Gradient Regularized Newton Boosting Trees with Global Convergence

Nikita Zozoulenko[1,*,†] **Daniel Falkowski**[2,*] **Thomas Cass**[1] **Lukas Gonon**[1,3]

[1]Imperial College London [2]Stockholm University [3]University of St. Gallen

*Equal contribution [†]Corresponding author: n.zozoulenko23@imperial.ac.uk

Abstract

Gradient Boosting Decision Trees (GBDTs) dominate tabular machine learning, with modern implementations like XGBoost, LightGBM, and CatBoost being based on Newton boosting: a second-order descent step in the space of decision trees. Despite its empirical success, the global convergence of Newton boosting is poorly understood compared to first-order boosting. In this paper, we introduce Restricted Newton Descent, a framework for convex optimization with Newton's method on Hilbert spaces with inexact iterates, based on the concepts of cosine angle and weak gradient edge. Within this framework, we recover Newton boosting with GBDTs and classical finite-dimensional theory as special cases. We first prove that vanilla Newton boosting achieves a linear rate of convergence for smooth, strongly convex losses that satisfy a Hessian-dominance condition. To handle general convex losses with Lipschitz Hessians, we extend a recent gradient regularized Newton scheme to the restricted weak learner setting, establishing the first global $\mathcal{O}(\frac{1}{k^2})$ convergence rate for second-order GBDTs. This scheme minimally modifies the classical algorithm by introducing an adaptive ℓ_2-regularization term proportional to the square root of the gradient norm at each iteration. In numerical experiments, we show that this scheme converges while vanilla Newton boosting may diverge.

1 Introduction

Gradient Boosting Decision Trees (GBDTs) are powerful machine learning models, excelling in tabular settings such as regression, classification, and learning to rank over neural baselines [9, 13, 23, 31, 32]. Modern implementations such as XGBoost [3], LightGBM [19], and CatBoost [30] are based on the Newton boosting algorithm, which updates a decision tree ensemble by performing a Newton descent step in the space of decision trees, derived via a second order Taylor expansion with respect to the training data [3]. However, despite the widespread use of GBDTs, the theoretical mechanisms of Newton boosting remain poorly understood, with most prior work focused on first-order gradient boosting [4, 10, 12, 14, 20–22].

While Newton's method in $\mathbb{R}^N$ offers fast quadratic local convergence, it lacks global convergence guarantees for general strictly convex losses, even when combined with line search [17]. Recently, the convex optimization literature has seen a resurgence of interest in regularized Newton schemes that achieve global $\mathcal{O}(\frac{1}{k^2})$ convergence rates. This was first established for the Cubic Regularized Newton (CRN) method [25, 27]; however, CRN requires solving an expensive subproblem at each iteration, making it unfit for GBDTs. More recently, the Gradient Regularized Newton (GRN) method was proposed, achieving the same $\mathcal{O}(\frac{1}{k^2})$ rate as CRN [7, 24]. GRN uses an adaptive ℓ_2-regularization term proportional to the square root of the gradient norm at each iteration. Crucially, this method avoids expensive subproblems and incurs negligible computational overhead compared to the standard Newton update. This discovery has sparked new research into the global convergence of second-order numerical schemes [6, 8, 15, 16]. A natural question arises: can these convergence

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

guarantees be extended to GBDTs and optimization with weak learners, providing convergence rates for second-order boosting methods used in practice?

Contributions: In this paper, we formulate Newton boosting as a convex optimization problem on a Hilbert space $\mathcal{H}$, where optimization is performed on a restricted set of permissible weak learners $\mathcal{F} \subset \mathcal{H}$. Our contributions are:

- We introduce the **Restricted Newton Descent** framework, based on the concepts of weak and exact iterates, cosine angles, and weak gradient edge. This recovers both Newton boosting with GBDTs and classical finite-dimensional convex optimization as special cases.
- For smooth strongly convex losses satisfying a Hessian-dominance condition, we prove that vanilla Restricted Newton Descent achieves a linear convergence rate. This includes practically relevant losses such as ℓ_2-regularized binary and categorical cross-entropy.
- For general convex losses with $2M$-Lipschitz Hessians, we generalize the GRN scheme of Mishchenko [24] to the Hilbert space setting with weak learners, establishing the first global $\mathcal{O}\big(\frac{1}{k^2}\big)$ convergence rate for second-order GBDTs, valid for all learning rates $\eta \leq 1$. If the loss is additionally strongly convex, a local linear rate is also obtained.
- We present numerical results showing that the gradient regularized Newton scheme converges while vanilla Newton boosting may diverge.

1.1 Related Literature

GBDTs: Gradient boosting models have traditionally been viewed as gradient descent in $\mathbb{R}^N$ on the training data, where N is the dataset size [11, 22]. Later works identified this process with gradient descent in the Hilbert space of square-integrable functions $\mathcal{L}^2(\widehat{\nu}_N)$, where $\widehat{\nu}_N$ is the empirical data distribution [14]; see also connections to gradient representation boosting in deep learning [28, 34, 38]. Recent theoretical work has shifted from the gradient perspective to studying boosting via coordinate descent in high-dimensional spaces, achieved by enumerating all possible decision trees [4, 20, 29].

Newton Boosting: Newton-like updates were first introduced by LogitBoost [11] for logistic loss, and later applied to GBDTs in a general context by XGBoost [3]. Empirical differences between first-order and Newton boosting were investigated in [33]. Theoretical analysis of Newton boosting remains limited compared to first-order methods; Notable work includes the LogitBoost analysis [35] using probability clamping, and later the SnapBoost [29] using coordinate descent.

Convergence Rates in Boosting: For second order boosting, both Sun et al. [35] and Parnell et al. [29] proved a linear rate in the strongly convex setting. In contrast, our work also covers general convex losses with Lipschitz Hessians. As for first order boosting, the earliest explicit convergence rate was proved in [1] with a slow $\mathcal{O}((\log k)^{-\frac{1}{2}})$ rate. For smooth convex losses, Sun et al. [35] proved a classical $\mathcal{O}(\frac{1}{k})$ rate. Lu et al. [21] incorporated Nesterov momentum to achieve an accelerated $\mathcal{O}(\frac{1}{k^2})$ rate, similar to our work. Linear rates for strongly convex losses were proven in [14] using gradient descent, and in [20] via randomized coordinate descent. Telgarsky [36] studied the primal-dual structure of boosting, obtaining a linear rate with non-standard constants.

2 Newton Boosting as Hilbert Space Convex Optimization

In this section, we recall Newton's method in the Hilbert space setting, and introduce the Restricted Newton Descent framework where Newton boosting with GBDTs is recovered as a special case.

2.1 Hilbert Space Newton's Method

Let $\mathcal{H}$ be a Hilbert space, and $L : \mathcal{H} \to \mathbb{R}$ be a convex, twice Fréchet differentiable loss function. We aim to minimize L with Hilbert space Newton's method with a fixed learning rate $\eta \in (0, 1]$. At each iteration k, we update an iterate $F_k = F_{k-1} + \eta f_k = F_0 + \sum_{t=1}^{k} \eta f_t$ by minimizing a second-order Taylor expansion of the loss:

$$L(F_k + f) \;\approx\; L(F_k) + \left\langle f, g_k + \frac{1}{2} H_k[f] \right\rangle \;=:\; L(F_k) + Q_k(f)$$

where $g_k = \nabla L(F_k)$ is the gradient, $H_k = \nabla^2 L(F_k)$ is the Hessian operator, and $Q_k(f)$ is the local

second-order approximation of L at the point F_k, that is after k iterations of the optimization scheme. Differentiation yields $\nabla Q_k(f) = g_k + H_k[f]$. Setting the derivative to 0 and solving for f gives the Newton update, assuming that H_k is invertible at the point g_k:

$$f_{k+1} := -H_k^{-1}[g_k] = \operatorname*{argmin}_{f \in \mathcal{H}} Q_k(f). \tag{1}$$

2.2 Restricted Newton Descent

In the context of gradient boosting, the exact update f_{k+1} in (1) is generally not attainable since weak learners, as opposed to exact Newton directions, are used to minimize the loss. Consequently, the ensemble F_k instead takes the form $F_k := F_0 + \sum_{t=1}^{k} \eta f_t^w$ for some notion of a weak iterate f_t^w. To formalize this, we introduce the **Restricted Newton Descent** framework (not to be confused with constrained optimization). This differs from standard Newton's method in two ways. Firstly, the gradient g_k and Hessian H_k are computed w.r.t. the sum of weak, rather than exact, iterates F_k. Secondly, we do not minimize $Q_k(f)$ over the entire space $f \in \mathcal{H}$, but rather restrict the weak update f_{k+1}^w to belong to some subset $f_{k+1}^w \in \mathcal{F} \subset \mathcal{H}$ of admissible weak learners. More specifically, given a family $\mathcal{F}$, we define the weak iterates f_{k+1}^w as any element

$$f_{k+1}^w \in \operatorname*{argmin}_{f \in \mathcal{F}} Q_k(f). \tag{2}$$

Note that f_{k+1}^w is not necessarily unique. We consider general families $\mathcal{F}$ satisfying the following scalability condition, first introduced in Lu et al. [21]:

Assumption 2.1. The family $\mathcal{F}$ is closed under scalar multiplication. That is, if $f \in \mathcal{F}$ then $af \in \mathcal{F}$ for all $a \in \mathbb{R}$.

Decision trees, commonly used in boosting, satisfy this assumption. If $\mathcal{F} = \mathcal{H}$, then the weak update f_{k+1}^w coincides with the exact update f_{k+1}, and standard Newton's method is recovered. The full algorithm is detailed in Algorithm 1, with the exact step in parentheses.

Algorithm 1 Vanilla Restricted Newton

Input: Loss $L : \mathcal{H} \to \mathbb{R}$. Learning rate η. Weak learner family $\mathcal{F}$. Initial value F_0.
for $k = 0, 1, 2, \ldots$ **do**
  $g_k,\ H_k \leftarrow \nabla L(F_k),\ \nabla^2 L(F_k)$
  $\left(\text{Exact target } f_{k+1} \leftarrow -H_k^{-1} g_k\right)$
  Weak learner $f_{k+1}^w \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} Q_k(f)$
  $F_{k+1} \leftarrow F_k + \eta f_{k+1}^w$
end for

Theoretical analysis of the optimization scheme now depends on the error between the chosen weak iterate f_{k+1}^w and the exact iterate f_{k+1}, as this difference introduces a component orthogonal to the weak update in the inner product induced by the quadratic objective. As we show later, different schemes rely on different error metrics. For vanilla Newton's method (Section 3), the relevant metric is the cosine angle between f_{k+1} and f_{k+1}^w in the induced Hessian norm, whereas for the globally convergent gradient regularized scheme (Section 4), the metric takes the form of a weak learner edge condition between the true gradient and the implied weak gradient.

2.3 GBDTs as Restricted Convex Optimization

In this section, we provide an overview of gradient boosting from a functional $\mathcal{L}^2(\nu)$ perspective [14, 34], in the context of Restricted Newton Descent. Let $(X, Y) \sim \nu$ be a random sample with its corresponding probability measure ν. Denote by $X \sim \nu_X$ the features in $\mathbb{R}^q$, with targets $Y \sim \nu_Y$ in $\mathbb{R}^K$, and $\nu_{Y|X=x}$ the conditional law. Let $\widehat{\nu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{(x_i, y_i)}$ be the empirical measure of ν with respect to a training dataset $\{x_i, y_i\}_{i=1}^N$ of size N. We work in the Hilbert space $\mathcal{L}^2(\nu_X)$ of square integrable functions defined on the feature space, with inner product $\langle f, g \rangle = \mathbb{E}_{\nu_X}[\langle f(X), g(X) \rangle_{\mathbb{R}^K}]$. When $\nu = \widehat{\nu}_N$, this space can be identified with $\mathbb{R}^{N \times K}$ using the inner product $\langle f, h \rangle = \frac{1}{N}\sum_{i=1}^{N} \langle f(x_i), h(x_i) \rangle_{\mathbb{R}^K}$. Let l be a twice differentiable convex loss function.

To derive the Newton boosting algorithm used by e.g. XGBoost and LightGBM as a special case of Restricted Newton Descent, we simply take $\mathcal{F}$ to be the set of decision trees up to a fixed depth, with base space $\mathcal{H} = \mathcal{L}^2(\widehat{\nu}_{X,N})$, and loss function $L(F) = \frac{1}{N}\sum_{i=1}^{N} l(F(x_i), y_i) = \mathbb{E}_{\widehat{\nu}_N}[l(F(X), Y)]$. In the space $\mathcal{L}^2(\nu_X)$, the Fréchet gradient and Hessian take the form

$$\nabla L(F)(x) = \mathbb{E}_{\nu_{Y|X=x}}\big[\partial_1 l(F(x), Y)\big], \tag{3}$$

$$\nabla^2 L(F)[f](x) = \mathbb{E}_{\nu_{Y|X=x}}\big[\partial_{11} l(F(x), Y) f(x)\big], \tag{4}$$

and are easily computed for empirical measures $\widehat{\nu}_N$ as partial derivatives and matrices with respect to $F(x_i)$. Write $g_{(i)} = \partial_1 l(F_k(x_i), y_i) \in \mathbb{R}^K$ and $h_{(i)} = \partial_{11} l(F_k(x_i), y_i) \in \mathbb{R}^{K\times K}$. Following the XGBoost derivation [3], the objective becomes to minimize

$$L(F_k + f) = \frac{1}{N}\sum_{i=1}^{N} l(F_k(x_i) + f(x_i), y_i) \approx L(F_k) + \frac{1}{N}\sum_{i=1}^{N}\left\langle f(x_i), g_{(i)} + \frac{1}{2}h_{(i)}f(x_i)\right\rangle_{\mathbb{R}^K}$$

which we see coincides with the Hilbert space derivation using Q_k. Furthermore, in boosting we minimize over the set of weak learners $f \in \mathcal{F}$ rather than over all $\mathcal{L}^2(\widehat{\nu}_N)$. Letting f be a decision tree of the form $f(x) = \sum_j w_j I_j(x)$ where $I_j(x) \in \{0, 1\}$ is a disjoint partition of the feature space, one finds that the optimal weights w_j minimizing $Q_k(f)$ are $w_j = -(\sum_{i\in I_j} h_{(i)})^{-1}(\sum_{i\in I_j} g_{(i)})$, coinciding with the classical XGBoost formula. See Section A.2 for more details and full derivation.

3 Restricted Newton for Hessian-dominated Losses

Before introducing the globally convergent scheme in Section 4, we first analyze the vanilla Newton update via Restricted Newton Descent. We prove for smooth strongly convex losses satisfying a Hessian-dominance condition, that vanilla Newton's method with weak learners converges globally with a linear rate. This analysis is inspired by the LogitBoost work [35]. However, our analysis uses a functional formulation on Hilbert spaces, which includes standard Newton's method as a special case and accommodates losses beyond logistic loss, rather than the matrix-based formulation of the work on LogitBoost. We also introduce a more direct, mathematically natural, trajectory-dependent definition of weak learner cosine angles. To ensure the Newton update is well-defined on $\mathcal{H}$, we make the following assumption:

Assumption 3.1. The equation $g_k = -H_k f$ admits a unique solution $f = f_{k+1} = -H_k^{-1} g_k$.

Assumption 3.1 is satisfied, for example, when L is strongly convex, or in most practical settings when $\mathcal{H}$ is a finite dimensional space with a strictly convex loss L. Note that this assumption is not required for the gradient regularized scheme in Section 4.

