Affine-invariant Cubic Newton with Weak Learners

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

Modern gradient boosting algorithms are based on Newton's method, which excels empirically but lacks global convergence guarantees. While recent work on Gradient-Regularized Newton (GRN) boosting achieved a global $\mathcal{O}(1/\gamma^{12}k^2)$ rate, its analysis was not able to exploit the natural Hessian geometry and suffers from a heavy γ^{12} dependence on a worst-case 'weak gradient edge' γ. In this paper, we tackle these theoretical drawbacks by generalizing the Affine-Invariant Cubic Newton (AICN) scheme to convex optimization with weak learners. Our analysis is based on average cosine values $\overline{\Theta}_k$ along the optimization path and directional improvements δ_k in the Hessian-induced norm. We show that the weak-AICN update reduces to a damped weak-Newton step with an explicit stepsize determined by δ_k. Under semi-strong self-concordance, we establish a global $\mathcal{O}(1/\overline{\Theta}_k^4 k^2)$ rate, significantly improving upon the weak-GRN bound. We further prove a local linear rate, global linear convergence for Hessian-dominated losses, cosine angle lower bounds, and present weak-AdaAICN, which adaptively tunes the cubic regularization. Finally, we demonstrate the broader applicability of this framework beyond gradient boosting by showing that Newton coordinate descent emerges naturally as a special case of optimization with weak learners.

1 Introduction

Modern gradient boosting decision tree (GBDT) libraries such as XGBoost [6], LightGBM [19], and CatBoost [35] use Newton's method to fit leaf weights. However, theoretical understanding of Newton boosting [12, 33, 37, 42] has remained limited in comparison to first-order gradient boosting [4, 7, 11, 12, 14, 21, 22, 24, 38]. This stems from the fact that Newton's method does not have global convergence guarantees, and can diverge when initialized far from the optimum, even for smooth, strictly convex losses [17, 23]. Recently [42] proved the first global $O(1/k^2)$ rate for Newton boosting by extending the Gradient-Regularized Newton (GRN) [9, 25] scheme to weak learners. They did this via the Restricted Newton Descent framework [42], which poses boosting as a special case of Hilbert space convex optimization with iterates restricted to a scalable set of weak learners.

While the weak-GRN scheme successfully globalizes Newton boosting, its convergence analysis was not able to take advantage of the natural Hessian geometry of the problem. The weak-GRN bounds rely on standard Lipschitz continuity of the Hessian, and the proofs are fundamentally based on bounding the growth of Euclidean gradient norms [42]. Because of this geometric mismatch, the analysis does not depend on any prior established definition of weak learner edge or cosine angle [14, 21, 37]. Instead, [42] introduced the "weak gradient edge" γ, which measures the similarity between the exact gradient and the implied weak gradient in the Euclidean norm. This results in a severe $\mathcal{O}(1/\gamma^{12}k^2)$ dependence in the convergence rate, rendering the bound highly pessimistic since γ can often be very small when using intentionally weak learners.

In this paper, we resolve these theoretical drawbacks by presenting a different global weak learner Newton scheme, namely Affine-Invariant Cubic Newton [16], generalized to weak learners. Here,

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

regularity is measured using semi-strong self-concordance, which is affine-invariant and allows for the analysis to take place entirely using local Hessian norms. This geometric shift allows us to measure convergence in terms of the cosine angle Θ_k between weak and exact Newton iterates. Furthermore, previous boosting papers defined their weak learner edges by taking infima over either the entire optimization sequence k [14, 37, 42], or over the set of all possible weak learners [21, 22, 33, 38]. We instead develop a novel recurrence analysis that allows our bounds to depend on the *average* cosine angle $\overline{\Theta}_k$ up to iteration k. By shifting away from a worst-case analysis, we theoretically guarantee that a single poorly-fitted weak learner does not invalidate the convergence of the whole scheme.

Contributions: Our main contributions are summarized as follows:

- **Affine-Invariant Weak Updates:** We formulate the weak-AICN algorithm within the Restricted Newton Descent framework. By introducing the directional improvement metric δ_k in the local Hessian norm, we prove that the weak-AICN update elegantly reduces to a damped weak-Newton step with an explicit, easily computable stepsize.
- **Global Convergence:** Under semi-strong self-concordance, we establish a global $\mathcal{O}(1/\overline{\Theta}_k^4 k^2)$ convergence rate. This drastically improves upon the $\mathcal{O}(1/\gamma^{12}k^2)$ rate of GRN, and transitions from a worst-case bound to an average-performance bound.
- **Linear Rates for Specific Loss Classes:** We prove that the weak-AICN scheme achieves a local linear convergence rate for general convex losses, and a global linear convergence rate for Hessian-dominated losses (such as cross entropy loss).
- **Applications to Coordinate Descent:** By taking the weak learner family to be the standard basis vectors, we show that Newton coordinate descent emerges naturally as a special case, allowing us to apply our weak learner convergence theory beyond simply gradient boosting, to classical coordinate optimization. We also present an adaptive variant of weak-AICN that dynamically tunes the cubic regularization without refitting the weak learners.
- **Tree-Angle Lower Bounds:** For depth-J decision trees, we prove explicit random lower bounds on tree alignment in a random configuration model, including a Beta-distributed stochastic lower bound and the expectation $E\left[(\widehat{\Theta}^{\text{prox}}_{n,p,J})^2\right] \geq \frac{\min(2^J,n)}{n}$.

1.1 Related Work

Boosting: Gradient boosting was first studied as gradient descent in $\mathbb{R}^N$ [12, 24], and later generalized to functional spaces [14]. The first boosting convergence rate of $\mathcal{O}((\log k)^{-\frac{1}{2}})$ was proved in [4], later improved to $\mathcal{O}(1/k)$ under smoothness [37], and to $\mathcal{O}(1/k^2)$ via Nesterov momentum [22]. Recently, the $\mathcal{O}(1/k^2)$ rate was also achieved by the weak-GRN boosting scheme [42]. For strongly convex losses, a linear rate was derived in [14], and later by [21] via coordinate descent by enumerating all possible weak learners. Newton boosting was first introduced by LogitBoost [12] for logistic loss and later popularized by XGBoost [6] for general losses. Theory for vanilla Newton boosting remains limited, with guarantees only in the strongly convex setting for small learning rates [33, 37, 42].

Newton's method in convex optimization: Newton's method has quadratic local convergence [28] but lacks global guarantees [17, 23], motivating regularized variants. Cubic regularized Newton achieves global $\mathcal{O}(1/k^2)$ rates under Lipschitz Hessians [26, 30], but requires solving a costly subproblem each iteration. GRN matches this rate without such subproblems via adaptive ℓ_2 regularization [9, 25]. More recently, affine-invariant approaches such as AICN [16] achieve the same global $\mathcal{O}(1/k^2)$ rate, together with local quadratic convergence, as a damped Newton method with explicit computable stepsize. For a broader historical overview, including earlier suboptimal rates and the development of self-concordance, we refer to [16, 25]. Finally, we want to mention [1, 2] which studies inexact (quasi-)Newton's method with inexact Hessians, rather than with weak learners.

2 Setup and Optimization with Exact and Weak Learners

In this paper we work in a general Hilbert space setting, following [42]. Let $(\mathcal{H}, \|\cdot\|)$ be a real Hilbert space, and $L : \mathcal{H} \to \mathbb{R}$ be a convex and twice differentiable loss function with positive definite Hessian. Let $\eta > 0$ be the learning rate (constant dampening factor). The goal of Newton's method is to produce a sequence of estimates F_k minimizing the loss $L(F_k)$. It does this iteratively

by minimizing a second order Taylor expansion of the loss at each iteration k:

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

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 second-order approximation of L centered on F_k. The Newton step is derived by differentiating Q_k and setting $0 = \nabla Q_k(f) = g_k + H_k f_{k+1}^{\mathrm{N}}$, yielding

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

The exact Newton iterates are then defined as $F_{k+1}^{\mathrm{N}} = F_k^{\mathrm{N}} + \eta f_{k+1}^{\mathrm{N}} = \sum_{t=1}^{k+1} \eta f_t^{\mathrm{N}}$, assuming $F_0^{\mathrm{N}} = 0$.

2.1 Restricted Newton Descent

In this paper, we work in the Restricted Newton Descent framework [42], where the updates f are restricted to a set of weak learners $f \in \mathcal{F} \subseteq \mathcal{H}$. The exact Newton update f_{k+1}^{N} was defined by minimizing $Q_k(f)$ over the whole space $\mathcal{H}$. The weak learner f_{k+1}^{wN} is instead defined as

$$f_{k+1}^{\mathrm{wN}} \in \operatorname*{argmin}_{f \in \mathcal{F}} Q_k(f), \qquad Q_k(\mathcal{F}) = \min_{f \in \mathcal{F}} Q_k(f). \tag{2}$$

Here $Q_k(\mathcal{F})$ is the minimally achievable weak learner quadratic Taylor loss. For boosting, $\mathcal{F}$ is typically the set of decision trees of a maximum depth. We use the following standard assumption:

Assumption 2.1 (Scalable weak learner family [22]). The weak-learner family $\mathcal{F} \subseteq \mathcal{H}$ is closed under scalar multiplication: if $f \in \mathcal{F}$ and $t \in \mathbb{R}$ then $tf \in \mathcal{F}$.

2.2 Hessian norms, cosine angle, and directional improvement

In this subsection we introduce the basic tools and definitions needed for our weak learner analysis. The notion of local Hessian norm and its dual norm, defined below, will play a central role.

Definition 2.2 (Hessian-induced norm and dual norm). We define the *Hessian-induced inner product and norm* as

$$\langle u, v \rangle_{H_k} := \langle u, H_k v \rangle, \qquad \|u\|_{H_k} := \langle u, u \rangle_{H_k}^{1/2} = \langle H_k u, u \rangle^{1/2},$$

with corresponding *dual norm* defined by

$$\|v\|_{H_k}^* := \|v\|_{H_k^{-1}} = \langle H_k^{-1} v, v \rangle^{1/2}.$$

Remark 2.3. In a general infinite-dimensional Hilbert space, the notation H_k^{-1} should be interpreted as requiring H_k to be *boundedly invertible* at the points of interest (equivalently, $H_k \succeq \mu I$ for some $\mu > 0$ at those points). In finite dimensions this is automatic from $H_k \succ 0$.

Next, we define the fundamental metrics governing our weak learner convergence bounds. In addition to the cosine angle Θ_k between exact and weak Newton directions [42], we introduce a new but related metric, denoted by δ_k. This quantity measures the directional alignment to the exact Newton direction in the local Hessian geometry.

Definition 2.4 (Cosine angle and directional improvement). Let f_{k+1}^{N} and f_{k+1}^{wN} be the exact and weak Newton directions, respectively, computed at iteration k. The weak learner *cosine angle* is defined as

$$\Theta_k := \frac{\langle f_{k+1}^{\mathrm{N}}, f_{k+1}^{\mathrm{wN}} \rangle_{H_k}}{\|f_{k+1}^{\mathrm{N}}\|_{H_k} \|f_{k+1}^{\mathrm{wN}}\|_{H_k}} = \operatorname{cosangle}_{\|\cdot\|_{H_k}}(f_{k+1}^{\mathrm{N}}, f_{k+1}^{\mathrm{wN}}). \tag{3}$$

For any nonzero $f \in \mathcal{H}$, define the *directional improvement* and *maximal directional improvement* as

$$\delta_k(f) := \frac{-\langle g_k, f \rangle}{\|f\|_{H_k}} = \left\langle f_{k+1}^{\mathrm{N}}, \frac{f}{\|f\|_{H_k}} \right\rangle_{H_k}, \qquad \delta_k(\mathcal{F}) = \max_{f \in \mathcal{F} \setminus \{0\}} \delta_k(f). \tag{4}$$

Before we introduce the weak-AICN scheme, we first need to present a key result related to the weak-Newton update which our analysis depends on. The following proposition shows that, instead

of minimizing the quadratic approximation $Q_k(f)$, we may equivalently maximize the directional improvement $\delta_k(f)$. We present the proof in Appendix A. This differs substantially from the approach in [16] as a tool for AICN analysis, as well as the weak-Newton results of [42] which only required elementary properties of cosine angles.

Proposition 2.5 (Q-δ duality). For any non-zero direction $\phi^* \in \mathcal{F}$ with $\delta_k(\phi^*) \geq 0$, we have

$$\phi^* \in \operatorname*{argmax}_{f \in \mathcal{F}\setminus\{0\}} \delta_k(f) \quad \Longleftrightarrow \quad \frac{\delta_k(\phi^*)}{\|\phi^*\|_{H_k}} \phi^* \in \operatorname*{argmin}_{f \in \mathcal{F}} Q_k(f).$$

In particular, $Q_k(\mathcal{F}) = -\frac{1}{2}\delta_k(\mathcal{F})^2$. Moreover, every minimizer $f_{k+1}^{\mathrm{wN}} \in \operatorname{argmin}_{f\in\mathcal{F}} Q_k(f)$ satisfies

$$-\langle g_k, f_{k+1}^{\mathrm{wN}} \rangle = \|f_{k+1}^{\mathrm{wN}}\|_{H_k}^2, \qquad \|f_{k+1}^{\mathrm{wN}}\|_{H_k} = \delta_k(\mathcal{F}).$$

If $f_{k+1}^{\mathrm{wN}} \neq 0$, then also

$$\delta_k(f_{k+1}^{\mathrm{wN}}) = \|f_{k+1}^{\mathrm{wN}}\|_{H_k} = \delta_k(\mathcal{F}).$$

In particular, we have the identities $\Theta_k \|g_k\|_{H_k}^* = \delta_k(\mathcal{F})$ and $\Theta_k = \frac{\|f_{k+1}^{\mathrm{wN}}\|_{H_k}}{\|f_{k+1}^{\mathrm{wN}}\|_{H_k}}$.

3 AICN with Weak Learners

We are now ready to define the weak-AICN scheme. Rather than simply minimizing the local quadratic Taylor bound Q_k, global methods like Cubic Regularized Newton (CRN) [26, 30] and AICN [16] instead add an additional cubic term to the minimization step, resulting in a *global bound* to be minimized. CRN measured this extra cubic term in the Euclidean norm, resulting in an optimization procedure which required solving an inner expensive subproblem at each iteration. AICN instead added a cubic term using the Hessian-induced norm, which resulted in the AICN update reducing to a damped Newton step. Below, we define our weak-AICN update:

Definition 3.1 (Weak-AICN). The weak-AICN iterates F_k with smoothness constant C_{est} and fixed learning rate $\eta > 0$ are defined recursively via the cubic model

$$F_{k+1} := F_k + \eta f_{k+1}^{\mathrm{wAICN}}, \qquad f_{k+1}^{\mathrm{wAICN}} = \operatorname*{argmin}_{f\in\mathcal{F}} M_k(f),$$

where $F_0 \in \mathcal{H}$ is a given initial value and

$$M_k(f) := \langle g_k, f\rangle + \frac{1}{2}\|f\|_{H_k}^2 + \frac{C_{\mathrm{est}}}{6}\|f\|_{H_k}^3 = Q_k(f) + \frac{C_{\mathrm{est}}}{6}\|f\|_{H_k}^3.$$

The fundamental objects in our analysis are the quadratic model Q_k, the cubic model M_k, the directional improvement δ_k and the weak learner cosine angle Θ_k. Using only Definition 3.1 and Proposition 2.5, we are able to prove the following theorem, which says that weak-AICN reduces to a damped weak-Newton step, with an explicit stepsize governed by the quantity $C_{\mathrm{est}}\delta_k(f_{k+1}^{\mathrm{wN}})$. The resulting algorithm is summarized in Algorithm 1. All proofs are deferred to Appendix B.

