate improves to linear convergence under additional smoothness assumptions on the weak learners, and convergence extends to relaxed and adaptive rescaling variants. Combining the empirical bound with uniform convergence arguments, we also bound the population multicalibration error for the log-loss under additional conditions on the weak learners, which yields a prescription for the number of rounds as a function of the sample size. Experiments on real-world datasets support our theory, clarify the regimes in which the method achieves fast convergence, and show that early stopping is needed on held-out data.

1 Introduction

Multicalibration has emerged as a rigorous criterion for trustworthy machine learning, requiring predictions to match true outcome expectations not only globally, but across a virtually unlimited class of (potentially overlapping) subpopulations (Hebert-Johnson et al., 2018; Haghtalab et al., 2023; Pfisterer et al., 2021). While originally applied for algorithmic fairness, multicalibration has proven relevant for general model reliability, improving (1) robustness against distribution shifts (Kim et al., 2022; Wu et al., 2024) and (2) performance in downstream tasks ranging from ranking to uncertainty quantification (Yang et al., 2024; Detommaso et al., 2024; Dwork et al., 2022).

To achieve this in practice, a new class of multicalibration gradient boosting methods has recently been developed (Tax et al., 2026; Jin et al., 2025). Unlike traditional post-processing methods that rely on grid discretization, these algorithms iteratively refine a predictor by training an ensemble of weak learners on residuals, dynamically augmenting the feature space at each round with the predictions from the previous round. Such an approach has shown significant empirical success: it requires no manual group specification, scales to large datasets and, as highlighted by MCGrad (Tax et al., 2026), can be deployed in production environments at web scale.

Classical versus Multicalibration Gradient Boosting. Classical gradient boosting fits weak learners on a fixed feature space to the residuals of the current ensemble; it is a greedy functional gradient method on a fixed convex objective over the span of a fixed weak-learner class, with well-understood convergence (e.g., Lu and Mazumder, 2020). Multicalibration gradient boosting differs in two ways: it has two loops (each outer round runs a full gradient boosting machine), and the weak learners

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of round $t+1$ take the current prediction $f_t(x)$ as an additional input, so the directions available to the algorithm change from round to round. The outer loop is thus not gradient descent on a fixed objective over a fixed function class. With an exact oracle and the squared or log-loss, its fixed points are exactly the predictors that are empirically multicalibrated with respect to the weak learners evaluated at their own predictions (Section 3). Gradient-boosted trees suit this task: with the current prediction as an extra feature, tree splits find the score and feature regions where the model is miscalibrated without manual group specification, repeated rounds correct miscalibration that shifts between regions, and fast implementations with logit rescaling, a minimum Hessian per leaf, and early stopping keep training and inference cheap (Tax et al., 2026); early stopping is also what our analysis and experiments recommend (Sections 5–6).

Despite this empirical adoption, the theory that supports multicalibration gradient boosting methods remains under-explored. Existing guarantees are either single-round and require a loss-saturation condition (Jin et al., 2025), or bound empirical multicalibration error by the prediction gap without proving it vanishes (Tax et al., 2026). Because the predictions at the t-th round become a feature input for round $t+1$, the optimization target shifts at every round, making existing convergence analysis for standard gradient boosting inapplicable for this setting. Thus, fundamental computational questions remain unanswered: does the sequence of predictors actually converge? Does the empirical multicalibration error converge to zero, and if so, how fast is this convergence? Can we guarantee stability when rescaling updates? And does the guarantee carry over to the population, and how many rounds should be run? Such computational guarantees establish algorithm correctness and guide hyperparameter selection.

In this paper, we address the aforementioned computational questions, i.e., identifying conditions under which the empirical multicalibration error converges to zero and, furthermore, establishing convergence rate bounds. Our analysis is focused on studying the convergence of the gap between successive prediction values, as it implies convergence for the empirical multicalibration error (Section 3). Our main convergence result accommodates both classification and regression tasks under different loss functions. We prove that for the standard algorithm, the gap between successive prediction values converges to zero with a decay rate of $O(1/\sqrt{T})$ (in the minimum over rounds), where T is the number of rounds, provided that the boosting oracle is sufficiently accurate in returning a nearly-optimal ensemble of weak learners. We provide further results for regression task with squared-error loss function, under the assumption of exact boosting oracle. Under smoothness assumptions on the weak learners, we show a linear convergence rate (i.e., the prediction gap shrinks by a factor $\kappa < 1$ per round when the projector onto the weak learners varies slowly enough). For the squared-error loss, we also extend our analysis to practical variants used to mitigate overfitting. We show that the $O(1/\sqrt{T})$ rate holds even when using relaxed rescaling weights $w_t \in (0, 1]$ that converge to 1 with $\sum_t (1 - w_t) < \infty$. We analyze an adaptive variant whose rescaling weight is set in each round by a line search on the training loss, and show that it keeps the loss non-increasing and, for bounded weak learners, drives the empirical multicalibration error to zero at an $O(1/\sqrt{T})$ rate in the minimum over rounds. Beyond the training set, we bound the *population* multicalibration error via uniform convergence over the class of T-round predictors (Section 5): for the log-loss and weak learners that are bounded and Lipschitz in the prediction feature, it is of order $(p/n)^{1/6}$ (up to constants and logarithmic factors) when T grows like $(n/p)^{1/3}$. We present empirical results validating our theory on the training set and on held-out data (Section 6).

2 Related Work

Foundations of Multicalibration. Multicalibration (Hebert-Johnson et al., 2018) is connected to outcome indistinguishability (Dwork et al., 2021), omniprediction (Gopalan et al., 2022a; Hu et al., 2023), boosting for regression (Globus-Harris et al., 2023), and swap-agnostic learning (Gopalan et al., 2023); related notions include multiaccuracy (Kim et al., 2019), low-degree multicalibration (Gopalan et al., 2022b), HappyMap (Deng et al., 2023), moment multicalibration (Jung et al., 2021), and multivalidity (Gupta et al., 2022; Jung et al., 2023) (Appendix E). Our work is different in providing the first finite-time convergence rate guarantees for recently proposed multi-round multicalibration gradient boosting algorithms (Tax et al., 2026) and considering both classification and regression tasks.

Analysis of Gradient Boosting. Classical gradient boosting is well studied, e.g., computational guarantees for (randomized) gradient boosting machines (Lu and Mazumder, 2020) and convergence and consistency of boosting with early stopping (Zhang and Yu, 2005). These analyses assume fixed features, i.e., a single round of our algorithm; we analyze the outer rounds, in which the prediction is fed back as a feature.

Multicalibration via Gradient Boosting. Jin et al. (2025) proposed a tree-ensemble method trained on features augmented with the current prediction and showed that approximate multicalibration can be achieved in a single round under a restrictive loss saturation condition. Tax et al. (2026) introduced MCGrad, a recursive GBDT method deployed at web-scale, and bounded the empirical multicalibration error by a function of the gap between successive predictions, without establishing that this gap converges to zero. Existing approaches thus lack convergence guarantees for arbitrarily many rounds (Appendix E). Regarding generalization, Tax et al. (2026) give no population guarantee, and the finite-sample analysis of Jin et al. (2025) concerns a single round under the loss saturation condition; it is therefore not directly comparable with our population bound for the multi-round algorithm (Theorem 5.1), which requires no such condition.

3 Preliminaries and Problem Setup

Multicalibration and Notation. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be data points taking values in $\mathbb{R}^d \times \mathbb{R}$, assumed to be random i.i.d. from a distribution D. We denote the vector of targets as $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and a predictor as $f : \mathbb{R}^d \to \mathbb{R}$, with $f := (f(x_1), \ldots, f(x_n))^\top \in \mathbb{R}^n$ also denoting the prediction vector when the context is unambiguous.

Let $\mathcal{B}$ be a finite set of functions (weak learners), where $b_j : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ for each $b_j \in \mathcal{B}$. The cardinality of $\mathcal{B}$ is denoted by p. A key feature of multicalibration boosting methods is that predictors interact with other predictions: intuitively, $b_j(x, f)$ represents a predicted response for a given feature vector x and an auxiliary prediction feature f. For any prediction vector $f \in \mathbb{R}^n$ (where f_i corresponds to x_i), we define the matrix-valued function $B(f) \in \mathbb{R}^{n \times p}$ such that the j-th column is $b_j(f) := (b_j(x_1, f_1), \ldots, b_j(x_n, f_n))^\top$.

Note that the set of weak learners can be arbitrarily large but finite. This assumption allows analysis in a finite-dimensional space and covers common weak learners such as regression trees and parametric models with floating-point parameters.

The multicalibration errors of a predictor f with respect to the hypothesis class of functions $\mathcal{B}$ are defined as the vector $\mathcal{E}(f) \in \mathbb{R}^p$ with components $\epsilon_j = \mathbb{E}_{(x,y)\sim D}[(y - \phi(f(x)))b_j(x, f(x))]$ for $j = 1, \ldots, p$, where ϕ is the link function (identity for squared loss; sigmoid for log-loss). The predictor f is multicalibrated with respect to the hypothesis class $\mathcal{B}$ if $\mathcal{E}(f) = 0$. This can be extended naturally to the relaxed notion of α-multicalibration, which requires $\|\mathcal{E}(f)\| \leq \alpha$ for some norm function $\|\cdot\|$. The multicalibration error can also be seen as the correlation between the residuals and the weak learners evaluated at the current predictions.

Similarly, the empirical multicalibration errors are defined as the sample means over the given dataset, which can be written as:

$$\widehat{\mathcal{E}}(f) := \frac{1}{n} B(f)^\top (\mathbf{y} - \phi(f)). \tag{1}$$

We say that a loss $L(y, f)$ has a *canonical gradient* (with respect to the link ϕ) if $\partial L(y, f)/\partial f = \phi(f) - y$ for all y and f. This holds for the squared loss ($\phi = \mathrm{id}$) and for the log-loss written in terms of the logit f (ϕ the sigmoid). For such losses, the empirical multicalibration error is the gradient of the training loss V (defined below) projected onto the weak learners: $\widehat{\mathcal{E}}(f) = -\frac{1}{n} B(f)^\top \nabla V(f)$.

Factorized Hypothesis Classes. A special type of the set $\mathcal{B}$ consists of functions that admit the following factorization. Let $\mathcal{H} = \{h : \mathbb{R}^d \to \mathbb{R}\}$ and $\mathcal{G} = \{g : \mathbb{R} \to \mathbb{R}\}$ be finite sets of functions with cardinalities m and k, respectively. For every $b \in \mathcal{B}$, we have $b(x, u) = h(x)g(u)$ for some $g \in \mathcal{G}$ and $h \in \mathcal{H}$. For this type of function, the matrix $B(f)$ can be expressed as follows. Let $h(x) = (h_1(x), \ldots, h_m(x))^\top$ and $g(u) = (g_1(u), \ldots, g_k(u))^\top$. Define (using boldface to disambiguate from the scalar b above):

$$\mathbf{b}(x, u) = h(x) \otimes g(u) \in \mathbb{R}^p \tag{2}$$

where $\otimes$ denotes the Kronecker product[1] and $p = mk$, and let

$$B(f) = (\mathbf{b}(x_1, f_1), \ldots, \mathbf{b}(x_n, f_n))^\top. \tag{3}$$

The Multicalibration Boosting Framework. We analyze the class of multicalibration gradient boosting algorithms (Tax et al., 2026; Jin et al., 2025). Unlike standard boosting, these methods dynamically update the feature space at every round by including the model's current prediction. For a loss $L : \mathbb{R}^2 \to \mathbb{R}$, let $V(f) := \sum_{i=1}^n L(y_i, f(x_i))$ denote the training loss. The procedure is formalized in Algorithm 1: at round t, the algorithm fits an ensemble of weak learners $h_t = B(f_t)\theta_t$, whose coefficients $\theta_t \in \mathbb{R}^p$ (approximately) minimize $\theta \mapsto V(f_t + B(f_t)\theta)$, with features augmented by f_t (equivalent to fitting residuals $y - f_t$ for squared loss).

Algorithm 1 Multicalibration Gradient Boosting

Require: Initial predictor f_0, data $(X, \mathbf{y})$, step size $\eta \in (0, 1]$, number of rounds $T > 0$.

1: **for** $t = 0, \ldots, T-1$ **do**
2: 1. Compute the matrix $B(f_t)$ using the current predictions f_t.
3: 2. Fit the ensemble $h_t = B(f_t)\theta_t$, with coefficients $\theta_t \in \mathbb{R}^p$ that (approximately) minimize $\theta \mapsto V(f_t + B(f_t)\theta)$.
4: 3. Update: $f_{t+1} = w_t(f_t + \eta h_t)$.
5: **end for**

The candidate predictor at round t is rescaled by a weight w_t to obtain the predictor for round $t+1$. By default $w_t = 1$ (no rescaling). We also consider relaxed rescaling weights following a fixed schedule and adaptive rescaling weights depending on prediction values, which we analyze for the squared-error loss; see Section 4.2.2 for details.

Abstracting from the inner boosting iterations, the algorithm induces the following discrete-time dynamical system:

$$f_{t+1} = w_t(f_t + \eta B(f_t)\theta_t) \tag{4}$$

where $\theta_t \in \mathbb{R}^p$ are the coefficients assigned to weak learners in round t by the (possibly inexact) boosting oracle.

For the squared-error loss function and an exact oracle, the dynamical system in (4) can be written as follows:

$$f_{t+1} = w_t(f_t + \eta A(f_t)(\mathbf{y} - f_t)), \tag{5}$$

where $A(f) := B(f)B(f)^+$ is the orthogonal projector onto the column space of $B(f)$, and $B(f)^+$ denotes the Moore-Penrose inverse of $B(f)$. The term $A(f_t)(\mathbf{y} - f_t)$ represents the best approximation of the residuals using the hypothesis class $\mathcal{B}$ conditioned on the current predictions.

Why the Algorithm Works, and the Prediction Gap. Each round (approximately) minimizes the loss over combinations of weak learners evaluated at the current predictions. For losses with a canonical gradient, such as the squared and log-losses, the first-order optimality condition of this problem is

$$B(f_t)^\top \big(\mathbf{y} - \phi(f_t + B(f_t)\theta_t)\big) = 0,$$

i.e., the candidate $f_t + B(f_t)\theta_t$ has zero empirical multicalibration error with respect to the weak learners evaluated at the *previous* predictions f_t (a similar argument appears in Tax et al., 2026); this is exact for an exact oracle and approximate otherwise. As the candidate differs from f_t by $\eta^{-1}(f_{t+1} - f_t)$, it remains to show that this gap shrinks, since $\|\widehat{\mathcal{E}}(f_t)\|$ is bounded by a multiple of $\|f_{t+1} - f_t\|$ plus an oracle noise floor (Eq. (6)). The gap evolves according to (4)–(5), analyzed in Section 4. With an exact oracle and $w_t = 1$, f_t is a fixed point ($h_t = 0$) exactly when $B(f_t)^\top \nabla V(f_t) = 0$, i.e., for losses with a canonical gradient, exactly when $\widehat{\mathcal{E}}(f_t) = 0$. Every rate on the gap thus translates into a rate for $\widehat{\mathcal{E}}$.

4 Main Results

We show convergence of (4) for general smooth losses and, for the squared-error loss, linear rates and extensions to rescaling. Full proofs are in Appendix A.

[1] For any two vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^k$, $u \otimes v = (u_1 v^\top, \ldots, u_m v^\top)^\top \in \mathbb{R}^{mk}$.

Proof Techniques and What Is New. Standard arguments are flagged after each result; the parts specific to multicalibration boosting are: (i) the reduction of the empirical multicalibration error to the prediction gap via first-order optimality of each round's subproblem (Section 3), which turns a multicalibration question into the convergence of a dynamical system whose feature matrix $B(f_t)$ changes every round; (ii) pseudoinverse perturbation theory (Stewart, 1977) to control how fast the projector onto the weak-learner span moves, giving checkable conditions for linear convergence (Lemmas 4.3–4.4); (iii) an explicit noise floor that separates whether the outer dynamical system converges from whether the inner boosting oracle is accurate (Theorem 4.1, Eq. (6)); and (iv) a discussion, with an empirical check, of where the Lipschitz assumptions strain against hard-split trees (Sections 4.2.1 and 6). For the population bound, the new ingredient is a covering argument through the T rounds for this self-referential structure (Section 5).

We introduce the following assumptions on the training loss V of Section 3:

(A1) **Smoothness.** $V(f)$ is L-smooth on the segment $\{\tau f_t + (1-\tau)f_{t+1} : 0 \le \tau \le 1\}$ for all $t = 0, \ldots, T-1$.

(A2) **Strong convexity.** $V(f)$ is m-strongly convex on $\{f_t + \tau h_t : 0 \le \tau \le 1\}$ for all $t = 0, \ldots, T-1$, where $h_t = B(f_t)\theta_t$.

(A3) **Inexact boosting oracle.** At each step t, for some $\epsilon_t \ge 0$, the weights θ_t satisfy

$$V(f_t + B(f_t)\theta_t) \le \min_{\theta \in \mathbb{R}^p} V(f_t + B(f_t)\theta) + \epsilon_t.$$

For the case of regression with squared-error loss function, i.e., when $L(y, f) = \frac{1}{2}(y-f)^2$, smoothness and strong convexity hold globally for every f with $L = 1$ and $m = 1$.

For the case of binary classification with log-loss, $V(f)$ is globally smooth with $L = 1/4$, but its strong convexity vanishes as $\|f\|_\infty \to \infty$, so (A2) requires $\max_{0 \le \tau \le 1} \|f_t + \tau h_t\|_\infty$ to be bounded for all t (details in Appendix A.13); this holds for any fixed T if no round's weak learners (quasi-)separate the labels (Lemma A.11).

The boosting oracle is said to be exact when condition (A3) holds with $\epsilon_t = 0$. In this case, in each round t, the boosting machine outputs a linear combination of weak learners $B(f_t)\theta_t$, where $B(f) \in \mathbb{R}^{n \times p}$ is the matrix of weak learner evaluations defined in Eq. (3), and θ_t is the optimal coefficient vector $\theta_t \in \mathbb{R}^p$ that minimizes the loss function, i.e.,

$$\theta_t := \arg\min_{\theta \in \mathbb{R}^p} \Big\{ L_t(\theta) := \sum_{i=1}^{n} L\Big(y_i, f_t(x_i) + \sum_{j=1}^{p} b_j(x_i, f_t(x_i))\theta_j\Big)\Big\}.$$

The exact boosting oracle assumption is justified by fast convergence of gradient boosting machines; e.g., in Lu and Mazumder (2020), it was shown that, under suitable conditions, gradient boosted machines converge exponentially fast for any smooth, strongly-convex loss function. In this case, the approximation error parameter ϵ_t can be controlled by the number of boosting iterations in round t. Within a round the features $(x, f_t(x))$ are fixed, so classical guarantees for the inner gradient boosting machine apply: ϵ_t decays geometrically (strongly convex losses) or as $O(1/M_t)$ (smooth losses) in the number M_t of inner iterations (Lu and Mazumder, 2020); Appendix B discusses the conditions and what they imply for the assumptions of Theorem 4.1.

4.1 Computational Guarantees

Our key result shows conditions for multicalibration boosting algorithms to have approximately non-increasing training loss, diminishing gap between successive predictions, and convergence of empirical multicalibration error to zero.

Theorem 4.1 (Convergence with No Rescaling)**.** *Consider the dynamical system in Eq. (4) with constant unit rescaling weights ($w_t = 1$) and the step size $\eta \le m/(2L)$. Then:*

1. ***Approximately Non-increasing Training Loss:*** *under conditions (A2)–(A3),*

$$V(f_{t+1}) \le V(f_t) + \epsilon_t.$$

2. ***Diminishing Gap between Successive Predictions:*** *Under conditions (A1)-(A3), (a) if $\sum_{t=0}^{\infty} \epsilon_t < \infty$, then the gap between successive predictions satisfies $\|f_{t+1} - f_t\|_2 \to 0$, and*

(b) the gaps between successive predictions are bounded as

$$\frac{1}{T}\sum_{t=0}^{T-1}\|f_{t+1}-f_t\|^2 \le \frac{4\eta}{m}\left(\frac{V(f_0)}{T}+\frac{\eta}{T}\sum_{t=0}^{T-1}\epsilon_t\right).$$

Consequently,

$$\min_{0\le t\le T-1}\|f_{t+1}-f_t\| \le \sqrt{\frac{4\eta V(f_0)}{m}}\,\frac{1}{\sqrt{T}}+\sqrt{\frac{4\eta^2}{m}\,\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_t}.$$

For the case of the exact boosting oracle, the loss is non-increasing, and the second term in the bound on the minimum gap between successive predictions is equal to zero; the step-size condition then relaxes to $\eta \le m/L$, and the constant 4 improves to 2 (Appendix A.1). When $\epsilon_t = \epsilon$ is constant, the bound converges to the noise floor $2\eta\sqrt{\epsilon/m}$. When $\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_t \to 0$ as $T \to \infty$, the bound still vanishes. Specifically, when $\sum_{t=0}^{T-1}\epsilon_t = O(1)$, the bound is still $O(1/\sqrt{T})$.