3.1 Cosine Angle

Recall that the iterates of Restricted Newton Descent are derived by minimizing $Q_k(f) = \left\langle f, g_k + \frac{1}{2}H_k f\right\rangle$ restricted to $f \in \mathcal{F}$ at each iteration k, yielding $F_k = \sum_{t=1}^{k} \eta f_t^w$. In order to analyze the restricted Newton scheme, we need a way to measure the error between the exact direction f_{k+1} and the weak learner f_{k+1}^w. It turns out that the critical quantity determining the performance of this scheme, is the cosine angle between f_{k+1} and f_{k+1}^w measured in the Newton geometry induced by the Hessian. For this, we use the following norm definition.

Definition 3.2. Let $A : \mathcal{H} \to \mathcal{H}$ be a self-adjoint positive operator. We define the inner product induced by A as

$$\langle h, f\rangle_A := \langle h, Af\rangle. \tag{5}$$

When $A = H_k$, we call this the **Hessian-induced inner product**, with induced norm $\|\cdot\|_{H_k}$.

The following lemma motivates the use of the Hessian-induced norm and naturally leads to the definition of the cosine angle between weak and exact iterates.

Lemma 3.3. Minimizing $Q_k(f)$ is equivalent to finding a closest element $f \in \mathcal{F}$ to the exact Newton direction f_{k+1} in the Hessian-induced norm $\|\cdot\|_{H_k}$.

Proof. Using Definition 3.2 and the exact Newton update $f_{k+1} = -H_k^{-1} g_k$, we can write

$$\|f - f_{k+1}\|_{H_k}^2 = \|f\|_{H_k}^2 - 2\langle f, -g_k\rangle + \|f_{k+1}\|_{H_k}^2 = 2Q_k(f) + \|f_{k+1}\|_{H_k}^2. \qquad \square$$

Definition 3.4. Consider the iterates produced by vanilla Newton's method. At each iteration k, we define the Hessian-induced cosine angle $\Theta_k \in [0, 1]$ between the exact update f_{k+1} and the weak learner f_{k+1}^w as

$$\Theta_k := \frac{\langle f_{k+1}, f_{k+1}^w\rangle_{H_k}}{\|f_{k+1}\|_{H_k}\|f_{k+1}^w\|_{H_k}}, \qquad \Theta := \inf_k \Theta_k. \tag{6}$$

In contrast to [35], which uses a weak learnability assumption specific to binary classification, more recent prior work uses definitions similar in spirit to ours, but defines them by taking suprema and infima over all admissible weak learner directions and target directions [20, 21, 29]; see also [14]. Our definition is simpler and more direct, capturing the specific two vectors f_{k+1}^w and f_{k+1} determining the cosine angle. Using this definition, we immediately obtain the following exact algebraic properties connecting the weak and exact Newton iterates with the gradient g_k, directly via the cosine angle Θ_k. See Section C for proof. These identities and related results provide precise control over the weak learner approximation errors, enabling us to adapt more traditional convex optimization techniques to establish convergence rates for boosting.

Proposition 3.5. If $\mathcal{F}$ is closed under scalar multiplication, then the weak and exact Newton directions satisfy

1. $\|f_{k+1} - f_{k+1}^w\|_{H_k}^2 = (1 - \Theta_k^2)\|f_{k+1}\|_{H_k}^2$,
2. $\|f_{k+1}^w\|_{H_k}^2 = \Theta_k^2\|f_{k+1}\|_{H_k}^2 = -\langle g_k, f_{k+1}^w\rangle$

3.2 Decrease in Loss for Vanilla Restricted Newton

The restricted Newton scheme guarantees a strict decrease in loss at each iteration, forming the basis of our convergence analysis. This relies on standard regularity assumptions used in convex optimization: L being S-smooth and μ-strongly convex, meaning the Hessian satisfies $\mu I \preceq \nabla^2 L(F) \preceq SI$. See Section B for more details of assumptions and standard inequalities that follow. Additionally, we introduce the following definition of a Hessian-dominated loss that will be useful in our analysis.

Definition 3.6. Assume $L_\star := \inf_{F\in\mathcal{H}} L(F) > -\infty$. We say that L is **Hessian-dominated** if there exists a constant $c > 0$ such that the weak iterates F_k and exact Newton direction f_{k+1} satisfy

$$\|f_{k+1}\|_{H_k}^2 \geq c\big(L(F_k) - L_\star\big) \qquad \text{for all } k.$$

Definition 3.6 can be viewed as a Polyak–Łojasiewicz condition [18] stated in the local Hessian norm along the optimization trajectory. This definition also generalizes the condition derived for LogitBoost in [35, Theorem 4] to general losses using the language of Hessian norms and weak iterates. Examples of Hessian-dominated losses include binary and categorical cross entropy, both with constant $c = 1$, as proved in Section E. Their ℓ_2-regularized versions are also Hessian-dominated; see Lemmas E.7 and E.8. We are now ready to state our main convergence result for vanilla Restricted Newton Descent.

Theorem 3.7 (Global Rate I). If L is S-smooth, μ-strongly convex, and $\mathcal{F}$ closed under scalar multiplication, then for any learning rate $\eta \in (0, \frac{2\mu}{S})$ we have

$$L(F_k) - L(F_{k+1}) \geq \eta\left(1 - \frac{\eta S}{2\mu}\right)\Theta^2\|f_{k+1}\|_{H_k}^2. \tag{7}$$

Consequently, if L is Hessian-dominated, then the optimality gap $L(F_k) - L_\star$ converges linearly with contraction factor $1 - \rho$, where $\rho = c\Theta^2\eta(1 - \frac{\eta S}{2\mu})$:

$$L(F_{k+1}) - L_\star \leq (1 - \rho)\big(L(F_k) - L_\star\big). \tag{8}$$

Remark 3.8. Our bound compares favourably to SnapBoost [29], which proved a linear rate with contraction factor $1-\Theta_{\text{snap}}^2\mu^2/S^2$ for the fixed learning rate $\eta = \mu/S$. With the same learning rate, our result gives the contraction factor $1-c\Theta^2\mu/(2S)$. Thus, when the Hessian-dominance constant c is independent of the condition number S/μ, our bound has better dependence. Furthermore, we achieve essentially the same rate as [35], which uses probability clamping and a change-of-Hessian lemma to establish a linear rate for logistic loss. Our theorem instead uses smoothness, strong convexity, and Hessian dominance.

While cross entropy loss itself is not strongly convex, this issue is typically addressed in two ways. First, adding ℓ_2-regularization to the loss function yields a strongly convex loss. Second, probability values can be clipped to $[\epsilon, 1 - \epsilon]$ for some $\epsilon > 0$, which implicitly implies strong convexity. An alternative approach to Theorem 3.7 specific to trees is to clip the leaf values rather than probabilities,

in combination with a *change of Hessian lemma*, similar to [35, Lemma 9]. However, we believe our approach offers a more concise theoretical treatment.

4 Gradient Regularized Restricted Newton

Despite the fast quadratic local convergence of Newton's method in finite-dimensional convex optimization, the method does not have global convergence guarantees for general smooth convex functions [17]. A classical example of this is $L(x) = \sqrt{1+x^2}$ defined for $x \in \mathbb{R}$, see Section D. However, cubic regularized Newton [27] and the more recent gradient regularized Newton [7, 24] achieve a global $O(\frac{1}{k^2})$ convergence rate. In this section, we extend the latter scheme to Newton boosting with weak learners within our framework of Restricted Newton Descent.

Instead of minimizing the local quadratic approximation $Q_k(f)$ at each iteration k, gradient regularized Newton's method minimizes the following regularized objective:

$$Q_k^{reg}(f) \;=\; \left\langle f, g_k + \frac{1}{2}H_k[f] \right\rangle + \frac{\lambda_k}{2}\|f\|^2 \;=\; Q_k(f) + \frac{\lambda_k}{2}\|f\|^2,$$

where $\lambda_k \in (0, \infty)$ is a scalar regularization parameter. For losses with $2M$-Lipschitz continuous Hessians, we will allow values $\lambda_k \geq \sqrt{M\|g_k\|}$. Minimizing $Q_k^{reg}(f)$ yields the exact regularized update

$$f_{k+1} = -(H_k + \lambda_k I)^{-1}[g_k]. \tag{9}$$

In other words, gradient regularized Newton is simply an ℓ_2-regularized Newton step where the ℓ_2-regularization is adaptive at each iteration, being proportional to the square root of the gradient norm $\sqrt{\|g_k\|}$. Similar to Section 3, we consider Restricted Newton Descent using weak learners rather than the exact update. The full gradient regularized scheme is detailed in Algorithm 2, with the implicit exact step denoted in parentheses.

Algorithm 2 Gradient Regularized Restricted Newton

Input: Loss $L : \mathcal{H} \to \mathbb{R}$ with $2M$-Lipschitz Hessian. Learning rate η. Weak learner family $\mathcal{F}$. Initial F_0.
for $k = 0, 1, 2, \ldots$ **do**
  $g_k,\ H_k \leftarrow \nabla L(F_k),\ \nabla^2 L(F_k)$
  Choose $\lambda_k \geq \sqrt{M\|g_k\|}$
  $\big($Exact target $f_{k+1} \leftarrow -(H_k + \lambda_k I)^{-1} g_k\big)$
  Weak learner $f_{k+1}^w \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} Q_k^{reg}(f)$
  $F_{k+1} \leftarrow F_k + \eta f_{k+1}^w$
end for

4.1 Convergence of Restricted GRN

We impose a standard regularity condition on the loss L in terms of the Lipschitz continuity of its Hessian, as is common for regularized Newton methods [7, 24, 25, 27].

Assumption 4.1. The Hessian $\nabla^2 L : \mathcal{H} \to \mathcal{L}(\mathcal{H})$ is $2M$-Lipschitz continuous for some constant $M > 0$:

$$\|\nabla^2 L(f) - \nabla^2 L(g)\| \leq 2M\|f - g\| \qquad \forall f, g \in \mathcal{H}.$$

The following proposition bounds the change in loss and guarantees strict decrease for $0 < \eta < \sqrt{3}$ whenever the weak update is nonzero. From this point on, we assume that $\mathcal{F}$ is closed under scalar multiplication. All proofs for this section are deferred to Section C.

Proposition 4.2. If L is convex with $2M$-Lipschitz Hessian, with learning rate $\eta \in (0, 2]$, then the iterates of Algorithm 2 satisfy

$$L(F_k) - L(F_{k+1}) \geq \left(\eta - \frac{\eta^3}{3}\right)\lambda_k \|f_{k+1}^w\|^2. \tag{10}$$

The bound (10) guarantees strict decrease in loss for $0 < \eta < \sqrt{3}$ whenever $f_{k+1}^w \neq 0$. For $\sqrt{3} \leq \eta < 2$, the inequality remains valid but does not guarantee strict descent.

We now require a way to quantify how well the weak updates f_{k+1}^w approximate the exact updates f_{k+1} in the gradient regularized scheme. Previously in Theorem 3.7 for vanilla Newton, the decrease

in loss was determined by the Hessian-induced norm $\|\cdot\|_{H_k}$. In contrast, for the regularized scheme the decrease in loss is governed by the quantity $\lambda_k\|f_{k+1}^w\|$. As such, the cosine angle from Definition 3.4 is no longer well suited for theoretical analysis of GRN, since the scheme and its convergence proof are fundamentally driven by gradient growth rather than directional alignment. Consequently, we introduce the notion of an implied weak gradient g_k^w, together with a corresponding weak gradient edge condition. This condition essentially states that the weak gradient remains within a relative radius of the true gradient, at each iteration k.

Definition 4.3. We define the **implied weak gradient** as

$$g_k^w := -(H_k + \lambda_k I) f_{k+1}^w. \tag{11}$$

We say that the iterates f_{k+1}^w have **weak gradient edge** $\gamma = \inf_k \gamma_k$ if there exists real numbers $0 < \gamma_k \leq 1$ satisfying

$$\|g_k^w - g_k\|^2 \leq (1 - \gamma_k^2)\|g_k\|^2. \tag{12}$$

The following result relates the gradient at consecutive iterations and serves as the primary tool for our convergence analysis.

Proposition 4.4 (Gradient Growth). If the iterates of *Algorithm* 2 have weak gradient edge γ, with learning rate $\eta \in (0, 1]$, then

$$\|g_{k+1}\| \leq \left(1 + \eta^2 + \eta\sqrt{1-\gamma^2}\right)\|g_k\|. \tag{13}$$

We now present our main results: establishing the global and local convergence rates of gradient regularized Newton's method with weak learners. Despite the inexact updates inherent to boosting and Restricted Newton Descent, Algorithm 2 retains the classical $\mathcal{O}(\frac{1}{k^2})$ rate, matching first-order momentum boosting [21] without the requirement of fitting two trees per iteration. In comparison, standard first-order boosting without momentum has a global $\mathcal{O}(1/k)$ rate. We use the following standard definition from the cubic Newton literature [24, 27].

Definition 4.5. L has a **bounded initial sublevel set** if a finite optimum $F^* \in \operatorname{argmin}_{F\in\mathcal{H}} L(F)$ exists, and if there exists a constant $\Lambda > 0$ such that $L(F) \leq L(F_0) \implies \|F - F^*\| \leq \Lambda$.

Theorem 4.6 (Global Rate II). Assume L is convex with $2M$-Lipschitz Hessian, a bounded initial sublevel set with $\|F - F^*\| \leq \Lambda$, and weak gradient edge γ. Let $\lambda_k = C\sqrt{M\|g_k\|}$ for some $C \geq 1$. For any learning rate $\eta \in (0, 1]$, the gradient regularized Newton scheme has global convergence rate

$$L(F_k) - L(F^*) \leq \mathcal{O}\left(\max\left\{\frac{C^2 M\Lambda^3}{\eta^2 k^2 \gamma^{12}},\ \Lambda\exp\left(-k\eta\gamma^2\left(\frac{\gamma^2}{2} + 2 - 2\eta\right)/16\right)\right\}\right).$$

The proof proceeds by analyzing two regimes based on the fraction of iterates where the gradient fails to contract sufficiently, determined by η and γ. In the first regime, we derive a recurrence of the form $\alpha_{k+1} \leq \alpha_k - \tau\alpha_k^{3/2}$ for some $\tau > 0$, yielding a $\mathcal{O}(\frac{1}{\eta^2\gamma^{12}k^2})$ rate. In the second regime, the cumulative contraction dominates, resulting in a linear rate. The global rate is governed by the maximum of these two bounds; see Section C for the full proof.

Theorem 4.7 (Local Rate). Let $\lambda_k = C\sqrt{M\|g_k\|}$ for some $C \geq 1$, and let the learning rate be $\eta \in (0, 1]$. If L is μ-strongly convex with $2M$-Lipschitz Hessian, weak gradient edge γ, and if there exists a $k_0 \geq 0$ such that $\|g_{k_0}\| \leq \frac{\mu^2}{4C^2M}\left(\frac{1-\sqrt{1-\gamma^2}}{1+\sqrt{1-\gamma^2}}\right)^2$, then for all $k \geq k_0$ we have

$$\|g_{k+1}\| \leq \left(1 - \frac{\eta}{4}\left(1 - \sqrt{1-\gamma^2}\right)\right)\|g_k\| \tag{14}$$

$$\leq \left(1 - \frac{\eta\gamma^2}{8}\right)\|g_k\|. \tag{15}$$

Hence the iterates of Algorithm 2 converge locally linearly.

Remark 4.8 (On Greedy Tree Construction). The results in Sections 3 and 4 are stated, for simplicity, for the same weak learner family $\mathcal{F}$ at each iteration k. In practice, libraries such as

XGBoost, LightGBM, and CatBoost use greedy decision tree construction rather than finding a globally optimal tree for the quadratic minimization. Our proofs require only optimality of the selected weak update along its scalar multiples, rather than global optimality over the full tree class. Thus, the results also apply to iteration-dependent families selected by greedy tree construction, provided the leaf values exactly minimize the corresponding quadratic objective for the chosen structure. The quantities Θ_k and γ_k capture the approximation quality of the resulting updates for vanilla and gradient regularized Newton, respectively, including the effect of greedy split finding.