Theorem 3.2 (w-AICN is damped w-Newton). The weak-AICN iterates f_{k+1}^{wAICN} are of the form

$$f_{k+1}^{\mathrm{wAICN}} = \alpha_k\, f_{k+1}^{\mathrm{wN}},$$

with explicit stepsize determined by the quantity $C_{\mathrm{est}}\delta_k$, where $\delta_k := \delta_k(\mathcal{F}) = \|f_{k+1}^{\mathrm{wN}}\|_{H_k}$:

$$\alpha_k := \frac{-1+\sqrt{1+2C_{\mathrm{est}}\delta_k}}{C_{\mathrm{est}}\delta_k} = \frac{2}{1+\sqrt{1+2C_{\mathrm{est}}\delta_k}}.$$

Moreover, we have $0 < \alpha_k \leq 1$, the identity $\frac{1}{2}C_{\mathrm{est}}\delta_k\alpha_k^2 = 1-\alpha_k$, and that $\alpha_k \uparrow 1$ as $\delta_k \downarrow 0$.

Unlike exact AICN [16, Theorem 1], which defines its dampening using the dual gradient norm $\|g_k\|_{H_k}^*$, our weak learner formulation reveals that the directional improvement δ_k is the true driving mechanism of the dampening. The exact case can be recovered via the identity $\Theta_k\|g_k\|_{H_k}^* = \delta_k$ from Proposition 2.5. Crucially, Theorem 3.2 is purely algebraic and requires no extra regularity assumptions on the loss, allowing us to derive a bound on the cubic model, presented in Proposition 3.3.

Algorithm 1 Weak-AICN

Input: Loss $L : \mathcal{H} \to \mathbb{R}$. Learning rate η. Weak learner family $\mathcal{F}$. Cubic regularization C_{est}. Initial value F_0.
for $k = 0, 1, 2, \ldots$ **do**
  $g_k,\ H_k \leftarrow \nabla L(F_k),\ \nabla^2 L(F_k)$
  $f_{k+1}^{\text{wN}} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} Q_k(f)$
  $\delta_k,\ \alpha_k \leftarrow \sqrt{|\langle g_k, f_{k+1}^{\text{wN}} \rangle|},\ \frac{2}{1+\sqrt{1+2C_{\text{est}}\delta_k}}$
  $F_{k+1} \leftarrow F_k + \eta \alpha_k f_{k+1}^{\text{wN}}$
end for

Proposition 3.3 (Weak Improvement Θ_k^2 Factor). Let α_k be the weak-AICN dampening, and write $M_k(\mathcal{F}) := \min_{f \in \mathcal{F}} M_k(f)$. We have the following weak learner bound:

$$M_k(\mathcal{F}) \;=\; \frac{\alpha_k(\alpha_k - 4)}{6} \|f_{k+1}^{\text{wN}}\|_{H_k}^2 \;\le\; \Theta_k^2\, M_k(\mathcal{H}). \tag{5}$$

3.1 Semi-strong Self-concordance and Weak Contracting-point Inequality

To establish our global convergence guarantees, we must impose regularity assumptions on the loss function. In the second-order optimization literature, global methods such as Cubic Regularized Newton [26, 30] and Gradient-Regularized Newton (GRN) [9, 25, 42] typically assume that the Hessian operator is Lipschitz continuous. In this paper, we instead adopt the regularity assumption of the classical AICN method [16], assuming the loss L satisfies *semi-strong self-concordance*. This allows us to bound the variation of the Hessian strictly within the local geometry induced by the Hessian itself, and has the advantage that our bounds are affine-invariant with no dependence on choice of basis. We formally state this definition below in the language of Hessian norms.

Definition 3.4 (Semi-strong self-concordance [16]). For a bounded linear operator $A : \mathcal{H} \to \mathcal{H}$, define the **Hessian-induced affine-invariant operator norm** as

$$\|A\|_{\text{aff-op},H_k} \;:=\; \sup_{u \neq 0} \frac{\|Au\|_{H_k^{-1}}}{\|u\|_{H_k}}.$$

Equivalently, $\|A\|_{\text{aff-op},H_k} = \left\|H_k^{-1/2}\, A\, H_k^{-1/2}\right\|_{\text{op}}$, where $\|\cdot\|_{\text{op}}$ is the standard operator norm. The convex loss L is said to be **semi-strongly self-concordant (SSC)** with constant $C_{\text{ssc}} > 0$ if for all $F \in \mathcal{H}$ and all $f \in \mathcal{H}$, writing $H_F := \nabla^2 L(F)$,

$$\|\nabla^2 L(F+f) - \nabla^2 L(F)\|_{\text{aff-op},H_F} \;\le\; C_{\text{ssc}} \|f\|_{H_F}.$$

Assuming SSC yields the following bound on the second-order Taylor error, which most subsequent analysis will depend on. A direct consequence is that the cubic model $M_k(f)$ being minimized by AICN is a global bound, rather than a local bound like the quadratic approximation $Q_k(f)$.

Proposition 3.5 (Affine-invariant cubic Taylor bound [16]). Assume L is SSC. Fix $F \in \mathcal{H}$ and write $H_F := \nabla^2 L(F)$. Then for all $f \in \mathcal{H}$,

$$\left|L(F+f) - L(F) - Q_F(f)\right| \;\le\; \frac{C_{\text{ssc}}}{6} \|f\|_{H_F}^3. \tag{6}$$

By evaluating this global upper bound at the weak-AICN update, we are able to prove that the weak-AICN update successfully contracts the objective function toward any arbitrary comparison point $Y \in \mathcal{H}$. Crucially, the strength of this contraction is scaled by the squared weak learner cosine angle Θ_k^2 and learning rate η.

Proposition 3.6 (Weak Contracting-point Inequality). Let the loss L be SSC and consider weak-AICN iterates with $C_{\text{est}} \ge C_{\text{ssc}}$ and learning rate $0 < \eta \le 1$. Then for every $Y \in \mathcal{H}$,

$$L(F_{k+1}) \;\le\; (1 - \eta\Theta_k^2) L(F_k) + \eta\Theta_k^2 \left(L(Y) + \frac{C_{\text{est}}}{3} \|Y - F_k\|_{H_k}^3 \right). \tag{7}$$

3.2 Global and Local Convergence of AICN with Weak Learners

Before we state our convergence theorems, we first present two recurrence lemmas used in our analysis. These lemmas allow us to improve the weak learner bounds from a *worst-case* analysis to an *average performance* analysis. Prior works define their weak learner edges either as an infimum over all iterations k [14, 37, 42], or as an infimum over all possible weak learners [21, 22, 33, 38]. This is problematic, since while weak learner conditions are generally stable on a rolling-window basis, they can be noisy from iteration to iteration, with worst-case values often close to 0. By using these recurrences, our final convergence rates depend on the path-averaged cosine angle instead of the infimum, proving that a single poorly-fitted weak learner does not degrade the theoretical guarantee of the whole optimization process.

Lemma 3.7. Let Δ_k be a non-negative sequence such that $\Delta_{k+1} \le \Delta_k - c_k \Delta_k^{\frac{3}{2}}$ for a sequence of constants $c_k \ge 0$. Then for all $k \ge 1$ with $\sum_{i=0}^{k-1} c_i > 0$

$$\Delta_k \le \frac{4}{\left(2c_0 + \sum_{i=0}^{k-1} c_i\right)^2} \le \frac{4}{\overline{c_k}^2 k^2}, \tag{8}$$

where $\overline{c_k}$ is the average $\overline{c_k} := \frac{1}{k}\sum_{i=0}^{k-1} c_i$.

Lemma 3.8. Let $\{\Delta_k\}_{k\ge 0}$ be a non-negative sequence such that $\Delta_{k+1} \le (1 - c_k)\,\Delta_k$ for a sequence of constants $c_k \in [0, 1]$. Then for all $k \ge 1$,

$$\Delta_k \;\le\; \Delta_0 \prod_{i=0}^{k-1}(1 - c_i) \;\le\; \Delta_0 \exp\left(-\sum_{i=0}^{k-1} c_i\right) \;\le\; \Delta_0 e^{-k\overline{c_k}}, \tag{9}$$

where $\overline{c_k} := \frac{1}{k}\sum_{i=0}^{k-1} c_i$.

We impose the following boundedness assumption which is standard for affine-invariant and self-concordance analysis, including for proving the global $\mathcal{O}(1/k^2)$ rate for the exact AICN scheme [16]. Note that this assumption automatically is satisfied if L is strongly convex and smooth, or strongly convex with Lipschitz continuous Hessian (see Remark B.3).

Assumption 3.9 (Bounded local radius). Let $F^* = \operatorname{argmin}_{F\in\mathcal{H}} L(F)$ and assume the weak-AICN iterates satisfy a uniform bound

$$\sup_{k\ge 0} \|F_k - F^*\|_{H_k} \;\le\; D < \infty.$$

Applying our recurrence lemmas to the weak contracting-point inequality (Proposition 3.6), we are able to establish the following global convergence of weak-AICN.

Theorem 3.10 (Global convergence). Let L be SSC and consider weak-AICN iterates F_k with $C_{\text{est}} \ge C_{\text{ssc}}$ and learning rate $\eta \in (0, 1]$. Define $\Delta_k := L(F_k) - L(F^*)$ and let k_0 be the first iteration such that $\Delta_{k_0} \le C_{\text{est}} D^3$. Define the path-averaged cosine edges $\overline{\Theta_k^2} := \frac{1}{k}\sum_{i=0}^{k-1} \Theta_i^2$ and $\overline{\Theta_{k_0,k}^2} := \frac{1}{k-k_0}\sum_{i=k_0}^{k-1} \Theta_i^2$. Weak-AICN exhibits an initial global linear phase, followed by a global $\mathcal{O}(1/k^2)$ phase. If $\Delta_0 \le C_{\text{est}} D^3$, then $k_0 = 0$. If $k_0 > 1$, the length of the initial phase is bounded by $k_0 < \frac{3}{2\eta\overline{\Theta_{k_0-1}^2}} \ln\left(\frac{\Delta_0}{C_{\text{est}} D^3}\right) + 1$. The convergence rates are:

1. Initial Linear Regime ($1 \le k < k_0$): $\quad \Delta_k \le \Delta_0 \exp\left(-\frac{2}{3}\eta k \overline{\Theta_k^2}\right)$.
2. Global $\mathcal{O}(1/k^2)$ Regime ($k > k_0$): $\quad \Delta_k \le \frac{9\, C_{\text{est}}\, D^3}{\eta^2 (\overline{\Theta_{k_0,k}^2})^2 (k-k_0)^2}$.

Next, we establish the local convergence of weak-AICN. While the global regime is governed by the minimization of the cubic upper bound, the local regime is instead characterized by the contraction of the gradient in the dual Hessian norm. Due to the extra error introduced by the weak learners, we are only able to prove a local linear rate—similar to prior boosting works [14, 21, 33, 37, 42]—rather than the quadratic rates found in the classical finite-dimensional convex optimization literature [16]. Our local rate's dependence on the squared edge Θ_k^2 is consistent with prior work referenced above.

Theorem 3.11 (Local Linear Rate). Let the loss L be SSC with learning rate $\eta \in (0, 1]$ and $C_{\text{est}} \geq C_{\text{ssc}}$. If the weak-AICN iterates satisfy the local condition $\|g_t\|_{H_t}^* \leq \frac{\Theta_t}{6C_{\text{est}}}$ for $t = 0, \ldots, k-1$, then the gradient contracts in the dual Hessian norm:

$$\|g_{t+1}\|_{H_{t+1}}^* \leq (1 - \rho_t)\|g_t\|_{H_t}^*, \tag{10}$$

where $\rho_t = \frac{1}{4}\eta\alpha_t\Theta_t^2$. Consequently, we obtain a local linear rate of convergence via the bound

$$L(F_k) - L(F^*) \;\leq\; \|g_k\|_{H_k}^* D \;\leq\; \|g_0\|_{H_0}^* D \exp\left(-k\overline{\rho_k}\right), \tag{11}$$

where $\overline{\rho_k} := \frac{1}{k}\sum_{t=0}^{k-1} \rho_t = \frac{\eta}{4k}\sum_{t=0}^{k-1} \alpha_t\Theta_t^2$ is the path-averaged contraction rate.

3.3 Global Convergence of weak-AICN for Hessian-dominated Losses

Weak-AICN can obtain global linear rates for the class of Hessian-dominated losses, a property recently introduced in the Restricted Newton Descent framework [42] motivated by the LogitBoost analysis [37]. The canonical example of such a loss is binary and categorical cross entropy with $c = 1$. Note that Hessian dominance alone does not imply SSC; in particular, unregularized binary log-loss is not even self-concordant [3]. However, ℓ_2-regularized cross entropy satisfies both Hessian dominance and SSC, since it is strongly convex and smooth with a Lipschitz continuous Hessian.

Definition 3.12 (Hessian-dominance [42]). Assume $L_\star := \inf_{F\in\mathcal{H}} L(F) > -\infty$. We say that the loss sequence $L(F_k)$ is **Hessian-dominated** if there exists a constant $c > 0$ such that $\|f_{k+1}^{\text{N}}\|_{H_k}^2 \geq c\big(L(F_k) - L_\star\big)$ for all k.

Theorem 3.13 (Linear weak-AICN rate for Hessian-dominated loss). Let L be SSC and consider weak-AICN iterates with $C_{\text{est}} \geq C_{\text{ssc}}$ and learning rate $\eta \in (0, 1]$. If L further satisfies Definition 3.12, then the weak-AICN iterates satisfy

$$L(F_{k+1}) - L_\star \leq (1 - \rho_k)\big(L(F_k) - L_\star\big), \tag{12}$$

where $\rho_k = c\eta\Theta_k^2\left(1 - \frac{\eta}{2}\right)\alpha_k$.

This result is notable because the analogous linear rate for vanilla Newton boosting typically requires a restricted learning rate $\eta \in (0, \frac{2\mu}{S})$ under strict μ-strong convexity and S-smoothness assumptions [42]. In contrast, Theorem 3.13 holds for all learning rates $\eta \leq 1$, trading strong convexity for a sufficiently large self-concordance constant $C_{\text{est}} \geq C_{\text{ssc}}$.