Proof idea (standard first-order argument). Strong convexity (A2) with near-optimality (A3) upper-bounds $\nabla V(f_t)^\top h_t$, and smoothness (A1) upper-bounds the loss at the shrunk step in terms of it; the unknown gradient term cancels and telescoping bounds $\sum_t \|f_{t+1}-f_t\|^2$ (Appendix A.1). This is the usual descent-lemma sandwich; what is specific here is that only per-round properties of the subproblem are used, although $B(f_t)$ changes at every round.

The convergence rate bounds on the gap between successive predictions imply convergence rate bounds on the empirical multicalibration error. For a loss with a canonical gradient (Section 3), i.e., $\nabla V(f) = \phi(f) - \mathbf{y}$, and V globally L-smooth (e.g., the squared loss, $\phi = \mathrm{id}$ and $L = 1$, and the log-loss, $L = 1/4$), with an inexact oracle and unit-valued rescaling weights, we have (Lemma A.1):

$$\|\widehat{\mathcal{E}}(f_t)\|_2 \le \frac{1}{n}\|B(f_t)\|_2\left(\frac{L}{\eta}\|f_{t+1}-f_t\|_2+\sqrt{2L\epsilon_t}\right). \tag{6}$$

The term $\sqrt{2L\epsilon_t}$ constitutes a noise floor: even after the prediction gap vanishes, the multicalibration error can be no better than $O\big(\|B(f_t)\|_2\sqrt{L\epsilon_t}/n\big)$. For such losses, a group-normalized multicalibration error is also bounded by the gap (Appendix A.3).

This theorem provides the first theoretical justification for the success of boosting algorithms like MCGrad (Tax et al., 2026). It guarantees that, despite the changing feature space, the algorithm converges to a limit set of points and effectively minimizes the empirical multicalibration error. In subsequent sections we provide further convergence results for the case of exact boosting oracle and regression tasks with the squared-error loss function.

4.2 Additional Results for Regression with Squared-Error Loss and Exact Boosting Oracles

4.2.1 When is Convergence Fast?

While Theorem 4.1 guarantees sublinear convergence, empirical insights (see Section 6) often suggest faster (linear) convergence. Therefore, we investigate the structural properties of the hypothesis class $\mathcal{B}$ required to allow such faster convergence rates.

Theorem 4.2 (Linear Convergence under Smoothness)**.** *Assume $A(f)$ is Lipschitz continuous with constant L_A along the trajectory $\{f_t\}$ and let $\kappa := 1-\eta+\eta L_A\|\mathbf{y}-f_0\|_2$. Then, for every $t \ge 1$:*

$$\|f_{t+1}-f_t\|_2 \le \kappa\|f_t-f_{t-1}\|_2.$$

As a result, if $\kappa < 1$, then the gap converges to zero linearly (geometrically fast).

Theorem 4.2 applies to weak learners for which $A(f)$ is Lipschitz continuous. For standard decision-tree-based weak learners, $A(f)$ is in general not Lipschitz continuous due to discontinuities introduced by hard splits. Natural examples of Lipschitz continuous weak learners include soft trees (replacing hard splits with sigmoid gating functions), linear models, and smooth kernel-based weak learners. Despite this lack of strict Lipschitz continuity, in Section 6 we observe geometric decay of the prediction gap when training with hard-split GBDTs. We interpret this as evidence that large ensembles approximately recover the smoothness regime required by Theorem 4.2 (Appendix A.4).

Is Lipschitz Continuity Essential? Only the linear rate of Theorem 4.2 uses the Lipschitz condition on $A(f)$; Theorems 4.1, 4.5, and 4.6 do not. The proof uses it only between consecutive iterates, $\|A(f_t) - A(f_{t-1})\|_2 \leq L_A \|f_t - f_{t-1}\|_2$, a weaker trajectory-level condition that hard-split trees still need not satisfy.

The Lipschitz continuity of the projector $A(f)$ is a non-trivial condition. It holds for smooth weak learners under some spectral properties of matrix $B(f)$ as shown next.

Lemma 4.3 (Smoothness of $A(f)$)**.** *Assume that for some $\mathcal{F} \subseteq \mathbb{R}^n$, for every $f \in \mathcal{F}$, $B(f)$ is Lipschitz continuous with constant L_B, $B(f)$ has full column rank with smallest singular value at least $\delta > 0$, and $\|B(f)\|_2 \leq M$, for some $M > 0$. Then, $A(f)$ is Lipschitz continuous on $\mathcal{F}$ with constant:*

$$L_A \leq \frac{2}{\delta}\left(1 + c\,\frac{M^2}{\delta^2}\right) L_B \quad \textit{where } c = \sqrt{2}.$$

The condition on the smallest singular value requires the weak learners to be sufficiently diverse on the training data, and the bound M on the largest requires their combined output to stay bounded; an intuitive interpretation is in Appendix A.5.

For the factorized functions common in multicalibration (defined in Section 3), we can explicitly bound the Lipschitz constant L_B.

Lemma 4.4 (Smoothness of $B(f)$ for Factorized Learners)**.** *Let $\mathcal{B}$ consist of factorized functions $b(x, u) = h(x)g(u)$. If every $g \in \mathcal{G}$ is L_G-Lipschitz, then $B(f)$ is Lipschitz (so, under the conditions of Lemma 4.3, is $A(f)$). Specifically, the Lipschitz constant L_B satisfies:*

$$L_B \leq \big(\max_i \|h(x_i)\|_2\big)\sqrt{k}\, L_G.$$

This lemma links the abstract stability of the algorithm to the concrete smoothness of the weak learners ($g(u)$) and the diversity of the features ($H(X)$). This aligns with the intuition that "sharp" variations in weak learner behavior (e.g., unstable decision tree splits) might harm fast convergence.

Proof ideas. Theorem 4.2 is a direct calculation (Appendix A.4). Lemma 4.3 is the more technical step: it combines a classical perturbation bound for Moore–Penrose pseudoinverses (Stewart, 1977) with careful norm bookkeeping (Appendix A.5); Lemma 4.4 uses the Kronecker structure (Appendix A.6).

4.2.2 Relaxed and Adaptive Rescaling Strategies

In practice, the simple additive updates ($w_t = 1$) can lead to overfitting. Practitioners often employ some rescaling of w_t to regularize the predictor. Here, we show that convergence guarantees are robust to such modifications.

Relaxed Rescaling. We first consider the case where the weights w_t approach 1 asymptotically.

Theorem 4.5 (Convergence with Relaxed Weights)**.** *Assume that rescaling weights $w_t \in (0, 1]$ satisfy $\sum_t (1 - w_t) < \infty$. Then, (1) the training loss converges to a limit point, (2) the asymptotic multicalibration (whenever $\|B(f_t)\|_2$ is bounded along the trajectory) and (3) the $O(1/\sqrt{T})$ convergence rate of the gap $\|f_{t+1} - f_t\|_2$ established in Theorem 4.1 still hold. Specifically, the following holds:*

$$\min_{0 \leq t \leq T-1} \|f_{t+1} - f_t\|_2 \leq \frac{\sqrt{\eta}\|\mathbf{y} - f_0\|_2 + \gamma}{\sqrt{T}}$$

where γ is a constant depending on η, $\mathbf{y}$, f_0 and $\sum_t (1 - w_t)$.

Relaxing the weights thus does not break the convergence results, provided $w_t \in (0, 1]$ and $\sum_t (1 - w_t) < \infty$. The non-increasing loss property of Theorem 4.1 is not guaranteed to hold in the relaxed case; a weaker property that does hold, and the behavior of the constant γ, are discussed in Appendix A.7.

Adaptive Rescaling. A more aggressive strategy is to optimize w_t at each step. We define the adaptive weight $\omega(f_t)$ as the minimizer of $\|\mathbf{y} - \omega\varphi(f_t)\|_2$, where $\varphi(f_t) = f_t + \eta A(f_t)(\mathbf{y} - f_t)$ is the unscaled update direction. This yields the closed form

$$\omega(f_t) = \frac{\mathbf{y}^\top \varphi(f_t)}{\|\varphi(f_t)\|_2^2}. \tag{7}$$

This transforms the system into $f_{t+1} = \omega(f_t)\varphi(f_t)$, effectively projecting $\mathbf{y}$ onto the 1D subspace of the current update.

The following result shows that monotonicity and asymptotic multicalibration are preserved when using adaptive rescaling weights instead of unit-valued rescaling weights.

Theorem 4.6 (Convergence with Adaptive Rescaling)**.** *Assume that the rescaling weights are adaptive according to Eq. (7) (with $f_{t+1} = 0$ if $\varphi(f_t) = 0$). Then, (1) the non-increasing property of the loss and (2) the asymptotic multicalibration established in Theorem 4.1 still hold; more precisely, $\min_{0\le t\le T-1} \|A(f_t)(\mathbf{y} - f_t)\|_2 \le \|\mathbf{y} - f_0\|_2/\sqrt{\eta(2-\eta)T}$ and, if $M_B := \sup_t \|B(f_t)\|_2 < \infty$, then $\widehat{\mathcal{E}}(f_t) \to 0$ and $\min_{0\le t\le T-1} \|\widehat{\mathcal{E}}(f_t)\|_2 \le (M_B/n)\|\mathbf{y} - f_0\|_2/\sqrt{\eta(2-\eta)T}$. In addition, (3) for $t \ge 1$ the rescaling weights are such that $\omega(f_t) \ge 0$, i.e., the angle between $\mathbf{y}$ and $\varphi(f_t)$ is acute or right.*

Appendix A.9 gives a local quadratic rate under an additional condition on the weak learners.

Proof ideas (more intricate). Theorem 4.5 uses a Lyapunov function with correction terms for the rescaling drift and a Cauchy-sequence argument (Appendix A.7); for Theorem 4.6, comparing the adaptive step with the unscaled one gives a per-round decrease $\eta(2-\eta)\|A(f_t)(\mathbf{y} - f_t)\|_2^2$ that telescopes (Appendix A.8).

5 Population Multicalibration Error and Generalization

We now bound the *population* error $\mathcal{E}(f_t)$ of Section 3 via uniform convergence and derive how T should scale with the sample size (proofs in Appendix C). We consider binary classification with log-loss (ϕ the sigmoid, $y \in \{0, 1\}$); see Remark 5.3 for other losses. Even for finite $\mathcal{B}$, a union bound over $\mathcal{B}$ does not suffice: $b_j(x, f_t(x))$ depends on f_t, which ranges over a data-dependent class built by T rounds, whose complexity must be controlled.

Assumptions for the Population Bound. (G1) The initial predictor f_0 is fixed in advance, independent of the sample. (G2) Weak learners are bounded and Lipschitz in the prediction feature: $b(x, u) = (b_1(x, u), \dots, b_p(x, u))$ satisfies $\|b(x, u)\|_2 \le B_2$ and $\|b(x, u) - b(x, u')\|_2 \le L_b|u - u'|$. (G3) The coefficients are bounded: $\|\theta_t\|_2 \le B_\theta$ for all t. The constants B_2, L_b, B_θ are fixed in advance (not chosen from the sample).

Write $\bar{V}(f_0) := V(f_0)/n$, $\bar{M} := \max_{t<T} \|B(f_t)\|_2/\sqrt{n}$ (note $\bar{M} \le B_2$), $\Gamma := 1 + \eta B_\theta L_b$, $K := 2\eta B_\theta(B_2/4 + L_b)$ (if the weak learners take a transformed prediction $\phi(f(x))$ as input, replace L_b by $L_b L_\phi$ in Γ and K), and

$$\Lambda := \log \frac{10}{\delta} + p \log(1 + 2n) + pT \log(1 + KnT) + \frac{p}{2} T(T-1) \log \Gamma.$$

Theorem 5.1 (Population multicalibration error)**.** *Consider log-loss with an exact oracle ($\epsilon_t = 0$), $w_t = 1$, $\eta \le m/(2L)$, and (A1)–(A2) with $L = 1/4$. Under (G1)–(G2), for $n \ge 16$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the training sample, if (G3) holds then the round $t^* \in \arg\min_{0\le t\le T-1} \|f_{t+1} - f_t\|_2$ satisfies*

$$\|\mathcal{E}(f_{t^*})\|_2 \;\le\; \underbrace{2L\bar{M}\sqrt{\frac{\bar{V}(f_0)}{\eta m}}\,\frac{1}{\sqrt{T}}}_{\textit{optimization}} \;+\; \underbrace{\bar{M}\sqrt{\frac{18\Lambda}{n-1}} + \frac{30 B_2 \Lambda}{n-1}}_{\textit{generalization}}.$$

Proof outline. (1) *Duality.* $\|\mathcal{E}(f) - \widehat{\mathcal{E}}(f)\|_2$ is a supremum over unit θ of population-minus-empirical means of $e_{\theta,f} = \theta^\top b(x, f(x))(y - \phi(f(x)))$. (2) *Cover of covers.* Their covering number is at most that of the T-round predictors times that of the unit-ball weak-learner class (Lemma C.3). (3) *Covering across rounds.* One boosting round is Γ-Lipschitz in sup norm, so a cover of the T-round predictors is built from one shrunk weak-learner cover per round; its logarithm is at most of order $\frac{p}{2} T^2 \log \Gamma$ (Lemma C.2). (4) *Uniform convergence.* The Maurer–Pontil bound holds simultaneously over the class, hence at the data-dependent pair (θ^*, f_{t^*}) (Lemma C.5). (5) *Optimization term.* Lemma A.1 with Theorem 4.1 bounds $\|\widehat{\mathcal{E}}(f_{t^*})\|_2$ (Appendix C.3).

Corollary 5.2 (Choice of T). *Let $\Gamma > 1$ be a constant and $p \max\{1, \log \Gamma\} \leq n$. Taking $T \asymp (n/(p \log \Gamma))^{1/3}$ gives, with probability at least $1 - \delta$,*

$$\|\mathcal{E}(f_{t^*})\|_2 = \tilde{O}\big((p \log \Gamma / n)^{1/6}\big),$$

where $\tilde{O}$ hides the constants in Theorem 5.1 and logarithms of $n, 1/\delta$.

The bound is a bias–variance trade-off: T should grow only like $(n/p)^{1/3}$ (compare Section 6). We do not claim $1/6$ is tight[2].

Remark 5.3 (Scope). *The result is for unit weights (relaxed and adaptive rescaling are not covered) and does not apply directly to our experiments, which use squared loss and an f_0 fit on the same data; see Appendix C.5 for extensions.*

6 Numerical Results

6.1 Experimental Setup

Data and Hyperparameters. We evaluate the algorithms on five standard regression datasets summarized in Table 1; see Appendix D.1 for more details. For all experiments, we use a Random Forest regressor (100 trees, max depth 5) as the initial predictor f_0. The weak learners used in the boosting steps are 100 regression trees (max depth 3), implemented via scikit-learn's gradient boosting (learning rate 0.1).

MCE Computation. The multicalibration error is estimated as a sample mean of $\|\widehat{\mathcal{E}}(f)\|_2$ over $N_{\text{tree}} = 100$ random regression trees of depth $h = 3$, a class broader than the one used in training (details in Appendix D.2).

Rescaling. We consider three rescaling strategies: UNIT sets $w_t = 1$, RELAXED uses the power-law decay $w_t = 1 - (t+2)^{-3}$, and ADAPTIVE optimizes the weights via Eq. (7) in each round. We run the algorithms for $T = 20$ rounds with shrinkage $\eta = 0.5$, which satisfies Theorem 4.1's step-size condition $\eta \leq m/(2L) = 1/2$ for the squared loss.

6.2 Experimental Results

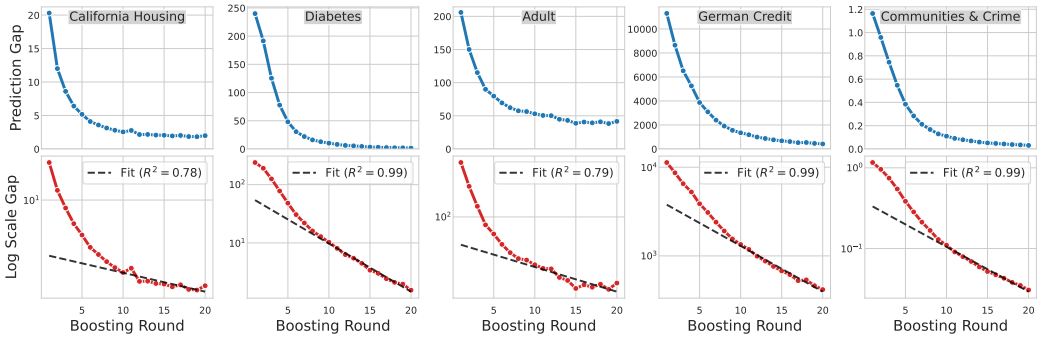

Figure 1: The prediction gap $\|f_{t+1} - f_t\|_2$ for all multicalibration boosting rounds $t \leq 20$. The bottom row shows the corresponding plots on a log–lin scale, along with the best fitting line (black) and its R^2 score.

Convergence with Unit Weights. Figure 1 shows a clear overall decay of the prediction gap toward zero on all datasets, with small per-round oscillations ($\leq 10\%$) consistent with Theorem 4.1 bounding the minimum and average gap. On a semi-log scale, a linear fit to the tail gives $R^2 \in [0.78, 0.99]$; the lower values (California Housing, Adult) reflect a sublinear plateau, consistent with hard-split GBDTs only approximately satisfying the condition of Theorem 4.2 (discussion in Appendix D.3).

[2]The exponent is set by the variance term: a weak learner with non-vanishing mean square has a column of norm $\asymp \sqrt{n}$, so $\|B(f_t)\|_2 \asymp \sqrt{n}$ and $\bar{M} = O(1)$. Treating $\|B(f_t)\|_2$ as bounded in n would suggest a faster rate, but with B_2 fixed this holds only for weak learners supported on sets of probability $O(1/n)$.

Convergence with Rescaling Strategies. On the training set, under all three strategies the training MSE decreases monotonically and the empirical MCE decays smoothly towards 0, often by orders of magnitude (Figure 3 and discussion in Appendix D.4).

Generalization and Overfitting. We also examine held-out data, averaging over 20 seeds with a fresh random 80/20 split each (protocol in Appendix D.5). Figure 2 plots the average test MCE with ± 1 standard-deviation bands.

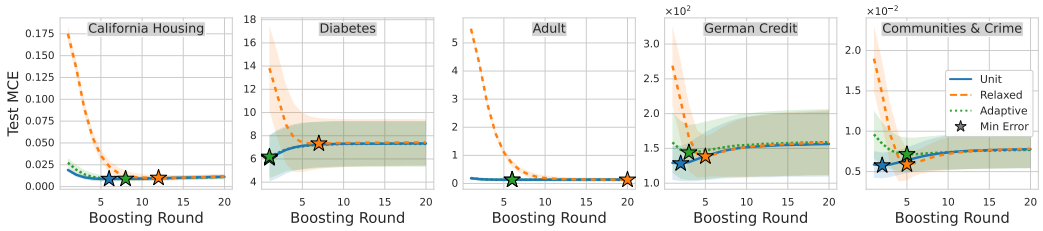

Figure 2: Average test Multicalibration Error (MCE) per dataset, with star markers indicating the optimal stopping point for each strategy.

Training versus test. Unlike the monotonic trend observed on the training set in Figure 3 (Appendix D.4), the test error shows the classic bias-variance trade-off. In datasets such as *California Housing* and *German Credit*, we observe a distinct U-shaped evolution, where the error decreases initially before facing overfitting. This confirms that while the dynamical system is stable, regularization via early stopping is necessary in practice. After its minimum, the test MCE rises (by up to about a third on the smaller datasets) and then plateaus, qualitatively as in Theorem 5.1.

Dependence on the sample size. Under Unit weights, the three datasets with $n \leq 2{,}000$ reach their minimum test MCE within two rounds, while the two with $n > 20{,}000$ do so around round 6; the other strategies show the same ordering. This is in line with Corollary 5.2 (the usable number of rounds grows with n), though with five datasets it is only suggestive.

Role of the weights. **Unit** and **Adaptive** typically reach their optimum within the first 1-8 rounds but can overfit almost immediately (most evident on *German Credit*), whereas **Relaxed** acts as an implicit regularizer that shifts the optimal stopping point later (around round 12 on *California Housing*, 5 on *German Credit*, and beyond $T = 20$ on *Adult*; Appendix D.5). However, Relaxed's minimum test MCE is not lower than Unit's (it is higher on *Diabetes* and *German Credit*), and the bands overlap on the smaller datasets: its benefit is robustness to the stopping time, not better generalization. A theory of this effect is left to future work; in practice, T should be chosen by validation.

7 Conclusion

In this work, we have bridged the gap between the empirical success of multicalibration gradient boosting and its theoretical foundations. While algorithms like MCGrad have been deployed at scale, their dynamic nature – where predictions recursively define the feature space – previously lacked rigorous analysis. By modeling this process as a deterministic dynamical system, we provided the first computational guarantees for this class of algorithms.

Specifically, we proved an $O(1/\sqrt{T})$ rate for the prediction gap and hence for the empirical multicalibration error (up to an $O(\sqrt{\epsilon})$ noise floor) for classification and regression; for the squared loss, linear convergence under smoothness and diversity of the weak learners, and $O(1/\sqrt{T})$ rates under relaxed and adaptive rescaling; and, for the log-loss, a population multicalibration bound of order $\tilde{O}((p \log \Gamma / n)^{1/6})$ (Theorem 5.1). Our experiments validate the training-set results and show that, on held-out data, the number of rounds must be controlled, in line with the trade-off in Theorem 5.1.