5 Practical Considerations

In this section, we discuss when the regularity assumptions of Theorems 3.7, 4.6 and 4.7 are satisfied in practice for boosting, and how the GRN scheme can be implemented into existing tree boosting algorithms with negligible computational overhead.

5.1 Regularity of the Boosting Objective

Our convergence analysis relies on standard convex optimization assumptions, such as Lipschitz-continuous Hessians, smoothness, and strong convexity. Using the same notation as Section 2.3, because the boosting empirical risk $L(F) = \frac{1}{N}\sum_{i=1}^{N} l(F(x_i), y_i)$ is constructed from a pre-specified element-wise loss l (e.g., MSE, cross entropy, Charbonnier), one can prove that the regularity of l carries over to L in the boosting space $\mathcal{H} = \mathcal{L}^2(\widehat{\nu}_{X,N})$. We formalize this below in Proposition 5.1. This result ensures that if the base loss l is well-behaved, the boosting objective is as well. Moreover, because l is chosen prior to training, its Lipschitz constant M_0 can be computed analytically.

Proposition 5.1. Assume that the inputs $x_1, \dots, x_N$ are pairwise distinct. If the element-wise loss $l(\cdot, y_i)$ has an M_0-Lipschitz Hessian in $\mathbb{R}^K$ uniformly across all fixed targets y_i, then the empirical risk L has an $M_0\sqrt{N}$-Lipschitz Hessian in the boosting space $\mathcal{L}^2(\widehat{\nu}_{X,N})$. Additionally, if l is S-smooth, μ-strongly convex, or Hessian-dominated with constant c, then L also is with the same constant values.

5.2 Implementing GRN for GBDTs

The GRN scheme is very simple to implement in practice and only differs from vanilla Newton boosting by an extra additive ℓ_2-regularization term based on the gradient norm. Both methods use the exact same split finding, gain calculation, and leaf fitting formulas. The only difference lies in the effective ℓ_2-regularization used. Standard tree boosting implementations like XGBoost and LightGBM use static ℓ_2-regularization in the tree-fitting objective. GRN, on the other hand, uses an adaptive regularization term λ_k given by

$$\lambda_k = \sqrt{M\|g_k\|}.$$

An additional static term λ_{base} can be added in practice, giving $\lambda_k = \lambda_{\text{base}} + \sqrt{M\|g_k\|}$, but our convergence guarantees are stated without this additional term. Using the notation of Section 2.3, the exact gradient norm in the boosting space is computed from the sample gradients $g_{(i)}$ as $\|g_k\| = \left(\frac{1}{N}\sum_{i=1}^{N} \|g_{(i)}\|_{\mathbb{R}^K}^2\right)^{1/2}$. Because updating this adaptive ℓ_2 term only requires a single vector norm calculation, and the main computational bottleneck for GBDT algorithms is greedy split finding, the computational overhead of GRN is negligible. Note, however, that our code implementation scales the regularization added to each aggregated leaf Hessian by the number of samples in that leaf, as suggested by our theory and derivations; see Remark A.1 for more details.

6 Numerical Experiments

In this section, we empirically validate our theoretical findings regarding the convergence of boosting schemes and the behaviour of weak learner edges. The purpose of this section is not to prove or disprove *state of the art* predictive performance of the GRN boosting scheme, but rather to supplement the theory presented in earlier sections regarding training and convergence. We implement first-order gradient boosting, Newton boosting, and the Gradient Regularized Newton (GRN) boosting scheme in JAX [2], following the standard depth-wise greedy

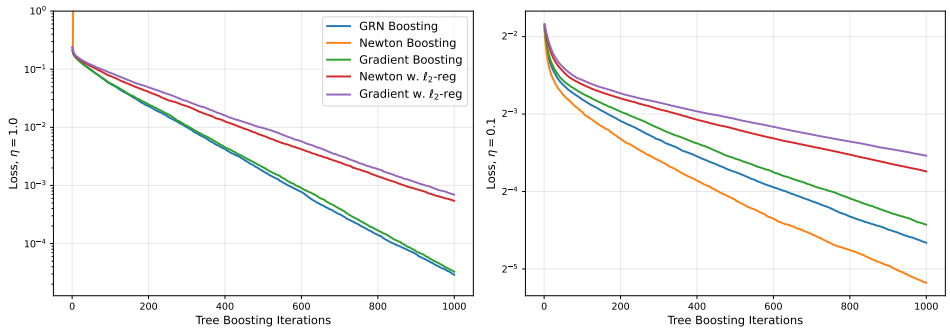

Figure 1: Charbonnier train loss on the Wine Quality UCI dataset, for different gradient boosting schemes and learning rates η. Vanilla Newton boosting diverges for $\eta = 1$.

tree construction strategy used by XGBoost [3]. The exact code used to produce these experiments is included in the supplementary material; an updated repository is available at `https://github.com/nikitazozoulenko/global-newton-boosting`.

6.1 Charbonnier loss divergence and convergence

The Charbonnier loss, defined as $l(\widehat{y}, y) = \sqrt{1 + (y - \widehat{y})^2} - 1$, is a classical example of a loss function where Newton's method can diverge (see Section D). We investigate if this also is true for boosting, using the Wine Quality UCI dataset [5] with depth 4 trees. For GRN we set $M = 0.1$, and otherwise use a static ℓ_2-regularization of 1.0 when explicitly stated. Here these terms were tuned empirically rather than chosen from the global Hessian-Lipschitz bound in Proposition 5.1. Figure 1 compares the training loss of Newton boosting, first-order gradient boosting, and GRN for different learning rates η. For $\eta = 1$, vanilla Newton boosting diverges, whereas GRN decreases the fastest. GRN initially tracks the trajectory of fixed ℓ_2-regularized Newton boosting but eventually outperforms it once the gradient norm decreases sufficiently. For the smaller value $\eta = 0.1$, Vanilla Newton boosting minimizes the loss most rapidly in this scenario, but is only able to converge in this small learning rate setting.

6.2 Weak learner conditions and tree depth

To better understand the theoretical constants governing convergence, we train models on a subset of the Higgs dataset [37] with 10000 data points using BCE loss with ℓ_2-regularization $\lambda_{\text{base}} = 0.1$ for non-GRN models, and $M = 0.01$ for GRN, giving roughly equivalent regularization at initialization. We use a learning rate $\eta = 0.1$ for all models. We compute the cosine angle Θ_k and weak learner edge γ_k in Fig. 2 at every iteration k, via (6) and (12), by instantiating $f^w_{k+1}, f_{k+1}, g^w_k$, and g_k as vectors in $\mathbb{R}^N$. As expected, increasing tree depth results in stronger weak learners, reflected in higher values for Θ_k and γ_k. We observe that these metrics are high during the initial iterations of boosting and eventually plateau at positive values. As optimization progresses, fitting the diminishing residuals becomes increasingly difficult for the weak learners, naturally leading to lower weak learner edges. The trajectories of the cosine angle Θ_k and weak gradient edge γ_k are remarkably similar, which is somewhat unexpected given that they are defined in different geometries. Conceptually, γ_k (Definition 4.3) can be viewed as an approximate $(H_k + \lambda_k I)^2$-induced cosine angle between f^w_{k+1} and f_{k+1}, in contrast to the standard H_k-induced cosine angle Θ_k. If the error $f_{k+1} - f^w_{k+1}$ were orthogonal to f^w_{k+1} in the $(H_k + \lambda_k I)^2$-induced inner product, then γ_k would equal the cosine between the weak and exact updates in that inner product. However, this orthogonality is not guaranteed.

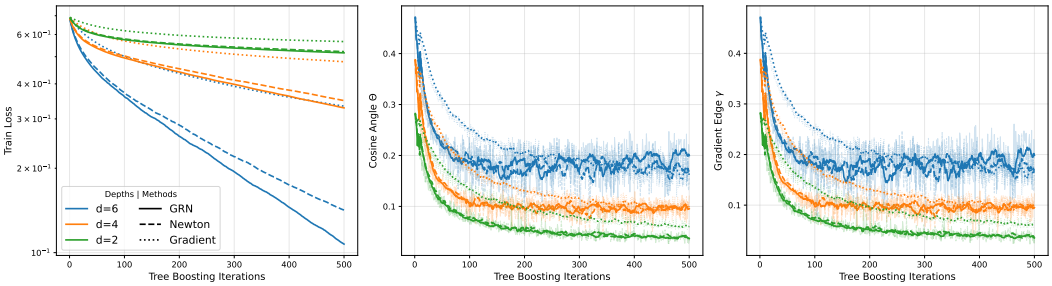

Figure 2: Binary cross entropy, cosine angle Θ_k, and weak gradient edge γ_k computed at each boosting iteration k, for varying tree depths d. The exact values are noisy between iterations, but the rolling window average (highlighted in bold) remains stable.

Conclusion, Limitations, and Future Work

By posing boosting as convex optimization, we introduced the Restricted Newton Descent framework to establish convergence rates for general weak learners, recovering Newton decision tree boosting as a special case. Using the cosine angle Θ_k, we proved that vanilla Newton boosting achieves a linear convergence rate for Hessian-dominated losses. For general convex losses with Lipschitz Hessians, we analyzed a novel gradient regularized Newton (GRN) boosting scheme, establishing the first global $\mathcal{O}(1/k^2)$ convergence rate for second-order GBDTs, via the weak gradient edge γ. We also demonstrated that vanilla Newton boosting may diverge on such losses while the GRN scheme converges.

However, there are some limitations to this work. The global theoretical guarantee comes with a heavy γ^{12} worst-case dependence in the bound. This dependency is likely an intrinsic artifact of the GRN scheme being fundamentally based on Euclidean gradient growth, rather than relying solely on the local Hessian geometry. While this does not cause any problems in the classical exact theory [24] (corresponding to $\gamma = 1$), the additional error terms introduced by the weak learners are difficult to control in this setting. Furthermore, because our work is primarily theoretical and aims to bridge the gap between GBDTs and traditional convex optimization, direct performance benchmarking against highly optimized C++ libraries falls outside our scope. Finally, while global Newton schemes are necessary to obtain global convergence guarantees, the empirical difference between the vanilla and GRN schemes can become minor at small learning rates.

Our results suggest several promising directions for future research. First, given that finite-dimensional cubic and gradient regularized Newton can be accelerated to $\mathcal{O}(1/k^3)$, investigating whether such acceleration can be realized in the boosting setting is a natural next step. Second, it would be interesting to study the Restriction operation for other optimization schemes to determine which convergence properties are retained under weak updates, and see if the dependence on γ^{12} can be improved. Finally, investigating the possibility of superlinear rates with weak learners and establishing generalization bounds for Newton boosting remain open problems. Future work will also explore other applications of Restricted Newton Descent beyond the gradient boosting setting.

Acknowledgements

NZ has been supported by the Roth Scholarship at Imperial College London. DF was supported by the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine via Alexander Berglund's 2024 Göran Gustafsson Prize. The work of TC was supported in part by UK Research and Innovation (UKRI) through the Engineering and Physical Sciences Research Council (EPSRC) via Programme Grants [Grant No. UKRI1010: High order mathematical and computational infrastructure for streamed data that enhance contemporary generative and large language models]. We acknowledge computational resources and support provided by the Imperial College Research Computing Service (DOI: `10.14469/hpc/2232`). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[1] Peter J. Bickel, Ya'acov Ritov, and Alon Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7(25):705–732, 2006.

[2] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

[3] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

[4] Corinna Cortes, Mehryar Mohri, and Dmitry Storcheus. Regularized gradient boosting. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

[6] Nikita Doikov. Minimizing quasi-self-concordant functions by gradient regularization of Newton method. *Mathematical Programming*, Aug 2025. ISSN 1436-4646. doi: 10.1007/s10107-025-02270-9.

[7] Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with Bregman distances. *Mathematical Programming*, 204(1):1–25, Mar 2024. ISSN 1436-4646. doi: 10.1007/s10107-023-01943-7.

[8] Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-universal regularized Newton method. *SIAM Journal on Optimization*, 34(1):27–56, 2024. doi: 10.1137/22M1519444.

[9] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[10] Robert M. Freund, Paul Grigas, and Rahul Mazumder. A new perspective on boosting in linear regression via subgradient optimization and relatives. *The Annals of Statistics*, 45(6):2328–2364, 2017.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, 28 (2):337 – 407, 2000.

[12] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[13] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[14] Alexander Grubb and J. Andrew Bagnell. Generalized boosting algorithms for convex optimization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1209–1216, Madison, WI, USA, 2011. Omnipress.

[15] Slavomír Hanzely. Sketch-and-project meets Newton method: Global $O(1/k^2)$ convergence with low-rank updates. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3205–3213. PMLR, 03–05 May 2025.

[16] Slavomír Hanzely, Dmitry Kamzolov, Dmitry Pasechnyuk, Alexander Gasnikov, Peter Richtárik, and Martin Takáč. A damped Newton method achieves global $\mathcal{O}\left(\frac{1}{k^2}\right)$ and local quadratic convergence rate. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25320–25334. Curran Associates, Inc., 2022.

[17] Florian Jarre and Philippe L. Toint. Simple examples for the failure of Newton's method with line search for strictly convex minimization. *Mathematical Programming*, 158(1):23–34, Jul 2016. ISSN 1436-4646. doi: 10.1007/s10107-015-0913-2.

[18] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer International Publishing, 2016.

[19] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Haihao Lu and Rahul Mazumder. Randomized gradient boosting machine. *SIAM Journal on Optimization*, 30(4):2780–2808, 2020. doi: 10.1137/18M1223277.

[21] Haihao Lu, Sai Praneeth Karimireddy, Natalia Ponomareva, and Vahab Mirrokni. Accelerating gradient boosting machines. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 516–526. PMLR, 26–28 Aug 2020.

[22] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[23] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C., Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[24] Konstantin Mishchenko. Regularized Newton method with global $\mathcal{O}(1/k^2)$ convergence. *SIAM Journal on Optimization*, 33(3):1440–1462, 2023. doi: 10.1137/22M1488752.

[25] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, Mar 2008. ISSN 1436-4646. doi: 10.1007/s10107-006-0089-x.

[26] Yurii Nesterov. *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.

[27] Yurii Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, Aug 2006. ISSN 1436-4646. doi: 10.1007/s10107-006-0706-8.

[28] Atsushi Nitanda and Taiji Suzuki. Functional gradient boosting based on residual network perception. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3819–3828. PMLR, 10–15 Jul 2018.

[29] Thomas Parnell, Andreea Anghel, Małgorzata Łazuka, Nikolas Ioannou, Sebastian Kurella, Peshal Agarwal, Nikolaos Papandreou, and Haralampos Pozidis. SnapBoost: A heterogeneous boosting machine. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11166–11177. Curran Associates, Inc., 2020.

[30] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6639–6649, Red Hook, NY, USA, 2018. Curran Associates Inc.

[31] Zhen Qin, Le Yan, Honglei Zhuang, Yi Tay, Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, and Marc Najork. Are neural rankers still outperformed by gradient boosted decision trees? In *International Conference on Learning Representations*, 2021.

[32] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022. ISSN 1566-2535.

[33] Fabio Sigrist. Gradient and Newton boosting for classification and regression. *Expert Systems with Applications*, 167:114080, 2021. ISSN 0957-4174.

[34] Arun Suggala, Bingbin Liu, and Pradeep Ravikumar. Generalized boosting. In *Advances in Neural Information Processing Systems*, volume 33, pages 8787–8797. Curran Associates, Inc., 2020.

[35] Peng Sun, Tong Zhang, and Jie Zhou. A convergence rate analysis for LogitBoost, MART and their variant. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1251–1259, Bejing, China, 22–24 Jun 2014. PMLR.

[36] Matus Telgarsky. A primal-dual convergence analysis of boosting. *Journal of Machine Learning Research*, 13(20):561–606, 2012.