4 Applications of Weak-AICN

In this section, we present two application areas of weak-AICN and optimization with weak learners.

4.1 Newton boosting

The primary application of weak learner analyses is gradient tree boosting. Given a training dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ and targets y_i, we work in the Hilbert space $\mathcal{H}_X = \mathcal{L}^2(\widehat{\nu}_n)$ of square-integrable functions equipped with the empirical measure $\widehat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. The boosting objective is to minimize the empirical risk $L(F) = \frac{1}{n}\sum_{i=1}^n l(F(x_i), y_i)$, where l is the data-point-wise loss. This space can be identified with $\mathbb{R}^n$, equipped with the inner product $\langle f, h\rangle_X = \frac{1}{n}\sum_{i=1}^n f(x_i)h(x_i)$ and induced norm $\|f\|_X^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2$. When the weak learner family $\mathcal{F}$ is restricted to the set of depth-J decision trees $\mathcal{T}_J$, the classical Newton boosting updates used by libraries like XGBoost are recovered [42]. Since weak-AICN is simply damped weak-Newton, implementing our scheme into existing methodologies is very simple, requiring only computing the stepsize α_k in Theorem 3.2. Next, we show that the regularity assumptions required for weak-AICN carry over naturally to L from the base loss l. See Appendix C for proof.

Theorem 4.1 (SSC inheritance for empirical risks). Assume that the inputs $x_1, \ldots, x_n$ are pairwise distinct and that, for each target y_i, the scalar function $l(\cdot, y_i)$ is semi-strongly self-concordant with constant C_l. Then the empirical risk $L(F) = \frac{1}{n}\sum_{i=1}^n l(F(x_i), y_i)$ is semi-strongly self-concordant on $\mathcal{L}^2(\widehat{\nu}_n)$ with constant $C_{\text{ssc}}(L) \leq \sqrt{n}\, C_l$.

To achieve convergence via Theorem 3.10, the weak learners must maintain a non-zero cosine angle Θ_k along the optimization path. Unlike the weak gradient edge γ used in weak-GRN's [42] bounds, which is difficult to verify theoretically, the strict positivity of Θ_k follows directly from the identity $\Theta_k \|g_k\|_{H_k}^* = \delta_k(\mathcal{F})$ established in Proposition 2.5. This relationship guarantees that any fitted decision tree achieving a strictly positive improvement in loss automatically has $\Theta_k > 0$. The real difficulty, however, lies in quantitative control. Establishing universal, deterministic lower bounds for Θ_k over all possible datasets is impossible in general. For instance, consider the parity function $y_i = (-1)^{\sum_j (x_i)_j}$ on the Boolean hypercube $X = \{0,1\}^p$; any decision tree of depth $J < p$ leaves at least one coordinate free per leaf, outputting exactly zero on average and yielding a worst-case cosine angle of exactly $\Theta_k = 0$. Because deterministic worst-case bounds are fatally pessimistic for general tree weak learners, analyzing tree alignment requires shifting from a worst-case perspective to a probabilistic one.

Since the exact Hessian-induced cosine angle Θ_k is presently too hard to analyze directly, we instead study a random proxy. Concretely, assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}([0,1]^p)$ and let Q be any metric model which, conditional on X, samples a positive definite self-adjoint operator $A : \mathcal{H}_X \to \mathcal{H}_X$. Write

$$\langle f, h \rangle_A := \langle f, Ah \rangle_X, \qquad S_A(\mathcal{H}_X) := \{f \in \mathcal{H}_X : \|f\|_A = 1\},$$

and draw $U_A \mid (X, A) \sim \text{Unif}(S_A(\mathcal{H}_X))$. We define the random depth-J metric proxy by

$$\widehat{\Theta}_{n,p,J}^{\text{prox},\mathsf{Q}} := \max_{g \in \mathcal{T}_J(X) \setminus \{0\}} \frac{\langle U_A, g \rangle_A}{\|g\|_A}.$$

The identity-metric proxy is recovered by taking $A = I$ deterministically. The point of this construction is to give a tractable no-signal baseline for random tree alignment in a realized Hessian geometry. The following theorem gives a lower bound for this quantity, uniformly over the metric model Q, which is dyadic with respect to tree depth J. See Appendix F for the geometric setup and proof.

Theorem 4.2 (Dyadic tree-angle lower bounds). Let $n \geq 2$, $p \geq 1$, and $J \geq 0$, and let Q be any metric model as above. If $2^J < n$, then there exists a random variable

$$Z_J \sim \text{Beta}\left(2^{J-1}, \frac{n - 2^J}{2}\right)$$

such that $(\widehat{\Theta}_{n,p,J}^{\text{prox},\mathsf{Q}})^2 \geq Z_J$ almost surely. In particular, $\mathbb{E}\left[(\widehat{\Theta}_{n,p,J}^{\text{prox},\mathsf{Q}})^2\right] \geq \frac{\min(2^J, n)}{n}$ and $\mathbb{E}[\widehat{\Theta}_{n,p,J}^{\text{prox},\mathsf{Q}}] \geq \frac{\Gamma\left(\frac{2^J+1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma(2^{J-1})\Gamma\left(\frac{n+1}{2}\right)}$. If $2^J \geq n$, then $\widehat{\Theta}_{n,p,J}^{\text{prox},\mathsf{Q}} = 1$ almost surely.

Remark 4.3. Theorem H.1 in the Appendix gives an alternative, potentially simpler and more intuitive, signal-plus-noise formulation of Theorem 4.2. There, a tree-aligned signal plus isotropic Gaussian noise in the Hessian geometry yields the same squared-angle expectation lower bound.

4.2 Newton Coordinate Descent

To the best of our knowledge, prior theoretical frameworks for optimization with weak learners have exclusively been used for gradient boosting. However, the Restricted Newton Descent framework is fundamentally space-agnostic. By simply changing the choice of the base space $\mathcal{H}$ and the weak learner family $\mathcal{F}$, we can recover classical algorithms from finite-dimensional convex optimization.

Coordinate descent is a classical optimization paradigm [40], with modern analyses spanning randomized updates [27, 36] and second-order subspace methods [10, 13, 32]. In our framework, let the base space be $\mathcal{H} = \mathbb{R}^d$ equipped with the standard basis $\{e_1, \ldots, e_d\}$. If we let the weak learner family be the set of all scaled basis vectors, $\mathcal{F} = \{ce_i : c \in \mathbb{R}, 1 \leq i \leq d\}$, we recover Newton Coordinate Descent. To compute the corresponding weak update, one has to maximize the directional improvement $\delta_k(\varphi)$ over the basis vectors. For a single coordinate direction $\pm e_i$, this evaluates to

$$\delta_k(\pm e_i) = \frac{\mp\langle g_k, e_i \rangle}{\|e_i\|_{H_k}} = \frac{\mp (g_k)_i}{\sqrt{(H_k)_{ii}}}. \tag{13}$$

Therefore, finding the optimal weak direction reduces to selecting the coordinate i^* that maximizes $i^* = \text{argmax}_i \, |(g_k)_i| / \sqrt{(H_k)_{ii}}$. This is a one-dimensional Newton analogue of the Gauss–Southwell–Lipschitz selection rule [31], with $(H_k)_{ii}$ acting as the local curvature. By Proposition 2.5,

the resulting weak Newton step is exactly a 1D scalar Newton step along the chosen coordinate:

$$f_{k+1}^{\mathrm{wN}} \;=\; \frac{|\delta_k(e_{i^*})|}{\|e_{i^*}\|_{H_k}}\,\mathrm{sign}(-(g_k)_{i^*})\,e_{i^*} \;=\; -\frac{(g_k)_{i^*}}{(H_k)_{i^*i^*}}e_{i^*}. \tag{14}$$

Calculating this update requires only the gradient and the diagonal of the Hessian, avoiding the $\mathcal{O}(d^3)$ cost of full matrix inversion. Furthermore, our general weak learner angle machinery immediately yields the following lower bound for the coordinate descent cosine angle. See Appendix D for proof.

Proposition 4.4. For Newton Coordinate Descent ($\mathcal{F} = \{ce_i : c \in \mathbb{R}, 1 \le i \le d\}$), the weak learner cosine angle is deterministically lower bounded by

$$\Theta_k \ge \frac{1}{\sqrt{d\kappa_k}}, \tag{15}$$

where d is the dimension and $\kappa_k = \lambda_{\max}(H_k)/\lambda_{\min}(H_k)$ is the condition number of the Hessian.

Consequently, applying weak-AICN with coordinate weak learners immediately yields a globally convergent, damped Newton Coordinate Descent scheme. This algorithm shares some connections with a recent low-rank and sketched subspace AICN method [15]. However, our method arises here organically as a special case of optimization with weak learners, inheriting all theoretical results from the weak-AICN framework. In the finite-dimensional convex optimization literature, it is common to use adaptive versions of algorithms where regularity constants such as C_{est} are adjusted dynamically during optimization [25]. We present one such modification to weak-AICN in Appendix D.3.

4.3 Supplementary Experiments

We provide three sets of experiments, with full experimental details in the appendix. The goal of these experiments is not to claim state-of-the-art predictive performance, but rather to empirically validate and supplement theoretical findings presented in this paper. All code was run on a single CPU, and is available at `github.com/nikitazozoulenko/global-newton-boosting`. Our experiments in this paper include:

1. **Adaptive weak-AICN (Appendix D.3):** We empirically validate our adaptive modification of the (weak-)AICN algorithm. Enabled by the simplification of the cubic model bound in Proposition 3.3, this adaptive scheme dynamically tunes the regularity constant without requiring expensive inner-loop subproblems or re-evaluations of the Hessian or gradient.
2. **Tree Boosting Convergence and Weak Learner Edges (Appendix E):** Following the experimental setup of [42], we compare weak-AICN with weak-GRN, vanilla Newton boosting, and first-order boosting, using decision trees as weak learners. We demonstrate the non-convergence of vanilla Newton boosting under the Charbonnier loss, compare weak learner edges on the HIGGS dataset, and analyze their relationship with tree depth. Figure 1 compares the Hessian cosine angle Θ_k used by weak-AICN with the weak gradient edge γ_k used by weak-GRN. These quantities are comparable as cosine-like weak learner edges defined in different geometries, and take similar empirical values. This supports the comparison of our global $\mathcal{O}(1/(\overline{\Theta}_k^4 k^2))$ rate with weak-GRN's $\mathcal{O}(1/(\gamma^{12}k^2))$ rate, showing substantially better dependence on weak learner alignment. Furthermore, our bound uses a path-averaged cosine angle, whereas weak-GRN uses the worst-case edge $\gamma = \inf_t \gamma_t$. The full experimental setup and metric definitions are provided in Appendix E.
3. **Random Tree Angle Proxy Experiments (Appendix G):** We evaluate the dyadic tree-angle lower bounds and structured-signal hypotheses in relation to Theorem 4.2.

Limitations

While weak-AICN is simple to implement as a damped weak-Newton step, and improves the currently best known weak learner edge dependence in the global $\mathcal{O}(1/k^2)$ rate from γ^{12} to $\overline{\Theta}_k^4$, this geometric shift comes with a trade-off. Because weak-AICN operates entirely within the local Hessian geometry, it can struggle with ill-conditioned problems if static ℓ_2 regularization is not added, unlike weak-GRN which handles such problems seamlessly via adaptive gradient-based ℓ_2 penalties. Furthermore,

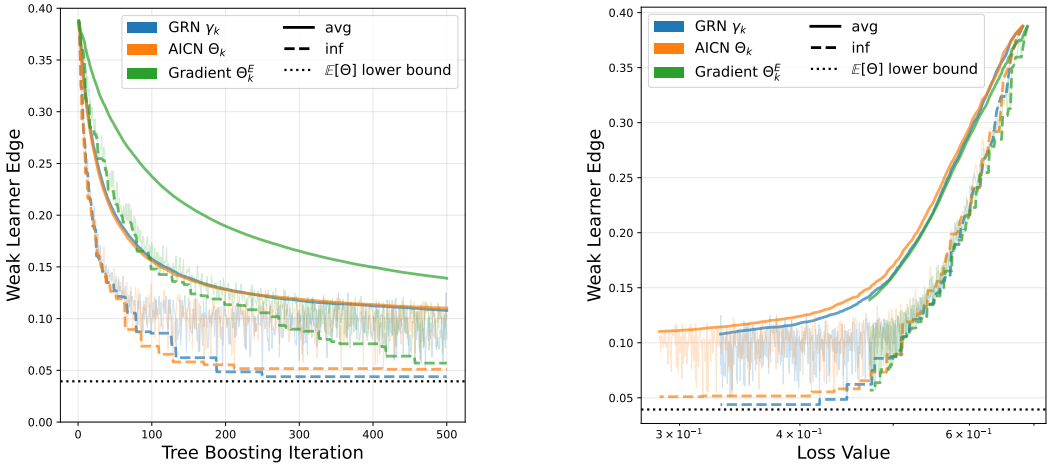


Figure 1: Comparison of weak learner edges for tree boosting on HIGGS ($n = 10000,\ J = 4$): Hessian cosine angle Θ_k for weak-AICN, weak gradient edge γ_k for weak-GRN, and Euclidean cosine angle Θ_k^E for first-order boosting. **Left:** Edges against boosting iterations. **Right:** Edges against training loss. Solid lines show arithmetic running averages; dashed lines show running infima. The dotted line shows the proxy expectation lower bound from Theorem 4.2 as a reference value.

in practical gradient boosting applications with small learning rates, there is often little empirical difference between vanilla Newton boosting and globally convergent variants. Nevertheless, global schemes remain strictly necessary to theoretically guarantee convergence. In this context, our weak-AICN analysis helps explain why libraries like XGBoost perform well in practice, as our theory guarantees that the standard Newton boosting algorithm will safely converge provided the learning rate η does not exceed the theoretically derived weak-AICN dampening factor α_k.

Conclusion

In this paper, we introduced weak-AICN, a globally convergent second-order optimization scheme with weak learners. By analysing directional improvements δ_k in the local Hessian norm, we proved that the weak-AICN update reduces to a damped weak-Newton step with an explicit, easily computable stepsize, without requiring any regularity assumptions on the loss function. Building upon this update and measuring regularity through semi-strong self-concordance, we established an affine-invariant global $\mathcal{O}(1/\overline{\Theta}_k^4 k^2)$ convergence rate. This significantly improves the weak learner edge dependence compared to prior best-known bounds for the weak-GRN scheme, and transitions theoretical guarantees from a pessimistic worst-case infimum to an average-performance bound. We further established a local linear rate for general convex losses, and a global linear rate for Hessian-dominated losses. To address the impossibility of deterministic worst-case bounds for decision trees, we derived explicit Beta-distributed stochastic lower bounds on tree alignment in a random configuration model, including an expectation bound that improves dyadically with tree depth. Finally, by showing that Newton coordinate descent emerges naturally when restricting the weak learner family to standard basis vectors, we demonstrated that this framework extends beyond gradient boosting, connecting our work and the field to the broader convex optimization literature.