Limitations and Future Work. Key limitations include: (A2) fails for the log-loss on weak-learner-separable training sets; the linear rate requires a Lipschitz $A(f)$ with well-conditioned $B(f)$, only approximately satisfied by hard-split GBDT ensembles; and the population bound (Theorem 5.1) covers the log-loss with an exact oracle and unit weights under (G1)–(G3), and so does not apply directly to our experiments. Natural directions include tightness of the bounds, faster rates under broader conditions, infinite weak-learner classes, and the extensions in Appendix C.5.

Acknowledgments and Disclosure of Funding

Milan Vojnović performed this work in part while a visiting scientist with Meta. We thank Nastaran Okati for early discussions about multicalibration in the regression setting.

References

Zhun Deng, Cynthia Dwork, and Linjun Zhang. HappyMap: A Generalized Multicalibration Method. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 41:1–41:23, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Gianluca Detommaso, Martin Bertran Lopez, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in LLMs. In *Forty-first International Conference on Machine Learning*, 2024.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Beyond Bernoulli: Generating random outcomes that cannot be distinguished from nature. In *International Conference on Algorithmic Learning Theory*, pages 342–380. PMLR, 2022.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11459–11492. PMLR, 23–29 Jul 2023.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:21, Dagstuhl, Germany, 2022a. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-217-4.

Parikshit Gopalan, Michael P. Kim, M. Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Annual Conference on Computational Learning Theory*, 2022b.

Parikshit Gopalan, Michael Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 39936–39956. Curran Associates, Inc., 2023.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, pages 81:1–81:24, 2022.

Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 Jul 2018.

Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. In *International Conference on Machine Learning*, pages 13497–13527. PMLR, 2023.

Hongyi Henry Jin, Zijun Ding, Dung Daniel Ngo, and Steven Wu. Discretization-free multicalibration through loss minimization over tree ensembles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2634–2678. PMLR, 15–19 Aug 2021.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *The Eleventh International Conference on Learning Representations*, 2023.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

Haihao Lu and Rahul Mazumder. Randomized gradient boosting machine. *SIAM Journal on Optimization*, 30(4):2780–2808, 2020.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Florian Pfisterer, Christoph Kern, Susanne Dandl, Matthew Sun, Michael P. Kim, and Bernd Bischl. mcboost: Multi-calibration boosting for R. *Journal of Open Source Software*, 6(64):3453, 2021.

Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

G. W. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review*, 19(4):634–646, 1977.

Niek Tax, Lorenzo Perini, Fridolin Linder, Daniel Haimovich, Dima Karamshuk, Nastaran Okati, Milan Vojnovic, and Pavlos Athanasios Apostolopoulos. MCGrad: Multicalibration at web scale. In *Proceedings of ACM KDD (Applied Data Science Track)*, 2026.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13:217–232, 1973.

Jiayun Wu, Jiashuo Liu, Peng Cui, and Steven Z Wu. Bridging multicalibration and out-of-distribution generalization beyond covariate shift. *Advances in Neural Information Processing Systems*, 37:73036–73078, 2024.

Shuai Yang, Hao Yang, Zhuang Zou, Linhe Xu, Shuo Yuan, and Yifan Zeng. Deep ensemble shape calibration: Multi-field post-hoc calibration in online advertising. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6117–6126, 2024.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

A Proofs and Additional Results for Section 4

This appendix contains the proofs of the results of Section 4 (Appendices A.1–A.8), including the two bounds on the empirical multicalibration error in terms of the prediction gap (Appendices A.2–A.3). It also contains three additional results: a sufficient condition for quadratic convergence under adaptive rescaling (Appendix A.9), a hybrid boosting algorithm with a preconditioned update (Appendices A.10–A.12), and the existence of the exact oracle for the log-loss (Appendix A.13).

A.1 Proof of Theorem 4.1

Overview. The proof bounds the change of the training loss in each round from two sides. Strong convexity (A2), together with the near-optimality of the oracle (A3), gives an upper bound on the directional derivative $\nabla V(f_t)^\top h_t$ along the full step h_t. Smoothness (A1) bounds the loss after the shrunk step $f_{t+1} = f_t + \eta h_t$ in terms of the same directional derivative and a quadratic term. Combining the two bounds eliminates the directional derivative and shows that each round decreases the loss by an amount proportional to $\|f_{t+1} - f_t\|^2$, up to the oracle error $\eta\epsilon_t$; summing over rounds, these decreases telescope. The argument is standard in first-order optimization. What matters in our setting is that (A1)–(A3) need to hold only round by round, along the segments visited by the algorithm, even though the matrix $B(f_t)$ changes from one round to the next.

Approximately Non-Increasing Loss. Since $\theta = 0$ is a feasible choice in each round's subproblem, $\min_\theta V(f_t + B(f_t)\theta) \le V(f_t)$. Combined with (A3), this gives a bound on the loss after the full step $h_t = B(f_t)\theta_t$, that is, with $\eta = 1$,

$$V(f_t + h_t) \le V(f_t) + \epsilon_t. \tag{8}$$

To obtain a bound for the actual update $f_{t+1} = f_t + \eta h_t$, we use that V is convex on the segment between f_t and $f_t + h_t$, which holds by (A2). Since $\eta \in (0, 1]$,

$$V(f_{t+1}) = V\big((1-\eta)f_t + \eta(f_t + h_t)\big) \le (1-\eta)V(f_t) + \eta V(f_t + h_t) \le V(f_t) + \eta\epsilon_t \le V(f_t) + \epsilon_t.$$

This proves the first claim of the theorem.

Per-Round Decrease. Let $\Delta_t := f_{t+1} - f_t = \eta h_t$. By the smoothness assumption (A1), the loss at the new iterate is at most its first-order approximation plus a quadratic term,

$$V(f_{t+1}) \le V(f_t) + \nabla V(f_t)^\top \Delta_t + \frac{L}{2}\|\Delta_t\|^2, \tag{9}$$

and by the strong convexity assumption (A2), the loss after the full step is at least its first-order approximation plus a quadratic term,

$$V(f_t + h_t) \ge V(f_t) + \nabla V(f_t)^\top h_t + \frac{m}{2}\|h_t\|^2. \tag{10}$$

Rearranging (10) gives an upper bound on the directional derivative along h_t,

$$\nabla V(f_t)^\top h_t \le V(f_t + h_t) - V(f_t) - \frac{m}{2}\|h_t\|^2, \tag{11}$$

and bounding the difference $V(f_t + h_t) - V(f_t)$ by (8) yields

$$\nabla V(f_t)^\top h_t \le \epsilon_t - \frac{m}{2}\|h_t\|^2. \tag{12}$$

On the other hand, writing $\Delta_t = \eta h_t$ in (9) gives

$$V(f_{t+1}) \le V(f_t) + \eta \nabla V(f_t)^\top h_t + \frac{\eta^2 L}{2}\|h_t\|^2. \tag{13}$$

Substituting (12) into (13) eliminates the directional derivative, and using $\|h_t\|^2 = \|\Delta_t\|^2/\eta^2$ we obtain

$$V(f_{t+1}) \le V(f_t) + \eta\epsilon_t - \frac{\eta m}{2}\|h_t\|^2 + \frac{\eta^2 L}{2}\|h_t\|^2 = V(f_t) + \eta\epsilon_t - \frac{m - \eta L}{2\eta}\|\Delta_t\|^2,$$

or, equivalently,

$$\frac{m - \eta L}{2\eta}\|\Delta_t\|^2 \le V(f_t) - V(f_{t+1}) + \eta\epsilon_t. \tag{14}$$

Under the step-size condition $\eta \le m/(2L)$ we have $m - \eta L \ge m/2$, so the coefficient on the left-hand side of (14) is at least $m/(4\eta)$, and

$$\frac{m}{4\eta}\|\Delta_t\|^2 \le V(f_t) - V(f_{t+1}) + \eta\epsilon_t. \tag{15}$$

Summing over Rounds. Summing (15) over $t = 0, \dots, T-1$, the loss differences telescope to $V(f_0) - V(f_T) \le V(f_0)$, since the loss is non-negative. Hence

$$\frac{m}{4\eta}\sum_{t=0}^{T-1}\|\Delta_t\|^2 \le V(f_0) + \eta\sum_{t=0}^{T-1}\epsilon_t. \tag{16}$$

Dividing by T and multiplying by $4\eta/m$ bounds the average squared gap,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\Delta_t\|^2 \le \frac{4\eta}{m}\left(\frac{V(f_0)}{T} + \frac{\eta}{T}\sum_{t=0}^{T-1}\epsilon_t\right).$$

Since the smallest squared gap is at most the average, taking square roots gives

$$\min_{0\le t\le T-1}\|\Delta_t\| \le \sqrt{\frac{4\eta}{m}\left(\frac{V(f_0)}{T} + \frac{\eta}{T}\sum_{t=0}^{T-1}\epsilon_t\right)},$$

and the bound on the minimum prediction gap stated in the theorem follows from $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$.

Finally, if $\sum_{t=0}^{\infty}\epsilon_t < \infty$, the right-hand side of (16) stays bounded as $T \to \infty$. Hence $\sum_{t=0}^{\infty}\|f_{t+1} - f_t\|^2 < \infty$, and in particular $\|f_{t+1} - f_t\| \to 0$. □

Exact Oracle: A Sharper Bound. Suppose $\epsilon_t = 0$, so that θ_t minimizes $g(\theta) := V(f_t + B(f_t)\theta)$. Then $\nabla g(\theta_t) = B(f_t)^\top \nabla V(f_t + h_t) = 0$, hence $\nabla V(f_t + h_t)^\top h_t = \theta_t^\top \nabla g(\theta_t) = 0$. The function $\tau \mapsto V(f_t + \tau h_t)$ is $m\|h_t\|^2$-strongly convex on $[0,1]$ by (A2), so its derivative increases by at least $m\|h_t\|^2$ between $\tau = 0$ and $\tau = 1$,

$$\nabla V(f_t)^\top h_t \le \nabla V(f_t + h_t)^\top h_t - m\|h_t\|^2 = -m\|h_t\|^2.$$

Substituting into (13),

$$V(f_{t+1}) \le V(f_t) - \eta\Big(m - \frac{\eta L}{2}\Big)\|h_t\|^2 = V(f_t) - \frac{m - \eta L/2}{\eta}\|\Delta_t\|^2.$$

If $\eta \le m/L$, then $m - \eta L/2 \ge m/2$, and summing over t as before gives

$$\sum_{t=0}^{T-1}\|f_{t+1} - f_t\|^2 \le \frac{2\eta V(f_0)}{m}, \qquad \min_{0\le t\le T-1}\|f_{t+1} - f_t\| \le \sqrt{\frac{2\eta V(f_0)}{m}}\,\frac{1}{\sqrt{T}}.$$

□

Compared with Theorem 4.1, the admissible step sizes double and the constant halves; for the squared loss ($m = L = 1$), the bound holds for every $\eta \in (0, 1]$.

A.2 A Bound on the Multicalibration Error

Overview. Writing $n\widehat{\mathcal{E}}(f_t) = -B(f_t)^\top \nabla V(f_t)$, we add and subtract the gradient of the round's subproblem at θ_t: that gradient is small by the (approximate) first-order optimality of the boosting subproblem, and the remainder is bounded by the prediction gap via smoothness. This is the bridge from the gap to the multicalibration error (Section 3), and it gives Eq. (6).

Lemma A.1. *Let the loss have a canonical gradient (Section 3), i.e., $\nabla V(f) = \phi(f) - \mathbf{y}$, with V globally L-smooth, let $w_t = 1$ and $\eta \in (0, 1]$, and assume (A3). Then*

$$\|\widehat{\mathcal{E}}(f_t)\|_2 \le \frac{\|B(f_t)\|_2}{n}\left(\frac{L}{\eta}\|f_{t+1} - f_t\|_2 + \sqrt{2L\epsilon_t}\right).$$

For the squared-error loss, $L = 1$.

Proof. Write $B = B(f_t)$ and $h_t = B\theta_t$, so $f_{t+1} - f_t = \eta h_t$. Since $\nabla V(f) = \phi(f) - \mathbf{y}$, the definition of the empirical multicalibration error (Eq. (1)) gives $n\widehat{\mathcal{E}}(f_t) = B^\top(\mathbf{y} - \phi(f_t)) = -B^\top \nabla V(f_t)$. Let $g(\theta) := V(f_t + B\theta)$, the objective of the round's subproblem, so $\nabla g(\theta) = B^\top \nabla V(f_t + B\theta)$. Adding and subtracting $\nabla g(\theta_t)$,

$$n\widehat{\mathcal{E}}(f_t) = -\nabla g(\theta_t) + B^\top \big(\nabla V(f_t + h_t) - \nabla V(f_t)\big).$$

The Hessian of g is $B^\top \nabla^2 V(\cdot) B$, so g is $L\|B\|_2^2$-smooth, and by (A3), $g(\theta_t) \le \min_\theta g(\theta) + \epsilon_t$. For a smooth function bounded below, suboptimality in value controls the gradient norm, so

$$\|\nabla g(\theta_t)\|_2^2 \le 2L\|B\|_2^2 \big(g(\theta_t) - \min_\theta g(\theta)\big) \le 2L\|B\|_2^2\, \epsilon_t.$$

By L-smoothness of V and sub-multiplicativity of the spectral norm,

$$\big\|B^\top \big(\nabla V(f_t + h_t) - \nabla V(f_t)\big)\big\|_2 \le \|B\|_2\, L\, \|h_t\|_2 = \frac{L}{\eta}\|B\|_2\, \|f_{t+1} - f_t\|_2.$$

Combining the two bounds by the triangle inequality and dividing by n proves the claim, for every $\eta \in (0, 1]$. For the squared loss, $\phi = \mathrm{id}$ and $\nabla^2 V = I$, so $L = 1$. □

A.3 Another Bound on the Multicalibration Error

Overview. If a group violates multicalibration, adding one more weak learner reduces the loss by an amount controlled by the violation (Lemma A.2, via the descent lemma). With an exact oracle, that achievable decrease is in turn bounded by the squared gap between successive predictions (Proposition A.3, via convexity and the first-order optimality condition of the round).

In this subsection, we bound a group-normalized empirical multicalibration error in terms of the gap between successive predictions. The bound holds for factorized weak learners $b(x, u) = h(x)g(u)$ with group indicators $h : \mathbb{R}^d \to \{0, 1\}$ and interval indicators $g : \mathbb{R} \to \{0, 1\}$, and for losses with a canonical gradient (Section 3), $\partial L(y, f)/\partial f = \phi(f) - y$. It generalizes the bound in Tax et al. (2026), which is for the log-loss.

For any given function $\psi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, let

$$\mathbb{E}_n[\psi(X, Y)] := \frac{1}{n}\sum_{i=1}^n \psi(x_i, y_i), \qquad \mathbb{E}_n[\psi(X, Y) \mid h(X) = 1] := \frac{\mathbb{E}_n[h(X)\psi(X, Y)]}{\mathbb{E}_n[h(X)]}.$$

Assumptions on the Loss. We assume that $L(y, f)$ is L-smooth in its second argument, $|L'(y, a) - L'(y, b)| \le L|a - b|$ for all y, and has a canonical gradient (Section 3),

$$L'(y, f) := \frac{d}{df} L(y, f) = \phi(f) - y.$$

These assumptions are satisfied by the squared loss ($\phi = \mathrm{id}$, $L = 1$) and by the log-loss (ϕ the sigmoid, $L = 1/4$). Proposition A.3 additionally uses convexity of $L(y, \cdot)$, which holds for both.

Group-Normalized Multicalibration. For $h \in \mathcal{H}$, $g \in \mathcal{G}$ and a predictor f, define

$$\Delta_{h,g}(f) := \Big|\mathbb{E}_n\big[h(X)\, g(f(X))\, (Y - \phi(f(X)))\big]\Big|, \qquad \tau_h(f)^2 := \mathbb{E}_n\big[h(X)\, v(X)\big],$$

where $v(x) \ge 0$ is a variance proxy, $v(x) = \phi(f(x))(1 - \phi(f(x)))$ for binary classification and $v(x) = \sigma^2(x)$ for regression. We say f is α-multicalibrated with respect to $(\mathcal{H}, \mathcal{G})$ if $\Delta_{h,g}(f) \le \alpha\, \tau_h(f)$ for all $h \in \mathcal{H}$, $g \in \mathcal{G}$. Let

$$C(f)^2 := \max_{h \in \mathcal{H},\, g \in \mathcal{G}} \frac{\mathbb{E}_n[g(f(X)) \mid h(X) = 1]}{\mathbb{E}_n[v(X) \mid h(X) = 1]},$$

where the maximum is over pairs with $\mathbb{E}_n[h(X)g(f(X))] > 0$, and let

$$D(f) := \mathbb{E}_n[L(Y, f(X))] - \inf_{\theta \in \mathbb{R}^p} \mathbb{E}_n\Big[L\Big(Y, f(X) + \sum_{j=1}^p \theta_j b_j(X, f(X))\Big)\Big]$$

be the largest loss decrease achievable by one exact boosting step from f.

Lemma A.2. *Assume the loss is L-smooth with a canonical gradient, and that $b = h \otimes g \in \mathcal{B}$ for all $h \in \mathcal{H}$, $g \in \mathcal{G}$. Then, for every predictor f, $h \in \mathcal{H}$, and $g \in \mathcal{G}$,*

$$\Delta_{h,g}(f)^2 \le 2L\,\mathbb{E}_n[h(X)g(f(X))]\,D(f), \qquad \text{and hence} \qquad \Delta_{h,g}(f) \le \sqrt{2L\,D(f)}\;C(f)\,\tau_h(f).$$

In particular, f is α-multicalibrated with $\alpha = \sqrt{2L}\,C(f)\sqrt{D(f)}$.

Proof. Fix h, g and let $s := \mathbb{E}_n[h(X)g(f(X))]$ and $c := \mathbb{E}_n[h(X)g(f(X))(Y - \phi(f(X)))]$, so $\Delta_{h,g}(f) = |c|$. If $s = 0$, then $c = 0$ and there is nothing to prove; otherwise let $\lambda := c/(Ls)$ and $\tilde{h}(x) := \lambda\, h(x)g(f(x))$, which is a multiple of the weak learner $b(x, f(x))$. By L-smoothness (the descent lemma) and the canonical gradient, for every (x, y),

$$L(y, f(x) + \tilde{h}(x)) \le L(y, f(x)) - (y - \phi(f(x)))\,\tilde{h}(x) + \frac{L}{2}\tilde{h}(x)^2.$$

Averaging, and using that h and g take values in $\{0, 1\}$ so that $\mathbb{E}_n[\tilde{h}(X)^2] = \lambda^2 s$,

$$D(f) \ge \mathbb{E}_n[L(Y, f(X))] - \mathbb{E}_n[L(Y, f(X) + \tilde{h}(X))] \ge \lambda c - \frac{L}{2}\lambda^2 s = \frac{c^2}{2Ls}.$$

This is the first claim. For the second, $s/\tau_h(f)^2 = \mathbb{E}_n[g(f(X)) \mid h(X) = 1]/\mathbb{E}_n[v(X) \mid h(X) = 1] \le C(f)^2$, so $c^2 \le 2L\,D(f)\,C(f)^2\,\tau_h(f)^2$. □

Lemma A.2 requires no condition such as $L < 1$; in particular it covers the squared loss ($L = 1$).