[37] Daniel Whiteson. HIGGS. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5V312.

[38] Nikita Zozoulenko, Thomas Cass, and Lukas Gonon. Random feature representation boosting. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80900–80932. PMLR, 13–19 Jul 2025.

A Discussions on Losses and Gradient Boosting Trees

In this appendix, we provide explicit calculations for the gradients and Hessians of common loss functions used in boosting. We also provide the full derivation of the optimal weight update for decision trees within the Restricted Newton Descent framework, recovering the standard XGBoost formula.

A.1 Common loss functions used in boosting

We denote the prediction for the i-th data point as $u_i = F(x_i)$. The gradient and Hessian with respect to the prediction are denoted by $g_{(i)} \in \mathbb{R}^K$ and $h_{(i)} \in \mathbb{R}^{K\times K}$, respectively.

Mean Squared Error (MSE): For the quadratic loss $l(\widehat{y}, y) = \frac{1}{2}(\widehat{y} - y)^2$, the gradient and Hessian are given by:

$$\begin{aligned} g_{(i)} &= \widehat{y}_i - y_i, \\ h_{(i)} &= 1. \end{aligned}$$

Binary Cross Entropy (BCE): For binary classification, let $y_i \in \{0, 1\}$ be the label and $u_i \in \mathbb{R}$ be the logit. The loss is defined as $l(u_i, y_i) = -[y_i \log \sigma(u_i) + (1 - y_i) \log(1 - \sigma(u_i))]$, where $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function. Using the chain rule, one finds that the derivatives are

$$\begin{aligned} g_{(i)} &= \sigma(u_i) - y_i, \\ h_{(i)} &= \sigma(u_i)(1 - \sigma(u_i)). \end{aligned}$$

Categorical Cross Entropy (CCE): In the multi-class setting ($K > 2$), let $y_i \in \{0, 1\}^K$ be the one-hot encoded label and $u_i \in \mathbb{R}^K$ be the vector of logits. The loss is $l(u_i, y_i) = -\sum_{k=1}^K (y_i)_k \log(p_i)_k$, where $p_i = \mathrm{softmax}(u_i)$ is the vector of class probabilities. The gradient and Hessian are

$$\begin{aligned} g_{(i)} &= p_i - y_i, \\ h_{(i)} &= \mathrm{diag}(p_i) - p_i p_i^\top, \end{aligned}$$

where $\mathrm{diag}(p_i)$ is the diagonal matrix with elements of p_i on the diagonal, and $p_i p_i^\top$ denotes the vector outer product.

A.2 Newton Tree Boosting Derivation

Let $f(x)$ be a decision tree defined by $f(x) = \sum_j w_j I_j(x)$, where $\{I_j\}_{j=1}^L$ is a disjoint partition of the feature space (i.e., the leaves of the tree) and $w_j \in \mathbb{R}^K$ are the leaf values. The second-order approximation of the loss, $Q(f)$, is given by

$$\begin{aligned} Q(f) &= \frac{1}{N} \sum_{i=1}^N \left\langle f(x_i), g_{(i)} + \frac{1}{2} h_{(i)} f(x_i) \right\rangle_{\mathbb{R}^K} \\ &= \frac{1}{N} \sum_{i=1}^N \left\langle f(x_i), g_{(i)} \right\rangle_{\mathbb{R}^K} + \frac{1}{2N} \sum_{i=1}^N \left\langle f(x_i), h_{(i)} f(x_i) \right\rangle_{\mathbb{R}^K}. \end{aligned}$$

We analyze the linear and quadratic terms separately. Substituting the definition of the tree $f(x)$, we find that

$$\begin{aligned} \sum_{i=1}^N \left\langle f(x_i), g_{(i)} \right\rangle_{\mathbb{R}^K} &= \sum_{i=1}^N \left\langle \sum_j w_j I_j(x_i), g_{(i)} \right\rangle_{\mathbb{R}^K} \\ &= \sum_j \left\langle w_j, \sum_{i=1}^N I_j(x_i) g_{(i)} \right\rangle_{\mathbb{R}^K} \\ &= \sum_j \left\langle w_j, \sum_{i \in I_j} g_{(i)} \right\rangle_{\mathbb{R}^K}. \end{aligned}$$

For the quadratic term, we use the property that the regions are disjoint, implying $I_j(x)I_k(x) = \delta_{jk}I_j(x)$, where δ_{jk} is the Kronecker delta.

$$\begin{aligned}
\sum_{i=1}^{N} \left\langle f(x_i), h_{(i)} f(x_i) \right\rangle_{\mathbb{R}^K} &= \sum_{i=1}^{N} \left\langle \sum_j w_j I_j(x_i), h_{(i)} \sum_k w_k I_k(x_i) \right\rangle_{\mathbb{R}^K} \\
&= \sum_{i=1}^{N} \sum_j \sum_k I_j(x_i) I_k(x_i) \left\langle w_j, h_{(i)} w_k \right\rangle_{\mathbb{R}^K} \\
&= \sum_{i=1}^{N} \sum_j I_j(x_i) \left\langle w_j, h_{(i)} w_j \right\rangle_{\mathbb{R}^K} \\
&= \sum_j \sum_{i \in I_j} \left\langle w_j, h_{(i)} w_j \right\rangle_{\mathbb{R}^K} \\
&= \sum_j \left\langle w_j, \left(\sum_{i \in I_j} h_{(i)} \right) w_j \right\rangle_{\mathbb{R}^K}.
\end{aligned}$$

The total objective decomposes into a sum of independent objectives for each leaf j. Differentiating with respect to a specific weight w_j, we obtain the local quadratic

$$q(w_j) = \left\langle w_j, \sum_{i \in I_j} g_{(i)} \right\rangle + \frac{1}{2} \left\langle w_j, \left(\sum_{i \in I_j} h_{(i)} \right) w_j \right\rangle.$$

Taking the gradient with respect to w_j yields

$$\nabla q(w_j) = \sum_{i \in I_j} g_{(i)} + \left(\sum_{i \in I_j} h_{(i)} \right) w_j.$$

Setting the gradient to zero gives the optimal weight update

$$w_j^* = - \left(\sum_{i \in I_j} h_{(i)} \right)^{-1} \sum_{i \in I_j} g_{(i)}.$$

Substituting w_j^* back into the objective (and defining $G_j = \sum_{i \in I_j} g_{(i)}$ and $H_j = \sum_{i \in I_j} h_{(i)}$), we recover the standard score used in Newton boosting and XGBoost:

$$Q(f^*) = \frac{1}{N} \sum_j q(w_j^*) = -\frac{1}{2N} \sum_j \left\langle H_j^{-1} G_j, G_j \right\rangle_{\mathbb{R}^K}.$$

Remark A.1. The ℓ_2-regularization used in standard implementations like XGBoost and LightGBM differs slightly from the one suggested by our theory. These libraries add a fixed scalar λ to the sum of sample Hessians within each leaf, yielding the update

$$w_j^* = -(H_j + \lambda I)^{-1} G_j.$$

By contrast, our framework regularizes the weak learner in the Hilbert space $L^2(\widehat{\nu}_N)$. In this space, the regularization modifies the Hessian operator in the quadratic model pointwise, which is equivalent to replacing each $h_{(i)}$ with $h_{(i)} + \lambda I$, giving the update

$$w_j^* = -(H_j + |I_j| \lambda I)^{-1} G_j.$$

Thus, in our formulation of boosting the theoretically consistent regularization scales proportionally with the number of samples routed to each leaf. From here, there are two ways to proceed. One can regularize only the quadratic step, $Q_k(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$, or instead regularize the loss itself, $L_\mu(F) = L(F) + \frac{\mu}{2}\|F\|_{\mathcal{H}}^2$. The latter modifies both the sample gradients and Hessians:

$$g_{(i)}^\mu = g_{(i)} + \mu F_k(x_i), \qquad h_{(i)}^\mu = h_{(i)} + \mu I.$$

If the GRN boosting scheme from Section 4 is applied to this regularized loss, its adaptive regularization must also use the modified gradient $g_k^\mu = g_k + \mu F_k$, giving $\lambda_k = \sqrt{M\|g_k^\mu\|}$. Our numerical experiments in Section 6 use the former approach: we regularize only the quadratic step, as is standard in boosting libraries, but with the leaf-size scaling described above. However, if regularization is intended to make the loss strongly convex in the space $L^2(\widehat{\nu}_N)$, it must instead be applied to the loss itself.

B Some Classical Optimization Results

We collect several standard smoothness and Lipschitz-Hessian definitions and results from optimization theory [26, 27], and present them in the Hilbert space setting.

Definition B.1. A function $L : \mathcal{H} \to \mathbb{R}$ is said to be *Fréchet differentiable* if there exists a mapping $\nabla L : \mathcal{H} \to \mathcal{H}$ such that for all $h, f \in \mathcal{H}$

$$L(h + f) = L(h) + \langle f, \nabla L(h) \rangle + o(\|f\|).$$

The element $\nabla L(h)$ is called the *gradient* at h. A function $L : \mathcal{H} \to \mathbb{R}$ is said to be *twice Fréchet differentiable* if L is Fréchet differentiable, and if for all $h \in \mathcal{H}$ there exists a bounded, self-adjoint linear operator $\nabla^2 L(h) : \mathcal{H} \to \mathcal{H}$ such that

$$L(h + f) = L(h) + \left\langle f, \nabla L(h) + \frac{1}{2}\nabla^2 L(h)[f] \right\rangle + o(\|f\|^2),$$

for all $f \in \mathcal{H}$. The map $\nabla^2 L(h)$ is called the *Fréchet Hessian* (or second derivative) of L at h.

Definition B.2 (S-smooth loss). Let $\mathcal{H}$ be a real Hilbert space. A Fréchet differentiable function $L : \mathcal{H} \to \mathbb{R}$ is called S-*smooth* if its gradient is Lipschitz continuous with constant S, i.e.,

$$\|\nabla L(f) - \nabla L(g)\| \leq S\|f - g\| \quad \forall f, g \in \mathcal{H}.$$

Equivalently, if L is twice Fréchet differentiable, then its Hessian satisfies

$$\|\nabla^2 L(h)\|_{\mathrm{op}} \leq S \quad \forall h \in \mathcal{H},$$

For convex L, this is equivalent to $\nabla^2 L(h) \preceq SI$.

Definition B.3 (μ-strongly convex loss). A function $L : \mathcal{H} \to \mathbb{R}$ is μ-*strongly convex* for some $\mu > 0$ if one of the following equivalent conditions holds:

1. **(Quadratic lower bound):** The function $h \mapsto L(h) - \frac{\mu}{2}\|h\|^2$ is convex.
2. **(First-order condition):** For all $f, g \in \mathcal{H}$,
$$L(f) \geq L(g) + \langle \nabla L(g), f - g \rangle + \frac{\mu}{2}\|f - g\|^2.$$
3. **(Gradient monotonicity):** For all $f, g \in \mathcal{H}$,
$$\langle \nabla L(f) - \nabla L(g), f - g \rangle \geq \mu\|f - g\|^2.$$
4. **(Second-order condition):** If L is twice Fréchet differentiable, then
$$\nabla^2 L(h) \succeq \mu I \quad \forall h \in \mathcal{H}.$$

Definition B.4. The Fréchet Hessian $\nabla^2 L : \mathcal{H} \to \mathcal{L}(\mathcal{H})$ is $2M$-Lipschitz continuous if there exists a constant $M > 0$ such that

$$\|\nabla^2 L(f) - \nabla^2 L(g)\| \leq 2M\|f - g\| \qquad \forall f, g \in \mathcal{H}.$$

Lemma B.5. If the loss L is S-smooth, then

$$L(F_k + f) \leq L(F_k) + \langle g_k, f \rangle + \frac{S}{2}\|f\|^2. \tag{16}$$

Proof. Using Taylor's theorem with integral remainder, we express the expansion of the loss as

$$L(F_k + f) = L(F_k) + \langle g_k, f \rangle + \int_0^1 (1 - t)\langle f, \nabla^2 L(F_k + tf) f \rangle \, dt. \tag{17}$$

The definition of S-smoothness implies that the Hessian is bounded in operator norm by S, i.e., $\nabla^2 L(\cdot) \preceq SI$. Consequently, the quadratic form is bounded by $\langle f, \nabla^2 L(\cdot) f \rangle \leq S\|f\|^2$. Substituting this into the integral yields $\int_0^1 (1 - t)S\|f\|^2 \, dt = \frac{S}{2}\|f\|^2$, completing the proof. □

Lemma B.6. If $\nabla^2 L$ is $2M$-Lipschitz, then for all $f, g \in \mathcal{H}$

1. $L(f) \leq L(g) + \langle \nabla L(g), f - g \rangle + \frac{1}{2}\langle \nabla^2 L(g)(f-g), f-g \rangle + \frac{M}{3}\|f-g\|^3$
2. $\|\nabla L(f) - \nabla L(g) - \nabla^2 L(g)(f-g)\| \leq M\|g-f\|^2$.

Proof. Let $h = f - g$. We begin with the Fundamental Theorem of Calculus and expand the terms:

$$\begin{aligned}
L(f) - L(g) - \langle \nabla L(g), h \rangle &= \int_0^1 \langle \nabla L(g+th) - \nabla L(g), h \rangle \, dt \\
&= \int_0^1 \left\langle \int_0^t \nabla^2 L(g+sh)[h] \, ds, h \right\rangle dt \\
&= \int_0^1 \int_0^t \langle h, \nabla^2 L(g+sh)[h] \rangle \, ds \, dt \\
&= \int_0^1 \int_0^t \langle h, \nabla^2 L(g)[h] \rangle \, ds \, dt + \int_0^1 \int_0^t \langle h, (\nabla^2 L(g+sh) - \nabla^2 L(g))[h] \rangle \, ds \, dt \\
&= \frac{1}{2}\langle h, \nabla^2 L(g)[h] \rangle + \int_0^1 \int_0^t \langle h, (\nabla^2 L(g+sh) - \nabla^2 L(g))[h] \rangle \, ds \, dt \\
&\leq \frac{1}{2}\langle h, \nabla^2 L(g)[h] \rangle + \int_0^1 \int_0^t \|h\| \cdot \|(\nabla^2 L(g+sh) - \nabla^2 L(g))[h]\| \, ds \, dt \\
&\leq \frac{1}{2}\langle h, \nabla^2 L(g)[h] \rangle + \int_0^1 \int_0^t \|\nabla^2 L(g+sh) - \nabla^2 L(g)\|_{op} \|h\|^2 \, ds \, dt \\
&\leq \frac{1}{2}\langle h, \nabla^2 L(g)[h] \rangle + \int_0^1 \int_0^t (2Ms\|h\|)\|h\|^2 \, ds \, dt \\
&= \frac{1}{2}\langle h, \nabla^2 L(g)[h] \rangle + 2M\|h\|^3 \int_0^1 \frac{t^2}{2} \, dt \\
&= \frac{1}{2}\langle h, \nabla^2 L(g)[h] \rangle + \frac{M}{3}\|h\|^3.
\end{aligned}$$

Rearranging the terms yields the desired first inequality. For inequality 2, we express the gradient difference as an integral of the Hessian.

$$\begin{aligned}
\|\nabla L(f) - \nabla L(g) - \nabla^2 L(g)[h]\| &= \left\| \int_0^1 \nabla^2 L(g+th)[h] \, dt - \int_0^1 \nabla^2 L(g)[h] \, dt \right\| \\
&= \left\| \int_0^1 (\nabla^2 L(g+th) - \nabla^2 L(g))[h] \, dt \right\| \\
&\leq \int_0^1 \|(\nabla^2 L(g+th) - \nabla^2 L(g))[h]\| \, dt \\
&\leq \int_0^1 \|\nabla^2 L(g+th) - \nabla^2 L(g)\|_{op} \|h\| \, dt \\
&\leq \int_0^1 (2Mt\|h\|)\|h\| \, dt \\
&= 2M\|h\|^2 \int_0^1 t \, dt \\
&= M\|h\|^2.
\end{aligned}$$

This completes the proof. □

C Convergence Rate Proofs and Related Technical Lemmas

In this Appendix, we gather all proofs of our theoretical results presented in the main paper body.