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] Artem Agafonov, Dmitry Kamzolov, Pavel Dvurechensky, Alexander Gasnikov, and Martin Takáč. Inexact tensor methods and their application to stochastic convex optimization. *Optimization Methods and Software*, 39(1):42–83, 2024. doi: 10.1080/10556788.2023.2261604.

[2] Artem Agafonov, Vladislav Ryspayev, Samuel Horváth, Alexander Gasnikov, Martin Takáč, and Slavomír Hanzely. Simple stepsize for quasi-Newton methods with global convergence guarantees, 2025. URL https://arxiv.org/abs/2508.19712.

[3] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010. doi: 10.1214/09-EJS521.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Nikita Doikov and Peter Richtárik. Randomized block cubic Newton method. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1290–1298. PMLR, 10–15 Jul 2018.

[11] 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.

[12] 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.

[13] Robert M. Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: randomized subspace Newton. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[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 $\mathcal{O}(1/k^2)$ 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] Dmitry Kamzolov, Artem Agafonov, Dmitry Pasechnyuk, Alexander Gasnikov, and Martin Takáč. OPTAMI: Global superlinear convergence of high-order methods. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Cpr6Wv2tfr.

[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] Tor Lattimore and András György. A second-order method for stochastic bandit convex optimisation. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2067–2094. PMLR, 12–15 Jul 2023.

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

[22] 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.

[23] Walter F. Mascarenhas. On the divergence of line search methods. *Computational & Applied Mathematics*, 26:129–169, 2007.

[24] 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.

[25] 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.

[26] 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.

[27] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001.

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

[29] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186(1):157–183, Mar 2021. ISSN 1436-4646. doi: 10.1007/s10107-019-01449-1.

[30] 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.

[31] Julie Nutini, Mark Schmidt, Issam H. Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, page 1632–1641. JMLR.org, 2015.

[32] Julie Nutini, Issam Laradji, and Mark Schmidt. Let's make block coordinate descent converge faster: faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

[33] 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.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[35] 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.

[36] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, Apr 2014. ISSN 1436-4646. doi: 10.1007/s10107-012-0614-z.

[37] 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.

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

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

[40] Stephen J. Wright. Coordinate descent algorithms. *Math. Program.*, 151(1):3–34, June 2015. ISSN 0025-5610. doi: 10.1007/s10107-015-0892-3.

[41] Qian Yu, Yining Wang, Baihe Huang, Qi Lei, and Jason Lee. Stochastic zeroth-order optimization under strongly convexity and Lipschitz hessian: Minimax sample complexity. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 99564–99600. Curran Associates, Inc., 2024. doi: 10.52202/079017-3159.

[42] Nikita Zozoulenko, Daniel Falkowski, Thomas Cass, and Lukas Gonon. Gradient regularized Newton boosting trees with global convergence. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=9MjVpTTm7u`.

A Proofs for Weak Learner Direction Selection

To prove Proposition 2.5, we first study the behaviour of the quadratic model $Q_k(f)$ along a single ray $t \mapsto ft$, $t \in \mathbb{R}$. Because our weak learner family $\mathcal{F}$ is closed under scalar multiplication, minimizing over $\mathcal{F}$ is equivalent to minimizing over all rays generated by directions in $\mathcal{F}$:

Lemma A.1 (Ray Minimization of the Quadratic Model). Fix any nonzero direction $f \in \mathcal{H}$. The minimum of the quadratic model $Q_k(tf)$ over the ray $\{tf : t \in \mathbb{R}\}$ is given by

$$\min_{t\in\mathbb{R}} Q_k(tf) = -\frac{1}{2}\delta_k(f)^2, \quad \text{attained at} \quad t^*(f) = \frac{\delta_k(f)}{\|f\|_{H_k}}. \tag{16}$$

Proof. Let $h(t) := Q_k(tf) = t\langle g_k, f\rangle + \frac{1}{2}t^2\|f\|_{H_k}^2$. This is a scalar quadratic function in t. Because $\|f\|_{H_k}^2 > 0$, it is strictly convex and its unique global minimum is achieved where the derivative is zero:

$$h'(t) = \langle g_k, f\rangle + t\|f\|_{H_k}^2 = 0 \qquad \Longrightarrow \qquad t^* = \frac{-\langle g_k, f\rangle}{\|f\|_{H_k}^2} = \frac{\delta_k(f)}{\|f\|_{H_k}}, \tag{17}$$

where the last equality follows from the definition $\delta_k(f) = \frac{-\langle g_k, f\rangle}{\|f\|_{H_k}}$. Substituting t^* back into the objective yields the minimal value

$$h(t^*) = \left(\frac{-\langle g_k, f\rangle}{\|f\|_{H_k}^2}\right)\langle g_k, f\rangle + \frac{1}{2}\left(\frac{-\langle g_k, f\rangle}{\|f\|_{H_k}^2}\right)^2 \|f\|_{H_k}^2 \tag{18}$$

$$= -\frac{\langle g_k, f\rangle^2}{\|f\|_{H_k}^2} + \frac{1}{2}\frac{\langle g_k, f\rangle^2}{\|f\|_{H_k}^2} = -\frac{1}{2}\left(\frac{-\langle g_k, f\rangle}{\|f\|_{H_k}}\right)^2 = -\frac{1}{2}\delta_k(f)^2. \tag{19}$$

□

With this lemma established, the proof of Proposition 2.5 follows naturally.

Proof of Proposition 2.5. Since $\mathcal{F}$ is closed under scalar multiplication, we can rewrite

$$Q_k(\mathcal{F}) = \min_{f\in\mathcal{F}} Q_k(f) = \min_{\phi\in\mathcal{F}\setminus\{0\}} \left(\min_{t\in\mathbb{R}} Q_k(t\phi)\right).$$

Applying Lemma A.1 to the inner minimization gives

$$Q_k(\mathcal{F}) = \min_{\phi\in\mathcal{F}\setminus\{0\}} \left(-\frac{1}{2}\delta_k(\phi)^2\right) = -\frac{1}{2}\sup_{\phi\in\mathcal{F}\setminus\{0\}} \delta_k(\phi)^2.$$

Because $\delta_k(-\phi) = -\delta_k(\phi)$ and $\mathcal{F}$ is closed under sign, we have

$$\sup_{\phi\in\mathcal{F}\setminus\{0\}} \delta_k(\phi)^2 = \delta_k(\mathcal{F})^2.$$

Therefore,

$$Q_k(\mathcal{F}) = -\frac{1}{2}\delta_k(\mathcal{F})^2.$$

Now fix a nonzero direction $\phi^* \in \mathcal{F}$ with $\delta_k(\phi^*) \geq 0$.

If $\phi^* \in \operatorname{argmax}_{f\in\mathcal{F}\setminus\{0\}} \delta_k(f)$, then $\delta_k(\phi^*) = \delta_k(\mathcal{F})$, and Lemma A.1 yields

$$Q_k\left(\frac{\delta_k(\phi^*)}{\|\phi^*\|_{H_k}}\phi^*\right) = -\frac{1}{2}\delta_k(\phi^*)^2 = -\frac{1}{2}\delta_k(\mathcal{F})^2 = Q_k(\mathcal{F}).$$

Hence

$$\frac{\delta_k(\phi^*)}{\|\phi^*\|_{H_k}}\phi^* \in \operatorname*{argmin}_{f\in\mathcal{F}} Q_k(f).$$

Conversely, suppose

$$f^* := \frac{\delta_k(\phi^*)}{\|\phi^*\|_{H_k}}\phi^* \in \operatorname*{argmin}_{f\in\mathcal{F}} Q_k(f).$$

Then

$$-\frac{1}{2}\delta_k(\phi^*)^2 = Q_k(f^*) = Q_k(\mathcal{F}) = -\frac{1}{2}\delta_k(\mathcal{F})^2.$$

Thus $\delta_k(\phi^*)^2 = \delta_k(\mathcal{F})^2$. Since both quantities are nonnegative,

$$\delta_k(\phi^*) = \delta_k(\mathcal{F}),$$

so

$$\phi^* \in \operatorname*{argmax}_{f\in\mathcal{F}\setminus\{0\}} \delta_k(f).$$

Finally, let $f_{k+1}^{\mathrm{wN}} \in \operatorname{argmin}_{f\in\mathcal{F}} Q_k(f)$. If $f_{k+1}^{\mathrm{wN}} = 0$, then $Q_k(\mathcal{F}) = 0$, hence $\delta_k(\mathcal{F}) = 0$, and the stated identities are immediate. Assume now that $f_{k+1}^{\mathrm{wN}} \neq 0$. Since the whole ray $\{tf_{k+1}^{\mathrm{wN}} : t \in \mathbb{R}\}$ lies in $\mathcal{F}$, the scalar function $t \mapsto Q_k(tf_{k+1}^{\mathrm{wN}})$ is minimized at $t = 1$. By Lemma A.1, its unique minimizer is

$$t^* = \frac{\delta_k(f_{k+1}^{\mathrm{wN}})}{\|f_{k+1}^{\mathrm{wN}}\|_{H_k}},$$

so $t^* = 1$, i.e.

$$\delta_k(f_{k+1}^{\mathrm{wN}}) = \|f_{k+1}^{\mathrm{wN}}\|_{H_k}.$$

Using the value formula from Lemma A.1 together with $Q_k(\mathcal{F}) = -\frac{1}{2}\delta_k(\mathcal{F})^2$, we obtain

$$-\frac{1}{2}\|f_{k+1}^{\mathrm{wN}}\|_{H_k}^2 = -\frac{1}{2}\delta_k(f_{k+1}^{\mathrm{wN}})^2 = Q_k(f_{k+1}^{\mathrm{wN}}) = Q_k(\mathcal{F}) = -\frac{1}{2}\delta_k(\mathcal{F})^2.$$

Hence

$$\|f_{k+1}^{\mathrm{wN}}\|_{H_k} = \delta_k(\mathcal{F}),$$

and therefore also

$$\delta_k(f_{k+1}^{\mathrm{wN}}) = \delta_k(\mathcal{F}).$$

Thus,

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

Combining the above one obtains that the cosine angle equivalently takes the form

$$\Theta_k = \frac{\langle f_{k+1}^{\mathrm{N}}, f_{k+1}^{\mathrm{wN}}\rangle_{H_k}}{\|f_{k+1}^{\mathrm{N}}\|_{H_k}\,\|f_{k+1}^{\mathrm{wN}}\|_{H_k}} = \frac{\delta_k(\mathcal{F})}{\|f_{k+1}^{\mathrm{N}}\|_{H_k}} = \frac{\|f_{k+1}^{\mathrm{wN}}\|_{H_k}}{\|f_{k+1}^{\mathrm{N}}\|_{H_k}}.$$

Finally, using the definition of the dual norm applied to the gradient $g_k = -H_k f_{k+1}^{\mathrm{N}}$, and the previous cosine angle identity, one obtains

$$\|g_k\|_{H_k}^* = \|H_k^{-\frac{1}{2}} H_k f_{k+1}^{\mathrm{N}}\| = \|H_k^{\frac{1}{2}} f_{k+1}^{\mathrm{N}}\| = \|f_{k+1}^{\mathrm{N}}\|_{H_k} = \frac{\delta_k(\mathcal{F})}{\Theta_k}$$

This completes the proof. □

B Weak-AICN Proofs

B.1 Proof of Weak-AICN is Damped weak-Newton

Proof of Theorem 3.2. If $\delta_k(\mathcal{F}) = 0$, then Proposition 2.5 gives $Q_k(\mathcal{F}) = 0$. Since $0 \in \mathcal{F}$, $Q_k(0) = 0$, and

$$M_k(f) = Q_k(f) + \frac{C_{\text{est}}}{6}\|f\|_{H_k}^3 \geq 0 = M_k(0), \tag{20}$$

for all $f \in \mathcal{F}$, M_k is also minimized at 0. In this case every weak Newton minimizer is $f_{k+1}^{\text{wN}} = 0$, and $\alpha_k = 1$.

Assume now that $\delta_k(\mathcal{F}) > 0$, and fix a direction $f \in \mathcal{F} \setminus \{0\}$. Since $\mathcal{F}$ is scalable, $-f \in \mathcal{F}$. For any $t \geq 0$,

$$M_k(tf) = -\delta_k(f)\, t\|f\|_{H_k} + \frac{1}{2}t^2\|f\|_{H_k}^2 + \frac{C_{\text{est}}}{6}t^3\|f\|_{H_k}^3, \tag{21}$$

and $M_k(tf) - M_k(-tf) = -2\,\delta_k(f)\, t\|f\|_{H_k}$. Therefore, for each fixed radius $t \geq 0$, the smaller of the two values $M_k(tf)$ and $M_k(-tf)$ is attained at the sign for which $\delta_k(f) \geq 0$. Thus minimizing M_k over the scalable family $\mathcal{F}$ is equivalent to first orienting the direction so that $\delta_k(f) \geq 0$, and then minimizing over $t \geq 0$. Let

$$r := t\|f\|_{H_k} \geq 0, \qquad m_\delta(r) := -\delta r + \frac{1}{2}r^2 + \frac{C_{\text{est}}}{6}r^3. \tag{22}$$

For ease of notation, we will minimize over r instead of t. Along an oriented direction f with $\delta_k(f) \geq 0$, we have $M_k(tf) = m_{\delta_k(f)}(r)$. For fixed $\delta \geq 0$,

$$m_\delta'(r) = -\delta + r + \frac{C_{\text{est}}}{2}r^2, \qquad m_\delta''(r) = 1 + C_{\text{est}}r, \tag{23}$$

so m_δ is strictly convex on $[0, \infty)$. Its unique minimizer on $[0, \infty)$ is $r^*(0) = 0$ when $\delta = 0$, and for $\delta > 0$ it is the unique positive solution of $m_\delta'(r) = 0$, namely

$$r^*(\delta) = \frac{-1 + \sqrt{1 + 2C_{\text{est}}\delta}}{C_{\text{est}}} = \frac{2\delta}{1 + \sqrt{1 + 2C_{\text{est}}\delta}}.$$

Define

$$\phi(\delta) := \min_{r \geq 0} m_\delta(r) = m_\delta(r^*(\delta)).$$

We claim that ϕ is strictly decreasing on $[0, \infty)$. If $0 < \delta_1 < \delta_2$ and $r_1 := r^*(\delta_1)$, then $r_1 > 0$ and

$$\phi(\delta_2) \leq m_{\delta_2}(r_1) = m_{\delta_1}(r_1) - (\delta_2 - \delta_1)r_1 = \phi(\delta_1) - (\delta_2 - \delta_1)r_1 < \phi(\delta_1).$$

Also $\phi(0) = 0$, while for every $\delta > 0$,

$$\phi(\delta) = m_\delta(r^*(\delta)) < m_\delta(0) = 0 = \phi(0).$$

Thus ϕ is strictly decreasing on $[0, \infty)$.