Proposition A.3. *Assume the conditions of Lemma A.2: the loss is L-smooth with a canonical gradient, and $b = h \otimes g \in \mathcal{B}$ for all $h \in \mathcal{H}$, $g \in \mathcal{G}$. Assume, in addition, that (i) the loss is convex in its second argument, (ii) the rescaling weight at round T is $w_T = 1$, and (iii) the oracle is exact at round T, i.e., θ_T minimizes $\theta \mapsto \mathbb{E}_n[L(Y, f_T(X) + \sum_j \theta_j b_j(X, f_T(X)))]$. Then f_T is α-multicalibrated with*

$$\alpha \le \sqrt{2}\,\frac{L}{\eta}\,C(f_T)\,\sqrt{\mathbb{E}_n\big[(f_{T+1}(X) - f_T(X))^2\big]}.$$

Proof. Let $h_T(x) := \sum_j \theta_{T,j} b_j(x, f_T(x))$, so $f_{T+1} - f_T = \eta h_T$ and, by exactness, $D(f_T) = \mathbb{E}_n[L(Y, f_T(X))] - \mathbb{E}_n[L(Y, f_T(X) + h_T(X))]$. By convexity, $L(y, f_T(x)) - L(y, f_T(x) + h_T(x)) \le -L'(y, f_T(x))\,h_T(x)$. The first-order optimality condition of θ_T gives $\mathbb{E}_n[L'(Y, f_T(X) + h_T(X))\,b_j(X, f_T(X))] = 0$ for all j, hence $\mathbb{E}_n[L'(Y, f_T(X) + h_T(X))\,h_T(X)] = 0$. Therefore, by L-smoothness,

$$\begin{aligned} D(f_T) &\le \mathbb{E}_n\big[\big(L'(Y, f_T(X) + h_T(X)) - L'(Y, f_T(X))\big)h_T(X)\big] \\ &\le L\,\mathbb{E}_n[h_T(X)^2] = \frac{L}{\eta^2}\mathbb{E}_n\big[(f_{T+1}(X) - f_T(X))^2\big]. \end{aligned}$$

Lemma A.2 completes the proof. □

Interpretation of the Bound. The bound depends on two quantities: $C(f_T)$, which depends inversely on the average variance proxy within groups, and the squared gap between successive predictors f_{T+1} and f_T, which vanishes as the predictions stabilize. With an inexact oracle, $D(f_T)$ additionally depends on the oracle's suboptimality, which can be reduced by increasing the number of inner boosting iterations. Note also that, since $g \le 1$,

$$C(f)^2 \le \frac{1}{\min_{h \in \mathcal{H}} \mathbb{E}_n[v(X) \mid h(X) = 1]},$$

which is finite if and only if the variance proxy has positive average within every group. For binary classification this reduces to $\mathbb{E}_n[\phi(f(X))(1 - \phi(f(X))) \mid h(X) = 1] > 0$ for every $h \in \mathcal{H}$, i.e., predictions are not degenerate within any group; this holds whenever $f(x) \in [-R, R]$ for some constant $R > 0$, and if $\phi(f)$ takes values in $[\epsilon, 1 - \epsilon]$ then $C(f)^2 \le 1/(\epsilon(1 - \epsilon)) \le 2/\epsilon$. For regression it requires a positive variance proxy within every group.

A.4 Proof of Theorem 4.2

Overview. The proof is a direct calculation. Because each round's boosting update is an orthogonal projection of the residual, the new gap $f_{t+1} - f_t$ can be written as $(1 - \eta)$ times the previous gap

plus a term that involves the change $A(f_t) - A(f_{t-1})$ of the projector. The Lipschitz continuity of A bounds this change by a multiple of the previous gap, which gives the one-step contraction with factor κ.

A Property of the Exact Step. For the squared loss with an exact oracle, the update is $f_{t+1} = f_t + \eta A(f_t)(\mathbf{y} - f_t)$, where $A(f_t)$ is the orthogonal projection onto the column space of $B(f_t)$. Applying $A(f_t)$ to the new residual and using $A(f_t)^2 = A(f_t)$ gives $A(f_t)(\mathbf{y} - f_{t+1}) = (1-\eta)A(f_t)(\mathbf{y} - f_t)$, and multiplying by η, for every $t \geq 0$,

$$\eta A(f_t)(\mathbf{y} - f_{t+1}) = (1 - \eta)(f_{t+1} - f_t). \tag{17}$$

In words, after a full step ($\eta = 1$) the new residual is orthogonal to the column space of $B(f_t)$, and after a shrunk step a fraction $1 - \eta$ of the boosting update remains.

A Recursion for the Gap. Let $t \geq 1$. We add and subtract $\eta A(f_{t-1})(\mathbf{y} - f_t)$ in the update, then use (17) for round $t-1$, and finally the update of round $t-1$ in the form $\mathbf{y} - f_t = (I - \eta A(f_{t-1}))(\mathbf{y} - f_{t-1})$. This gives

$$\begin{aligned} f_{t+1} - f_t &= \eta A(f_t)(\mathbf{y} - f_t) \\ &= \eta(A(f_t) - A(f_{t-1}))(\mathbf{y} - f_t) + \eta A(f_{t-1})(\mathbf{y} - f_t) \\ &= \eta(A(f_t) - A(f_{t-1}))(\mathbf{y} - f_t) + (1 - \eta)(f_t - f_{t-1}) \\ &= \eta(A(f_t) - A(f_{t-1}))(I - \eta A(f_{t-1}))(\mathbf{y} - f_{t-1}) + (1 - \eta)(f_t - f_{t-1}). \end{aligned}$$

The Contraction. We bound the first term using the Lipschitz continuity of A along the trajectory, $\|I - \eta A(f_{t-1})\|_2 \leq 1$, and $\|\mathbf{y} - f_{t-1}\|_2 \leq \|\mathbf{y} - f_0\|_2$, which holds because the residual norm is non-increasing for the squared loss with an exact oracle (Theorem 4.1). This gives

$$\|f_{t+1} - f_t\|_2 \leq (1 - \eta + \eta L_A \|\mathbf{y} - f_0\|_2)\|f_t - f_{t-1}\|_2 = \kappa \|f_t - f_{t-1}\|_2,$$

which is the claimed bound. If $\kappa < 1$, iterating gives $\|f_{t+1} - f_t\|_2 \leq \kappa^t \|f_1 - f_0\|_2$, so the gap converges to zero linearly. □

Hard-Split Trees. In Section 6 we observe geometric decay of the prediction gap with hard-split GBDTs, although they do not satisfy the Lipschitz condition of Theorem 4.2. We interpret this as evidence that large ensembles average over many independent split decisions, so the column space of $B(f_t)$ varies smoothly with f_t along the trajectory even though individual trees do not, approximately recovering the smoothness regime required by Theorem 4.2 without literally satisfying its pointwise condition.

A.5 Proof of Lemma 4.3

Overview. Lemma 4.3 follows from a perturbation bound for the projector $A(f) = B(f)B(f)^+$ in terms of the matrix $B(f)$ (Lemma A.4 below), combined with the Lipschitz continuity of $B(f)$. To prove Lemma A.4, we write $A(u) - A(v)$ as a sum of three terms. In two of them the difference $B(u) - B(v)$ appears directly; the third involves the difference of the pseudoinverses of the Gram matrices, which reduces to the difference of the pseudoinverses $B(u)^+ - B(v)^+$ and is controlled by a classical perturbation bound for pseudoinverses. The singular-value bounds enter through $\|B(f)^+\|_2 \leq 1/\delta$ and $\|B(f)\|_2 \leq M$.

Interpretation of the Conditions. The conditions of Lemma 4.3 have a simple interpretation if $B(f_t)$ is viewed as a design matrix with n data points and p features. The condition that $B(f_t)$ has full column rank with smallest singular value at least $\delta > 0$ requires the weak learners to be sufficiently diverse on the training data: no weak learner can be well approximated by a linear combination of the others, uniformly along the trajectory f_t. The condition that the largest singular value of $B(f_t)$ is at most M requires the combined output of the weak learners to remain bounded. Since

$$\sigma_{\max}(B(f_t))^2 \leq \operatorname{tr}\big(B(f_t)^\top B(f_t)\big) = \sum_{j=1}^{p} \|b_j(f_t)\|_2^2,$$

a sufficient condition is that each weak learner is bounded on the training data, $\|b_j(f_t)\|_2 \leq M/\sqrt{p}$ for all j.

Lemma A.4. *Assume that for some $u, v \in \mathbb{R}^n$, $B(u)$ and $B(v)$ have full column rank with smallest singular values at least $\delta > 0$ and for some $M > 0$, $\|B(u)\|_2, \|B(v)\|_2 \le M$. Then,*

$$\|A(u) - A(v)\|_2 \le \frac{2}{\delta}\left(1 + c\,\frac{M^2}{\delta^2}\right)\|B(u) - B(v)\|_2$$

where $c = \sqrt{2}$.

Proof. Write $B_u = B(u)$, $B_v = B(v)$, $A_u = A(u) = B_u B_u^+$ and $A_v = A(v) = B_v B_v^+$. Throughout, we use two facts about a matrix B with full column rank and smallest singular value at least δ. First, $\|B^+\|_2 = 1/\sigma_{\min}(B) \le 1/\delta$. Second, $B^\top B$ is invertible and $B^+ = (B^\top B)^{-1}B^\top$, which gives the identities

$$B^+ = (B^\top B)^+ B^\top, \qquad (B^+)^\top = B(B^\top B)^+, \qquad (B^\top B)^+ = B^+(B^+)^\top.$$

Decomposing the Difference of Projectors. Adding and subtracting $B_v B_u^+$,

$$A_u - A_v = (B_u - B_v)B_u^+ + B_v(B_u^+ - B_v^+).$$

To make the difference $B_u - B_v$ appear also in the second term, we express the pseudoinverses through the Gram matrices. By the first identity, and adding and subtracting $(B_v^\top B_v)^+ B_u^\top$,

$$B_u^+ - B_v^+ = \big((B_u^\top B_u)^+ - (B_v^\top B_v)^+\big)B_u^\top + (B_v^\top B_v)^+(B_u - B_v)^\top.$$

Multiplying from the left by B_v and using the second identity, $B_v(B_v^\top B_v)^+ = (B_v^+)^\top$, we arrive at

$$A_u - A_v = (B_u - B_v)B_u^+ + B_v\big((B_u^\top B_u)^+ - (B_v^\top B_v)^+\big)B_u^\top + (B_v^+)^\top(B_u - B_v)^\top.$$

Bounding the Three Terms. Since $\|B_u^+\|_2, \|B_v^+\|_2 \le 1/\delta$, the first and the last term each have norm at most $\|B_u - B_v\|_2/\delta$. Since $\|B_u\|_2, \|B_v\|_2 \le M$, the middle term has norm at most $M^2\|(B_u^\top B_u)^+ - (B_v^\top B_v)^+\|_2$. Hence

$$\|A_u - A_v\|_2 \le \frac{2}{\delta}\|B_u - B_v\|_2 + M^2\|(B_u^\top B_u)^+ - (B_v^\top B_v)^+\|_2.$$

It remains to bound the difference of the pseudoinverses of the Gram matrices. By the third identity, and adding and subtracting $B_u^+(B_v^+)^\top$,

$$(B_u^\top B_u)^+ - (B_v^\top B_v)^+ = B_u^+(B_u^+ - B_v^+)^\top + (B_u^+ - B_v^+)(B_v^+)^\top,$$

whose norm is at most $(2/\delta)\|B_u^+ - B_v^+\|_2$. Therefore,

$$\|A_u - A_v\|_2 \le \frac{2}{\delta}\big(\|B_u - B_v\|_2 + M^2\|B_u^+ - B_v^+\|_2\big).$$

Bounding the Difference of Pseudoinverses. We use a classical perturbation bound for the Moore–Penrose pseudoinverse of matrices of equal rank, due to Wedin (1973) (see Stewart, 1977, Theorem 3.4): for two matrices of the same size and the same rank, the distance between their pseudoinverses is at most a constant times the distance between the matrices, multiplied by the norms of the two pseudoinverses. For the spectral norm and matrices of full column rank, the constant is $c = \sqrt{2}$ (and $c = 1$ for square matrices). In other words, on matrices of fixed rank whose smallest nonzero singular value is bounded away from zero, the pseudoinverse is a Lipschitz function. Here B_u and B_v both have full column rank p, so

$$\|B_u^+ - B_v^+\|_2 \le c\,\|B_u^+\|_2\,\|B_v^+\|_2\,\|B_u - B_v\|_2 \le \frac{c}{\delta^2}\|B_u - B_v\|_2.$$

Substituting this bound into the previous display gives

$$\|A_u - A_v\|_2 \le \frac{2}{\delta}\left(1 + \frac{cM^2}{\delta^2}\right)\|B_u - B_v\|_2,$$

which proves the lemma. □

Proof of Lemma 4.3 Let $u, v \in \mathcal{F}$. The assumptions of Lemma 4.3 imply those of Lemma A.4, and the Lipschitz continuity of $B(f)$ on $\mathcal{F}$ gives $\|B(u) - B(v)\|_2 \le L_B\|u - v\|_2$. Hence, by Lemma A.4,

$$\|A(u) - A(v)\|_2 \le \frac{2}{\delta}\left(1 + c\,\frac{M^2}{\delta^2}\right)L_B\,\|u - v\|_2,$$

which is the claimed bound on L_A. □

A.6 Proof of Lemma 4.4

Overview. For factorized weak learners, every row of $B(f)$ is a Kronecker product $h(x_i) \otimes g(f_i)$. This allows us to write $B(f)$ as the product of a block-diagonal matrix built from the vectors $g(f_i)$ and a matrix that does not depend on f. Reordering the columns of $B(f)$ gives an analogous factorization with the roles of h and g exchanged, in which the dependence on f is isolated in the single matrix $G(f) = (g(f_1), \dots, g(f_n))^\top$. The Lipschitz constant of $B(f)$ is then the Lipschitz constant of $G(f)$, which follows from that of the functions g, multiplied by the size of the features $h(x_i)$. The same factorizations also bound $\|B(f)\|_2$, which provides the constant M in Lemma 4.3.

Setup and Notation. Recall from Section 3 that for factorized weak learners $b(x, u) = h(x)g(u)$, with $h(x) = (h_1(x), \dots, h_m(x))^\top$ and $g(u) = (g_1(u), \dots, g_k(u))^\top$, the i-th row of $B(f)$ is $\big(h(x_i) \otimes g(f_i)\big)^\top$ and $p = mk$; see (2) and (3). Let

$$H(X) = (h(x_1), \dots, h(x_n))^\top \in \mathbb{R}^{n\times m} \quad \text{and} \quad G(f) = (g(f_1), \dots, g(f_n))^\top \in \mathbb{R}^{n\times k},$$

and, for a matrix with rows $z_1^\top, \dots, z_n^\top$, let $\|\cdot\|_{2,\infty}$ denote the largest Euclidean norm of a row; for instance, $\|H(X)\|_{2,\infty} = \max_i \|h(x_i)\|_2$. We also use the block-diagonal matrices

$$D_g(f) = \text{block-diag}\big(g(f_1)^\top, \dots, g(f_n)^\top\big) \in \mathbb{R}^{n\times nk},$$
$$D_h(X) = \text{block-diag}\big(h(x_1)^\top, \dots, h(x_n)^\top\big) \in \mathbb{R}^{n\times nm}.$$

Since $D_g(f)D_g(f)^\top = \text{diag}(\|g(f_1)\|_2^2, \dots, \|g(f_n)\|_2^2)$, the spectral norm of such a matrix is the largest norm of its blocks: $\|D_g(f)\|_2 = \|G(f)\|_{2,\infty}$, and likewise $\|D_h(X)\|_2 = \|H(X)\|_{2,\infty}$.

Two Factorizations of $B(f)$. By the mixed-product property of the Kronecker product, $g(f_i)^\top\big(h(x_i)^\top \otimes I_k\big) = h(x_i)^\top \otimes g(f_i)^\top$, which is the i-th row of $B(f)$. Hence

$$B(f) = D_g(f)\,\big(H(X) \otimes I_k\big).$$

In the same way, $h(x_i)^\top\big(g(f_i)^\top \otimes I_m\big) = g(f_i)^\top \otimes h(x_i)^\top$, which is the i-th row of $B(f)$ with its entries reordered. Hence the matrix

$$R(f) := D_h(X)\,\big(G(f) \otimes I_m\big)$$

satisfies $R(f) = B(f)P$ for a permutation matrix P that depends only on m and k, not on f. Since multiplication by a permutation matrix does not change the spectral norm, $\|R(f)\|_2 = \|B(f)\|_2$ and $\|R(f) - R(f')\|_2 = \|B(f) - B(f')\|_2$ for all $f, f' \in \mathbb{R}^n$.

Bounds on the Norm of $B(f)$. Using $\|H(X) \otimes I_k\|_2 = \|H(X)\|_2$ and $\|G(f) \otimes I_m\|_2 = \|G(f)\|_2$, the two factorizations give

$$\|B(f)\|_2 \le \|G(f)\|_{2,\infty}\|H(X)\|_2 \qquad \text{and} \qquad \|B(f)\|_2 \le \|H(X)\|_{2,\infty}\|G(f)\|_2.$$

Either bound can serve as the constant M in Lemma 4.3.

Lipschitz Continuity of $B(f)$. In the second factorization, only $G(f)$ depends on f, so

$$R(f) - R(f') = D_h(X)\big((G(f) - G(f')) \otimes I_m\big),$$

and therefore

$$\|B(f) - B(f')\|_2 = \|R(f) - R(f')\|_2 \le \|H(X)\|_{2,\infty}\|G(f) - G(f')\|_2.$$

It remains to bound the change in $G(f)$. If every g_j is L_G-Lipschitz, then, bounding the spectral norm by the Frobenius norm,

$$\|G(f) - G(f')\|_2 \le \|G(f) - G(f')\|_F = \sqrt{\sum_{j=1}^{k}\sum_{i=1}^{n}\big(g_j(f_i) - g_j(f_i')\big)^2}$$
$$\le \sqrt{\sum_{j=1}^{k}\sum_{i=1}^{n} L_G^2 (f_i - f_i')^2} = L_G\sqrt{k}\,\|f - f'\|_2.$$

Combining the last two displays,

$$\|B(f) - B(f')\|_2 \le \Big(\max_i \|h(x_i)\|_2\Big)\sqrt{k}\,L_G\,\|f - f'\|_2,$$

that is, $L_B \le (\max_i \|h(x_i)\|_2)\sqrt{k}\,L_G$, as claimed. Under the conditions of Lemma 4.3, $A(f)$ is then Lipschitz as well, with the constant L_A given there. □

A.7 Proof of Theorem 4.5

Overview. The proof follows the same pattern as for unit weights, with the rescaling treated as a perturbation. We track the training loss $V(f) = \frac{1}{2}\|\mathbf{y} - f\|_2^2$ along the iterates. As for unit weights, each round decreases it by a multiple of $\|A(f_t)(\mathbf{y} - f_t)\|_2^2$, but now up to a perturbation ξ_t caused by the rescaling. Because the residuals stay bounded, $|\xi_t|$ is at most a constant times $1 - w_t$, so the perturbations are summable under $\sum_t (1 - w_t) < \infty$. This gives $A(f_t)(\mathbf{y} - f_t) \to 0$, hence convergence of the gap and, by a Cauchy-sequence argument, convergence of the loss. Asymptotic multicalibration then follows from the first-order optimality condition of each round's subproblem. For the rate, we modify V by subtracting the accumulated effect of the rescaling, which yields a function that is non-increasing along the iterates; summing its decreases and applying the Cauchy–Schwarz inequality bounds $\sum_t \|f_{t+1} - f_t\|_2^2$.

Discussion. The theorem shows that relaxing the weights for regularization does not break the convergence results, provided the weights $w_t \in (0, 1]$ converge to 1 fast enough that $\sum_t (1-w_t) < \infty$. Without this condition the conclusions can fail: for instance, a constant weight $w_t = w < 1$ shrinks the predictor towards zero in every round, so the iterates need not approach a multicalibrated predictor.

The non-increasing loss property of Theorem 4.1 need not hold in the relaxed case. A weaker property does hold: the proof below shows that $\|\mathbf{y} - f_t\|_2 \le \rho_t \|\mathbf{y} - f_0\|_2 + (1 - \rho_t)\|\mathbf{y}\|_2$, where $\rho_t := \prod_{s=0}^{t-1} w_s$. For unit weights, $\rho_t \equiv 1$, and the bound says that the residual never exceeds its initial value. Under $\sum_t (1 - w_t) < \infty$, the products ρ_t decrease to a positive limit $1 - \epsilon$ with $\epsilon \in [0, 1)$, so that $\|\mathbf{y} - f_t\|_2 \le \|\mathbf{y} - f_0\|_2 + \epsilon \max\{0, \|\mathbf{y}\|_2 - \|\mathbf{y} - f_0\|_2\}$ for all t: the training loss can exceed its initial value only by an amount of order ϵ.

In the convergence rate bound, γ is increasing in $C_w := \sum_t (1 - w_t)$ and in $\tilde{C}_w := \sum_t (1 - w_t)^2$, and $\gamma = 0$ for unit weights. In that case, the bound coincides with the exact-oracle bound of Theorem 4.1 for the squared loss (Appendix A.1), $\sqrt{\eta}\|\mathbf{y} - f_0\|_2/\sqrt{T}$.

Setting. Throughout, the loss is the squared loss, the oracle is exact, and the update (5) reads $f_{t+1} = w_t\big(f_t + \eta A(f_t)(\mathbf{y} - f_t)\big)$ with $\eta \in (0, 1]$. We repeatedly use that $A(f_t)$ is an orthogonal projection, so that $\|I - \eta A(f_t)\|_2 \le 1$ and $(\mathbf{y} - f_t)^\top A(f_t)(\mathbf{y} - f_t) = \|A(f_t)(\mathbf{y} - f_t)\|_2^2$.

Residual Recursion and the Training Loss. Subtracting (5) from $\mathbf{y}$ gives the recursion for the residual,

$$\mathbf{y} - f_{t+1} = (1 - w_t)\mathbf{y} + w_t(I - \eta A(f_t))(\mathbf{y} - f_t). \tag{18}$$

Expanding the square of (18), and using $\|(I - \eta A(f_t))(\mathbf{y} - f_t)\|_2^2 = \|\mathbf{y} - f_t\|_2^2 - \eta(2 - \eta)\|A(f_t)(\mathbf{y} - f_t)\|_2^2$, which holds because $A(f_t)$ is an orthogonal projection, we obtain the change of the training loss in one round,

$$V(f_{t+1}) - V(f_t) = -w_t^2 \eta \left(1 - \frac{1}{2}\eta\right) \|A(f_t)(\mathbf{y} - f_t)\|_2^2 + \xi_t, \tag{19}$$

where the perturbation due to the rescaling is

$$\xi_t = -\frac{1}{2}(1 - w_t^2)\|\mathbf{y} - f_t\|_2^2 + \frac{1}{2}(1 - w_t)^2\|\mathbf{y}\|_2^2 + w_t(1 - w_t)\,\mathbf{y}^\top(I - \eta A(f_t))(\mathbf{y} - f_t).$$

For unit weights, $\xi_t = 0$, and (19) shows that the loss decreases in every round.