C.1 Vanilla Restricted Newton Proofs

Proof of Proposition 3.5. Since $\mathcal{F}$ is closed under scalar multiplication, the objective $Q_k(\alpha f_{k+1}^w)$ is minimized at $\alpha = 1$. Differentiating the equivalent objective (see Lemma 3.3) $\phi(\alpha) := \frac{1}{2}\|\alpha f_{k+1}^w - f_{k+1}\|_{H_k}^2$ with respect to α gives the derivative $\phi'(\alpha) = \langle f_{k+1}^w, \alpha f_{k+1}^w - f_{k+1}\rangle_{H_k}$. Setting this to zero at $\alpha = 1$ yields $\langle f_{k+1}, f_{k+1}^w\rangle_{H_k} = \|f_{k+1}^w\|_{H_k}^2$. Substituting this into the definition of Θ_k, we obtain

$$\Theta_k = \frac{\|f_{k+1}^w\|_{H_k}^2}{\|f_{k+1}\|_{H_k}\|f_{k+1}^w\|_{H_k}} = \frac{\|f_{k+1}^w\|_{H_k}}{\|f_{k+1}\|_{H_k}}, \tag{18}$$

which gives the first equality in the second item. The first item then follows immediately from expanding the norm using inner products. The remaining equality in the second item follows from $-\langle g_k, f_{k+1}^w\rangle = \langle H_k f_{k+1}, f_{k+1}^w\rangle = \langle f_{k+1}, f_{k+1}^w\rangle_{H_k}$. □

Proof of Theorem 3.7. Starting with the standard smoothness upper bound $L(F + f) \leq L(F) + \langle g, f\rangle + \frac{S}{2}\|f\|^2$ from Lemma B.5, we substitute the weak update $f = \eta f_{k+1}^w$. Using strong convexity to write $\|u\|^2 \leq \frac{1}{\mu}\|u\|_{H_k}^2$, and the cosine angle properties from Proposition 3.5, we obtain

$$L(F_k) - L(F_{k+1}) \geq -\eta\langle g_k, f_{k+1}^w\rangle - \frac{S\eta^2}{2}\|f_{k+1}^w\|^2 \tag{19}$$

$$\geq \eta\|f_{k+1}^w\|_{H_k}^2 - \frac{S\eta^2}{2\mu}\|f_{k+1}^w\|_{H_k}^2 \tag{20}$$

$$= \left(\eta - \frac{S\eta^2}{2\mu}\right)\Theta_k^2\|f_{k+1}\|_{H_k}^2. \tag{21}$$

Using $\Theta_k \geq \Theta$ and applying Definition 3.6, we obtain

$$\begin{aligned} L(F_k) - L(F_{k+1}) &\geq \eta\left(1 - \frac{\eta S}{2\mu}\right)\Theta^2\|f_{k+1}\|_{H_k}^2 \\ &\geq \rho\big(L(F_k) - L_\star\big). \end{aligned}$$

Rearranging and subtracting $L_\star$ gives

$$L(F_{k+1}) - L_\star \leq (1 - \rho)\big(L(F_k) - L_\star\big).$$

Iterating yields

$$L(F_k) - L_\star \leq (1 - \rho)^k\big(L(F_0) - L_\star\big),$$

proving linear convergence of the optimality gap. □

C.2 Properties of Gradient Regularization

We begin by establishing some basic properties the gradient regularized scheme. In the subsequent analysis, define $K_k := H_k + \lambda_k I$.

Lemma C.1. The exact regularized update (9) satisfies

$$\lambda_k f_{k+1} = -\left(g_k + H_k[f_{k+1}]\right). \tag{22}$$

Proof. Multiply the Newton update (9) by $(H_k + \lambda_k I)$. □

Lemma C.2. If L is convex, then for all $u \in \mathcal{H}$

$$\begin{aligned} \|u\|_{K_k}^2 &:= \langle (H_k + \lambda_k I)u, u\rangle \\ &= \lambda_k\|u\|^2 + \langle u, H_k u\rangle \\ &\geq \lambda_k\|u\|^2. \end{aligned}$$

Furthermore, if the weak learner family $\mathcal{F}$ is closed under scalar multiplication, then the iterates of Algorithm 2 satisfy

1. $\|f_{k+1}^w\|_{K_k}^2 = -\langle g_k, f_{k+1}^w \rangle$,
2. $\lambda_k \|f_{k+1}^w\| \leq \|g_k\|$.

Proof. The first result follows by convexity. For the rest, let

$$\phi(t) := Q_k^{reg}(t f_{k+1}^w) = \frac{1}{2} t^2 \|f_{k+1}^w\|_{K_k}^2 + t\langle g_k, f_{k+1}^w \rangle.$$

Since $\mathcal{F}$ is scalable, ϕ is minimized at $t = 1$. Therefore, $0 = \phi'(1) = \|f_{k+1}^w\|_{K_k}^2 + \langle g_k, f_{k+1}^w \rangle$. As for the last inequality, this follows by the first inequality and Cauchy–Schwarz:

$$\lambda_k \|f_{k+1}^w\|^2 \leq \|f_{k+1}^w\|_{K_k}^2 = -\langle g_k, f_{k+1}^w \rangle \leq \|g_k\| \|f_{k+1}^w\|.$$

□

Proof of Proposition 4.2. Write $s = f_{k+1}^w$. By Lemmas B.6 and C.2,

$$\begin{aligned} L(F_{k+1}) - L(F_k) &\leq \eta \langle g_k, s \rangle + \frac{\eta^2}{2} \langle H_k s, s \rangle + \frac{M\eta^3}{3} \|s\|^3 \\ &= -\left(\eta - \frac{\eta^2}{2}\right) \|s\|_{K_k}^2 - \frac{\eta^2}{2} \lambda_k \|s\|^2 + \frac{M\eta^3}{3} \|s\|^3. \end{aligned}$$

Since $\lambda_k \|s\| \leq \|g_k\|$ and $\lambda_k^2 \geq M\|g_k\|$, we have

$$M\|s\|^3 \leq \frac{M\|g_k\|}{\lambda_k} \|s\|^2 \leq \lambda_k \|s\|^2.$$

Consequently,

$$\begin{aligned} L(F_k) - L(F_{k+1}) &\geq \left(\eta - \frac{\eta^2}{2}\right) \|s\|_{K_k}^2 + \left(\frac{\eta^2}{2} - \frac{\eta^3}{3}\right) \lambda_k \|s\|^2 \\ &\geq \left(\eta - \frac{\eta^3}{3}\right) \lambda_k \|s\|^2, \end{aligned}$$

where the last inequality uses $\|s\|_{K_k}^2 \geq \lambda_k \|s\|^2$ and $\eta - \eta^2/2 \geq 0$ for $\eta \in (0, 2]$. □

For ease of exposition, we restate some results of the relevant lemmas and theorems from the main body of the paper.

Lemma C.3. If the iterates of Algorithm 2 have weak gradient edge γ, with learning rate $\eta \in (0, 1]$, then the gradient growth is bounded by

$$\|g_{k+1}\| \leq \left(1 - \eta(1 - \sqrt{1 - \gamma^2})\right) \|g_k\| + (\eta + \eta^2) \lambda_k \|f_{k+1}^w\| \tag{23}$$

$$\leq \left(1 + \eta^2 + \eta\sqrt{1 - \gamma^2}\right) \|g_k\|. \tag{24}$$

Proof. Using Eq. (11), and Lemmas B.6 and C.1, we can write

$$g_{k+1} = g_k + \eta H_k f_{k+1}^w + E \tag{25}$$

$$= g_k + \eta(-g_k^w - \lambda_k f_{k+1}^w) + E \tag{26}$$

$$= (1 - \eta) g_k + \eta(g_k - g_k^w) - \eta \lambda_k f_{k+1}^w + E \tag{27}$$

where $\|E\| \leq M\eta^2 \|f_{k+1}^w\|^2$. Hence using the bounds in Lemma C.2 we obtain that

$$\|g_{k+1}\| = \|(1 - \eta) g_k + \eta(g_k - g_k^w) - \eta \lambda_k f_{k+1}^w + E\| \tag{28}$$

$$\leq (1 - \eta)\|g_k\| + \eta\|g_k - g_k^w\| + \eta\lambda_k \|f_{k+1}^w\| + M\eta^2 \|f_{k+1}^w\|^2 \tag{29}$$

$$\leq (1 - \eta)\|g_k\| + \eta\sqrt{1 - \gamma^2}\|g_k\| + \eta\lambda_k \|f_{k+1}^w\| + \eta^2 \lambda_k \|f_{k+1}^w\| \tag{30}$$

$$= \left(1 - \eta(1 - \sqrt{1 - \gamma^2})\right) \|g_k\| + (\eta + \eta^2) \lambda_k \|f_{k+1}^w\| \tag{31}$$

$$\leq \left(1 + \eta^2 + \eta\sqrt{1 - \gamma^2}\right) \|g_k\|. \tag{32}$$

□

C.3 Proof of Global Convergence for the Gradient Regularized Scheme with Weak Learners

The following global convergence proof of the gradient regularized Newton scheme generalizes the proof of Mishchenko [24] to the setting of Restricted Newton Descent with a constant learning rate $\eta \in (0, 1]$ and weak learners. The primary challenge lies in handling the additional error terms introduced by the learning rate and the inexact weak iterates f_{k+1}^w, which we handled in our previous Lemmas in this paper. Additionally, we adapt the original proof's fixed subsequence condition and fraction to depend on both the learning rate η and weak gradient edge γ.

Theorem C.4. Assume L is convex with $2M$-Lipschitz Hessian, a bounded initial sublevel set with $\|F - F^*\| \leq \Lambda$, and weak gradient edge γ. Let the regularization be chosen as $\lambda_k = C\sqrt{M\|g_k\|}$ for some $C \geq 1$. For any learning rate $\eta \in (0, 1]$, the gradient regularized Newton scheme has the global convergence rate

$$L(F_k) - L(F^*) \leq \max\left\{A(k, \Lambda, \eta, \gamma, g_0), \quad B(k, \Lambda, \eta, \gamma, M, C)\right\}. \tag{33}$$

where

$$A(k, \Lambda, \eta, \gamma, g_0) = \underbrace{\Lambda\|g_0\| \exp\left(-\frac{k\eta}{8}(1 - \sqrt{1-\gamma^2})(3 - \sqrt{1-\gamma^2} - 2\eta)\right)}_{\text{Linear Regime}}, \tag{34}$$

and

$$B(k, \Lambda, \eta, \gamma, M, C) = \underbrace{\frac{1024C^2M\Lambda^3}{\eta^2(1 - \sqrt{1-\gamma^2})^6(2+k)^2}}_{\mathcal{O}\left(\frac{1}{k^2}\right) \text{ Regime}} \frac{(1+\eta)^4}{(1 - \frac{1}{3}\eta^2)^2}. \tag{35}$$

Proof. By Proposition 4.2 we have that $L(F_k) \leq L(F_{k-1}) \leq ... \leq L(F_0)$, implying that the iterates F_k remain within the sublevel set defined by F_0. Consequently, $\|F_k - F^*\| \leq \Lambda$ for all k. By convexity we have the bound

$$L(F_k) - L(F^*) \leq \langle g_k, F_k - F^* \rangle \leq \|g_k\|\Lambda. \tag{36}$$

Let $\mathcal{I}_\infty := \{t \in \mathbb{Z}_{\geq 0} : \|g_{t+1}\| \geq (1 - \eta a)\|g_t\|\}$ be the set of indices where the gradient does not contract sufficiently, for some $0 < a < 1$ to be specified later. Consider first any $k \in \mathcal{I}_\infty$. It follows from Lemma C.3 that

$$(1 - \eta a)\|g_k\| \leq \|g_{k+1}\| \leq \left(1 - \eta(1 - \sqrt{1-\gamma^2})\right)\|g_k\| + (\eta + \eta^2)\lambda_k\|f_{k+1}^w\| \tag{37}$$

Rearranging the terms we obtain that

$$\lambda_k\|f_{k+1}^w\| \geq \frac{1 - \sqrt{1-\gamma^2} - a}{1+\eta}\|g_k\| \tag{38}$$

For the next argument we require the bound (38) to be positive, hence we choose

$$a := \frac{1 - \sqrt{1-\gamma^2}}{2}. \tag{39}$$

By (38) and Proposition 4.2 we have

$$L(F_k) - L(F_{k+1}) \geq \left(1 - \frac{1}{3}\eta^2\right)\eta\lambda_k\|f_{k+1}^w\|^2 \tag{40}$$

$$\geq \left(1 - \frac{1}{3}\eta^2\right)a^2\frac{\eta}{(1+\eta)^2}\frac{\|g_k\|^2}{\lambda_k} \tag{41}$$

$$= \left(1 - \frac{1}{3}\eta^2\right)a^2\frac{\eta}{(1+\eta)^2}\frac{\|g_k\|^{\frac{3}{2}}}{C\sqrt{M}} \tag{42}$$

$$\geq \left(1 - \frac{1}{3}\eta^2\right)\frac{a^2}{C\sqrt{M\Lambda^3}}\frac{\eta}{(1+\eta)^2}\Big(L(F_k) - L(F^*)\Big)^{\frac{3}{2}} \tag{43}$$

$$= \tau\Big(L(F_k) - L(F^*)\Big)^{\frac{3}{2}} \tag{44}$$

where $\tau = \left(1 - \frac{1}{3}\eta^2\right) \frac{a^2}{C\sqrt{M\Lambda^3}} \frac{\eta}{(1+\eta)^2}$. If this recursion held at every iteration, the $\mathcal{O}(\frac{1}{k^2})$ rate would follow by Lemma C.5. Instead, we use the subsequence technique of [24].

Fix any $k \geq 1$ and define

$$\mathcal{I}_k := \mathcal{I}_\infty \cap \{0, \dots, k-1\}, \qquad m := |\mathcal{I}_k|.$$

We consider the cases $m \geq pk$ and $m < pk$, for some fraction $0 < p < 1$ to be specified later.

Case $|\mathcal{I}_k| \geq pk$: Here $m \geq 1$. Enumerate the selected indices as $t_0 < \cdots < t_{m-1}$ and define

$$\alpha_j := L(F_{t_j}) - L(F^*) \quad (0 \leq j < m), \qquad \alpha_m := L(F_k) - L(F^*).$$

By monotonicity of the loss and the decrease bound at each selected index, for $0 \leq j < m$ we have

$$\begin{aligned} \alpha_{j+1} &\leq L(F_{t_j+1}) - L(F^*) && (45)\\ &= \alpha_j + L(F_{t_j+1}) - L(F_{t_j}) && (46)\\ &\leq \alpha_j - \tau\alpha_j^{3/2}. && (47) \end{aligned}$$

If $\alpha_m = 0$, the desired bound is immediate. Otherwise, applying Lemma C.5 to these m steps gives

$$\begin{aligned} L(F_k) - L(F^*) &= \alpha_m && (48)\\ &\leq \frac{4}{\tau^2(m+2)^2} && (49)\\ &\leq \frac{4}{\tau^2 p^2 (k+2)^2} && (50)\\ &= \frac{4C^2 M\Lambda^3}{\eta^2 a^4 p^2 (k+2)^2} \frac{(1+\eta)^4}{(1-\frac{1}{3}\eta^2)^2}. && (51) \end{aligned}$$

The second inequality uses $m \geq pk$ and $p < 1$, which imply $m + 2 \geq p(k+2)$. Consequently, this case gives a global $\mathcal{O}(\frac{1}{k^2})$ rate.