Consequently, since $\phi(\delta_k(f)) = \min_{t \geq 0} M_k(tf)$, minimizing M_k over $\mathcal{F}$ is equivalent to maximizing $\delta_k(f)$. By Proposition 2.5, $\delta_k(f)$ is maximized by the weak Newton direction f_{k+1}^{wN}, which additionally satisfies $\|f_{k+1}^{\text{wN}}\|_{H_k} = \delta_k(f_{k+1}^{\text{wN}}) = \delta_k(\mathcal{F})$. Therefore the minimizer of M_k is

$$r^*(\delta_k)\,\frac{f_{k+1}^{\text{wN}}}{\|f_{k+1}^{\text{wN}}\|_{H_k}} = \alpha_k f_{k+1}^{\text{wN}}, \qquad \alpha_k := \frac{r^*(\delta_k)}{\delta_k} = \frac{2}{1 + \sqrt{1 + 2C_{\text{est}}\delta_k}}.$$

Finally, as for the α_k^2 identity, substituting $r = \alpha_k \delta_k$ into the optimality condition

$$-\delta_k + r + \frac{C_{\text{est}}}{2} r^2 = 0$$

gives

$$\frac{C_{\text{est}} \delta_k}{2} \alpha_k^2 = 1 - \alpha_k.$$

This concludes the proof. □

B.2 Proof of Weak Improvement Θ_k^2 Factor

Proof of Proposition 3.3. By Theorem 3.2 we can write

$$M_k(f_{k+1}^{\text{wAICN}}) = \langle g_k, f_{k+1}^{\text{wAICN}} \rangle + \frac{1}{2} \|f_{k+1}^{\text{wAICN}}\|_{H_k}^2 + \frac{C_{\text{est}}}{6} \|f_{k+1}^{\text{wAICN}}\|_{H_k}^3 \tag{24}$$

$$= -\alpha_k \|f_{k+1}^{\text{wN}}\|_{H_k}^2 + \frac{\alpha_k^2}{2} \|f_{k+1}^{\text{wN}}\|_{H_k}^2 + \frac{C_{\text{est}} \alpha_k^3}{6} \|f_{k+1}^{\text{wN}}\|_{H_k}^3 \tag{25}$$

$$= \alpha_k \|f_{k+1}^{\text{wN}}\|_{H_k}^2 \left(-1 + \frac{\alpha_k}{2} + \frac{C_{\text{est}} \alpha_k^2}{6} \|f_{k+1}^{\text{wN}}\|_{H_k} \right) \tag{26}$$

$$= \alpha_k \|f_{k+1}^{\text{wN}}\|_{H_k}^2 \left(-1 + \frac{\alpha_k}{2} + \frac{1 - \alpha_k}{3} \right) \tag{27}$$

$$= \frac{\alpha_k(\alpha_k - 4)}{6} \|f_{k+1}^{\text{wN}}\|_{H_k}^2 \tag{28}$$

where we also used Proposition 2.5 for the inner product with the gradient, and Theorem 3.2 for the $C_{\text{est}} \alpha_k^2$ identity. This proves the first part. Next, define

$$\alpha(s) := \frac{2}{1 + \sqrt{1 + 2C_{\text{est}} s}} \in (0, 1], \qquad \psi(s) := \frac{\alpha(s)\big(4 - \alpha(s)\big)}{6} s^2. \tag{29}$$

Since $s \mapsto \alpha(s)$ is decreasing and $a \mapsto a(4-a)$ is increasing on $(0, 1]$, the prefactor $\frac{\alpha(s)(4-\alpha(s))}{6}$ is decreasing in s. Hence for any $t \leq s$,

$$\psi(t) = \frac{\alpha(t)\big(4 - \alpha(t)\big)}{6} t^2 \;\geq\; \frac{\alpha(s)\big(4 - \alpha(s)\big)}{6} (t)^2 = \left(\frac{t}{s} \right)^2 \psi(s).$$

Applying this with the choices

$$s := \|f_{k+1}^N\|_{H_k}, \qquad t := \|f_{k+1}^{\text{wN}}\|_{H_k},$$

yields

$$\psi(t) \geq \Theta_k^2 \psi(s).$$

Here we used that $t \leq s$ and

$$\Theta_k = \frac{\langle f_{k+1}^N, f_{k+1}^{\text{wN}} \rangle_{H_k}}{\|f_{k+1}^N\|_{H_k} \|f_{k+1}^{\text{wN}}\|_{H_k}} = \frac{\|f_{k+1}^{\text{wN}}\|_{H_k}}{\|f_{k+1}^N\|_{H_k}} = \frac{t}{s},$$

which follows from Proposition 2.5. Since $M_k(\mathcal{F}) = -\psi(t)$ and $M_k(\mathcal{H}) = -\psi(s)$, this yields

$$M_k(\mathcal{F}) \leq \Theta_k^2 M_k(\mathcal{H}),$$

as claimed. □

B.3 Proof of Weak Contracting-point Inequality

Proof of Proposition 3.6. By the upper Taylor bound (6) from Proposition 3.5 applied to the weak-AICN update, we obtain

$$L(F_{k+1}) \leq L(F_k) + M_k(\eta f_{k+1}^{\text{wAICN}}) \tag{30}$$

$$\leq L(F_k) + \eta M_k(f_{k+1}^{\text{wAICN}}) \tag{31}$$

$$\leq L(F_k) + \eta \Theta_k^2 M_k(f_{k+1}^{AICN}) \tag{32}$$

$$= (1 - \eta\Theta_k^2) L(F_k) + \eta \Theta_k^2 \left(L(F_k) + \min_{f \in \mathcal{H}} M_k(f) \right), \tag{33}$$

where the second inequality follows by Jensen's inequality (sum of linear and convex terms), and the third by Proposition 3.3. Additionally, applying the lower Taylor bound (6) at $f = Y - F_k$ yields

$$L(F_k) + Q_k(Y - F_k) \leq L(Y) + \frac{C_{\text{ssc}}}{6}\|Y - F_k\|_{H_k}^3.$$

Adding the cubic penalty $\frac{C_{\text{est}}}{6}\|Y - F_k\|_{H_k}^3$ to both sides to form $M_k(Y - F_k)$ and applying the assumption $C_{\text{ssc}} \leq C_{\text{est}}$, we obtain

$$\begin{aligned} L(F_k) + M_k(Y - F_k) &\leq L(Y) + \frac{C_{\text{ssc}} + C_{\text{est}}}{6}\|Y - F_k\|_{H_k}^3 \\ &\leq L(Y) + \frac{C_{\text{est}}}{3}\|Y - F_k\|_{H_k}^3. \end{aligned} \tag{34}$$

Inserting (34) into (33) yields the conclusion. □

B.4 Proof of Recurrence Lemmas

Proof of Lemma 3.7. If any $\Delta_i = 0$, all subsequent terms vanish and the bound is immediate for those terms. It therefore suffices to prove the bound for k such that $\Delta_0, \ldots, \Delta_k > 0$. The proof is based on a telescoping sum argument. We start by writing the difference between two consecutive terms as

$$\frac{1}{\sqrt{\Delta_{k+1}}} - \frac{1}{\sqrt{\Delta_k}} = \frac{\sqrt{\Delta_k} - \sqrt{\Delta_{k+1}}}{\sqrt{\Delta_k \Delta_{k+1}}} \tag{35}$$

$$= \frac{\Delta_k - \Delta_{k+1}}{\sqrt{\Delta_k \Delta_{k+1}}(\sqrt{\Delta_k} + \sqrt{\Delta_{k+1}})}. \tag{36}$$

Substituting the recurrence $\Delta_k - \Delta_{k+1} \geq c_k \Delta_k^{3/2}$ into the numerator gives

$$\frac{1}{\sqrt{\Delta_{k+1}}} - \frac{1}{\sqrt{\Delta_k}} \geq \frac{c_k \Delta_k^{3/2}}{\sqrt{\Delta_k \Delta_{k+1}}(\sqrt{\Delta_k} + \sqrt{\Delta_{k+1}})}. \tag{37}$$

Because $c_k \geq 0$ and $\Delta_k \geq 0$ by assumption, the sequence is non-increasing ($\Delta_{k+1} \leq \Delta_k$), and we obtain that

$$\frac{1}{\sqrt{\Delta_{k+1}}} - \frac{1}{\sqrt{\Delta_k}} \geq \frac{c_k \Delta_k^{3/2}}{\sqrt{\Delta_k^2}(\sqrt{\Delta_k} + \sqrt{\Delta_k})} = \frac{c_k}{2}. \tag{38}$$

Summing these terms from $i = 0$ to $k - 1$ yields the telescoping sum

$$\frac{1}{\sqrt{\Delta_k}} \geq \frac{1}{\sqrt{\Delta_0}} + \frac{1}{2}\sum_{i=0}^{k-1} c_i. \tag{39}$$

We must also bound the initial term Δ_0. To this end, observe that the recurrence at $k = 0$ implies $0 \leq \Delta_1 \leq \Delta_0 - c_0 \Delta_0^{\frac{3}{2}}$. Rearranging this inequality yields $1/\sqrt{\Delta_0} \geq c_0$. Inserting this into the summation leads to

$$\frac{1}{\sqrt{\Delta_k}} \geq c_0 + \frac{1}{2}\sum_{i=0}^{k-1} c_i = \frac{1}{2}\left(2c_0 + \sum_{i=0}^{k-1} c_i\right). \tag{40}$$

The result follows immediately by squaring and inverting the inequality. □

Proof of Lemma 3.8. Recursively applying the hypothesis $\Delta_{i+1} \leq (1 - c_i)\Delta_i$ for $i = k - 1, \ldots, 0$ immediately yields

$$\Delta_k \leq \Delta_0 \prod_{i=0}^{k-1}(1 - c_i). \tag{41}$$

Using the standard inequality $1 - x \leq e^{-x}$, valid for all real x, we bound each term in the product:

$$\prod_{i=0}^{k-1}(1 - c_i) \leq \prod_{i=0}^{k-1} e^{-c_i} = \exp\left(-\sum_{i=0}^{k-1} c_i\right). \tag{42}$$

Substituting the average $\overline{c_k}$ into the exponent completes the proof. □

B.5 Proof of Global Convergence with Weak Learners

Proof of Theorem 3.10. Let $\Delta_k := L(F_k) - L(F^*) \geq 0$. We apply (7) with

$$Y := (1-\beta)F_k + \beta F^*, \qquad \beta \in [0,1].$$

By convexity,

$$L(Y) \leq (1-\beta)L(F_k) + \beta L(F^*) = L(F^*) + (1-\beta)\Delta_k.$$

Also, $Y - F_k = \beta(F^* - F_k)$, so $\|Y - F_k\|_{H_k} \leq \beta D$ by (3.9). Thus (7) implies

$$\Delta_{k+1} \leq (1 - \eta\Theta_k^2\beta)\Delta_k + \eta\Theta_k^2 \frac{C_{\text{est}}}{3}\beta^3 D^3. \tag{43}$$

First, note that $\Delta_{k+1} \leq \Delta_k$ which follows from setting $\beta = 0$ above. Next, we choose

$$\beta := \min\left(1, \sqrt{\frac{\Delta_k}{C_{\text{est}} D^3}}\right).$$

Case 1: Linear Regime ($\beta = 1$): Since Δ_k is monotone, β will remain equal to one until the condition $\Delta_k < C_{\text{est}} D^3$ is met, where it will then remain strictly less than one. In the case $\beta = 1$, we have that $\Delta_k \geq C_{\text{est}} D^3$. Factoring Δ_k out of (43) yields

$$\Delta_{k+1} \leq \left(1 - \eta\Theta_k^2 + \eta\Theta_k^2 \frac{C_{\text{est}} D^3}{3\Delta_k}\right)\Delta_k \leq \left(1 - \frac{2}{3}\eta\Theta_k^2\right)\Delta_k.$$

Applying Lemma 3.8 to this multiplicative recurrence gives the linear rate:

$$\Delta_k \leq \Delta_0 \exp\left(-\frac{2}{3}\eta k \overline{\Theta_k^2}\right).$$

By definition, k_0 is the first iteration where $\Delta_{k_0} \leq C_{\text{est}} D^3$. If $k_0 > 1$, then $\Delta_{k_0-1} > C_{\text{est}} D^3$. Evaluating our linear rate at $k = k_0 - 1$ and chaining these inequalities yields

$$C_{\text{est}} D^3 < \Delta_{k_0-1} \leq \Delta_0 \exp\left(-\frac{2}{3}\eta(k_0-1)\overline{\Theta_{k_0-1}^2}\right).$$

Dividing by Δ_0, taking the logarithm of both sides, and rearranging for k_0 gives the stated bound on the length of the initial phase.

Case 2: Global Regime ($\beta < 1$): Once $\Delta_{k_0} \leq C_{\text{est}} D^3$, monotonicity implies $\Delta_k \leq C_{\text{est}} D^3$ for all $k \geq k_0$. Thus for $k \geq k_0$, we have $\beta = \sqrt{\frac{\Delta_k}{C_{\text{est}} D^3}}$. Substituting this β into (43) yields

$$\Delta_{k+1} \leq \Delta_k - \frac{2\eta\Theta_k^2}{3\sqrt{C_{\text{est}} D^3}}\Delta_k^{3/2}.$$

Let $\gamma_t := \Delta_{k_0+t}$ for $t \geq 0$. The sequence satisfies $\gamma_{t+1} \leq \gamma_t - c_t\gamma_t^{3/2}$ with sequence constants $c_t = \frac{2\eta\Theta_{k_0+t}^2}{3\sqrt{C_{\text{est}} D^3}}$. Applying Lemma 3.7 directly gives

$$\gamma_t \leq \frac{4}{\overline{c_t}^2 t^2} = \frac{9\, C_{\text{est}}\, D^3}{\eta^2(\overline{\Theta_{k_0,k_0+t}^2})^2 t^2}.$$

Substituting $t = k - k_0$ back into the inequality concludes the proof. □

Remark B.1. If $\eta = 1$ and $\mathcal{F} = \mathcal{H}$, then $f_{k+1}^{\text{wN}} = f_{k+1}^{\text{N}}$ and $\Theta_k = 1$. Applying (43) at $k = 0$ with $\beta = 1$ yields $\Delta_1 \leq \frac{1}{3}C_{\text{est}} D^3 < C_{\text{est}} D^3$. This implies that $k_0 \leq 1$ when not using weak learners.