Bounded Residuals and Summable Perturbations. Taking norms in (18) and using $\|I - \eta A(f_t)\|_2 \le 1$,

$$\|\mathbf{y} - f_{t+1}\|_2 \le (1 - w_t)\|\mathbf{y}\|_2 + w_t\|\mathbf{y} - f_t\|_2.$$

Since $w_t \in (0, 1]$, the right-hand side is a convex combination, and by induction $\|\mathbf{y} - f_t\|_2 \le \rho_t\|\mathbf{y} - f_0\|_2 + (1 - \rho_t)\|\mathbf{y}\|_2$ with $\rho_t = \prod_{s<t} w_s$. In particular, for all $t \ge 0$,

$$\|\mathbf{y} - f_t\|_2 \le R := \max\{\|\mathbf{y} - f_0\|_2, \|\mathbf{y}\|_2\}, \qquad \|f_t\|_2 \le R_f := \|\mathbf{y}\|_2 + R.$$

Each of the three terms of ξ_t is therefore at most a constant times $1 - w_t$: using $1 - w_t^2 \le 2(1 - w_t)$, $(1 - w_t)^2 \le 1 - w_t$, and, by the Cauchy–Schwarz inequality, $|\mathbf{y}^\top(I - \eta A(f_t))(\mathbf{y} - f_t)| \le \|\mathbf{y}\|_2 R$, we get

$$|\xi_t| \le C_\xi(1 - w_t), \qquad C_\xi := R^2 + \tfrac{1}{2}\|\mathbf{y}\|_2^2 + \|\mathbf{y}\|_2 R.$$

Hence $\sum_{t=0}^{\infty} |\xi_t| \le C_\xi C_w < \infty$.

Convergence of the Gap and of the Loss. Summing (19) over $t = 0, \dots, T-1$ and using $V(f_T) \geq 0$,

$$\eta \left(1 - \frac{1}{2}\eta\right) \sum_{t=0}^{T-1} w_t^2 \|A(f_t)(\mathbf{y} - f_t)\|_2^2 \leq V(f_0) + \sum_{t=0}^{T-1} |\xi_t|.$$

The right-hand side is bounded uniformly in T, so the series $\sum_t w_t^2 \|A(f_t)(\mathbf{y} - f_t)\|_2^2$ converges. Since $w_t \to 1$, it follows that $A(f_t)(\mathbf{y} - f_t) \to 0$. The difference of successive predictions is

$$f_{t+1} - f_t = w_t \eta A(f_t)(\mathbf{y} - f_t) - (1 - w_t) f_t,$$

so $\|f_{t+1} - f_t\|_2 \leq w_t \eta \|A(f_t)(\mathbf{y} - f_t)\|_2 + (1 - w_t)\|f_t\|_2$. Both terms tend to zero, the first as just shown and the second because $\|f_t\|_2 \leq R_f$ and $w_t \to 1$. Hence $f_{t+1} - f_t \to 0$.

To show that the loss converges, let $s < t$. Summing (19) from s to $t-1$,

$$|V(f_t) - V(f_s)| \leq \eta \left(1 - \frac{1}{2}\eta\right) \sum_{i=s}^{t-1} w_i^2 \|A(f_i)(\mathbf{y} - f_i)\|_2^2 + \sum_{i=s}^{t-1} |\xi_i|.$$

Both series on the right-hand side converge, so their tails tend to zero, and $\{V(f_t)\}$ is a Cauchy sequence. Hence $V(f_t)$ converges to a limit V^*. This proves part (1) of the theorem.

Asymptotic Multicalibration. Since the oracle is exact, the update (5) can be written as

$$f_{t+1} = w_t\big(f_t + \eta B(f_t)\theta(f_t)\big), \tag{20}$$

where $\theta(f_t)$ minimizes the round's subproblem $L_t(\theta) = \frac{1}{2}\|\mathbf{y} - f_t - B(f_t)\theta\|_2^2$. The first-order optimality condition of this subproblem is

$$B(f_t)^\top \big(\mathbf{y} - f_t - B(f_t)\theta(f_t)\big) = 0. \tag{21}$$

By (21), $n\widehat{\mathcal{E}}(f_t) = B(f_t)^\top(\mathbf{y} - f_t) = B(f_t)^\top B(f_t)\theta(f_t)$, and by (20), $B(f_t)\theta(f_t) = \big(f_{t+1} - f_t + (1 - w_t)f_t\big)/(\eta w_t)$. Therefore, for any weights $0 < w_t \leq 1$,

$$\widehat{\mathcal{E}}(f_t) = \frac{1 - w_t}{\eta w_t n} B(f_t)^\top f_t + \frac{1}{\eta w_t n} B(f_t)^\top (f_{t+1} - f_t).$$

If $\|B(f_t)\|_2$ is bounded along the trajectory then, since also $\|f_t\|_2 \leq R_f$, there are constants $C_1, C_2 > 0$ such that

$$\|\widehat{\mathcal{E}}(f_t)\|_2 \leq C_1 \frac{1 - w_t}{w_t} + C_2 \frac{1}{w_t} \|f_{t+1} - f_t\|_2.$$

Since $w_t \to 1$ and $f_{t+1} - f_t \to 0$, both terms tend to zero, so $\widehat{\mathcal{E}}(f_t) \to 0$. This proves part (2).

A Non-Increasing Lyapunov Function. For the rate, we need a quantity that decreases in every round. We first record an identity for the change of the squared residual norm.

Lemma A.5. *The difference of the successive residual norms satisfies, for every $t \geq 0$,*

$$\|\mathbf{y} - f_t\|_2^2 - \|\mathbf{y} - f_{t+1}\|_2^2 = w_t\eta(2 - w_t\eta)\|A(f_t)(\mathbf{y} - f_t)\|_2^2 + (1 - w_t)^2\|f_t\|_2^2 - 2(1 - w_t) f_t^\top (\mathbf{y} - f_{t+1}).$$

Proof. Rearranging (5), $\mathbf{y} - f_{t+1} = (I - w_t\eta A(f_t))(\mathbf{y} - f_t) + (1 - w_t)f_t$. Expanding the square,

$$\|\mathbf{y} - f_{t+1}\|_2^2 = \|(I - w_t\eta A(f_t))(\mathbf{y} - f_t)\|_2^2 + 2(1 - w_t) f_t^\top (I - w_t\eta A(f_t))(\mathbf{y} - f_t) + (1 - w_t)^2\|f_t\|_2^2.$$

Since $A(f_t)$ is an orthogonal projection, the first term equals $\|\mathbf{y} - f_t\|_2^2 - w_t\eta(2 - w_t\eta)\|A(f_t)(\mathbf{y} - f_t)\|_2^2$. In the cross term, we substitute $(I - w_t\eta A(f_t))(\mathbf{y} - f_t) = (\mathbf{y} - f_{t+1}) - (1 - w_t)f_t$, which turns it into $2(1 - w_t) f_t^\top (\mathbf{y} - f_{t+1}) - 2(1 - w_t)^2\|f_t\|_2^2$. Collecting terms proves the identity. □

The last term in Lemma A.5 has no definite sign, which is why the loss need not decrease. We therefore add the accumulated cross terms back and consider

$$U_t := \frac{1}{2}\|\mathbf{y} - f_t\|_2^2 - \sum_{s=0}^{t-1} (1 - w_s) f_s^\top (\mathbf{y} - f_{s+1}) + C,$$

where the constant $C \geq 0$ is chosen below so that $U_t \geq 0$. By Lemma A.5, the cross terms cancel in $U_t - U_{t+1}$, and

$$U_t - U_{t+1} = w_t\eta \left(1 - \frac{1}{2}w_t\eta\right) \|A(f_t)(\mathbf{y} - f_t)\|_2^2 + \frac{1}{2}(1 - w_t)^2\|f_t\|_2^2 \geq 0,$$

so U_t is non-increasing along the iterates.

To choose C, it suffices that every cross term is bounded,

$$|f_t^\top(\mathbf{y} - f_{t+1})| \leq D \quad \text{for all } t \geq 0, \tag{22}$$

since then $\left|\sum_{s=0}^{t-1}(1 - w_s)f_s^\top(\mathbf{y} - f_{s+1})\right| \leq DC_w$, and $C := DC_w$ makes $U_t \geq 0$. To verify (22), we use the Cauchy–Schwarz inequality, $\|f_t\|_2 \leq \|\mathbf{y}\|_2 + \|\mathbf{y} - f_t\|_2$, and the residual bound R,

$$|f_t^\top(\mathbf{y} - f_{t+1})| \leq (\|\mathbf{y}\|_2 + \|\mathbf{y} - f_t\|_2)\|\mathbf{y} - f_{t+1}\|_2 \leq (\|\mathbf{y}\|_2 + R)R. \tag{23}$$

Hence (22) holds with $D := (\|\mathbf{y}\|_2 + R)R = \max\{2\|\mathbf{y}\|_2^2, (\|\mathbf{y}\|_2 + \|\mathbf{y} - f_0\|_2)\|\mathbf{y} - f_0\|_2\}$.

Convergence Rate. We split each step into a boosting part and a rescaling part,

$$f_{t+1} - f_t = a_t + b_t, \qquad a_t := w_t\eta A(f_t)(\mathbf{y} - f_t), \qquad b_t := -(1 - w_t)f_t.$$

The boosting part is controlled by the decrease of U_t. Since $w_t\eta \leq \eta \leq 1$, we have $w_t\eta/(1 - w_t\eta/2) \leq 2\eta$, and therefore

$$\|a_t\|_2^2 = \frac{w_t\eta}{1 - w_t\eta/2} \cdot w_t\eta \left(1 - \frac{1}{2}w_t\eta\right) \|A(f_t)(\mathbf{y} - f_t)\|_2^2 \leq 2\eta(U_t - U_{t+1}).$$

Summing, and using $U_T \geq 0$ and $U_0 = \frac{1}{2}\|\mathbf{y} - f_0\|_2^2 + DC_w$,

$$\sum_{t=0}^{T-1} \|a_t\|_2^2 \leq 2\eta U_0 = \eta\|\mathbf{y} - f_0\|_2^2 + 2\eta DC_w.$$

The rescaling part is controlled by the summability of the weights: $\|b_t\|_2 \leq (1 - w_t)R_f$, so

$$\sum_{t=0}^{T-1} \|b_t\|_2^2 \leq R_f^2 \sum_{t=0}^{\infty}(1 - w_t)^2 = R_f^2\tilde{C}_w, \qquad \tilde{C}_w \leq C_w.$$

Finally, expanding $\|a_t + b_t\|_2^2$ and applying the Cauchy–Schwarz inequality to the sum of the cross terms,

$$\sum_{t=0}^{T-1} \|a_t + b_t\|_2^2 \leq \sum_{t=0}^{T-1} \|a_t\|_2^2 + 2\sqrt{\sum_{t=0}^{T-1} \|a_t\|_2^2}\sqrt{\sum_{t=0}^{T-1} \|b_t\|_2^2} + \sum_{t=0}^{T-1} \|b_t\|_2^2.$$

Combining the three bounds,

$$\sum_{t=0}^{T-1} \|f_{t+1} - f_t\|_2^2 \leq \eta\|\mathbf{y} - f_0\|_2^2 + 2\eta DC_w \\ + 2\sqrt{\eta\|\mathbf{y} - f_0\|_2^2 + 2\eta DC_w}\,\sqrt{R_f^2\tilde{C}_w} + R_f^2\tilde{C}_w =: K.$$

Since the smallest squared gap is at most the average,

$$\min_{0 \leq t \leq T-1} \|f_{t+1} - f_t\|_2 \leq \sqrt{K}\,\frac{1}{\sqrt{T}},$$

and $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ gives $\sqrt{K} \leq \sqrt{\eta}\|\mathbf{y} - f_0\|_2 + \gamma$ with

$$\gamma^2 = 2\eta DC_w + 2\sqrt{\eta\|\mathbf{y} - f_0\|_2^2 + 2\eta DC_w}\,\sqrt{R_f^2\tilde{C}_w} + R_f^2\tilde{C}_w.$$

Since γ is increasing in $\tilde{C}_w$ and $\tilde{C}_w \leq C_w$, γ can be bounded in terms of η, $\mathbf{y}$, f_0 and $C_w = \sum_t(1 - w_t)$ alone, as stated in the theorem. This proves part (3). □

A.8 Proof of Theorem 4.6

Overview. Monotonicity is immediate because $\omega(f_t)$ minimizes the residual over the line spanned by $\varphi(f_t)$. Comparing the adaptive step with the unscaled step $\varphi(f_t)$ gives a per-round decrease of the squared residual by $\eta(2-\eta)\|A(f_t)(\mathbf{y}-f_t)\|_2^2$, which telescopes and yields asymptotic multicalibration with an explicit rate, without any continuity assumption on $A(\cdot)$. Non-negativity of $\omega(f_t)$ for $t \geq 1$ follows from the fact that f_t is then an orthogonal projection of $\mathbf{y}$, together with the inequality between the arithmetic and geometric means (AM–GM), $ab \leq \frac{1}{2}(a^2+b^2)$ for $a, b \geq 0$.

Throughout, write $r_t := \mathbf{y} - f_t$ and $A_t := A(f_t)$. If $\varphi(f_t) = 0$, every ω gives the same point $\omega\varphi(f_t) = 0$; we then set $f_{t+1} = 0$ (and $\omega(f_t) := 0$).

Non-Increasing Loss. By definition, $\varphi(f_t) = f_t + \eta A(f_t)(\mathbf{y} - f_t) = f_t + \eta B(f_t)\theta(f_t)$, where $\theta(f_t)$ is the minimizer of $L_t(\theta) = \frac{1}{2}\|\mathbf{y} - f_t - B(f_t)\theta\|_2^2$ over $\theta \in \mathbb{R}^p$. Hence,

$$\|\mathbf{y} - \varphi(f_t)\|_2 \leq \|\mathbf{y} - f_t\|_2.$$

Since $\omega = 1$ is a feasible choice,

$$\min_{\omega} \|\mathbf{y} - \omega\varphi(f_t)\|_2 \leq \|\mathbf{y} - \varphi(f_t)\|_2,$$

and, by definition of $\omega(f_t)$, $\|\mathbf{y} - \omega(f_t)\varphi(f_t)\|_2 = \min_\omega \|\mathbf{y} - \omega\varphi(f_t)\|_2$. Since $f_{t+1} = \omega(f_t)\varphi(f_t)$, it follows that

$$\|\mathbf{y} - f_{t+1}\|_2 \leq \|\mathbf{y} - f_t\|_2.$$

(If $\varphi(f_t) = 0$, then $f_{t+1} = 0 = \varphi(f_t)$ and the same inequality holds.)

Multicalibration Error: Finite-Time Bound and Convergence. Since $\mathbf{y} - \varphi(f_t) = (I - \eta A_t)r_t$ and A_t is an orthogonal projection,

$$\|\mathbf{y} - \varphi(f_t)\|_2^2 = \|r_t\|_2^2 - 2\eta r_t^\top A_t r_t + \eta^2 \|A_t r_t\|_2^2 = \|r_t\|_2^2 - \eta(2-\eta)\|A_t r_t\|_2^2.$$

Combining with $\|r_{t+1}\|_2 = \min_\omega \|\mathbf{y} - \omega\varphi(f_t)\|_2 \leq \|\mathbf{y} - \varphi(f_t)\|_2$ (which also holds when $\varphi(f_t) = 0$) gives, for every $t \geq 0$,

$$\|r_{t+1}\|_2^2 \leq \|r_t\|_2^2 - \eta(2-\eta)\|A(f_t)r_t\|_2^2. \tag{24}$$

Summing (24) over $t = 0, \ldots, T-1$ and using $\|r_T\|_2 \geq 0$,

$$\sum_{t=0}^{T-1} \|A(f_t)(\mathbf{y} - f_t)\|_2^2 \leq \frac{\|\mathbf{y} - f_0\|_2^2}{\eta(2-\eta)},$$

so $A(f_t)(\mathbf{y} - f_t) \to 0$ and $\min_{0 \leq t \leq T-1} \|A(f_t)(\mathbf{y} - f_t)\|_2 \leq \|\mathbf{y} - f_0\|_2 / \sqrt{\eta(2-\eta)T}$. Finally, $B(f)^\top A(f) = B(f)^\top$ (as $A(f) = B(f)B(f)^+$ is the orthogonal projector onto the column space of $B(f)$), so

$$\|\widehat{\mathcal{E}}(f_t)\|_2 = \frac{1}{n}\|B(f_t)^\top A(f_t)(\mathbf{y} - f_t)\|_2 \leq \frac{\|B(f_t)\|_2}{n}\|A(f_t)(\mathbf{y} - f_t)\|_2 \to 0$$

whenever $\|B(f_t)\|_2$ is bounded along the trajectory (e.g., for bounded weak learners, $\|B(f)\|_2 \leq \sqrt{np}\max_{j,x,u}|b_j(x,u)|$). If $M_B := \sup_t \|B(f_t)\|_2 < \infty$, evaluating this bound at a round $t' \leq T-1$ that attains $\min_{0 \leq t \leq T-1} \|A(f_t)(\mathbf{y} - f_t)\|_2$ gives

$$\min_{0 \leq t \leq T-1} \|\widehat{\mathcal{E}}(f_t)\|_2 \leq \|\widehat{\mathcal{E}}(f_{t'})\|_2 \leq \frac{M_B}{n} \frac{\|\mathbf{y} - f_0\|_2}{\sqrt{\eta(2-\eta)T}}.$$

Non-Negative Rescaling. Let $t \geq 1$. If $\varphi(f_{t-1}) \neq 0$, then $f_t = \omega(f_{t-1})\varphi(f_{t-1})$ is the orthogonal projection of $\mathbf{y}$ onto the span of $\varphi(f_{t-1})$; if $\varphi(f_{t-1}) = 0$, then $f_t = 0$. In both cases $f_t^\top r_t = 0$, hence $\mathbf{y}^\top f_t = \|f_t\|_2^2$ and

$$\mathbf{y}^\top \varphi(f_t) = \mathbf{y}^\top f_t + \eta(f_t + r_t)^\top A_t r_t = \|f_t\|_2^2 + \eta\|A_t r_t\|_2^2 + \eta f_t^\top A_t r_t.$$

By the Cauchy–Schwarz inequality, $|f_t^\top A_t r_t| = |(A_t f_t)^\top A_t r_t| \leq \|f_t\|_2 \|A_t r_t\|_2$, and by the AM–GM inequality, $\|f_t\|_2\|A_t r_t\|_2 \leq \frac{1}{2}(\|f_t\|_2^2 + \|A_t r_t\|_2^2)$. Therefore

$$\mathbf{y}^\top \varphi(f_t) \geq \left(1 - \frac{\eta}{2}\right)\|f_t\|_2^2 + \frac{\eta}{2}\|A_t r_t\|_2^2 \geq 0$$

for any $\eta \in (0, 2]$ (in particular for $\eta \in (0, 1]$, the range used throughout the paper), so $\omega(f_t) = \mathbf{y}^\top \varphi(f_t)/\|\varphi(f_t)\|_2^2 \geq 0$ whenever $\varphi(f_t) \neq 0$. The same inequality shows that $\varphi(f_t) = 0$ with $t \geq 1$ is possible only if $f_t = 0$ and $A_t \mathbf{y} = 0$, i.e., only at an empirically multicalibrated fixed point. For $t = 0$ the claim can fail: with $f_0 = -\mathbf{y}$ and $A(f_0) = 0$, $\omega(f_0) = -1$. □

For completeness we record the spectral structure of the matrix behind this argument. For $t \geq 1$ with $\varphi(f_{t-1}) \neq 0$, write $M(f_{t-1}) = \tilde{\varphi}\tilde{\varphi}^\top$ with $\tilde{\varphi} = \varphi(f_{t-1})/\|\varphi(f_{t-1})\|_2$; then $f_t = M(f_{t-1})\mathbf{y}$ and $\varphi(f_t) = N(f_t)\mathbf{y}$ with $N(f_t) := \eta A_t + (I - \eta A_t)M(f_{t-1})$. The argument above does not need the following lemma.

Lemma A.6. *Let A be an orthogonal projection matrix of rank $r \geq 1$ and $M = vv^\top$ with $\|v\| = 1$, $Av \neq 0$ and $(I - A)v \neq 0$. Then the eigenvalues of*

$$\eta A + (I - \eta A)M$$

are the value 1 with multiplicity 1, $\eta\|(I - A)v\|_2^2$ with multiplicity 1, η with multiplicity $r - 1$, and the value 0 with multiplicity $n - r - 1$ (some of these values may coincide).