Case $|\mathcal{I}_k| < pk$: For indices $t \in \{0, \dots, k-1\} \setminus \mathcal{I}_k$, we have the contraction $\|g_{t+1}\| \leq (1 - \eta a)\|g_t\|$. For $t \in \mathcal{I}_k$ the gradient may grow, but it is bounded by Lemma C.3 via $\|g_{t+1}\| \leq \left(1 + \eta^2 + \eta\sqrt{1-\gamma^2}\right)\|g_t\|$. Combining both we obtain that

$$\begin{aligned} \|g_k\| &\leq \left(1 + \eta^2 + \eta\sqrt{1-\gamma^2}\right)^{|\mathcal{I}_k|} (1 - \eta a)^{k - |\mathcal{I}_k|} \|g_0\| && (52)\\ &\leq \left(1 + \eta^2 + \eta\sqrt{1-\gamma^2}\right)^{pk} (1 - \eta a)^{(1-p)k} \|g_0\|. && (53) \end{aligned}$$

Using the bound $1 + x \leq e^x$ and taking logarithms, we see that this scheme produces a contraction if

$$p\left(\eta^2 + \eta\sqrt{1-\gamma^2}\right) - (1-p)\eta a < 0. \tag{54}$$

Rearranging the above, we find that we require

$$p < \frac{a}{\eta + a + \sqrt{1-\gamma^2}}. \tag{55}$$

Substituting our choice $a = (1 - \sqrt{1-\gamma^2})/2$, the denominator simplifies to $\eta + (1 + \sqrt{1-\gamma^2})/2$. Since $\sqrt{1-\gamma^2} \leq 1$, this denominator is strictly less than $\eta + 1$. Therefore, for any $\eta \in (0, 1]$, the choice $p = \frac{a}{2}$ is sufficient. In general, one can show that a and p need to satisfy the inequality

$$\frac{p\left(\eta + \sqrt{1-\gamma^2}\right)}{1-p} < a < 1 - \sqrt{1-\gamma^2}. \tag{56}$$

for the proof in both cases above to be valid. With the choice $4p = 2a = 1 - \sqrt{1-\gamma^2}$, one finds through elementary algebra and the bound $1 + x \leq e^x$ that the gradient growth can be bounded by

$$\|g_k\| \leq \exp\left(-\frac{k\eta}{8}(1 - \sqrt{1-\gamma^2})(3 - \sqrt{1-\gamma^2} - 2\eta)\right)\|g_0\|, \tag{57}$$

yielding a linear rate of convergence.

Final bound: The final convergence rate is obtained by taking the maximum bound of the two cases above. Note that $1 - \sqrt{1-\gamma^2} = \frac{\gamma^2}{1+\sqrt{1-\gamma^2}} \geq \frac{\gamma^2}{2}$. Hence the global convergence rate in the first case is

$$\mathcal{O}\left(\frac{C^2 M \Lambda^3}{\eta^2 k^2 \gamma^{12}}\right), \tag{58}$$

while for the second case this is

$$\mathcal{O}\left(\exp\left(-k\eta\gamma^2\left(\frac{\gamma^2}{2} + 2 - 2\eta\right)/16\right)\right). \tag{59}$$

This concludes the proof. □

The following lemma used in the proof of Theorem C.4 is a standard result in the convex optimization literature, see e.g. [27]. We include the general proof here for completeness

Lemma C.5. If α_k is a decreasing sequence satisfying $0 \leq \alpha_{k+1} \leq \alpha_k - c\alpha_k^{\frac{3}{2}}$ for some $c > 0$, then for all $k \geq 0$:

$$\alpha_k \leq \frac{1}{\left(\frac{1}{\sqrt{\alpha_0}} + \frac{c}{2}k\right)^2} \leq \frac{4}{c^2(k+2)^2}. \tag{60}$$

Proof. Using the hypothesis $\alpha_k - \alpha_{k+1} \geq c\alpha_k^{3/2}$, consider the increment

$$\frac{1}{\sqrt{\alpha_{k+1}}} - \frac{1}{\sqrt{\alpha_k}} = \frac{\sqrt{\alpha_k} - \sqrt{\alpha_{k+1}}}{\sqrt{\alpha_k \alpha_{k+1}}} \tag{61}$$

$$= \frac{\alpha_k - \alpha_{k+1}}{\sqrt{\alpha_k \alpha_{k+1}}(\sqrt{\alpha_k} + \sqrt{\alpha_{k+1}})} \tag{62}$$

$$\geq \frac{c\alpha_k^{3/2}}{\sqrt{\alpha_k \alpha_{k+1}}(\sqrt{\alpha_k} + \sqrt{\alpha_{k+1}})}. \tag{63}$$

Since the sequence is decreasing ($\alpha_{k+1} \leq \alpha_k$), we have $\sqrt{\alpha_{k+1}} \leq \sqrt{\alpha_k}$, hence

$$\frac{1}{\sqrt{\alpha_{k+1}}} - \frac{1}{\sqrt{\alpha_k}} \geq \frac{c\alpha_k^{3/2}}{\sqrt{\alpha_k^2}(\sqrt{\alpha_k} + \sqrt{\alpha_k})} = \frac{c}{2}. \tag{64}$$

Summing from $i = 0$ to $k - 1$ yields a telescoping sum, from which we obtain that

$$\frac{1}{\sqrt{\alpha_k}} \geq \frac{1}{\sqrt{\alpha_0}} + \frac{ck}{2}. \tag{65}$$

Inverting and squaring both sides yields the inequality

$$\alpha_k \leq \frac{1}{\left(\frac{1}{\sqrt{\alpha_0}} + \frac{c}{2}k\right)^2}. \tag{66}$$

The final bound follows from the fact that $c\sqrt{\alpha_k} \leq 1$, which follows directly from the hypothesis $0 \leq \alpha_{k+1} \leq \alpha_k - c\alpha_k^{\frac{3}{2}}$. □

C.4 Proof of Local Convergence for GRN with weak learners

Proof of Theorem 4.7. We start by using the triangle inequality and the definition of the weak gradient edge γ and write

$$\|g_k^w\| \leq \|g_k\| + \|g_k^w - g_k\| \tag{67}$$

$$\leq \left(1 + \sqrt{1-\gamma^2}\right)\|g_k\| \tag{68}$$

$$:= C_\gamma \|g_k\|. \tag{69}$$

Since we assumed μ-strong convexity, we have that

$$\|f_{k+1}^w\| = \|(H_k + \lambda_k I)^{-1} g_k^w\| \tag{70}$$

$$\leq \frac{1}{\mu}\|g_k^w\| \tag{71}$$

$$\leq \frac{C_\gamma}{\mu}\|g_k\|. \tag{72}$$

We combine the above with Lemmas C.2 and C.3, specifically Eq. (29), to obtain that

$$\|g_{k+1}\| \leq (1-\eta)\|g_k\| + \eta\|g_k - g_k^w\| + \eta\lambda_k\|f_{k+1}^w\| + M\eta^2\|f_{k+1}^w\|^2 \tag{73}$$

$$\leq (1 - 2\eta + \eta C_\gamma)\|g_k\| + \eta\lambda_k\|f_{k+1}^w\| + M\eta^2\|f_{k+1}^w\|^2 \tag{74}$$

$$\leq (1 - 2\eta + \eta C_\gamma)\|g_k\| + \sqrt{C^2 M\eta^2 \frac{C_\gamma^2}{\mu^2}}\|g_k\|^{\frac{3}{2}} + C^2 M\eta^2 \frac{C_\gamma^2}{\mu^2}\|g_k\|^2. \tag{75}$$

Write $\rho = (1 - 2\eta + \eta C_\gamma)$ and $K = C^2 M\eta^2 \frac{C_\gamma^2}{\mu^2}$. We seek the condition for a strict decrease $\|g_{k+1}\| < \|g_k\|$. Substituting the above and dividing by $\|g_k\|$ we obtain that

$$\sqrt{K\|g_k\|} + K\|g_k\| < 1 - \rho. \tag{76}$$

Let $\Delta = 1 - \rho = \eta(1 - \sqrt{1-\gamma^2})$. Since $\eta \leq 1$, we observe that $\Delta \leq 1$. Let $y = \sqrt{K\|g_k\|}$. The condition becomes $y + y^2 < \Delta$. A sufficient condition for this to hold is $y \leq \frac{\Delta}{2}$, because this implies $y + y^2 \leq \frac{\Delta}{2} + \frac{\Delta^2}{4} = \frac{\Delta(2+\Delta)}{4}$. Since $\Delta \leq 1$, we have $(2 + \Delta) \leq 3$, and thus $y + y^2 \leq \frac{3}{4}\Delta < \Delta$. Expanding the condition $y \leq \frac{\Delta}{2}$ yields

$$\sqrt{\frac{C^2 M\eta^2 C_\gamma^2}{\mu^2}\|g_k\|} \leq \frac{\eta(1 - \sqrt{1-\gamma^2})}{2}. \tag{77}$$

Canceling η and rearranging for $\sqrt{\|g_k\|}$, we obtain

$$\sqrt{\|g_k\|} \leq \frac{\mu}{2C\sqrt{M}} \frac{1 - \sqrt{1-\gamma^2}}{C_\gamma} = \frac{\mu}{2C\sqrt{M}} \frac{1 - \sqrt{1-\gamma^2}}{1 + \sqrt{1-\gamma^2}}. \tag{78}$$

Squaring both sides gives the bound in the hypothesis. Under this condition, the sum of the higher order terms is bounded by $\frac{3}{4}\Delta\|g_k\| = \frac{3}{4}(1-\rho)\|g_k\|$. Plugging this back into the gradient bound we obtain that

$$\|g_{k+1}\| \leq \rho\|g_k\| + \frac{3}{4}(1-\rho)\|g_k\| = \frac{3+\rho}{4}\|g_k\|. \tag{79}$$

The conclusion follows since

$$1 - \sqrt{1-\gamma^2} = \frac{\gamma^2}{1 + \sqrt{1-\gamma^2}} \geq \frac{\gamma^2}{2} \tag{80}$$

whence

$$\frac{3+\rho}{4} = 1 - \frac{\eta}{4}\left(1 - \sqrt{1-\gamma^2}\right) \leq 1 - \frac{\eta\gamma^2}{8} \tag{81}$$

Since this bound preserves the gradient threshold in the hypothesis, induction from k_0 gives the claimed contraction for every $k \geq k_0$. □

Remark C.6. If $\eta = 1$ and $\gamma = 1$, the coefficient of the linear $\|g_k\|$ term vanishes, allowing for a superlinear rate with exponent $\frac{3}{2}$, recovering the classical result of Mishchenko [24]. However, when $\eta < 1$ or $\gamma < 1$, this term is non-zero, and we are only able to prove a linear rate. We conjecture that superlinear rates are generally unattainable in Newton schemes involving weak iterates, suggesting an interesting direction for future research.

D Divergence of Vanilla Newton on Smooth Convex Losses

This appendix records explicit smooth convex objectives for which the classical (unregularized) Newton step may fail to converge globally, even though the loss satisfies the Lipschitz-Hessian assumption used throughout Section 4 and Section C. These examples show that a fixed damping (constant learning rate) does not, by itself, globalize Newton's method, and motivate the adaptive gradient regularization used in Algorithm 2.

D.1 Newton, damped Newton, and gradient regularized Newton

Definition D.1 (Newton-type updates in $\mathbb{R}^d$). Let $L : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and assume $\nabla^2 L(x) \succ 0$ at the current iterate.

- The (classical) Newton update is
$$N(x) := x - (\nabla^2 L(x))^{-1} \nabla L(x).$$
- Given a fixed learning rate $\eta > 0$, the damped Newton update is
$$N_\eta(x) := x - \eta(\nabla^2 L(x))^{-1} \nabla L(x).$$

Assume moreover that $\nabla^2 L$ is $2M$-Lipschitz in operator norm:

$$\|\nabla^2 L(u) - \nabla^2 L(v)\|_{\mathrm{op}} \le 2M\|u - v\| \qquad \forall u, v \in \mathbb{R}^d.$$

Then the (exact) gradient regularized Newton (GRN) update is

$$\mathrm{GRN}_{M,\eta}(x) := x - \eta\big(\nabla^2 L(x) + \lambda(x) I\big)^{-1} \nabla L(x), \qquad \lambda(x) := \sqrt{M\|\nabla L(x)\|}.$$

D.2 A canonical 1D counterexample: $L(x) = \sqrt{1 + x^2}$

Proposition D.2 (A smooth convex loss with divergent Newton (1D)). Let $L : \mathbb{R} \to \mathbb{R}$ be given by $L(x) := \sqrt{1 + x^2}$. Then L is convex and smooth, and L'' is globally Lipschitz. The classical Newton iteration $x_{k+1} = N(x_k)$ satisfies the exact recursion

$$x_{k+1} = -x_k^3.$$

In particular, $|x_{k+1}| = |x_k|^3$, yielding cubic local convergence for $|x_0| < 1$ and divergence for $|x_0| > 1$.

Proof. We compute

$$L'(x) = \frac{x}{\sqrt{1 + x^2}}, \qquad L''(x) = \frac{1}{(1 + x^2)^{3/2}}, \qquad L^{(3)}(x) = -\frac{3x}{(1 + x^2)^{5/2}}.$$

In particular $L''(x) > 0$ for all x, hence L is convex. Since $L^{(3)}$ is bounded, L'' is globally Lipschitz. Moreover,

$$\frac{L'(x)}{L''(x)} = \frac{x}{\sqrt{1 + x^2}} (1 + x^2)^{3/2} = x(1 + x^2).$$

Thus the Newton update is

$$x_{k+1} = x_k - \frac{L'(x_k)}{L''(x_k)} = x_k - x_k(1 + x_k^2) = -x_k^3.$$

Finally, $|x_{k+1}| = |x_k|^3$ implies $x_k \to 0$ if $|x_0| < 1$ and $|x_k| \to \infty$ if $|x_0| > 1$. □

Proposition D.3 (Fixed damping does not globalize Newton). Fix any learning rate $\eta > 0$ and consider $x_{k+1} = N_\eta(x_k)$ for $L(x) = \sqrt{1 + x^2}$. Define

$$X(\eta) := \sqrt{\max\Big\{0, \frac{3}{\eta} - 1\Big\}}.$$

Then every nonzero initialization with $|x_0| \geq X(\eta)$ diverges and satisfies

$$|x_{k+1}| \geq 2|x_k| \qquad \forall k \geq 0.$$

Proof. From the computations above, $L'(x)/L''(x) = x(1+x^2)$, hence

$$x_{k+1} = x_k - \eta x_k(1+x_k^2) = x_k\big(1-\eta(1+x_k^2)\big).$$

If $|x_k| \geq X(\eta)$, then by definition $\eta(1+x_k^2) \geq 3$, so $|1-\eta(1+x_k^2)| \geq 2$ and therefore $|x_{k+1}| \geq 2|x_k|$. Induction yields divergence. □

D.3 A drift-based generator for smooth convex Newton counterexamples

The previous example is not accidental. We now record a general construction principle: prescribing a one-dimensional *Newton drift* $d(x)$ determines a convex objective L whose Newton iteration is exactly $x_{k+1} = x_k - \tilde{d}(x_k)$, where $\tilde{d}$ is the odd extension of d. This yields an infinite family of smooth convex losses on which classical Newton diverges.