B.6 Proof of Local Linear Convergence with Weak Learners

Lemma B.2. Write $\delta_k := \delta_k(f_{k+1}^{\mathrm{wN}})$. If $C_{\mathrm{est}}\|\eta f_{k+1}^{\mathrm{wAICN}}\|_{H_k} = C_{\mathrm{est}}\eta\alpha_k\delta_k < 1$ then

$$\left(1 - C_{\mathrm{est}}\eta\alpha_k\delta_k\right)^2 H_{k+1} \preceq H_k \preceq \frac{1}{\left(1 - C_{\mathrm{est}}\eta\alpha_k\delta_k\right)^2} H_{k+1} \tag{44}$$

Proof. Follows from [28, Theorem 5.1.7] applied to the weak-AICN update f_{k+1}^{wAICN} since semi-strong self-concordance implies self-concordance. □

Proof of Theorem 3.11. The weak-AICN update is $F_{k+1} = F_k + s_k$, where $s_k = \eta\alpha_k f_{k+1}^{\mathrm{wN}} = \eta f_{k+1}^{\mathrm{wAICN}}$. Using the fundamental theorem of calculus, we expand the gradient at the next weak-AICN iterate:

$$g_{k+1} = g_k + \int_0^1 \nabla^2 L(F_k + ts_k)s_k \, dt \tag{45}$$

$$= \underbrace{g_k + H_k s_k}_{v_{\mathrm{lin}}} + \underbrace{\int_0^1 \left(\nabla^2 L(F_k + ts_k) - H_k\right) s_k \, dt}_{e_k}. \tag{46}$$

We analyze the norms of v_{lin} and e_k separately.

1. Bound on the Linear Term v_{lin}: Recall that $g_k = -H_k f_{k+1}^{\mathrm{N}}$. Thus, $v_{\mathrm{lin}} = -H_k f_{k+1}^{\mathrm{N}} + \eta\alpha_k H_k f_{k+1}^{\mathrm{wN}}$. Taking squared dual norms we see that

$$\|v_{\mathrm{lin}}\|_{H_k}^{*2} = \| - f_{k+1}^{\mathrm{N}} + \eta\alpha_k f_{k+1}^{\mathrm{wN}}\|_{H_k}^2 \tag{47}$$

$$= \|f_{k+1}^{\mathrm{N}}\|_{H_k}^2 + \eta^2\alpha_k^2\|f_{k+1}^{\mathrm{wN}}\|_{H_k}^2 - 2\eta\alpha_k\langle f_{k+1}^{\mathrm{N}}, f_{k+1}^{\mathrm{wN}}\rangle_{H_k}. \tag{48}$$

Write $\delta_k = \delta_k(f_{k+1}^{\mathrm{wN}})$ for brevity. By Proposition 2.5 we have the identities $\langle f_{k+1}^{\mathrm{N}}, f_{k+1}^{\mathrm{wN}}\rangle_{H_k} = \delta_k^2$ and $\|f_{k+1}^{\mathrm{wN}}\|_{H_k} = \delta_k = \Theta_k\|f_{k+1}^{\mathrm{N}}\|_{H_k}$. Substituting these into the equation yields

$$\|v_{\mathrm{lin}}\|_{H_k}^{*2} = \left(\frac{\delta_k}{\Theta_k}\right)^2 + \eta^2\alpha_k^2\delta_k^2 - 2\eta\alpha_k\delta_k^2 \tag{49}$$

$$= \left(\frac{\delta_k}{\Theta_k}\right)^2 \left[1 - \Theta_k^2\eta\alpha_k(2 - \eta\alpha_k)\right] \tag{50}$$

$$= \|g_k\|_{H_k}^{*2} \left[1 - \Theta_k^2\eta\alpha_k(2 - \eta\alpha_k)\right]. \tag{51}$$

Taking square roots and using the bounds $\sqrt{1-x} \le 1 - \frac{x}{2}$ and $\eta\alpha_k \le 1$ gives

$$\|v_{\mathrm{lin}}\|_{H_k}^{*} = \|g_k\|_{H_k}^{*}\sqrt{1 - \Theta_k^2\eta\alpha_k(2 - \eta\alpha_k)} \tag{52}$$

$$\le \|g_k\|_{H_k}^{*}\left(1 - \Theta_k^2\eta\alpha_k\frac{2 - \eta\alpha_k}{2}\right) \tag{53}$$

$$\le \|g_k\|_{H_k}^{*}\left(1 - \frac{\Theta_k^2\eta\alpha_k}{2}\right). \tag{54}$$

2. Bound on the Error Term e_k: Using the definition of the Hessian-induced affine-invariant operator norm and semi-strong self-concordance we find that

$$\|e_k\|_{H_k}^{*} \le \int_0^1 \|\left(\nabla^2 L(F_k + ts_k) - H_k\right) s_k\|_{H_k}^{*} \, dt \tag{55}$$

$$\le \int_0^1 \|\nabla^2 L(F_k + ts_k) - H_k\|_{\mathrm{aff-op},H_k}\|s_k\|_{H_k} \, dt \tag{56}$$

$$\le \int_0^1 C_{\mathrm{est}}\|ts_k\|_{H_k}\|s_k\|_{H_k} \, dt \tag{57}$$

$$= \frac{C_{\mathrm{est}}}{2}\|s_k\|_{H_k}^2. \tag{58}$$

Substituting $\|s_k\|_{H_k} = \eta\alpha_k\delta_k$ and $\|g_k\|_{H_k}^* = \frac{\delta_k}{\Theta_k}$ gives the final bound

$$\|e_k\|_{H_k}^* \leq \|g_k\|_{H_k}^* \frac{C_{\text{est}}}{2}\eta^2\alpha_k^2\delta_k\Theta_k. \tag{59}$$

3. Change of Metric: We must measure the gradient g_{k+1} in the norm induced by H_{k+1}. Let $r_k := \|s_k\|_{H_k} = \eta\alpha_k\delta_k$. By Lemma B.2 we have for $C_{\text{est}}r_k < 1$ and all $u \in \mathcal{H}$ that

$$\|u\|_{H_{k+1}}^* \leq \frac{1}{1 - C_{\text{est}}r_k}\|u\|_{H_k}^*. \tag{60}$$

Applying this to $u = g_{k+1} = v_{\text{lin}} + e_k$ and substituting the above bounds yields

$$\|g_{k+1}\|_{H_{k+1}}^* \leq \frac{1}{1 - C_{\text{est}}\eta\alpha_k\delta_k}\left(\|v_{\text{lin}}\|_{H_k}^* + \|e_k\|_{H_k}^*\right) \tag{61}$$

$$\leq \frac{\|g_k\|_{H_k}^*}{1 - C_{\text{est}}\eta\alpha_k\delta_k}\left(1 - \frac{\Theta_k^2\eta\alpha_k}{2} + \frac{1}{2}C_{\text{est}}\delta_k\Theta_k\eta^2\alpha_k^2\right) \tag{62}$$

$$=: \|g_k\|_{H_k}^* \frac{\mathcal{N}}{\mathcal{D}}. \tag{63}$$

4. Contraction: Denote the numerator and denominator of the fraction multiplying $\|g_k\|_{H_k}^*$ by $\mathcal{N}$ and $\mathcal{D}$ respectively. We proceed by analyzing the difference $\mathcal{D} - \mathcal{N}$.

$$\mathcal{D} - \mathcal{N} = (1 - C_{\text{est}}\eta\alpha_k\delta_k) - \left(1 - \frac{\Theta_k^2\eta\alpha_k}{2} + \frac{1}{2}C_{\text{est}}\delta_k\Theta_k\eta^2\alpha_k^2\right) \tag{64}$$

$$= \frac{\Theta_k^2\eta\alpha_k}{2} - C_{\text{est}}\eta\alpha_k\delta_k - \frac{1}{2}C_{\text{est}}\delta_k\Theta_k\eta^2\alpha_k^2 \tag{65}$$

$$= \eta\alpha_k\left[\frac{\Theta_k^2}{2} - C_{\text{est}}\delta_k\left(1 + \frac{\eta\alpha_k\Theta_k}{2}\right)\right]. \tag{66}$$

Since $\eta, \alpha_k, \Theta_k \in (0, 1]$, we have $1 + \frac{\eta\alpha_k\Theta_k}{2} \leq \frac{3}{2}$. Substituting this bound and the hypothesis $\delta_k \leq \frac{\Theta_k^2}{6C_{\text{est}}}$ gives

$$\mathcal{D} - \mathcal{N} \geq \eta\alpha_k\left[\frac{\Theta_k^2}{2} - \frac{3}{2}C_{\text{est}}\delta_k\right] \tag{67}$$

$$\geq \eta\alpha_k\left[\frac{\Theta_k^2}{2} - \frac{3}{2}C_{\text{est}}\left(\frac{\Theta_k^2}{6C_{\text{est}}}\right)\right] \tag{68}$$

$$= \frac{1}{4}\eta\alpha_k\Theta_k^2. \tag{69}$$

Finally, since $\delta_k \leq \frac{1}{6C_{\text{est}}}$ and $\eta\alpha_k \leq 1$, the denominator satisfies $\mathcal{D} = 1 - C_{\text{est}}\eta\alpha_k\delta_k \geq 5/6$, and specifically $\mathcal{D} \leq 1$. Therefore,

$$\frac{\mathcal{N}}{\mathcal{D}} = 1 - \frac{\mathcal{D} - \mathcal{N}}{\mathcal{D}} \leq 1 - (\mathcal{D} - \mathcal{N}) \leq 1 - \frac{1}{4}\eta\alpha_k\Theta_k^2. \tag{70}$$

Setting $\rho_k = \frac{1}{4}\eta\alpha_k\Theta_k^2$ completes the proof of the one-step contraction.

5. Final Linear Convergence: The inner product on $\mathcal{H}$ can be bounded by the product of the Hessian-induced norm and its dual. For any $u, v \in \mathcal{H}$ we have by Cauchy-Schwarz inequality

$$\langle u, v\rangle = \langle H_k^{-1/2}u, H_k^{1/2}v\rangle \leq \|H_k^{-1/2}u\|\|H_k^{1/2}v\| = \|u\|_{H_k}^*\|v\|_{H_k}. \tag{71}$$

Applying this to g_k and $(F_k - F^*)$, after using the convexity of L, yields

$$L(F_k) - L(F^*) \leq \langle g_k, F_k - F^*\rangle \leq \|g_k\|_{H_k}^*\|F_k - F^*\|_{H_k} \leq \|g_k\|_{H_k}^* D, \tag{72}$$

where the final inequality follows from Assumption 3.9. Finally, applying Lemma 3.8 to the multiplicative recurrence $\|g_{k+1}\|_{H_{k+1}}^* \leq (1 - \rho_k)\|g_k\|_{H_k}^*$ gives the result. □

Remark B.3 (Sufficient conditions for the bounded local radius).
As previously discussed, a bounded-radius condition of the form Assumption 3.9 is also used in classical AICN analysis [16]. However, simple sufficient conditions for this bound follow from standard regularity conditions on L. Under the assumptions of Proposition 3.6, taking $Y = F_k$ gives $L(F_{k+1}) \le L(F_k)$. If L is additionally μ-strongly convex, then

$$\|F_k - F^*\|^2 \le R_0^2 := \frac{2}{\mu}\big(L(F_0) - L(F^*)\big).$$

If L is also S-smooth, then

$$\|F_k - F^*\|_{H_k}^2 \le S\|F_k - F^*\|^2 \le SR_0^2,$$

hence Assumption 3.9 holds with $D = \sqrt{S}\, R_0$. Alternatively, if its Hessian is M-Lipschitz, then

$$\|H_k\|_{\mathrm{op}} \le \|\nabla^2 L(F^*)\|_{\mathrm{op}} + M\|F_k - F^*\| \le \|\nabla^2 L(F^*)\|_{\mathrm{op}} + MR_0,$$

and hence

$$\|F_k - F^*\|_{H_k}^2 \le \|H_k\|_{\mathrm{op}}\|F_k - F^*\|^2 \le \big(\|\nabla^2 L(F^*)\|_{\mathrm{op}} + MR_0\big)R_0^2 =: D^2.$$

For examples of radius assumptions in stochastic zeroth-order and bandit convex optimization, see [20, 41]; note that these settings and results are not directly comparable to ours.

B.7 Proof of Linear Rate for Hessian-dominated Losses

Proof of Theorem 3.13. Write $\delta_k = \delta_k(f_{k+1}^{\mathrm{wN}})$. We start by applying the affine-invariant cubic bound (Proposition 3.5) to the weak AICN update, which by Theorem 3.2 takes the form $f_{k+1}^{\mathrm{wAICN}} = \alpha_k f_{k+1}^{\mathrm{wN}}$ gives

$$\begin{aligned}
L(F_k) - L(F_{k+1}) &\ge -Q_k(\eta f_{k+1}^{\mathrm{wAICN}}) - \frac{C_{\mathrm{est}}}{6}\|\eta f_{k+1}^{\mathrm{wAICN}}\|_{H_k}^3 \\
&= \eta\alpha_k\delta_k^2 - \frac{1}{2}(\eta\alpha_k)^2\delta_k^2 - \frac{C_{\mathrm{est}}}{6}(\eta\alpha_k)^3\delta_k^3 \\
&= \eta\alpha_k\delta_k^2\left(1 - \frac{1}{2}\eta\alpha_k - \eta^2\left[\frac{C_{\mathrm{est}}\delta_k}{6}\alpha_k^2\right]\right).
\end{aligned}$$

To simplify the bracketed term, we recall from Theorem 3.2 that α_k satisfies $\frac{C_{\mathrm{est}}\delta_k}{6}\alpha_k^2 = \frac{1-\alpha_k}{3}$. Substituting this back into the descent inequality, we obtain

$$\begin{aligned}
L(F_k) - L(F_{k+1}) &\ge \eta\alpha_k\delta_k^2\left(1 - \frac{1}{2}\eta\alpha_k - \eta^2\frac{1-\alpha_k}{3}\right) \\
&= \eta\alpha_k\delta_k^2\left(1 - \frac{\eta^2}{3} - \alpha_k\left[\frac{\eta}{2} - \frac{\eta^2}{3}\right]\right) \\
&\ge \eta\alpha_k\delta_k^2\left(1 - \frac{\eta}{2}\right) \\
&= \eta\alpha_k\Theta_k^2\|f_{k+1}^N\|_{H_k}^2\left(1 - \frac{\eta}{2}\right)
\end{aligned}$$

The final inequality follows because $\eta \in (0, 1]$ implies $\frac{\eta}{2} \ge \frac{\eta^2}{3}$, meaning the term in square brackets is non-negative. Since we subtract this term scaled by $\alpha_k < 1$, the expression is minimized as $\alpha_k \to 1$. Finally, applying the Hessian-dominance condition $\|f_{k+1}^N\|_{H_k}^2 \ge c\big(L(F_k) - L_\star\big)$ and rearranging yields the result. □

C Semi-Strong Self-Concordance in the Boosting Space

To prove the inheritance of the SSC property for empirical risks (Theorem 4.1), we first explicitly state the SSC condition for 1D scalar functions.