Proof. Let $N = \eta A + (I - \eta A)vv^\top$. Then $Nv = \eta Av + (I - \eta A)v = v$. For $w \in \mathrm{range}(A) \cap v^\perp$ (a subspace of dimension $r - 1$, since $Av \neq 0$), $Nw = \eta w$; for $w \in \mathrm{range}(A)^\perp \cap v^\perp$ (dimension $n - r - 1$, since $(I - A)v \neq 0$), $Nw = 0$. These subspaces and $\mathrm{span}(v)$ are independent and together span an N-invariant subspace of dimension $n - 1$, on which N has eigenvalues 1, η ($r - 1$ times), and 0 ($n - r - 1$ times). The remaining eigenvalue equals $\mathrm{tr}(N) - 1 - \eta(r - 1) = \eta r + 1 - \eta\|Av\|_2^2 - 1 - \eta(r - 1) = \eta\|(I - A)v\|_2^2$. □

A.9 A Sufficient Condition for Quadratic Convergence for Adaptive Rescaling

Overview. An exact expression for the residual after an adaptive step (Lemma A.8) shows that the step removes the component of $(I - \eta A(f_t))(\mathbf{y} - f_t)$ along $\mathbf{y}$ but not its orthogonal part. Local quadratic convergence therefore holds when the weak learners capture the residual up to second order, and not in general.

In the following theorem, we give a sufficient condition for local convergence of the training loss to zero at a quadratic rate.

Theorem A.7 (Quadratic Convergence Rate)**.** *Consider the adaptive rescaling variant for the squared-error loss with $\eta \in (0, 1]$ and $\mathbf{y} \neq 0$. Suppose that, for some $\kappa > 0$ and all t, the residual $r_t = \mathbf{y} - f_t$ satisfies the local realizability condition*

$$\left\|P_{\mathbf{y}}^\perp (I - \eta A(f_t))r_t\right\|_2 \leq \kappa\|r_t\|_2^2, \qquad P_{\mathbf{y}}^\perp := I - \mathbf{y}\mathbf{y}^\top/\|\mathbf{y}\|_2^2,$$

and that $\|\mathbf{y} - f_0\|_2 \leq \|\mathbf{y}\|_2/2$. Then $\|\mathbf{y} - f_{t+1}\|_2 \leq 2\kappa\|\mathbf{y} - f_t\|_2^2$ for all $t \geq 0$, so the residual converges to zero at a quadratic rate if, in addition, $\|\mathbf{y} - f_0\|_2 < 1/(2\kappa)$.

The condition holds, for instance, with $\eta = 1$ if $\|(I - A(f))(\mathbf{y} - f)\|_2 \leq \kappa\|\mathbf{y} - f\|_2^2$ near $\mathbf{y}$, i.e., if the weak learners at f represent the residual up to a second-order error. It is not implied by the other assumptions of the paper: since multicalibration does not require a zero residual, the residual of the adaptive iteration can converge to a nonzero limit, e.g., when the column space of $B(f)$ does not depend on f and $\mathbf{y}$ is not in the span of f_0 and this column space, which contains all iterates. Under the condition, the 1D line search of adaptive rescaling yields Newton-like local convergence.

Proof. We first establish an exact expression for the residual after an adaptive step.

Lemma A.8. *If $\varphi(f_t) \neq 0$, the residuals evolve according to*

$$r_{t+1} = P_{\varphi(f_t)}^\perp \mathbf{y}, \qquad \|r_{t+1}\|_2 = \frac{\|\mathbf{y}\|_2 \left\|P_{\mathbf{y}}^\perp (I - \eta A(f_t))r_t\right\|_2}{\|\varphi(f_t)\|_2}, \tag{25}$$

where $P_v^\perp := I - vv^\top/\|v\|_2^2$.

Proof. By Eq. (7), $f_{t+1} = \omega(f_t)\varphi(f_t)$ is the orthogonal projection of $\mathbf{y}$ onto the span of $\varphi(f_t)$, so $r_{t+1} = P_{\varphi(f_t)}^\perp \mathbf{y}$ and $\|r_{t+1}\|_2^2 = \|\mathbf{y}\|_2^2 - (\mathbf{y}^\top \varphi(f_t))^2/\|\varphi(f_t)\|_2^2$. Write $q := (I - \eta A(f_t))r_t$, so that $\varphi(f_t) = \mathbf{y} - q$. Then

$$\|\mathbf{y}\|_2^2\|\varphi(f_t)\|_2^2 - (\mathbf{y}^\top \varphi(f_t))^2 = \|\mathbf{y}\|_2^2\|q\|_2^2 - (\mathbf{y}^\top q)^2 = \|\mathbf{y}\|_2^2 \, \|P_{\mathbf{y}}^\perp q\|_2^2,$$

and dividing by $\|\varphi(f_t)\|_2^2$ gives the claim. □

By Theorem 4.6, $\|r_t\|_2 \le \|r_0\|_2 \le \|\mathbf{y}\|_2/2$ for all t. Since $\|q\|_2 = \|(I - \eta A(f_t))r_t\|_2 \le \|r_t\|_2$, we have $\|\varphi(f_t)\|_2 = \|\mathbf{y} - q\|_2 \ge \|\mathbf{y}\|_2 - \|r_t\|_2 \ge \|\mathbf{y}\|_2/2 > 0$. Lemma A.8 and the local realizability condition then give

$$\|r_{t+1}\|_2 \le \frac{\|\mathbf{y}\|_2\, \kappa \|r_t\|_2^2}{\|\mathbf{y}\|_2/2} = 2\kappa \|r_t\|_2^2.$$

If $\|r_0\|_2 < 1/(2\kappa)$, then $2\kappa\|r_{t+1}\|_2 \le (2\kappa\|r_t\|_2)^2$, so $2\kappa\|r_t\|_2 \le (2\kappa\|r_0\|_2)^{2^t} \to 0$. □

A.10 A Hybrid Boosting Algorithm

Overview. This subsection presents a variant of Algorithm 1 that adds, in every round, a fixed multiple of the residuals of a given predictor $\hat{y}$ to the boosting update. On the training sample, and for the squared loss with an exact oracle, the variant is an inexact preconditioned gradient descent on the training loss. Its preconditioner $\gamma I + A(f_t)$ is bounded away from zero even though $A(f_t)$ is a projection, and its inexactness is proportional to the error of $\hat{y}$. A general contraction bound for such iterations (Lemma A.9, proved in Appendix A.11) then gives linear convergence of the residual to a neighbourhood of zero whose size is proportional to $\|\mathbf{y} - \hat{y}\|_2$ (Proposition A.10, proved in Appendix A.12). Unlike Theorem 4.2, this requires no smoothness condition on the weak learners.

The Algorithm. We modify the update of Algorithm 1 to

$$\hat{f}_{t+1}(x) = \hat{f}_t(x) + \eta\big[\delta\hat{f}_t(x) + \gamma(\hat{y}(x) - \hat{f}_t(x))\big],$$

where $\hat{y}$ is a given predictor, $\gamma > 0$ is a fixed weight, and $\delta\hat{f}_t$ is the boosting update of round t; on the sample, $\delta\hat{f}_t = A(f_t)(\mathbf{y} - f_t)$ for the squared loss with an exact oracle. As before, f_t denotes the vector of values of $\hat{f}_t$ on the sample, and we write $\hat{y}$ also for the vector of values of $\hat{y}$ on the sample.

Interpretation. The difference from Algorithm 1 is the additional term $\gamma(\hat{y}(x) - \hat{f}_t(x))$. We may regard it as adding an extra predictor with fixed weight γ, whose predictions are the approximate residuals $\hat{y}(x) - \hat{f}_t(x)$; if $\hat{y}$ were a perfect predictor, these would be exactly the residuals $y(x) - \hat{f}_t(x)$. The weak learners can only correct the part of the residual that lies in the column space of $B(f_t)$, which is why $A(f_t)$, a projection, can be badly conditioned as a preconditioner. The extra predictor acts on the whole residual and thereby removes this ill-conditioning. The method can be seen as a hybrid form of boosting, as it mixes weak learners with a strong learner.

Preconditioned Gradient Descent Form. On the sample, with $V(f) = \frac{1}{2}\|\mathbf{y} - f\|_2^2$, the iterates satisfy

$$f_{t+1} = f_t - \eta\big[P(f_t)\nabla V(f_t) + \xi_t\big], \qquad P(f_t) = \gamma I + A(f_t), \quad \xi_t = \gamma(\mathbf{y} - \hat{y}). \tag{26}$$

Indeed, since $\nabla V(f_t) = f_t - \mathbf{y}$, we have $P(f_t)\nabla V(f_t) + \xi_t = -A(f_t)(\mathbf{y} - f_t) - \gamma(\hat{y} - f_t)$, which is minus the bracket in the update above. Thus (26) is a preconditioned gradient descent on V, made inexact by the term ξ_t, which vanishes when $\hat{y} = \mathbf{y}$.

Multicalibration Error. Since $A(f_t)$ is the orthogonal projection onto the column space of $B(f_t)$, we have $B(f_t)^\top A(f_t) = B(f_t)^\top$. Multiplying the update by $B(f_t)^\top$ therefore gives $B(f_t)^\top(f_{t+1} - f_t) = \eta B(f_t)^\top(\mathbf{y} - f_t) + \eta\gamma B(f_t)^\top(\hat{y} - f_t)$, and hence

$$\widehat{\mathcal{E}}(f_t) = \gamma\frac{1}{n}B(f_t)^\top(f_t - \hat{y}) + \frac{1}{\eta n}B(f_t)^\top(f_{t+1} - f_t).$$

As for Algorithm 1, the second term vanishes as the gap between successive predictions vanishes. The first term is a floor, proportional to γ and to the distance between the current predictor and $\hat{y}$, that is the price of the extra predictor.

A Contraction Bound. The following lemma gives a general convergence result for inexact preconditioned gradient descent of the form (26); we prove it in Appendix A.11.

Lemma A.9. *Assume that $V : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function that is L-smooth and μ-strongly convex, $P(f)$ is a symmetric positive definite matrix such that for some constants $0 < m \le M < \infty$, $mI \preceq P(f) \preceq MI$, and $(f - f^*)^\top P(f)\nabla V(f) \ge m(f - f^*)^\top \nabla V(f)$, where f^* is the minimizer of V. Then, for any step size $\eta \in (0, 2m\mu/(L^2M^2))$, the iterates of* (26) *satisfy*

$$\|f^* - f_{t+1}\|_2 \le \kappa\|f^* - f_t\|_2 + \eta\|\xi_t\|_2$$

where

$$\kappa = \sqrt{1 - 2\mu m\eta\left(1 - \frac{L^2M^2}{2\mu m}\eta\right)} < 1.$$

Linear Convergence of the Hybrid Algorithm. For the squared loss, $V(f) = \frac{1}{2}\|\mathbf{y} - f\|_2^2$ is L-smooth and μ-strongly convex with $L = \mu = 1$, and its minimizer is $f^* = \mathbf{y}$. Since $A(f)$ is an orthogonal projection, $P(f) = \gamma I + A(f)$ satisfies $\gamma I \preceq P(f) \preceq (1+\gamma)I$ and $(f - \mathbf{y})^\top P(f)(f - \mathbf{y}) \geq \gamma\|f - \mathbf{y}\|_2^2$. Lemma A.9 therefore applies with $m = \gamma$ and $M = 1 + \gamma$, and with $\|\xi_t\|_2 = \gamma\|\mathbf{y} - \hat{y}\|_2$ it gives the following result (the details are in Appendix A.12).

Proposition A.10. *Assume that the shrinkage parameter satisfies $\eta \in (0, 2\gamma/(1+\gamma)^2)$. Then, for all $t \geq 0$,*

$$\|\mathbf{y} - f_{t+1}\|_2 \leq \kappa\|\mathbf{y} - f_t\|_2 + \eta\gamma\|\mathbf{y} - \hat{y}\|_2$$

where

$$\kappa = \sqrt{1 - 2\gamma\eta\left(1 - \frac{(1+\gamma)^2}{2\gamma}\eta\right)} < 1.$$

Consequences and Choice of the Step Size. Unrolling the recursion of Proposition A.10 and summing the geometric series,

$$\|\mathbf{y} - f_t\|_2 \leq \kappa^t\|\mathbf{y} - f_0\|_2 + \frac{\eta\gamma}{1-\kappa}\|\mathbf{y} - \hat{y}\|_2.$$

The residual thus converges linearly, at a rate κ that does not depend on $A(f_t)$, to a neighbourhood of zero whose size is proportional to the error $\|\mathbf{y} - \hat{y}\|_2$ of the strong learner. If the step size is set to $\eta = c\gamma$ for some constant $0 < c < 1$, the condition of Proposition A.10 becomes $\gamma < \sqrt{2/c} - 1$. For small γ, the contraction factor is then $\kappa \approx 1 - c(1 - c/2)\gamma^2$, and the size of the neighbourhood is governed by $\eta\gamma/(1-\kappa) \approx 2/(2-c)$. A smaller weight γ therefore slows down the convergence, but does not enlarge the neighbourhood.

A.11 Proof of Lemma A.9

Overview. The proof is a standard contraction argument for inexact preconditioned gradient descent. We first bound the error after an *exact* step $f_t - \eta P(f_t)\nabla V(f_t)$, that is, with the inexactness ξ_t removed. Strong convexity of V, together with the lower bound on P, gives a descent term proportional to $\|f_t - f^*\|_2^2$, and smoothness of V, together with the upper bound on P, bounds the quadratic term. The inexactness then enters additively, by the triangle inequality.

Consequences of Smoothness and Strong Convexity. By the μ-strong convexity of V, for every f, f',

$$(f - f')^\top(\nabla V(f) - \nabla V(f')) \geq \mu\|f - f'\|_2^2,$$

and, by the L-smoothness of V,

$$\|\nabla V(f) - \nabla V(f')\|_2 \leq L\|f - f'\|_2.$$

Taking $f' = f^*$ and using $\nabla V(f^*) = 0$, we obtain, for every f,

$$(f - f^*)^\top\nabla V(f) \geq \mu\|f - f^*\|_2^2 \tag{27}$$

and

$$\|\nabla V(f)\|_2 \leq L\|f - f^*\|_2. \tag{28}$$

The Exact Step. Let $\delta_t := f_t - f^* - \eta P(f_t)\nabla V(f_t)$ be the error after an exact step. Expanding the square,

$$\|\delta_t\|_2^2 = \|f_t - f^*\|_2^2 - 2\eta(f_t - f^*)^\top P(f_t)\nabla V(f_t) + \eta^2\|P(f_t)\nabla V(f_t)\|_2^2.$$

We bound the two terms involving P in turn. For the cross term, the assumption $(f_t - f^*)^\top P(f_t)\nabla V(f_t) \geq m(f_t - f^*)^\top\nabla V(f_t)$, combined with (27), gives

$$(f_t - f^*)^\top P(f_t)\nabla V(f_t) \geq m\mu\|f_t - f^*\|_2^2.$$

For the quadratic term, $P(f_t)$ is symmetric with $P(f_t) \preceq MI$, so $\|P(f_t)\|_2 \le M$, which combined with (28) gives

$$\|P(f_t)\nabla V(f_t)\|_2 \le ML\|f_t - f^*\|_2.$$

Substituting both bounds,

$$\|\delta_t\|_2^2 \le (1 - 2\eta m\mu + \eta^2 M^2 L^2)\|f_t - f^*\|_2^2 = \kappa^2 \|f_t - f^*\|_2^2,$$

with $\kappa^2 = 1 - 2\eta m\mu + \eta^2 M^2 L^2 = 1 - 2\mu m\eta\big(1 - L^2 M^2 \eta/(2\mu m)\big)$. The step-size condition $\eta < 2m\mu/(L^2M^2)$ ensures that $\kappa^2 < 1$. Moreover, $\kappa^2 \ge (1 - \eta m\mu)^2 \ge 0$ because $m\mu \le ML$, so κ is well defined.

Adding the Inexactness. By (26), $f_{t+1} - f^* = \delta_t - \eta\xi_t$. By the triangle inequality,

$$\|f_{t+1} - f^*\|_2 \le \|\delta_t\|_2 + \eta\|\xi_t\|_2 \le \kappa\|f_t - f^*\|_2 + \eta\|\xi_t\|_2,$$

which proves the lemma. □

A.12 Proof of Proposition A.10

Overview. Proposition A.10 is the special case of Lemma A.9 for the squared loss and the preconditioner $P(f) = \gamma I + A(f)$; it remains to verify the assumptions of the lemma and to identify its constants.

The training loss $V(f) = \frac{1}{2}\|f - \mathbf{y}\|_2^2$ is L-smooth and μ-strongly convex with $\mu = L = 1$, and its minimizer is $f^* = \mathbf{y}$. Since $A(f)$ is an orthogonal projection, $P(f) = \gamma I + A(f)$ is symmetric with $\gamma I \preceq P(f) \preceq (1 + \gamma)I$, so $m = \gamma$ and $M = 1 + \gamma$. Since $\nabla V(f) = f - f^*$, the remaining assumption of Lemma A.9 reads

$$(f - f^*)^\top P(f)(f - f^*) \ge m\|f - f^*\|_2^2,$$

which holds because $mI \preceq P(f)$. With these constants, the step-size condition $\eta < 2m\mu/(L^2M^2)$ of Lemma A.9 becomes $\eta < 2\gamma/(1 + \gamma)^2$, its contraction factor κ becomes the one stated in Proposition A.10, and $\|\xi_t\|_2 = \gamma\|\mathbf{y} - \hat{y}\|_2$ gives the additive term $\eta\gamma\|\mathbf{y} - \hat{y}\|_2$. □

A.13 Existence of the Exact Oracle and Assumption (A2) for the Log-Loss

Overview. For the log-loss, a minimizer of the round's subproblem need not exist when the weak learners separate the labels. Lemma A.11 below gives the standard non-separation condition under which it exists and is unique; (A2) then holds for any fixed number of rounds.

Smoothness and Strong Convexity of the Log-Loss. For binary classification with the log-loss $L(y, f) = -y\log(\phi(f)) - (1-y)\log(1-\phi(f))$, where $\phi(f) := 1/(1+e^{-f})$ and labels take values in $\{0, 1\}$, the training loss $V(f)$ is globally smooth with $L = 1/4$, but its strong convexity vanishes as $\|f\|_\infty \to \infty$. Indeed,

$$\nabla^2 V(f) = \operatorname{diag}\big(\phi(f(x_1))(1 - \phi(f(x_1))), \ldots, \phi(f(x_n))(1 - \phi(f(x_n)))\big).$$

Hence, the smallest eigenvalue of the Hessian is $\lambda_{\min}(\nabla^2 V(f)) = \phi(\|f\|_\infty)(1 - \phi(\|f\|_\infty))$ which goes to 0 as $\|f\|_\infty \to \infty$. Assumption (A2) therefore holds only if $\max_{0\le\tau\le1} \|f_t + \tau h_t\|_\infty$ is bounded for all $t = 0, 1, \ldots, T-1$; Lemma A.11 gives a condition under which this is the case.

Lemma A.11. *Consider the log-loss and a round t, and write $B = B(f_t)$, $s_i = 2y_i - 1 \in \{-1, +1\}$. Suppose there is no $v \in \mathbb{R}^p \setminus \{0\}$ with $s_i(Bv)_i \ge 0$ for all i. Then B has full column rank, $\theta \mapsto V(f_t + B\theta)$ is strictly convex and coercive, and the exact oracle's minimizer θ_t exists and is unique. If this holds for every $t < T$, then (A2) holds with*

$$m = \min_{0\le t<T} \min_{0\le\tau\le1} \min_i \phi'\big(f_t(x_i) + \tau h_t(x_i)\big) > 0, \qquad \phi' = \phi(1 - \phi).$$

Conversely, if some $v \ne 0$ satisfies $s_i(Bv)_i > 0$ for all i, the infimum of $V(f_t + B\theta)$ is not attained.

Proof. If $Bv = 0$ for some $v \ne 0$, then $s_i(Bv)_i = 0$ for all i, contradicting the assumption, so B has full column rank. The Hessian in θ is $B^\top W B$ with $W = \operatorname{diag}(\phi'(\cdot)) \succ 0$, hence positive definite, so the function is strictly convex. For coercivity, note $L(1, z) = \log(1 + e^{-z}) \ge \max(0, -z)$ and $L(0, z) = \log(1 + e^z) \ge \max(0, z)$, i.e., $L(y_i, z) \ge \max(0, -s_i z)$. The function $c(v) :=$

$\max_i(-s_i(Bv)_i)$ is continuous and, by assumption, positive on the unit sphere, so $c(v) \geq c_0 > 0$ there. Hence, for $\theta \neq 0$,

$$V(f_t + B\theta) \geq \max_i \big(- s_i f_t(x_i) - s_i(B\theta)_i \big) \geq c_0\|\theta\|_2 - \max_i |f_t(x_i)| \to \infty \quad (\|\theta\|_2 \to \infty),$$

so a minimizer exists, and it is unique by strict convexity. Then $h_t = B\theta_t$ is finite, the segment $\{f_t + \tau h_t : 0 \leq \tau \leq 1\}$ is bounded, and $\nabla^2 V = \operatorname{diag}(\phi'(\cdot)) \succeq m_t I$ on it with $m_t > 0$; taking the minimum over $t < T$ gives (A2). For the converse, every term $L(y_i, f_t(x_i) + (B\theta)_i + \alpha(Bv)_i)$ is strictly decreasing in α, so $V(f_t + B(\theta + \alpha v)) < V(f_t + B\theta)$ for every θ and $\alpha > 0$, and no minimizer exists. □

B The Boosting Oracle and Practical Gradient Boosting

Within a round, the features $(x, f_t(x))$ are fixed, so the inner loop is a gradient boosting machine over a fixed dictionary of weak learners, and convergence results for such machines bound the oracle error ϵ_t of assumption (A3). By Lu and Mazumder (2020), in expectation over the random selection of weak learners, ϵ_t decays geometrically in the number M_t of inner iterations for smooth, strongly convex losses such as the squared loss, at a rate set by the minimal cosine angle of the weak learners (their Theorem 4.1). For smooth losses with bounded level sets, it decays as $O(1/M_t)$ (their Theorem 4.2; the log-loss needs a small ridge term). If the constants in these bounds are uniform over rounds, then $\sum_t \epsilon_t < \infty$ almost surely, provided $M_t \geq c \log t$ for a large enough constant c in the first case, and $\sum_t 1/M_t < \infty$ (for instance, $M_t = t^2$) in the second. A constant M_t gives the noise floor discussed after Theorem 4.1. Greedy tree fitting only approximates the exact weak-learner selection that these results assume.