Theorem D.4 (Lipschitz-Hessian drift-to-loss generator for 1D Newton dynamics). Let $d : [0,\infty) \to [0,\infty)$ be continuous and satisfy:

(D0) $d(0) = 0$ and $d(x) > 0$ for all $x > 0$;

(D1) there exist $\delta > 0$ and $C_0 < \infty$ such that

$$|d(x) - x| \leq C_0 x^2 \qquad \forall x \in [0,\delta].$$

Define, for $x > 0$,

$$\Phi(x) := \int_1^x \frac{ds}{d(s)}, \qquad L'_+(x) := a\, e^{\Phi(x)}\ (a > 0), \qquad L_+(x) := \int_0^x L'_+(t)\, dt,$$

and extend evenly by $L(x) := L_+(|x|)$. Let $\tilde{d} : \mathbb{R} \to \mathbb{R}$ be the odd extension $\tilde{d}(x) := \operatorname{sign}(x)\, d(|x|)$ with $\tilde{d}(0) = 0$. Then:

(i) L is even, convex, and $C^2(\mathbb{R})$, with $L''(x) > 0$ for $x \neq 0$. Moreover, for $x > 0$,

$$L'(x) = L'_+(x), \qquad L''(x) = \frac{L'_+(x)}{d(x)}.$$

In particular, L' is odd and L'' is even.

(ii) For every $x \neq 0$,

$$\frac{L'(x)}{L''(x)} = \tilde{d}(x),$$

and therefore the classical Newton update for L is exactly

$$N(x) = x - \frac{L'(x)}{L''(x)} = x - \tilde{d}(x).$$

(iii) If there exist $X_0 > 0$ and $\beta > 1$ such that

$$d(x) \geq (1+\beta)x \qquad \forall x \geq X_0,$$

then every Newton iterate initialized with $|x_0| \geq X_0$ diverges and satisfies

$$|x_{k+1}| \geq \beta |x_k| \qquad \forall k \geq 0.$$

(iv) **(Globally Lipschitz Hessian.)** Assume in addition that $d \in C^1((0,\infty))$ and that

$$C_1 := \sup_{x>0} \frac{|d'(x) - 1|}{d(x)} < \infty. \tag{82}$$

Assume moreover that there exist $\kappa > 1$ and $X_1 > 0$ such that

$$d(x) \geq \kappa x \qquad \forall x \geq X_1. \tag{83}$$

(For instance, Eq. (83) holds whenever the hypothesis in (iii) holds.) Then L'' is globally Lipschitz on $\mathbb{R}$, i.e. there exists $M < \infty$ such that

$$|L''(u) - L''(v)| \leq 2M|u - v| \qquad \forall u, v \in \mathbb{R}.$$

More precisely, for $x > 0$ we have

$$L^{(3)}(x) = \frac{L'_+(x)}{d(x)^2}\big(1 - d'(x)\big) = L''(x)\,\frac{1 - d'(x)}{d(x)}, \tag{84}$$

and hence $\sup_{x>0} |L^{(3)}(x)| < \infty$, so one may take $2M = \sup_{x\neq 0} |L^{(3)}(x)|$.

Proof. We split the argument into four parts.

Part A: construction and smoothness on $(0,\infty)$. By (D0), the map $s \mapsto 1/d(s)$ is continuous on compact subsets of $(0,\infty)$. Hence $\Phi \in C^1((0,\infty))$ with $\Phi'(x) = 1/d(x)$. Therefore $L'_+(x) = ae^{\Phi(x)}$ is $C^1((0,\infty))$ and

$$(L'_+)'(x) = ae^{\Phi(x)}\Phi'(x) = \frac{L'_+(x)}{d(x)}.$$

Thus $L_+ \in C^2((0,\infty))$, with

$$L''_+(x) = (L'_+)'(x) = \frac{L'_+(x)}{d(x)} > 0 \qquad (x > 0),$$

so L_+ is strictly convex on $(0,\infty)$.

Part B: behavior at 0 and C^2 even extension. Write $d(x) = x(1+\varepsilon(x))$ on $(0,\delta]$, where by (D1) we have $|\varepsilon(x)| \le C_0 x$, hence $\varepsilon(x) \to 0$ as $x \downarrow 0$. Then on $(0,\delta]$,

$$\frac{1}{d(x)} = \frac{1}{x}\cdot\frac{1}{1+\varepsilon(x)} = \frac{1}{x} + g(x),$$

where g is bounded on $(0,\delta]$ (since $(1+\varepsilon)^{-1} = 1 + O(\varepsilon)$). Integrating gives

$$\Phi(x) = \int_1^x \frac{ds}{d(s)} = \int_1^x \frac{ds}{s} + \int_1^x g(s)\,ds = \log x + C_* + o(1) \qquad (x \downarrow 0)$$

for some finite constant C_*. Exponentiating yields

$$L'_+(x) = ae^{\Phi(x)} = ae^{C_*}\,x(1+o(1)) \qquad (x \downarrow 0),$$

so $L'_+(x) \to 0$ and $L'_+(x)/x \to ae^{C_*} \in (0,\infty)$. Since $d(x) = x(1+o(1))$, we obtain

$$L''_+(x) = \frac{L'_+(x)}{d(x)} \to ae^{C_*} \in (0,\infty) \qquad (x \downarrow 0).$$

Define $L_+(0) := 0$, $L'_+(0) := 0$, and $L''_+(0) := \lim_{x\downarrow 0} L''_+(x)$; then $L_+ \in C^2([0,\infty))$.

Now define $L(x) = L_+(|x|)$. This L is even and $C^2(\mathbb{R})$ with $L'(x)$ odd and $L''(x)$ even, and for $x > 0$ we have $L'(x) = L'_+(x)$ and $L''(x) = L''_+(x) = L'_+(x)/d(x)$. Since $L''(x) \ge 0$ everywhere and $L''(x) > 0$ for $x \ne 0$, L is convex. This proves (i).

Part C: drift identity and divergence. For $x > 0$, by (i) we have $L'(x) = L'_+(x)$ and $L''(x) = L'_+(x)/d(x)$, hence

$$\frac{L'(x)}{L''(x)} = d(x).$$

By evenness of L'' and oddness of L', the ratio L'/L'' is odd, so for $x < 0$,

$$\frac{L'(x)}{L''(x)} = -\frac{L'(|x|)}{L''(|x|)} = -d(|x|) = \tilde{d}(x),$$

proving (ii). For the divergence statement (iii), if $x \ge X_0$, then $N(x) = x - d(x) \le -\beta x$, hence $|N(x)| \ge \beta|x|$, and similarly for $x \le -X_0$.

Part D: globally Lipschitz Hessian. Assume Eqs. (82) and (83). For $x > 0$, differentiating $L''(x) = L'_+(x)/d(x)$ and using $(L'_+)'(x) = L'_+(x)/d(x)$ from Part A gives Eq. (84). Hence, for $x > 0$,

$$|L^{(3)}(x)| \le \frac{L'_+(x)}{d(x)^2}\,|1 - d'(x)| \le C_1\frac{L'_+(x)}{d(x)} = C_1 L''(x).$$

It remains to show $\sup_{x>0} L''(x) < \infty$. By Part B, L'' extends continuously to $x = 0$ with $0 < L''(0) < \infty$. On $[0, X_1]$, continuity gives $\sup_{[0,X_1]} L'' < \infty$. For $x \ge X_1$, using Eq. (83),

$$\Phi(x) = \Phi(X_1) + \int_{X_1}^x \frac{ds}{d(s)} \le \Phi(X_1) + \int_{X_1}^x \frac{ds}{\kappa s} = \Phi(X_1) + \frac{1}{\kappa}\log\frac{x}{X_1},$$

so $L'_+(x) = ae^{\Phi(x)} \le C\,x^{1/\kappa}$ for some C. Therefore,

$$L''(x) = \frac{L'_+(x)}{d(x)} \le \frac{Cx^{1/\kappa}}{\kappa x} = \frac{C}{\kappa} x^{\frac{1}{\kappa}-1},$$

which is bounded for $x \ge X_1$ since $\kappa > 1$. Thus $\sup_{x>0} |L^{(3)}(x)| < \infty$, and since L'' is even, L'' is globally Lipschitz on $\mathbb{R}$. □

D.4 Explicit Counterexamples

Corollary D.5 (Elementary closed-form losses from explicit drifts). Applying Theorem D.4 with the following choices of d (and the corresponding choice of a) yields the stated closed-form losses:

1. If $d(x) = x(1+x)$ and $a = C/2$, then
$$L(x) = C\big(|x| - \ln(1+|x|)\big).$$
2. If $d(x) = x(1+x^2)$ and $a = C/\sqrt{2}$, then
$$L(x) = C\big(\sqrt{1+x^2} - 1\big).$$
3. Fix an integer $m \ge 3$. If
$$d(x) = \frac{1+x}{m-1}\Big((1+x)^{m-1} - 1\Big) \qquad \text{and} \qquad a = C\Big(1 - 2^{-(m-1)}\Big),$$
then
$$L(x) = C|x| - \frac{C}{m-2}\Big(1 - (1+|x|)^{-(m-2)}\Big).$$
4. If $d(x) = (1+x^2)\arctan x$ and $a = C(\pi/4)$, then
$$L(x) = C\Big(|x|\arctan|x| - \tfrac{1}{2}\ln(1+x^2)\Big).$$

In each case, L is even, convex, C^2 with globally Lipschitz Hessian, and the classical Newton method diverges from sufficiently large initializations.

Proof. For each item, one checks that the stated L has $L'(x)/L''(x) = \tilde{d}(x)$ for $x \ne 0$, and that $L'_+(1) = a$ matches the choice of a above; therefore Theorem D.4 constructs exactly the stated L. The regularity and divergence conclusions follow from the theorem. We indicate this in the case of $d(x) = x + x^3 = x(1+x^2)$. We compute, for $x > 0$,

$$\Phi(x) = \int_1^x \frac{ds}{s(1+s^2)} = \Big[\log s - \tfrac{1}{2}\log(1+s^2)\Big]_{s=1}^{s=x} = \log x - \tfrac{1}{2}\log(1+x^2) + \tfrac{1}{2}\log 2.$$

Therefore

$$e^{\Phi(x)} = \sqrt{2}\,\frac{x}{\sqrt{1+x^2}}, \qquad L'_+(x) = ae^{\Phi(x)} = \frac{x}{\sqrt{1+x^2}} \quad \text{when } a = 2^{-1/2}.$$

Integrating from 0 to x and using $L(0) = 0$ yields $L(x) = \sqrt{1+x^2} - 1$. Finally, by Theorem D.4(ii),

$$x_{k+1} = x_k - \tilde{d}(x_k) = x_k - (x_k + x_k^3) = -x_k^3,$$

as claimed. □

D.5 A boosting/Hilbert-space realization

Proposition D.6 (A separable empirical-risk objective with divergent Newton updates). Assume that the training inputs $x_1, \dots, x_N$ are pairwise distinct. Let $\mathcal{H} = \mathcal{L}^2(\hat{\nu}_N) \cong \mathbb{R}^N$ be the space of predictions on the training points, and consider the empirical risk

$$L(F) := \frac{1}{N} \sum_{i=1}^{N} \sqrt{1 + F(x_i)^2}.$$

Then L is convex and has $2M$-Lipschitz Hessian for some $M < \infty$, and the exact Newton update satisfies, coordinate-wise,

$$F_{k+1}(x_i) = -F_k(x_i)^3, \qquad i = 1, \dots, N.$$

In particular, if $|F_0(x_i)| > 1$ for some i, then $\|F_k\| \to \infty$.

Proof. The objective is separable across coordinates in $\mathbb{R}^N$. The coordinate-wise gradient and Hessian are those of $L(x) = \sqrt{1 + x^2}$, so Proposition D.2 applies componentwise. Lipschitzness of the Hessian follows because the scalar third derivative is bounded, hence the diagonal Hessian map is Lipschitz in operator norm. □

E Newton Boosting for Hessian-Dominated Losses

This appendix proves the claims stated around Definition 3.6 and Proposition 5.1: binary and categorical cross entropy are Hessian-dominated with constant $c = 1$, and pointwise Hessian-dominance implies Hessian-dominance of the empirical risk. For the unregularized cross-entropy losses, the pointwise infimum is zero, so the pointwise bounds below are already bounds in terms of the optimality gap.

E.1 A useful log inequality

Lemma E.1 (A log inequality). For all $p \in (0, 1]$,

$$\frac{1-p}{p} \;\geq\; -\log p. \tag{85}$$

Equivalently, for all $p \in [0, 1)$,

$$\frac{p}{1-p} \;\geq\; -\log(1-p). \tag{86}$$

Proof. Define $\phi(p) = \frac{1}{p} - 1 + \log p$ on $(0, 1]$. Then

$$\phi'(p) = -\frac{1}{p^2} + \frac{1}{p} = \frac{p-1}{p^2} \leq 0,$$

so ϕ is non-increasing. Since $\phi(1) = 0$, we have $\phi(p) \geq 0$ for all $p \in (0, 1]$, which is (85). Substituting $q = 1 - p$ gives (86). □

E.2 Binary cross entropy is Hessian-dominated

Proposition E.2 (Binary cross entropy is Hessian-dominated with $c = 1$). Let $y \in \{0, 1\}$ and $u \in \mathbb{R}$, and set $p = \sigma(u) = (1 + e^{-u})^{-1}$. Consider the binary cross entropy

$$l(u, y) \;=\; -\Big(y \log p + (1-y)\log(1-p)\Big).$$

Let $g(u, y) = \partial_u l(u, y)$ and $h(u) = \partial_{uu} l(u, y)$. Then the (scalar) Newton step $f(u, y) = -h(u)^{-1} g(u, y)$ satisfies

$$\|f(u, y)\|^2_{h(u)} := f(u, y)^2\, h(u) = \frac{g(u, y)^2}{h(u)} \;\geq\; l(u, y). \tag{87}$$

Hence binary cross entropy is pointwise Hessian-dominated with constant $c = 1$.

Proof. A direct computation gives

$$g(u, y) = p - y, \qquad h(u) = p(1-p) > 0.$$

Thus $\|f\|^2_h = g^2/h$. If $y = 1$, then $g = p - 1 = -(1-p)$, so

$$\frac{g^2}{h} = \frac{(1-p)^2}{p(1-p)} = \frac{1-p}{p}.$$

Also $l(u, 1) = -\log p$, and Lemma E.1 gives $\frac{1-p}{p} \geq -\log p = l(u, 1)$. If $y = 0$, then $g = p$, so

$$\frac{g^2}{h} = \frac{p^2}{p(1-p)} = \frac{p}{1-p}.$$

Also $l(u, 0) = -\log(1-p)$, and Lemma E.1 gives $\frac{p}{1-p} \geq -\log(1-p) = l(u, 0)$. This proves (87). □

Lemma E.3 (Softmax Hessian is positive definite on the zero-sum subspace). Let $p \in \mathbb{R}^K$ satisfy $p_k > 0$ and $\sum_{k=1}^K p_k = 1$, and define

$$H \;:=\; \mathrm{diag}(p) - pp^\top.$$

Let $V := \{v \in \mathbb{R}^K : \mathbf{1}^\top v = 0\}$. Then H maps V to V and is positive definite on V. In particular, the restriction $H : V \to V$ is invertible.