Lemma C.1 (One-dimensional SSC). Let $l : \mathbb{R} \to \mathbb{R}$ be twice differentiable with $l''(z) > 0$. The semi-strong self-concordance condition (Definition 3.4) is equivalent in one dimension to:

$$\frac{|l''(z+t) - l''(z)|}{l''(z)} \leq C_l \, |t| \sqrt{l''(z)} \qquad \forall z, t \in \mathbb{R}.$$

Proof. In one dimension, the Hessian operator reduces to the positive scalar $h_z = l''(z)$. The affine-invariant operator norm of a scalar a is therefore $\|a\|_{\text{aff-op},h_z} = \|h_z^{-1/2} a h_z^{-1/2}\|_{\text{op}} = |a|/h_z$. The Hessian norm of t is $\|t\|_{h_z} = |t|\sqrt{h_z}$. Substituting these into the definition of SSC yields the result. □

Proof of Theorem 4.1. We operate in the space $\mathcal{H} = \mathbb{R}^n$ equipped with the empirical inner product $\langle u, v \rangle_X = \frac{1}{n} \sum_{i=1}^n u_i v_i$. The empirical risk is $L(F) = \frac{1}{n} \sum_{i=1}^n l(F(x_i), y_i)$. For any iterate $F \in \mathcal{H}$ and direction $f \in \mathcal{H}$, the Hessian operator $H_F := \nabla^2 L(F)$ and its variation $\Delta := \nabla^2 L(F + f) - \nabla^2 L(F)$ act as diagonal linear operators on $\mathcal{H}$. Specifically, their representations in the standard basis are diagonal matrices with entries $d_i = l''(F(x_i), y_i)$ and $\Delta_i = l''(F(x_i) + f(x_i), y_i) - l''(F(x_i), y_i)$, respectively.

For any diagonal operator, the operator norm induced by $\langle \cdot, \cdot \rangle_X$ is simply the maximum absolute value of its entries. Therefore, the affine-invariant operator norm is

$$\|\Delta\|_{\text{aff-op},H_F} \;=\; \|H_F^{-1/2} \Delta H_F^{-1/2}\|_{\text{op}} \;=\; \max_{1 \leq i \leq n} \frac{|\Delta_i|}{d_i}.$$

Applying the 1D SSC condition (Lemma C.1) to the base loss $l(\cdot, y_i)$, we bound each component by

$$\frac{|\Delta_i|}{d_i} \;\leq\; C_l |f(x_i)| \sqrt{l''(F(x_i), y_i)} \;=\; C_l |f(x_i)| \sqrt{d_i}.$$

Taking the maximum over all $i = 1, \ldots, n$, we obtain

$$\|\Delta\|_{\text{aff-op},H_F} \leq C_l \max_{1 \leq i \leq n} \left(|f(x_i)| \sqrt{d_i} \right).$$

Finally, we relate this maximum to the local Hessian norm $\|f\|_{H_F}$. By definition, the squared Hessian norm is the empirical average $\|f\|_{H_F}^2 = \frac{1}{n} \sum_{i=1}^n d_i f(x_i)^2$. Since the maximum element of a set of non-negative numbers is bounded by their sum, we have

$$\max_{1 \leq i \leq n} \left(d_i f(x_i)^2 \right) \;\leq\; \sum_{i=1}^n d_i f(x_i)^2 \;=\; n \left(\frac{1}{n} \sum_{i=1}^n d_i f(x_i)^2 \right) \;=\; n \|f\|_{H_F}^2.$$

Taking the square root and substituting this into our operator norm bound yields

$$\|\Delta\|_{\text{aff-op},H_F} \leq \sqrt{n} C_l \|f\|_{H_F},$$

proving that the empirical risk L is semi-strongly self-concordant with constant $C_{\text{ssc}}(L) \leq \sqrt{n} C_l$. □

D Newton Coordinate Descent: Lower Bound Proofs and Experiments

D.1 Proof of Newton Coordinate Descent Cosine Angle Lower Bound

Proof of Proposition 4.4. Recall from Proposition 2.5 that the cosine angle can be expressed as the ratio of directional improvements $\Theta_k = \delta_k(f_{k+1}^{\mathrm{wN}})/\delta_k(f_{k+1}^{\mathrm{N}})$. We bound the numerator and denominator separately.

1. Lower bound on the weak improvement: For coordinate weak learners, the maximal directional improvement is achieved at the optimal coordinate i^*. Using the standard Rayleigh quotient bound $(H_k)_{ii} \le \lambda_{\max}(H_k)$, we have:

$$\delta_k(f_{k+1}^{\mathrm{wN}}) \;=\; \max_{1\le i\le d} \frac{|(g_k)_i|}{\sqrt{(H_k)_{ii}}} \;\ge\; \max_{1\le i\le d} \frac{|(g_k)_i|}{\sqrt{\lambda_{\max}(H_k)}} \;=\; \frac{\|g_k\|_\infty}{\sqrt{\lambda_{\max}(H_k)}}. \tag{73}$$

2. Upper bound on the exact improvement: By Proposition 2.5, the exact directional improvement is exactly the dual Hessian norm of the gradient. Using the eigenvalue bound $H_k^{-1} \preceq \lambda_{\min}(H_k)^{-1} I$ and the vector norm equivalence $\|g_k\|_2 \le \sqrt{d}\|g_k\|_\infty$, we obtain:

$$\delta_k(f_{k+1}^{\mathrm{N}}) \;=\; \sqrt{\langle g_k, H_k^{-1} g_k\rangle} \;\le\; \sqrt{\frac{1}{\lambda_{\min}(H_k)}\|g_k\|_2^2} \;\le\; \sqrt{\frac{d}{\lambda_{\min}(H_k)}}\|g_k\|_\infty. \tag{74}$$

3. Final Ratio: Substituting these two bounds into the cosine angle identity yields:

$$\Theta_k \;=\; \frac{\delta_k(f_{k+1}^{\mathrm{wN}})}{\delta_k(f_{k+1}^{\mathrm{N}})} \;\ge\; \frac{\|g_k\|_\infty/\sqrt{\lambda_{\max}(H_k)}}{\sqrt{d/\lambda_{\min}(H_k)}\|g_k\|_\infty} \;=\; \sqrt{\frac{\lambda_{\min}(H_k)}{d\lambda_{\max}(H_k)}} \;=\; \frac{1}{\sqrt{d\kappa_k}}. \tag{75}$$

This completes the proof. □

D.2 Algorithm for Newton Coordinate Descent

For completeness and ease of use, we detail the exact computation for the Newton coordinate weak learner below.

Algorithm 2 Computing f_{k+1}^{wN} for coordinate weak learners

1: **Input:** Diagonal of Hessian $\mathrm{diag}(H_k)$ and gradient g_k
2: $A \leftarrow g_k \odot g_k \oslash \mathrm{diag}(H_k)$
3: $i^* \leftarrow \mathrm{argmax}_{1\le i\le d} |A_i|$
4: $f_{k+1}^{\mathrm{wN}} \leftarrow -\frac{(g_k)_{i^*}}{(H_k)_{i^*i^*}} e_{i^*}$

D.3 Adaptive Weak-AICN Algorithm

In the classical finite-dimensional convex optimization literature, adaptive variants of Newton algorithms are often preferred in practice. Rather than relying on a fixed, global regularity constant (which can be difficult to compute or overly conservative), adaptive schemes dynamically estimate the local regularity during optimization. While the original authors of exact-AICN [16] did not explore such an algorithm, we provide an adaptive variant of weak-AICN below, which we call weak-AdaAICN.

Weak-AdaAICN operates using a backtracking line search that dynamically adjusts the local estimate C_k. At each iteration, if the empirical decrease in the loss does not match the decrease predicted by the cubic model, we double the estimate C_k and recompute the stepsize α_k. Crucially, this check is highly computationally efficient. Because Proposition 3.3 allows us to explicitly evaluate the minimum of the cubic model as $M_k(\alpha_k f_{k+1}^{\mathrm{wN}}) = \frac{\alpha_k(\alpha_k-4)}{6}\delta_k^2$, validating the descent condition requires only scalar operations and a single loss evaluation, completely bypassing the need to re-evaluate the Hessian or the weak learner.

From a theoretical standpoint, Proposition 3.5 guarantees that the required descent condition is satisfied whenever $C_k \ge C_{\mathrm{ssc}}$. As a result, the adaptive constant will never grow indefinitely and is bounded by $C_k \le 2\max(C_0, C_{\mathrm{ssc}})$.

Algorithm 3 Adaptive Weak AICN

Input: Convex loss L, Weak learner set $\mathcal{F}$, Learning rate $\eta \in (0, 1]$, Initial regularity constant $C_0 > 0$ and initial value F_0.
for $k = 0, 1, \dots$ **do**
  Compute gradient $g_k = \nabla L(F_k)$ and Hessian $H_k = \nabla^2 L(F_k)$.
  $f_{k+1}^{\mathrm{wN}} \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \left(\langle g_k, f \rangle + \frac{1}{2} \langle f, H_k f \rangle \right)$
  $\delta_k \leftarrow \sqrt{|\langle g_k, f_{k+1}^{\mathrm{wN}} \rangle|}$
  repeat
    Increase estimate $C_k \leftarrow 2C_k$
    $\alpha_k \leftarrow \frac{2}{1+\sqrt{1+2C_k\delta_k}}$
    $M_k \leftarrow \frac{\alpha_k(\alpha_k-4)}{6}\delta_k^2$
    $F_{test} \leftarrow F_k + \alpha_k f_{k+1}^{\mathrm{wN}}$
  until $L(F_{test}) \leq L(F_k) + M_k$
  Update $F_{k+1} \leftarrow F_k + \eta\alpha_k f_{k+1}^{\mathrm{wN}}$
  Decrease estimate $C_{k+1} \leftarrow C_k/4$
end for

D.4 Experimental Validation of weak-AdaAICN

We implement weak-AdaAICN as an OPTAMI-compatible [18] PyTorch optimizer [34] to validate its effectiveness against the non-adaptive AICN algorithm [16]. Following the experimental setup of the original AICN paper, we benchmark the methods on Nesterov's second-order lower bound function [29], with optional ℓ_2 regularization μ. For $x \in \mathbb{R}^d$, the objective is defined as

$$L(x) = \frac{1}{3}\sum_{i=1}^{d-1} |x_i - x_{i+1}|^3 + \frac{1}{3}|x_d|^3 - x_1 + \frac{\mu}{2}\|x\|_2^2. \tag{76}$$

For exactness and reproducibility, this objective can be efficiently implemented in PyTorch using the following snippet:

```python
import torch

def nesterov_general_lower_bounded_fun(x, p_method: int = 2, mu: float = 1e-3):
    """
    Args:
        x: Input tensor of shape (d,)
        p_method: The order of the optimization method (p).
                  e.g., Newton is p=2, Tensor is p>=3.
        mu: optional l2-reg for stability (not used by Nesterov).
    """
    power = p_method + 1
    diffs = torch.abs(x[:-1] - x[1:])
    chain_term = torch.sum(diffs**power) / power
    last_term = (torch.abs(x[-1])**power) / power
    linear_term = -x[0]
    l2_reg = (mu / 2) * torch.sum(x**2)
    return chain_term + last_term + linear_term + l2_reg
```

We evaluate the adaptive scheme in two distinct settings. First, we consider the exact (full-space) setting, recovering exact AICN by setting the weak learner family to the full space $\mathcal{F} = \mathcal{H} = \mathbb{R}^d$. Second, we consider Newton Coordinate Descent (as introduced in Section 4.2) by choosing the weak learner family to be the set of scaled standard basis vectors, $\mathcal{F}_{\mathrm{CD}} = \{ce_i : c \in \mathbb{R}, 1 \leq i \leq d\}$.

For both settings, we compare three specific optimizer configurations:

1. Non-adaptive AICN with an optimally tuned regularity constant (C_{best}), representing the smallest empirical constant that ensures a monotonic decrease in loss.
2. Non-adaptive AICN with a conservative constant choice ($C_{\mathrm{conservative}}$).
3. Our proposed AdaAICN scheme initialized with the conservative guess ($C_{\mathrm{conservative}}$).

For the exact full-space experiments, we set $C_{\mathrm{best}} = 81$ and $C_{\mathrm{conservative}} = 1000$. For the coordinate descent experiments, we set $C_{\mathrm{best}} = 16$ and $C_{\mathrm{conservative}} = 100$. We set the problem dimension

to $d = 100$ for the exact setting and $d = 20$ for the coordinate descent setting. All methods are initialized from the same random starting point. We set the regularization parameter to $\mu = 10^{-3}$ and learning rate $\eta = 1.0$.

The results, presented in Fig. 2 and Fig. 3, demonstrate that the adaptive variant significantly accelerates convergence across both settings, improving both the iteration complexity and the total wall-clock time. In the exact setting, AdaAICN achieves a $2\times$ reduction in wall-clock time and in the number of iterations required to reach the minimum value, rapidly improving the performance of the statically tuned C_{best} model without requiring costly hyperparameter sweeps. In the weak learner coordinate descent setting (Fig. 3), the difference is significant but smaller compared to the exact setting.

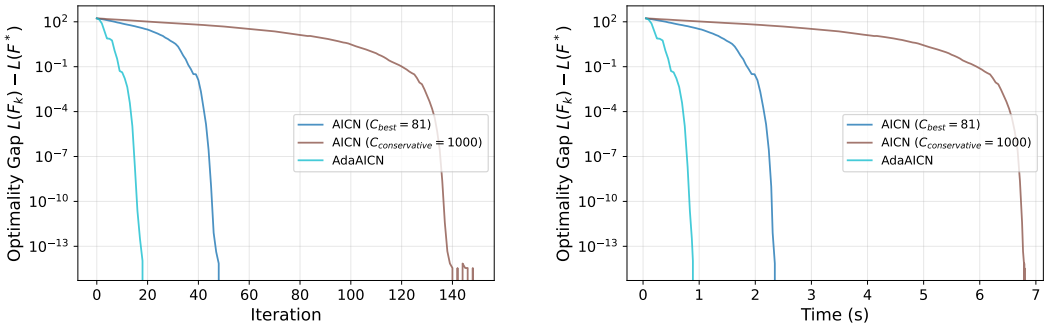


Figure 2: Performance of exact AdaAICN versus non-adaptive AICN on Nesterov's second-order lower bound function ($d = 100$). **Left:** Loss against the number of optimization iterations. **Right:** Loss against wall-clock time (seconds).

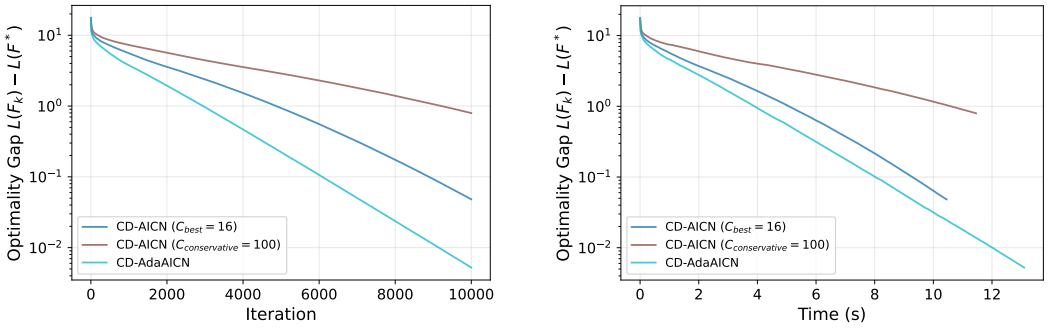


Figure 3: Performance of Newton coordinate descent AdaAICN versus non-adaptive AICN on Nesterov's second-order lower bound function ($d = 20$). **Left:** Loss against the number of optimization iterations. **Right:** Loss against wall-clock time (seconds).