C Proofs for Section 5

Overview. The proof of Theorem 5.1 has three parts. (i) *Empirical term:* Lemma A.1 (first-order optimality plus smoothness) and Theorem 4.1. (ii) *Covering numbers:* we bound the ℓ_∞ covering numbers of the class of T-round predictors by chaining one-round covers (Lemma C.2), and of the multicalibration-error class by a "cover of covers" (Lemma C.3); each is a standard covering argument, adapted to the fact that the weak learner at round t is evaluated at the previous prediction. (iii) *Uniform convergence and ℓ_2-duality:* the Maurer–Pontil bound gives a uniform bound over the class, and ℓ_2-duality turns it into a bound on the norm of the error vector. Throughout, ϕ is the sigmoid, so $|y - \phi(f)| \leq 1$ and ϕ is $1/4$-Lipschitz; we write $L_\phi = 1/4$ and $B_r = 1$ to indicate where these constants enter.

C.1 Covering Numbers

Overview. This subsection bounds the covering number of the class of error functions $e_{\theta,f}$ whose population-minus-empirical means appear in the proof of Theorem 5.1. The argument proceeds in three steps, one lemma each. Lemma C.1 covers the class of weak-learner combinations with bounded coefficients; this class is finite-dimensional, and its covering number grows like $(1/\varepsilon)^p$. Lemma C.2 covers the class of predictors that the algorithm can reach in T rounds, by propagating covers of the weak-learner combinations through the rounds; because the prediction is fed back as a feature, an approximation error made in an early round is amplified by the per-round Lipschitz constant Γ in every later round. Lemma C.3 combines the two to cover the class of error functions. The resulting bound enters the uniform convergence bound of Appendix C.2 through the quantity Λ.

Covering Numbers. For a class $\mathcal{G}$ of functions on a domain $\mathcal{Z}$ and a sample $z = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, let $\mathcal{G}(z) = \{(g(z_1), \ldots, g(z_n)) : g \in \mathcal{G}\} \subset \mathbb{R}^n$, and define

$$N_\infty(\varepsilon, \mathcal{G}, n) := \sup_{z \in \mathcal{Z}^n} N\big(\varepsilon, \mathcal{G}(z), \|\cdot\|_\infty\big),$$

where $N(\varepsilon, A, \|\cdot\|_\infty)$ is the smallest number of ε-balls in $\|\cdot\|_\infty$, *centred at points of* A, that cover A (the convention of Maurer and Pontil 2009). All covers constructed below have this property: their centres are value vectors of members of the class.

Weak-Learner Combinations. For $R > 0$ let

$$\mathcal{S}_R := \{(x, u) \mapsto \theta^\top b(x, u) : \theta \in \mathbb{R}^p, \|\theta\|_2 \leq R\}, \qquad \mathcal{H}_1 := \mathcal{S}_1.$$

Under (G2), every $h \in \mathcal{S}_R$ satisfies $|h(x,u)| \le RB_2$ and $|h(x,u) - h(x,u')| \le RL_b|u-u'|$.

Lemma C.1 (Covering the weak-learner combinations)**.** *Under (G2), for every $\varepsilon > 0$ and every n,*

$$N_\infty(\varepsilon, \mathcal{S}_R, n) \le \Big(1 + \frac{2RB_2}{\varepsilon}\Big)^p.$$

Proof. For $\|\theta\|_2, \|\tilde\theta\|_2 \le R$ and any (x,u), $|(\theta - \tilde\theta)^\top b(x,u)| \le B_2\|\theta - \tilde\theta\|_2$, uniformly over points. Hence an (ε/B_2)-cover of the ℓ_2-ball of radius R induces an ε-cover of $\mathcal{S}_R$ on any sample. A volumetric argument bounds the number of (ε/B_2)-balls, centred in the ball, needed to cover it by $(1 + 2RB_2/\varepsilon)^p$ (Vershynin, 2018); the centres correspond to members of $\mathcal{S}_R$. □

Predictors Reachable in T Rounds. Define the classes of predictors reachable in t rounds recursively by

$$\mathcal{F}_0 := \{f_0\}, \qquad \mathcal{F}_{t+1} := \big\{ f + \eta\, h(\cdot, f(\cdot)) : f \in \mathcal{F}_t,\ h \in \mathcal{S}_{B_\theta} \big\} \quad (t \ge 0),$$

where $\mathcal{F}_0$ is a single predictor by (G1). Under (G3) the iterates satisfy $f_t \in \mathcal{F}_t$, and $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ since $0 \in \mathcal{S}_{B_\theta}$.

Lemma C.2 (Covering the T-round predictors)**.** *Under (G1)–(G2), with $\Gamma = 1 + \eta B_\theta L_b$,*

$$N_\infty(\varepsilon, \mathcal{F}_T, n) \le \prod_{s=0}^{T-1} N_\infty\Big(\frac{\varepsilon}{T\eta\Gamma^{T-1-s}}, \mathcal{S}_{B_\theta}, n\Big).$$

Proof. Fix a sample $x_1, \dots, x_n$ and identify functions with their value vectors in $\mathbb{R}^n$. For $h \in \mathcal{S}_{B_\theta}$ let $G_h(f) := f + \eta\,(h(x_i, f_i))_{i \le n}$. Since h is $B_\theta L_b$-Lipschitz in its second argument, $\|G_h(f) - G_h(f')\|_\infty \le \Gamma\|f - f'\|_\infty$.

We construct, by induction on t, sets $C_t \subset \mathbb{R}^n$ with $|C_t| \le \prod_{s<t} N_s$, $N_s := N_\infty(\beta_s, \mathcal{S}_{B_\theta}, n)$, such that every $f_t \in \mathcal{F}_t$ is within $\varepsilon_t := \eta \sum_{s<t} \Gamma^{t-1-s}\beta_s$ of some element of C_t in $\|\cdot\|_\infty$. For $t = 0$, $C_0 = \{f_0\}$ and $\varepsilon_0 = 0$. For the step, for *each* $\tilde f \in C_t$ take a β_t-cover $\hat C(\tilde f)$ of $\mathcal{S}_{B_\theta}$ evaluated at the n points $(x_i, \tilde f_i)$; it exists with at most N_t elements because N_∞ is a supremum over all sample points. Set $C_{t+1} = \{G_{\tilde h}(\tilde f) : \tilde f \in C_t, \tilde h \in \hat C(\tilde f)\}$. Given $f_{t+1} = G_h(f_t)$ with $f_t \in \mathcal{F}_t$, pick $\tilde f \in C_t$ with $\|f_t - \tilde f\|_\infty \le \varepsilon_t$ and $\tilde h \in \hat C(\tilde f)$ within β_t of h at the points $(x_i, \tilde f_i)$. Then

$$\|f_{t+1} - G_{\tilde h}(\tilde f)\|_\infty \le \|G_h(f_t) - G_h(\tilde f)\|_\infty + \|G_h(\tilde f) - G_{\tilde h}(\tilde f)\|_\infty \le \Gamma\varepsilon_t + \eta\beta_t = \varepsilon_{t+1}.$$

By induction, every element of C_t is the value vector of a member of $\mathcal{F}_t$ (the centres $\tilde h$ are members of $\mathcal{S}_{B_\theta}$). Choosing $\beta_s = \varepsilon/(T\eta\Gamma^{T-1-s})$ gives $\varepsilon_T = \sum_{s<T} \varepsilon/T = \varepsilon$. □

The Error Class. The error functions and the class they form are

$$e_{\theta,f}(x,y) := \theta^\top b(x, f(x))\,\big(y - \phi(f(x))\big),$$
$$\mathcal{E}_1 := \big\{ e_{\theta,f} : \theta \in \mathbb{R}^p,\ \|\theta\|_2 \le 1,\ f \in \mathcal{F}_T \big\}.$$

Since $|\theta^\top b(x,u)| \le B_2$ for $\|\theta\|_2 \le 1$ and $|y - \phi(f(x))| \le B_r = 1$, every $e \in \mathcal{E}_1$ satisfies $|e| \le B_2$.

Lemma C.3 (Covering the error class)**.** *Under (G2), for every $\varepsilon > 0$,*

$$N_\infty(\varepsilon, \mathcal{E}_1, n) \le N_\infty\Big(\frac{\varepsilon}{2(B_2L_\phi + B_rL_b)}, \mathcal{F}_T, n\Big)\, N_\infty\Big(\frac{\varepsilon}{2B_r}, \mathcal{H}_1, n\Big).$$

Proof. Fix a sample $z_i = (x_i, y_i)$. Take an α-cover $\mathcal{F}_\alpha$ of $\mathcal{F}_T$ on $x_1, \dots, x_n$, and for each $\tilde f \in \mathcal{F}_\alpha$ a β-cover $\mathcal{C}(\tilde f)$ of $\mathcal{H}_1$ at the points $(x_i, \tilde f(x_i))$ (a "cover of covers"; it has at most $N_\infty(\beta, \mathcal{H}_1, n)$ elements). For (θ, f) choose $\tilde f$ with $\|f - \tilde f\|_\infty \le \alpha$ and then $\tilde h \in \mathcal{C}(\tilde f)$ with $|h_\theta(x_i, \tilde f_i) - \tilde h(x_i, \tilde f_i)| \le \beta$. For each i, adding and subtracting $h_\theta(x_i, f_i)(y_i - \phi(\tilde f_i))$,

$$\begin{aligned} |e_{\theta,f}(z_i) - \tilde h(x_i, \tilde f_i)(y_i - \phi(\tilde f_i))| &\le |h_\theta(x_i, f_i)|\,|\phi(f_i) - \phi(\tilde f_i)| \\ &\quad + |y_i - \phi(\tilde f_i)|\,|h_\theta(x_i, f_i) - \tilde h(x_i, \tilde f_i)| \\ &\le B_2L_\phi\alpha + B_r(L_b\alpha + \beta), \end{aligned}$$

using that h_θ is L_b-Lipschitz in its second argument. This is at most ε for $\alpha = \varepsilon/(2(B_2L_\phi + B_rL_b))$ and $\beta = \varepsilon/(2B_r)$, and the cover, whose elements are value vectors of members $e_{\tilde\theta,\tilde f}$ of $\mathcal{E}_1$, has at most $N_\infty(\alpha, \mathcal{F}_T, n)N_\infty(\beta, \mathcal{H}_1, n)$ elements. □

C.2 Uniform Convergence

Overview. This subsection turns the covering bounds of Appendix C.1 into a bound on the difference between the population and the empirical mean of the error functions $e_{\theta,f}$ that holds *simultaneously* for all $e_{\theta,f} \in \mathcal{E}_1$. Uniformity is essential here: in the proof of Theorem 5.1, the bound is applied at the pair (θ^*, f_{t^*}), which depends on the training sample, and a bound for a fixed pair would not suffice. We use the empirical Bernstein bound of Maurer and Pontil (2009) (Theorem C.4 below), which holds uniformly over a class of $[0,1]$-valued functions. Its leading term is proportional to the sample standard deviation of the function, rather than to its range, and the complexity of the class enters only through a covering number at scale $1/n$ on a sample of size $2n$. Lemma C.5 then applies this bound to the error class in two steps: it rescales the error functions, which take values in $[-B_2, B_2]$, to the interval $[0,1]$, and it bounds the resulting covering number by the lemmas of Appendix C.1, which yields the quantity Λ.

For a vector $Z = (Z_1, \dots, Z_n)$, we denote the sample variance by

$$V_n(Z) := \frac{1}{n(n-1)} \sum_{i<j} (Z_i - Z_j)^2.$$

Theorem C.4 (Theorem 6 of Maurer and Pontil 2009)**.** *Let $\mathcal{F}$ be a class of functions $g : \mathcal{Z} \to [0,1]$, $\delta \in (0,1)$, $n \geq 16$, and $\mathcal{M}(n) = 10\, N_\infty(1/n, \mathcal{F}, 2n)$. With probability at least $1-\delta$, for all $g \in \mathcal{F}$,*

$$\mathbb{E}_D[g] - \mathbb{E}_n[g] \leq \sqrt{\frac{18 V_n(g(Z)) \log(\mathcal{M}(n)/\delta)}{n}} + \frac{15 \log(\mathcal{M}(n)/\delta)}{n-1}.$$

Lemma C.5 (Uniform bound for the error class)**.** *Under (G1)–(G2), for $n \geq 16$ and $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $\|\theta\|_2 \leq 1$ and $f \in \mathcal{F}_T$,*

$$\mathbb{E}_D[e_{\theta,f}] - \mathbb{E}_n[e_{\theta,f}] \leq \sqrt{\frac{18 V_n(e_{\theta,f}(Z)) \Lambda}{n}} + \frac{30 B_2 \Lambda}{n-1}.$$

Proof. Apply Theorem C.4 to $\tilde{\mathcal{E}} = \{(e + B_2)/(2B_2) : e \in \mathcal{E}_1\}$, which is $[0,1]$-valued. The affine map scales $\|\cdot\|_\infty$-distances by $1/(2B_2)$, so $N_\infty(1/n, \tilde{\mathcal{E}}, 2n) = N_\infty(2B_2/n, \mathcal{E}_1, 2n)$, and V_n scales by $1/(2B_2)^2$. Multiplying the resulting inequality by $2B_2$ gives the stated form with $\log(\mathcal{M}(n)/\delta)$ in place of Λ. It remains to show $\log(\mathcal{M}(n)/\delta) \leq \Lambda$. By Lemma C.3 with $\varepsilon = 2B_2/n$, $\beta = \varepsilon/(2B_r) = B_2/n$ and Lemma C.1, $\log N_\infty(\beta, \mathcal{H}_1, 2n) \leq p \log(1+2n)$. With $\alpha = \varepsilon/(2(B_2/4 + L_b)) = B_2/(n(B_2/4 + L_b))$, Lemmas C.2 and C.1 give

$$\log N_\infty(\alpha, \mathcal{F}_T, 2n) \leq p \sum_{s=0}^{T-1} \log\Big(1 + \frac{2B_\theta B_2 T \eta \Gamma^{T-1-s}}{\alpha}\Big) = p \sum_{s=0}^{T-1} \log\big(1 + KnT\Gamma^{T-1-s}\big),$$

and since $1 + a\Gamma^k \leq \Gamma^k(1+a)$ for $\Gamma \geq 1$, the right-hand side is at most $pT\log(1+KnT) + \frac{p}{2}T(T-1)\log\Gamma$. Adding $\log 10$ and $\log(1/\delta)$ gives Λ. □

C.3 Proof of Theorem 5.1

Overview. The proof splits the population multicalibration error into the empirical error, which is bounded by the optimization analysis of Section 4, and the generalization gap, which is bounded by the uniform convergence bound of Lemma C.5 combined with ℓ_2-duality.

By the triangle inequality, $\|\mathcal{E}(f_{t^*})\|_2 \leq \|\widehat{\mathcal{E}}(f_{t^*})\|_2 + \|\mathcal{E}(f_{t^*}) - \widehat{\mathcal{E}}(f_{t^*})\|_2$.

Empirical Term. By Lemma A.1 with $\epsilon_t = 0$, $\|B(f_{t^*})\|_2 \leq \sqrt{n}\bar{M}$, and Theorem 4.1 (exact oracle) in the form $\|f_{t^*+1} - f_{t^*}\|_2 \leq \sqrt{4\eta V(f_0)/(mT)} = 2\sqrt{\eta n \bar{V}(f_0)/(mT)}$,

$$\|\widehat{\mathcal{E}}(f_{t^*})\|_2 \leq \frac{\sqrt{n}\bar{M}}{n} \cdot \frac{L}{\eta} \cdot 2\sqrt{\frac{\eta n \bar{V}(f_0)}{mT}} = 2L\bar{M}\sqrt{\frac{\bar{V}(f_0)}{\eta m}}\frac{1}{\sqrt{T}}.$$

Generalization Term. Note that $\theta^\top \mathcal{E}(f) = \mathbb{E}_D[e_{\theta,f}]$ and $\theta^\top \widehat{\mathcal{E}}(f) = \mathbb{E}_n[e_{\theta,f}]$ by linearity. If $\mathcal{E}(f_{t^*}) = \widehat{\mathcal{E}}(f_{t^*})$ there is nothing to prove; otherwise let $\theta^* = (\mathcal{E}(f_{t^*}) - \widehat{\mathcal{E}}(f_{t^*}))/\|\mathcal{E}(f_{t^*}) - \widehat{\mathcal{E}}(f_{t^*})\|_2$,

so that $\|\mathcal{E}(f_{t^*}) - \widehat{\mathcal{E}}(f_{t^*})\|_2 = \mathbb{E}_D[e_{\theta^*,f_{t^*}}] - \mathbb{E}_n[e_{\theta^*,f_{t^*}}]$ (this is the ℓ_2-duality $\|v\|_2 = \sup_{\|\theta\|_2 \le 1} \theta^\top v$). Although θ^* and f_{t^*} depend on the sample, (G3) gives $f_{t^*} \in \mathcal{F}_T$, and the event of Lemma C.5 is uniform over $\|\theta\|_2 \le 1$ and $f \in \mathcal{F}_T$, so it applies at (θ^*, f_{t^*}). Finally, since $|y - \phi| \le 1$,

$$V_n(e_{\theta^*,f_{t^*}}(Z)) \le \frac{1}{n-1} \sum_{i=1}^{n} e_{\theta^*,f_{t^*}}(z_i)^2 \le \frac{\|B(f_{t^*})\theta^*\|_2^2}{n-1} \le \frac{n\bar{M}^2}{n-1},$$

so $\sqrt{18 V_n \Lambda / n} \le \bar{M}\sqrt{18\Lambda/(n-1)}$. Adding the two terms proves the theorem. □

C.4 Proof of Corollary 5.2

Overview. The corollary follows by choosing T to balance the optimization term of Theorem 5.1, which decreases like $T^{-1/2}$, against the leading part of the generalization term, which grows like $T\sqrt{p \log \Gamma / n}$.

Let $s := p \log \Gamma / n \le 1$ and $T \asymp s^{-1/3}$ (so $T \ge 1$). The optimization term is of order $T^{-1/2} \asymp s^{1/6}$. For $\Gamma > 1$ constant, $\Lambda \lesssim pT^2 \log \Gamma + pT \log(KnT) + p \log n + \log(1/\delta)$, so $\sqrt{\Lambda/n}$ is at most of order $T\sqrt{p \log \Gamma / n} = s^{1/6}$ plus lower-order terms of order $(p/n)^{1/3}$ and $\sqrt{p \log n / n}$ (up to logarithms). The last term is of order $\Lambda/n \lesssim pT^2 \log \Gamma / n = s^{1/3}$ plus lower-order terms. Since $s \le 1$ and $p \le n$, all terms are $\tilde{O}(s^{1/6})$, with constants depending on Γ. □

C.5 Scope and Extensions of Theorem 5.1

Overview. This subsection discusses which assumptions of Theorem 5.1 can be relaxed, and at what cost. An inexact oracle and an initial predictor chosen from a class require only small changes to the proof; the squared loss requires bounded labels and changes the trade-off in T; rescaling, infinite weak-learner classes, and weak learners that violate (G2) are not covered and would require new arguments.

Inexact Oracle. With $\epsilon_t > 0$, Lemma A.1 and Theorem 4.1 add to the optimization term a noise floor of order $(\bar{M}/\sqrt{n})\big(2L\sqrt{\bar{\epsilon}/m} + \sqrt{2L\bar{\epsilon}}\big)$, where $\bar{\epsilon} = \max_t \epsilon_t$; the generalization term is unchanged.