Proof. First, for any $v \in \mathbb{R}^K$,

$$\mathbf{1}^\top H v = \mathbf{1}^\top \mathrm{diag}(p) v - \mathbf{1}^\top p (p^\top v) = p^\top v - (p^\top v) = 0,$$

so $Hv \in V$ whenever $v \in V$. Next, for any $v \in \mathbb{R}^K$,

$$v^\top H v = \sum_{k=1}^K p_k v_k^2 - \Big(\sum_{k=1}^K p_k v_k\Big)^2 = \sum_{k=1}^K p_k \big(v_k - \sum_{j=1}^K p_j v_j\big)^2 \;\geq\; 0.$$

Equality holds iff v_k is constant in k, i.e. $v \in \mathrm{span}\{\mathbf{1}\}$. Intersecting with V leaves only $v = 0$, hence $v^\top H v > 0$ for all nonzero $v \in V$. Therefore H is positive definite on V and invertible as $H : V \to V$. □

Proposition E.4 (Categorical cross entropy is Hessian-dominated with $c = 1$). Let $u \in \mathbb{R}^K$, $p = \mathrm{softmax}(u)$, and let $y = e_t$ be a one-hot label. Consider categorical cross entropy

$$l(u, y) \;=\; -\sum_{k=1}^K y_k \log p_k \;=\; -\log p_t.$$

Let $g(u, y) = \nabla_u l(u, y) = p - y$ and $H(u) = \nabla^2_{uu} l(u, y) = \mathrm{diag}(p) - pp^\top$. Work on the subspace $V = \{v : \mathbf{1}^\top v = 0\}$; then $g(u, y) \in V$ and by Lemma E.3 the Newton direction $f \in V$ solving

$$H(u) f \;=\; y - p \tag{88}$$

is unique. Moreover,

$$\|f\|^2_{H(u)} := f^\top H(u) f \;=\; \frac{1 - p_t}{p_t} \;\geq\; -\log p_t \;=\; l(u, y). \tag{89}$$

Hence categorical cross entropy is pointwise Hessian-dominated with constant $c = 1$.

Proof. Let $H = \mathrm{diag}(p) - pp^\top$. Define $s := p^\top f$. Expanding (88) yields

$$\mathrm{diag}(p) f - p(p^\top f) = y - p \quad\Longleftrightarrow\quad p_k(f_k - s) = y_k - p_k \qquad \forall k,$$

so (since $p_k > 0$) we have

$$f_k = s + \frac{y_k}{p_k} - 1.$$

With $y = e_t$, this becomes

$$f_t = s + \frac{1}{p_t} - 1, \qquad f_k = s - 1 \quad (k \neq t).$$

Using symmetry of H and (88),

$$\|f\|^2_H = f^\top H f = f^\top (y - p).$$

Compute

$$\begin{aligned} f^\top (y - p) &= f_t(1 - p_t) + \sum_{k \neq t} f_k (0 - p_k) \\ &= f_t(1 - p_t) - (s - 1) \sum_{k \neq t} p_k \\ &= \big(f_t - (s - 1)\big)(1 - p_t) = \left(\left(s + \frac{1}{p_t} - 1\right) - (s - 1)\right)(1 - p_t) \\ &= \frac{1 - p_t}{p_t}. \end{aligned}$$

Finally, Lemma E.1 gives $\frac{1-p_t}{p_t} \geq -\log p_t = l(u, y)$, proving (89). □

Remark E.5 (Uniqueness of the Newton step). On $\mathbb{R}^K$ one has $H(u)\mathbf{1} = 0$ (softmax is shift-invariant), so the Newton equation is not uniquely solvable without fixing a gauge. Restricting to the zero-sum subspace $V = \{v : \mathbf{1}^\top v = 0\}$ (or equivalently fixing one logit coordinate) yields uniqueness and satisfies Assumption 3.1. The decrement $\|f\|^2_{H(u)}$ is invariant to the choice of representative.

Remark E.6. Binary cross entropy corresponds to the special case $K = 2$ of categorical cross entropy (after identifying the scalar logit with the difference of the two softmax logits). Thus Proposition E.4 strictly generalizes Proposition E.2; the latter is included separately for clarity and because it avoids the gauge-fixing required in the multiclass setting.

Lemma E.7. Suppose $l(\cdot, y)$ is twice continuously differentiable, convex, and S-smooth. For $\mu > 0$, let

$$l_\mu(u, y) := l(u, y) + \frac{\mu}{2}\|u\|^2, \qquad l_{\mu,\star}(y) := \inf_v l_\mu(v, y).$$

Then

$$\nabla_1 l_\mu(u, y)^\top \big(\nabla_1^2 l_\mu(u, y)\big)^{-1} \nabla_1 l_\mu(u, y) \geq \frac{2\mu}{S + \mu}\big(l_\mu(u, y) - l_{\mu,\star}(y)\big).$$

Proof. The regularized loss is μ-strongly convex and $(S + \mu)$-smooth. Writing $g = \nabla_1 l_\mu(u, y)$ and $H = \nabla_1^2 l_\mu(u, y)$, we therefore have

$$g^\top H^{-1} g \geq \frac{\|g\|^2}{S + \mu} \geq \frac{2\mu}{S + \mu}\big(l_\mu(u, y) - l_{\mu,\star}(y)\big).$$

□

Binary and categorical cross-entropy are convex and smooth, so Lemma E.7 applies to both. Their regularized empirical risks are therefore Hessian-dominated by Lemma E.9. For binary cross entropy, we can prove the stronger bound $c = (1 + 4\mu)/(1 + 2\mu) > 1$, as shown in Lemma E.8.

Lemma E.8 (Regularized binary cross entropy). Let

$$l_\mu(u, y) = \log(1 + e^u) - yu + \frac{\mu}{2}u^2, \qquad l_{\mu,\star}(y) = \inf_v l_\mu(v, y),$$

where $y \in \{0, 1\}$ and $\mu > 0$. Then

$$\frac{(\partial_u l_\mu(u, y))^2}{\partial_{uu} l_\mu(u, y)} \geq \frac{1 + 4\mu}{1 + 2\mu}\big(l_\mu(u, y) - l_{\mu,\star}(y)\big).$$

Proof. Write $p = \sigma(u)$, $h = p(1 - p)$, $g = p - y + \mu u$, and $d = g/(h + \mu)$. Set

$$q = p - hd = \frac{hy + \mu(p - hu)}{h + \mu} \in (0, 1).$$

Indeed, $p - hu$ increases from 0 to 1, since $(p - hu)' = -uh' \geq 0$. Set $a = \log(q/(1 - q))$, so that $\partial_1 l(a, y) = q - y$. By convexity,

$$l_\mu(v, y) \geq l(a, y) + (q - y)(v - a) + \frac{\mu}{2}v^2.$$

The right-hand side is minimized at $v = (y - q)/\mu$, giving

$$l_{\mu,\star}(y) \geq l(a, y) - (q - y)a - \frac{(q - y)^2}{2\mu}.$$

Using the cross-entropy expressions, completing the square, and applying $\log x \le x - 1$, we obtain

$$\begin{aligned} l_\mu(u,y) - l_{\mu,\star}(y) &\le l(u,y) - l(a,y) + \frac{\mu}{2}u^2 + (q-y)a + \frac{(q-y)^2}{2\mu} \\ &= l(u,y) - l(a,y) - (q-y)(u-a) + \frac{(q-y+\mu u)^2}{2\mu} \\ &= q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p} + \frac{(q-y+\mu u)^2}{2\mu} \\ &\le q\left(\frac{q}{p}-1\right) + (1-q)\left(\frac{1-q}{1-p}-1\right) + \frac{(q-y+\mu u)^2}{2\mu} \\ &= \frac{(q-p)^2}{h} + \frac{(q-y+\mu u)^2}{2\mu} \\ &= \left(h + \frac{\mu}{2}\right) d^2, \end{aligned}$$

where $q - p = -hd$ and $q - y + \mu u = \mu d$. Since $g^2/(h+\mu) = (h+\mu)d^2$ and $h \le 1/4$, the claimed constant follows from

$$\frac{h+\mu}{h+\mu/2} \ge \frac{1+4\mu}{1+2\mu}.$$

□

E.3 From pointwise to empirical Hessian-dominance

Lemma E.9 (Pointwise ⇒ empirical Hessian-dominance). Let $\{(x_i, y_i)\}_{i=1}^N$ be a dataset with pairwise distinct inputs $x_1, \dots, x_N$, and define the empirical risk

$$L(F) = \frac{1}{N}\sum_{i=1}^N l(F(x_i), y_i),$$

with $F(x_i) \in \mathbb{R}^K$ (take $K = 1$ for binary). We work in the empirical Hilbert space $\mathcal{H} = \mathcal{L}^2(\hat{\nu}_N)$ with inner product

$$\langle A, B\rangle = \frac{1}{N}\sum_{i=1}^N \langle A(x_i), B(x_i)\rangle_{\mathbb{R}^K}.$$

Let $l_{\star,i} := \inf_u l(u, y_i) > -\infty$ and $L_\star := \inf_{F\in\mathcal{H}} L(F)$. At each iterate $F = F_k$, write $u_i := F(x_i)$ and let

$$g_{(i)} := \partial_1 l(u_i, y_i) \in \mathbb{R}^K, \qquad h_{(i)} := \partial_{11} l(u_i, y_i) \in \mathbb{R}^{K\times K}.$$

Assume there exists $c > 0$ such that for all i and k,

$$g_{(i)}^\top h_{(i)}^{-1} g_{(i)} \;\ge\; c\big(l(u_i, y_i) - l_{\star,i}\big), \tag{90}$$

where $h_{(i)}^{-1}$ is taken on the relevant space (for categorical CE one may restrict to V pointwise as in Proposition E.4). Let $g = \nabla L(F)$ and $H = \nabla^2 L(F)$, and let $f = -H^{-1}g$ be the exact Newton direction. Then

$$\|f\|_H^2 \;\ge\; c\big(L(F) - L_\star\big).$$

In particular, L is Hessian-dominated in the sense of Definition 3.6 with constant c.

Proof. In the empirical setting, the gradient and Hessian act pointwise:

$$g(x_i) = g_{(i)}, \qquad (H[f])(x_i) = h_{(i)} f(x_i).$$

Thus the Newton equation $H[f] = -g$ decouples over samples:

$$h_{(i)} f(x_i) = -g_{(i)} \quad \Rightarrow \quad f(x_i) = -h_{(i)}^{-1} g_{(i)}.$$

Therefore, using Definition 3.2,

$$\|f\|_H^2 = \langle f, H[f]\rangle = \frac{1}{N}\sum_{i=1}^N f(x_i)^\top h_{(i)} f(x_i) = \frac{1}{N}\sum_{i=1}^N g_{(i)}^\top h_{(i)}^{-1} g_{(i)}.$$

Applying (90) term-by-term yields

$$\|f\|_H^2 \geq \frac{c}{N}\sum_{i=1}^{N}\big(l(u_i,y_i)-l_{\star,i}\big) \geq c\big(L(F)-L_\star\big),$$

where the last inequality uses $L_\star \geq N^{-1}\sum_{i=1}^{N} l_{\star,i}$. This completes the proof. □

F Proof of Regularity Conditions in the Boosting Space

Recall from Section 2.3 that we identify the empirical boosting space $\mathcal{H} = \mathcal{L}^2(\widehat{\nu}_{X,N})$ with $\mathbb{R}^{N\times K}$, equipped with the inner product $\langle f,h\rangle = \frac{1}{N}\sum_{i=1}^{N}\langle f(x_i),h(x_i)\rangle_{\mathbb{R}^K}$. The induced norm is therefore $\|f\| = \big(\frac{1}{N}\sum_{i=1}^{N}\|f(x_i)\|_{\mathbb{R}^K}^2\big)^{1/2}$.

Proof of Proposition 5.1. First, we prove the claim regarding the Lipschitz continuity of the Hessian. Suppose the base loss $l(u,y)$ has a M_0-Lipschitz Hessian with respect to the prediction $u \in \mathbb{R}^K$, meaning $\|\nabla_{11}^2 l(u,y) - \nabla_{11}^2 l(v,y)\|_{\text{op}} \leq M_0\|u-v\|_{\mathbb{R}^K}$. In the gradient boosting space, we minimize the empirical risk $L(F) = \frac{1}{N}\sum_{i=1}^{N} l(F(x_i),y_i)$. Because the Hessian operator $\nabla^2 L(F)$ acts pointwise (i.e., it is block-diagonal over the N samples), its operator norm on $\mathcal{H}$ is exactly the maximum of the operator norms of the individual $K\times K$ blocks. Thus, for any $F,G\in\mathcal{H}$, we can bound the difference in Hessians as follows:

$$\begin{aligned}
\|\nabla^2 L(F) - \nabla^2 L(G)\|_{\text{op},\mathcal{H}} &= \max_{1\leq i\leq N}\|\nabla_{11}^2 l(F(x_i),y_i) - \nabla_{11}^2 l(G(x_i),y_i)\|_{\text{op},\mathbb{R}^K} \\
&\leq \max_{1\leq i\leq N} M_0\|F(x_i)-G(x_i)\|_{\mathbb{R}^K} \\
&\leq M_0\left(\sum_{i=1}^{N}\|F(x_i)-G(x_i)\|_{\mathbb{R}^K}^2\right)^{1/2} \\
&= M_0\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\|F(x_i)-G(x_i)\|_{\mathbb{R}^K}^2\right)^{1/2} \\
&= M_0\sqrt{N}\|F-G\|.
\end{aligned}$$

This proves that L has an $M_0\sqrt{N}$-Lipschitz Hessian in the space $\mathcal{H}$.

Next, we prove that smoothness and strong convexity are preserved with the same constants. The Hessian quadratic form in the boosting space is given by $\langle f,\nabla^2 L(F)[f]\rangle = \frac{1}{N}\sum_{i=1}^{N}\langle f(x_i),\nabla_{11}^2 l(F(x_i),y_i)f(x_i)\rangle_{\mathbb{R}^K}$. If the base loss l is μ-strongly convex and S-smooth, its Euclidean quadratic form is bounded pointwise:

$$\mu\|f(x_i)\|_{\mathbb{R}^K}^2 \leq \langle f(x_i),\nabla_{11}^2 l(F(x_i),y_i)f(x_i)\rangle_{\mathbb{R}^K} \leq S\|f(x_i)\|_{\mathbb{R}^K}^2.$$

Summing these bounds over $i=1,\dots,N$ and dividing by N directly yields $\mu\|f\|^2 \leq \langle f,\nabla^2 L(F)[f]\rangle \leq S\|f\|^2$. Finally, if l is pointwise Hessian-dominated, this property is similarly preserved for L with the same constant c via an identical pointwise summation argument, as explicitly proven in Lemma E.9 in the Appendix on Hessian dominance. □

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 and introduction clearly state the scope (theoretical mechanisms of Newton boosting) and our specific contributions (convergence rates for vanilla and gradient regularized Newton), which are subsequently proven and empirically validated.

   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: We explicitly discuss the limitations of our work in the section "Conclusion, Limitations, and Future Work".

   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: All theoretical results explicitly state their required assumptions (e.g., S-smoothness, Lipschitz continuous Hessians, a bounded initial sublevel set) in the main text, and complete formal proofs are provided in the Appendix.

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: We describe our implementation details, hyperparameter choices, and datasets in Section 6, and the full JAX source code to reproduce all experiments and figures is included in the supplementary material.

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 code used to produce the experiments is included in the supplementary material. An updated repository is available at `https://github.com/nikitazozoulenko/global-newton-boosting`. The datasets are publicly available from the UCI repository.

   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: We specify the tree depths, learning rates, regularization parameters, and dataset details used to validate our theoretical claims in Section 6 and in the supplementary material code.

   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: [No]

   Justification: Our experiments track exact training loss and geometric properties over optimization iterations to validate deterministic theoretical convergence properties, making statistical significance testing and error bars either unnecessary or not possible to include, partly since convex optimization is a deterministic rather than statistical process.

   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: [Yes]

   Justification: All experiments were performed on a single CPU.

   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: This research constitutes foundational theoretical work on optimization algorithms and fully conforms to the NeurIPS Code of Ethics.

   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: This paper strictly focuses on the mathematical foundations and convergence theory of gradient boosting and convex optimization theory, and thus does not have direct or specific societal impacts to discuss.

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 high-risk models, image generators, or new datasets that would necessitate safeguards.

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: The publicly available datasets (Wine Quality, HIGGS) and software libraries (JAX, XGBoost) used in our work are appropriately cited in the main text and bibliography.

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: [N/A]

    Justification: The paper does not introduce or release any new datasets or large-scale models.

    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 research does not involve crowdsourcing or human subjects.

    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 research does not involve human subjects.

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: Large Language Models were not used as a component of this paper.

    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.