Initial Predictor Chosen from a Class. If f_0 is selected from a class $\mathcal{F}_0$ using the sample, start the induction in the proof of Lemma C.2 with an ε_0-cover of $\mathcal{F}_0$ instead of $C_0 = \{f_0\}$. The recursion becomes $\varepsilon_T = \Gamma^T \varepsilon_0 + \eta \sum_{s<T} \Gamma^{T-1-s} \beta_s$; with $\varepsilon_0 = \varepsilon/(2\Gamma^T)$ and $\beta_s = \varepsilon/(2T\eta\Gamma^{T-1-s})$,

$$N_\infty(\varepsilon, \mathcal{F}_T, n) \le N_\infty\Big(\frac{\varepsilon}{2\Gamma^T}, \mathcal{F}_0, n\Big) \prod_{s=0}^{T-1} N_\infty\Big(\frac{\varepsilon}{2T\eta\Gamma^{T-1-s}}, \mathcal{S}_{B_\theta}, n\Big),$$

so Λ gains the term $\log N_\infty(\alpha/(2\Gamma^T), \mathcal{F}_0, 2n)$ (with α as in the proof of Lemma C.5) and K doubles.

Squared Loss. Lemma A.1 holds with $L = 1$. If $|y| \le B_y$, every $f \in \mathcal{F}_T$ satisfies $\|f\|_\infty \le \|f_0\|_\infty + T\eta B_\theta B_2$, so Lemmas C.1–C.5 apply with $L_\phi = 1$ and $B_r = B_y + \|f_0\|_\infty + T\eta B_\theta B_2$, with $B_2 B_r$ in place of B_2 in the rescaling of Lemma C.5, and with an extra factor B_r^2 in the variance bound of Appendix C.3. Because B_r grows with T, the trade-off of Corollary 5.2 changes; we have not worked out the resulting rate.

Rescaling. The analysis is for unit weights $w_t = 1$; relaxed and adaptive rescaling are not covered.

Infinite Weak-Learner Classes. If the weak learners form a ball of a reproducing kernel Hilbert space whose sup-norm entropy satisfies $\log N_\infty(\varepsilon) = O(\varepsilon^{-a})$ for some $a > 0$, this bound can replace Lemma C.1. In Lemma C.2, however, the per-round radii $\varepsilon/(T\eta\Gamma^{T-1-s})$ then enter polynomially rather than logarithmically, and the resulting bound on $\log N_\infty(\varepsilon, \mathcal{F}_T, n)$ grows like Γ^{aT}, exponentially in T rather than quadratically. Handling such classes requires an argument that does not compound covers across rounds; we leave this open.

Removing the Lipschitz Condition (G2). Weak learners that depend on the prediction only through a fixed partition of its range, i.e., piecewise constant in u (as trees that split the score at fixed thresholds), violate (G2). For such classes, a combinatorial argument that bounds the pseudo-dimension of the T-round class, as in analyses of level-set boosting (Globus-Harris et al., 2023), could replace the covering recursion and appears to give a complexity that grows only linearly in T for the squared loss.

We have not verified this extension in full, in particular for the log-loss, where the error functions are not polynomial in the parameters, and leave it to future work.

D Supplementary Material for the Experiments

This appendix describes the datasets (Appendix D.1) and the computation of the multicalibration error (Appendix D.2), and gives the full discussion of the experiments summarized in Section 6 (Appendices D.3–D.5).

D.1 Information About Datasets

Table 1: Statistics of the datasets including the count of raw features, the total dimensionality (d) after pre-processing, the dataset size (n).

Dataset	# Raw Features	# Features	# Data Points
California Housing	8	8	20640
Diabetes	10	10	442
Adult	13	101	32561
German Credit	10	48	1000
Communities and Crime	122	122	1994

California Housing. This dataset, from the StatLib repository, contains information about California districts and is used to predict the median house value. It includes data on housing characteristics, population, income, and geographic location, making it useful for regression tasks and geospatial analysis.

Diabetes. Sourced from the UCI Machine Learning Repository, the diabetes dataset focuses on predicting disease progression one year after baseline measurements. It includes personal and medical indicators such as age, BMI, blood pressure, and blood serum levels, and is commonly used for regression modeling in healthcare studies.

Adult (Census Income). Also from the UCI Machine Learning Repository, the Adult dataset aims to classify whether an individual earns more than \$50K per year based on census data. It incorporates demographic and employment-related attributes, making it suitable for binary classification and regression tasks and studies on fairness and social analytics. In our study, the prediction target is defined to be the sum `education-num` (a numeric encoding of the categorical education feature) and `hours-per-week` (the number of hours an individual works per week). Both `education-num` and `hours-per-week` are removed from the input features prior to training to avoid trivial leakage; the 13 raw features reported in Table 1 reflect the original 14 predictive attributes minus `education` (the redundant categorical version of `education-num`), with `education-num` and `hours-per-week` subsequently moved to the target.

German Credit. This dataset, available from the UCI Machine Learning Repository, is used to predict creditworthiness, labeling individuals as having good or bad credit risk. It includes financial, personal, and loan-related information, providing a standard benchmark for classification and credit scoring problems. In our study, the prediction target is `CreditAmount`, an attribute representing the amount of credit (loan) requested by the individual. This attribute is removed from the input features prior to training so that no component of the target appears as a covariate.

Communities and Crime. Sourced from the UCI Machine Learning Repository, the Communities and Crime dataset combines socio-economic, law enforcement, and crime rate statistics for US communities. It is typically used to predict violent crime rates based on community-level indicators, making it valuable for regression analysis and policy research. The prediction target is `ViolentCrimesPerPop`, which represents the violent crime rate per capita in a community.

D.2 MCE Computation

The multicalibration error of a predictor f is computed as a sample-mean estimator of $\|\widehat{\mathcal{E}}(f)\|_2$ evaluated over a random sample of N_{tree} regression trees with random node splits of depth h. In

our experiments, $N_{\text{tree}} = 100$ and $h = 3$. This random-tree class differs from the GBDT class used during training and is intentionally broader: evaluating MCE against this larger class stress-tests multicalibration beyond the training hypothesis space, providing a closer proxy to the population multicalibration error in Eq. (1) than evaluating on the training class itself.

D.3 Prediction Gap under Unit Weights

Theorem 4.1 guarantees that the gap $\|f_{t+1} - f_t\|_2$ decays at a rate of at least $O(1/\sqrt{T})$. Furthermore, Theorem 4.2 establishes that under smoothness conditions on the weak learners, this rate improves to linear (geometric decay).

To validate these bounds, Figure 1 shows the evolution of the prediction gap from two perspectives. The top row displays the gap on a linear scale, showing a clear overall decay toward zero across all datasets (with small per-round oscillations of $\leq 10\%$ in some cases, consistent with Theorem 4.1 bounding the *minimum* and *average* gap rather than the per-round gap), which empirically confirms the global convergence guarantees of Theorem 4.1.

To verify the stronger claim of linear convergence, the bottom row shows the same values on a semi-log scale. In this view, a linear trajectory implies geometric decay (i.e., $\|f_{t+1} - f_t\|_2 \propto \kappa^t$ for some $\kappa < 1$). We fit a linear regression to the tail of the log-gap sequence (dashed black lines) and observe near-linear behavior with $R^2 \in [0.78, 0.99]$. The lower R^2 on California Housing and Adult reflects a sublinear plateau, consistent with hard-split GBDTs only approximately satisfying the Lipschitz condition of Theorem 4.2.

D.4 Training MSE and MCE under the Rescaling Strategies

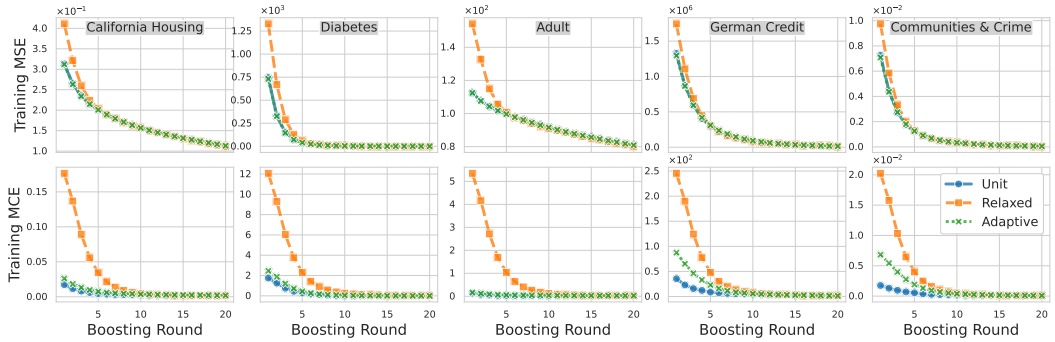

Figure 3: Evolution of the training MSE (top row) and training MCE (bottom row) over 20 boosting rounds across five datasets.

Theorem 4.6 guarantees that the training Mean Squared Error (MSE) is non-increasing under adaptive weights (under relaxed weights, Theorem 4.5 gives only the weaker bound discussed in Section 4.2.2 and Appendix A.7), and Theorems 4.5 and 4.6 guarantee that the empirical multicalibration error converges to zero, with an explicit rate for adaptive weights (Theorem 4.6). Thus, we investigate the reduction of the MSE and the Multicalibration Error (MCE) under the three considered rescaling strategies.

The top row of Figure 3 shows that every strategy yields a monotonic decrease in training MSE across all datasets. This confirms that the dynamic system remains stable regardless of the chosen rescaling. Interestingly, the **Relaxed** strategy (orange squares) shows a slightly slower initial convergence compared to **Unit** and **Adaptive** approaches; this is expected as the weights $w_t = 1 - (t+2)^{-3}$ are smaller in the early iterations, effectively damping the update step. On the other hand, the **Adaptive** strategy (green crosses) matches or exceeds the convergence speed of the unit baseline, showing that optimizing the rescaling weights w_t does not introduce instabilities.

While the MSE is a relevant loss, the main goal is to minimize the MCE. The bottom row of Figure 3 shows how the empirical MCE evolves over the rounds. Following our theoretical results, the plots show that the MCE has a consistent trend towards 0, often decreasing by some orders of magnitude

(e.g., German Credit and Diabetes). The empirical results show a smooth and stable decay of multicalibration error to zero.

D.5 Generalization

We examine how the theoretical convergence guarantees translate to unseen data; Figure 2 and its discussion are in Section 6. To make results more robust, we repeat the experiment for 20 different seeds and aggregate the test performance over the different runs. Each seed corresponds to a fresh random 80/20 train/test split (uniform random shuffle, no stratification, since all targets are continuous regression values). Figure 2 plots the average test MCE across iterations, with star markers indicating the optimal stopping point (minimum average error) for each strategy. Shaded bands around each curve indicate ± 1 standard deviation across the 20 seeds, so that the comparative claims in Section 6 can be assessed against random fluctuations.

Comparing the strategies shows a trade-off between efficiency and stability. The **Unit** and **Adaptive** strategies (blue and green lines) are highly efficient, typically reaching their optimal performance within the first 1-8 rounds (depending on the dataset). However, their aggressive updates can be harmful in some cases: this is most evident in the *German Credit* dataset, where they both overfit almost immediately. In contrast, the **Relaxed** strategy (orange dashed line) acts as an implicit regularizer. By reducing the early updates, it decays more gradually, shifting the optimal stopping point to later rounds (e.g., around round 12 in *California Housing* and round 5 in *German Credit*; on *Adult* it does not reach its minimum within $T = 20$). This makes the Relaxed strategy more robust to the choice of stopping time, reducing the risk of initial overfitting.

E Extended Related Work

This appendix complements the discussion of related work in Section 2.

Foundations of Multicalibration. Hebert-Johnson et al. (2018) introduced multicalibration, showing that a predictor can be calibrated across a potentially exponential family of intersecting subpopulations using an iterative postprocessing algorithm (HKRR). Dwork et al. (2021) later unified this line of work under outcome indistinguishability, placing multicalibration as a relaxation of indistinguishability from the true distribution. Related notions include multiaccuracy (Kim et al., 2019) and low-degree multicalibration (Gopalan et al., 2022b), which weaken calibration constraints to low-degree polynomial moments. Subsequent work suggested that multicalibration does not need to arise solely from post-hoc correction, but can instead emerge as a natural fixed point of empirical risk minimization (ERM). Gopalan et al. (2022a) introduced omniprediction, showing that multicalibrated predictors support optimal decision-making across a broad class of convex loss functions (Hu et al., 2023). Building on this perspective, Globus-Harris et al. (2023) established an equivalence between multicalibration and Bayes optimality under a swap-regret condition, leading to swap-agnostic learning (Gopalan et al., 2023). Globus-Harris et al. (2023) further connected multicalibration to boosting for squared-error regression, through repeated calls to a weak learner on subsets of the data distribution. Our work is different in providing the first finite-time convergence rate guarantees for recently proposed multi-round multicalibration gradient boosting algorithms (Tax et al., 2026) and considering both classification and regression tasks. Generalization guarantees for multicalibration have concerned discretized predictors (Hebert-Johnson et al., 2018; Globus-Harris et al., 2023) or uniform convergence over a given predictor class (Shabat et al., 2020); the class of predictors generated by multi-round boosting, in which the prediction is fed back as a feature, is data-dependent and is handled in Section 5 by a covering argument across rounds.

Multicalibration via Gradient Boosting. Motivated by these connections, recent work has explored achieving multicalibration directly within gradient boosting. These approaches fit residuals using ensembles of decision trees, where each boosting round is defined with respect to the predictor from the previous round. A key feature is that trees are trained on an extended input space that augments the original features with the current prediction. Jin et al. (2025) proposed such a method and showed that approximate multicalibration can be achieved in a single round under a restrictive loss saturation condition, which requires that further rounds yield only marginal improvement. Similarly, Tax et al. (2026) introduced MCGrad, a recursive GBDT method deployed at web-scale. They showed that the empirical multicalibration error can be bounded by a function of the gap between successive predictions, but did not establish that this gap converges to zero.

Despite strong empirical performance, existing gradient boosting approaches lack convergence guarantees for arbitrarily many rounds. Current guarantees either rely on single-round analysis under a loss saturation condition (Jin et al., 2025) or assume diminishing gaps between successive predictions (Tax et al., 2026). In contrast, we study the finite-time dynamics of multi-round boosting and derive convergence and rate guarantees.

Extensions Beyond the Mean. Although our analysis focuses on mean multicalibration, prior work has extended calibration guarantees to higher-order properties. The HappyMap framework (Deng et al., 2023) generalizes calibration targets to arbitrary statistics (e.g., quantiles), while Jung et al. (2021) introduced moment multicalibration to control variance. Gupta et al. (2022) and Jung et al. (2023) further developed multivalidity for constructing prediction intervals, primarily in online learning settings. Our results complement these advances by providing the first rigorous convergence rates for the batch boosting dynamics underlying practical multicalibration methods.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract claims (i) an $O(1/\sqrt{T})$ convergence rate for the prediction gap (which implies the same rate for empirical multicalibration error), (ii) linear convergence under smoothness assumptions on the weak learners, (iii) convergence guarantees for adaptive rescaling variants, and (iv) a bound on the population multicalibration error. These match Theorem 4.1, Theorem 4.2, Theorems 4.5 and 4.6, and Theorem 5.1, respectively. The introduction (Section 1) further qualifies the claims by specifying the inexact-oracle setting, the squared-loss specialization for the linear rate, and the experimental scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 ("Limitations and Future Work") explicitly states the main limitations: (i) assumption (A2) (strong convexity) breaks down on weak-learner-separable training sets, which is relevant for log-loss; (ii) the linear convergence rate of Theorem 4.2 requires Lipschitz continuity of the projector $A(f)$, which is only approximately satisfied by hard-split GBDT ensembles (also discussed in Section 4 after Theorem 4.2); (iii) our convergence guarantees are training-set only; and (iv) the population bound of Theorem 5.1 is for log-loss with unit weights and requires bounded coefficients, Lipschitz weak learners and an initial predictor independent of the sample (assumptions (G1)–(G3), Remark 5.3). The paper also notes (Section 4) that the step-size condition $\eta \leq m/(2L)$ is required by Theorem 4.1.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical contributions are stated as numbered theorems and lemmas (Theorem 4.1, 4.2, 4.5, 4.6, 5.1; Lemmas 4.3, 4.4 and A.1; Corollary 5.2; Theorem A.7). The assumptions (A1)–(A3) are stated explicitly at the start of Section 4, the assumptions (G1)–(G3) before Theorem 5.1, and additional structural assumptions (Lipschitz $A(f)$, factorised hypothesis class, etc.) are given at the point where they are used. Full proofs are provided in the Appendix: Appendix A.1 (Theorem 4.1), Appendix A.4 (Theorem 4.2), Appendix A.5 (Lemma 4.3), Appendix A.6 (Lemma 4.4), Appendix A.7 (Theorem 4.5), Appendix A.8 (Theorem 4.6), Appendix A.9 (Theorem A.7), Appendix A.2 (Lemma A.1), Appendix A.13 (Lemma A.11), and Appendix C (Theorem 5.1, Corollary 5.2). Theorem 4.2 additionally relies on the perturbation-theory result for Moore–Penrose pseudoinverses cited from Stewart [1977].

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 6 (Experimental Setup) specifies all hyperparameters needed to reproduce the experiments: initial predictor (Random Forest with 100 trees and max depth 5), weak learners (scikit-learn gradient boosting with 100 trees, max depth 3, learning rate 0.1), MCE estimation ($N_{\text{tree}} = 100$ random regression trees of depth $h = 3$), the three rescaling strategies (UNIT, RELAXED with $w_t = 1 - (t+2)^{-3}$, ADAPTIVE via Eq. (7)), $T = 20$ rounds, and shrinkage $\eta = 0.5$. The five datasets (Table 1) are publicly available standard benchmarks; Appendix D.1 describes the targets, preprocessing, and exclusions used. The implementation uses off-the-shelf scikit-learn components, so the procedure can be re-implemented in $\sim$50 lines. We commit to releasing our experimental code publicly upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The five datasets used (California Housing, Diabetes, Adult, German Credit, Communities and Crime) are all publicly available from UCI / scikit-learn (see Appendix D.1), and the algorithm is implemented entirely via scikit-learn's stock `RandomForestRegressor` and `GradientBoostingRegressor` with the hyperparameters listed in Section 6; Algorithm 1 fully specifies the boosting loop. We commit to releasing our experimental code publicly upon acceptance under a permissive open-source license, with a link included in the camera-ready version of the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 (Experimental Setup) lists all hyperparameters: number of rounds $T = 20$, shrinkage $\eta = 0.5$ (chosen to satisfy the step-size condition $\eta \leq m/(2L) = 1/2$ for squared loss), Random Forest initial predictor configuration, GBDT weak-learner configuration, MCE estimation parameters, and the three rescaling schemes. Appendix D.1 describes each dataset, its prediction target, and any preprocessing (e.g., for the Adult dataset the target is the sum of `education-num` and `hours-per-week`, with both removed from the input features). Appendix D.5 (Generalization) specifies the 80/20 train/test split protocol and the 20-seed averaging used for the test-set figure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The test-set generalization figure (Figure 2 in Section 6; protocol in Appendix D.5) reports averages over 20 random 80/20 train/test splits with shaded ± 1 standard-deviation bands, allowing the comparative claims about UNIT/RELAXED/ADAPTIVE to be assessed against random fluctuations. The training-set figures (Figure 1 in Section 6 and Figure 3 in Appendix D.4) are deterministic for a fixed seed because the training algorithm is deterministic and the entire training dataset is used in each round; for these the relevant "significance" is the goodness of fit of the geometric-decay tail, reported as R^2 for the log-gap regression in Figure 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We do not report compute usage explicitly. The experiments are intentionally lightweight: all five datasets fit comfortably in memory (the largest, Adult, has 32,561 rows), the weak learners are small scikit-learn GBDT ensembles, and a full $T = 20$-round run for one dataset and one rescaling strategy completes in a few minutes on a single CPU core; the entire experimental section can be reproduced on a standard laptop without GPU acceleration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is a theoretical convergence analysis of an existing algorithm class, evaluated on five widely-used public benchmark datasets. It introduces no new data collection, no human-subjects research, no model release, and no deployment. We have reviewed the NeurIPS Code of Ethics and confirm conformance.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper provides a finite-time convergence analysis of an existing class of algorithms (multicalibration gradient boosting); it does not introduce a new model, dataset, or deployment. The motivating problem of multicalibration is itself oriented toward fairness across subpopulations, so any downstream societal impact derives from the previously published algorithms whose convergence we analyze, not from the analysis itself. We do not foresee a path by which our theoretical results enable misuse beyond what is already feasible with the underlying algorithms.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any new data, models, or pretrained artifacts. The experimental section uses publicly available benchmark datasets and stock scikit-learn implementations.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All five datasets are publicly available standard benchmarks: California Housing (StatLib), Diabetes, Adult (Census Income), German Credit, and Communities and Crime (UCI Machine Learning Repository); their sources are described in Appendix D.1. The implementation uses scikit-learn (BSD 3-Clause). We use these assets within their stated terms of use; no scraped or restricted data is used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: While the paper does not release any new datasets or models, we will release the code artifact used for our experiments upon acceptance, under a permissive open-source license. It will be documented and linked in the camera-ready version.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments use pre-existing public benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core methods. The algorithms analyzed (multi-calibration gradient boosting and its rescaling variants), the theoretical results, and the experiments do not involve LLMs. Any use of LLMs by the authors during preparation of this manuscript was limited to writing, editing, and proof-checking assistance, which does not require declaration under the NeurIPS LLM policy.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.