E Supplementary Tree Boosting Experiments on Real-World Data

In this appendix, we compare the training loss trajectories of Newton boosting methods and study the behavior of the weak learner alignment metrics used by weak-GRN [42], weak-AICN, and vanilla first-order gradient boosting. This section parallels the analysis of [42], using the same losses and datasets as presented there. Our code builds upon the JAX [5] boosting implementation released alongside weak-GRN, adding the weak-AICN cubic regularization as an optional parameter within the Newton boosting class.

E.1 Charbonnier Loss: Non-convergence of Vanilla Newton Boosting

A classical example of the divergence of Newton's method is the Charbonnier loss function, $l(\widehat{y}, y) = \sqrt{1 + (y - \widehat{y})^2} - 1$. For this experiment, we use the UCI Wine Quality dataset [8] and fit decision trees of depth $J = 4$. We compare weak-AICN to weak-GRN [42], vanilla Newton boosting, and vanilla first-order gradient boosting across two learning rates: $\eta = 1.0$ and $\eta = 0.1$. We apply a static ℓ_2 regularization of $\lambda_{\text{base}} = 0.1$.

As discussed in Section 3, weak-AICN relies on the C_{est} parameter determining the affine-invariant cubic regularization used by the method. Weak-GRN on the other hand, uses a different mechanism to globalize Newton's method. It instead applies extra adaptive gradient-based ℓ_2 regularization equal to $\lambda_k = \sqrt{M\|g_k\|}$ at each iteration k, where g_k is the gradient and M is the Lipschitz constant of the Hessian operator. For these experiments, we select the smallest values of C_{est} and M that successfully prevent divergence for both algorithms, resulting in $C_{\text{est}} = 10$ and $M = 0.1$.

The training loss trajectories are presented in Fig. 4. Consistent with the findings in [42], we observe that vanilla Newton boosting fails to converge in the undamped setting ($\eta = 1.0$), while weak-GRN, weak-AICN, and first-order boosting all converge robustly. Under this fast learning rate, the three convergent methods perform comparably. When the learning rate is reduced to $\eta = 0.1$, vanilla Newton boosting successfully converges. For consistency, we maintain the same C_{est} and M parameters across both learning rate settings. Because weak-GRN dynamically adds ℓ_2 regularization and weak-AICN dynamically increases dampening, both methods converge slightly slower than the vanilla Newton update in the $\eta = 0.1$ regime. Notably, if the parameters C_{est} and M were set to 0, both weak-GRN and weak-AICN would reduce to vanilla Newton boosting. First-order boosting naturally lags behind the second-order methods, as it fails to leverage curvature information for the Charbonnier loss.

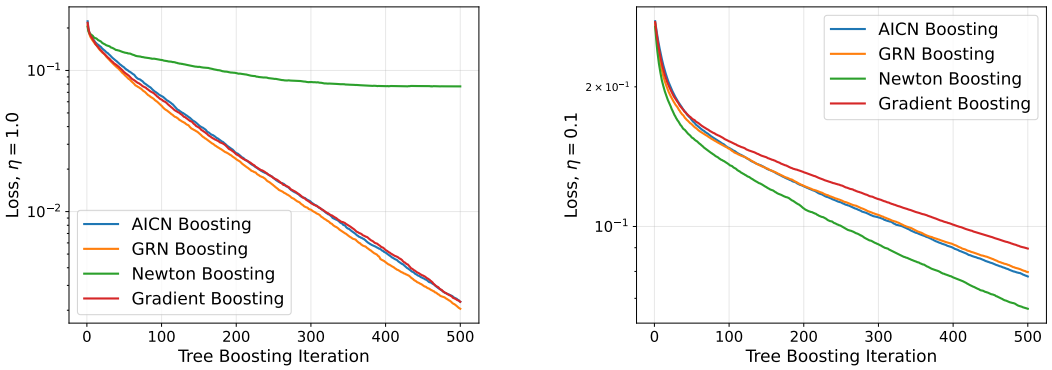


Figure 4: Training loss trajectories on the UCI Wine Quality dataset using the Charbonnier loss. We compare weak-AICN, weak-GRN, vanilla Newton boosting, and first-order gradient boosting at two learning rates: $\eta = 1.0$ (left) and $\eta = 0.1$ (right). Vanilla Newton boosting fails to converge without sufficient dampening ($\eta = 1.0$), while weak-AICN and weak-GRN converge robustly across both settings.

E.2 Binary Cross Entropy: Effects of Tree Depth and Weak Learner Edge

In this section, we compare the weak learner alignment metrics that govern the convergence of weak-AICN and weak-GRN. These metrics are the Hessian-induced cosine angle Θ_k (used by weak-AICN) and the weak gradient edge γ_k (used by weak-GRN [42]), defined below for the reader's convenience.

Definition E.1 (Weak Gradient Edge [42]). Let f_{k+1}^{wGRN} be the weak-GRN update, and let $\lambda_k = \lambda_{\mathrm{base}} + \sqrt{M\|g_k\|}$ be the effective weak-GRN regularization. The **implied weak gradient** is defined as

$$g_k^w := -(H_k + \lambda_k I) f_{k+1}^{\mathrm{wGRN}}. \tag{77}$$

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

$$\|g_k^w - g_k\|^2 \leq (1 - \gamma_k^2)\|g_k\|^2. \tag{78}$$

We also include comparisons against first-order gradient boosting [14], whose convergence relies on Euclidean cosine angles, denoted here as Θ_k^E. We define this cosine angle analogously to Definition 2.4, but evaluated using the standard Euclidean (Hilbert) norm rather than the Hessian-induced norm, and using first-order trees rather than second-order trees.

Following the experimental setup of [42], we minimize the binary cross entropy loss on a $10,000$-sample subset of the HIGGS classification dataset [39]. Because the binary cross entropy loss is well-behaved (specifically, Hessian-dominated per Definition 3.12), aggressive regularization is unnecessary; we therefore fix a small baseline penalty of $\lambda_{\mathrm{base}} = 1 \times 10^{-4}$. Throughout these experiments, we set the learning rate to $\eta = 0.1$, the tree depth to $J = 4$, and the global regularization constants to $C_{\mathrm{est}} = 1.0$ and $M = 0.01$. As in the original weak-GRN experiments [42], the static ℓ_2 term regularizes only the quadratic step: each sample Hessian is replaced by $h_{(i)} + \lambda_{\mathrm{base}} I$, while the sample gradients remain unchanged. This differs from adding ℓ_2-regularization to the loss itself, which modifies both the sample gradients and Hessians.

Fig. 5 visualizes the training loss trajectories, the weak-AICN dampening factor α_k and directional improvement δ_k, and the weak-GRN regularization λ_k. As optimization progresses, the weak-AICN dampening factor α_k gradually approaches 1.0. Recall that $\alpha_k = \frac{2}{1+\sqrt{1+2C_{\mathrm{est}}\delta_k}}$, where $\delta_k = \delta_k(f_{k+1}^{\mathrm{wN}})$ is the directional improvement of the weak Newton iterate. As the residual error becomes harder to fit during the later stages of training, δ_k shrinks, causing α_k to relax towards 1.0. This increasing difficulty to fit the residual is further illustrated in the weak learner edge comparison plots of Fig. 1. As optimization progresses, there is less signal and more noise in the residuals, causing the cosine angles and weak gradient edge to decrease. These plots additionally highlight the differences between the pessimistic worst-case edges (infimums) and the running averages computed from the start of optimization. On this dataset, both the Hessian cosine angle Θ_k and the weak gradient edge γ_k plateau at approximately 0.1 on average. In contrast, their respective infimums are near 0.05. This difference is critical for theoretical convergence bounds: our weak-AICN global rate relies on the path-averaged angle raised to the fourth power $\overline{\Theta}_k^4$, whereas the weak-GRN bound suffers from a twelfth-power dependence on the worst-case edge γ^{12}. Furthermore, the left plot of Fig. 1 shows that tracking weak learner alignment against iteration count can be visually misleading. Plotting them against the training loss (right plot) clarifies the underlying trend: evaluated at equivalent loss values, weak-AICN (Θ_k) maintains slightly higher weak learner edge compared to the other methods, though all three metrics are very similar overall.

Finally, in Fig. 6, we test the effect of tree depth ($J \in \{2, 4, 6\}$) on each weak learner edge, similar to [42]. We observe a clear monotonic relationship between tree depth and weak learner alignment. For all metrics, the weak learner edge is high during the initial phase of optimization before stabilizing. The stabilized running averages roughly equal 0.2 for $J = 6$, 0.1 for $J = 4$, and 0.05 for $J = 2$, effectively doubling every two additional levels of tree depth. This dyadic scaling (after squaring Θ_k) is consistent with our theoretical tree-angle proxy bounds presented in Theorem 4.2. The empirical angles along the actual optimization path remain above the no-signal reference values.

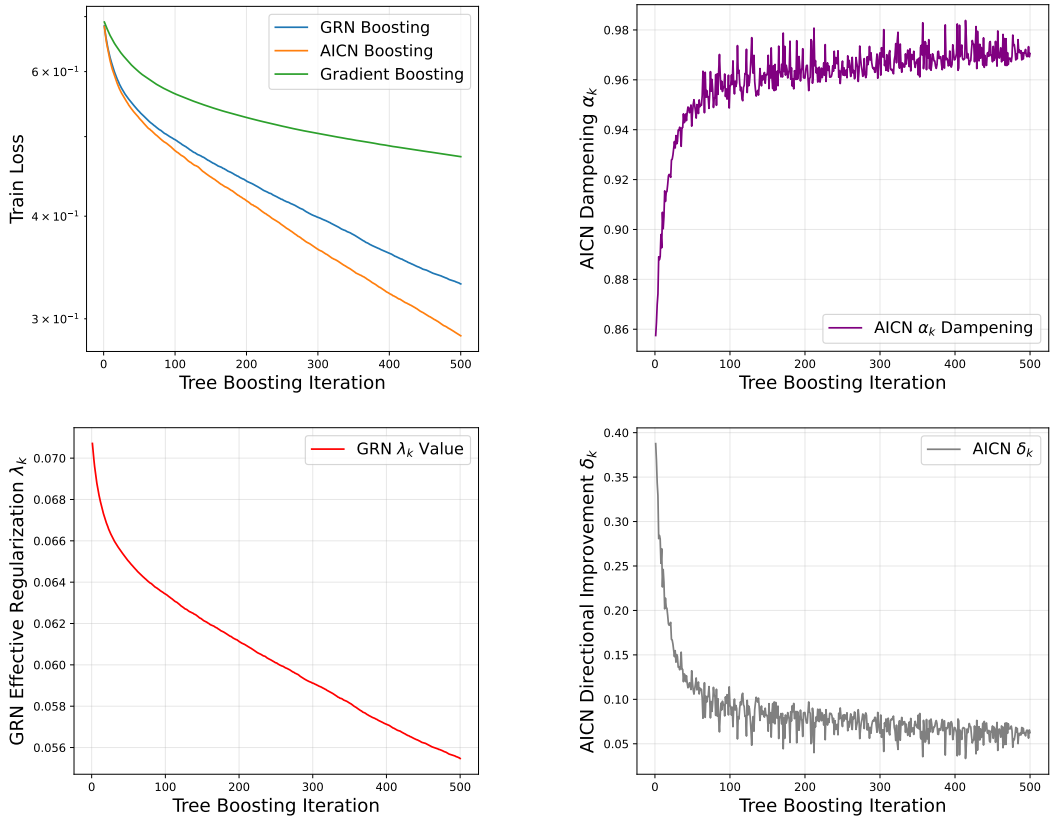


Figure 5: Convergence metrics on the HIGGS dataset using Binary Cross Entropy ($\eta = 0.1, J = 4$). **Top left:** Training loss convergence for weak-AICN, weak-GRN, and first-order gradient boosting. **Top right:** Evolution of the weak-AICN dampening factor α_k, which approaches 1.0 as optimization progresses. **Bottom left:** Evolution of the effective adaptive ℓ_2 regularization λ_k used by weak-GRN. **Bottom right:** Directional improvement δ_k of the weak Newton update used by weak-AICN.

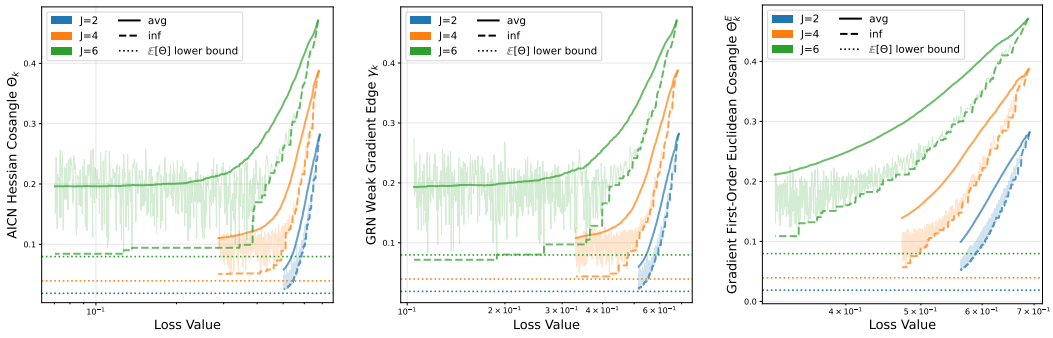


Figure 6: Effect of tree depth ($J \in \{2, 4, 6\}$) on weak learner alignment metrics on the HIGGS dataset. **Left:** Hessian-induced cosine angle Θ_k for weak-AICN. **Middle:** Weak gradient edge γ_k for weak-GRN. **Right:** Euclidean cosine angle Θ_k^E for first-order gradient boosting. Across all metrics, the stabilized average alignment roughly doubles with every two additional levels of tree depth. Dotted lines show the expectation lower bounds for the random tree-angle proxy in Theorem 4.2, included as reference values.

F Tree Weak Learners Cosine Angle Lower Bounds

We continue in the empirical boosting setting of Section 4.1. Thus $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^p$ is fixed,

$$\mathcal{H}_X := \mathcal{L}^2(\widehat{\nu}_n) \cong \mathbb{R}^X,$$

and $\mathcal{H}_X$ is equipped with the empirical inner product $\langle \cdot, \cdot \rangle_X$. Throughout this appendix, g_k, H_k, Q_k, and f_{k+1}^N are understood in this empirical space. We work with scalar outputs; the vector-valued case is analogous.

F.1 Fixed-sample tree-angle geometry

Fix $J \geq 0$. For a depth-J axis-aligned tree T, let $V_T \subset \mathcal{H}_X$ denote the subspace of functions on X that are constant on the leaves of T. Thus

$$\mathcal{T}_J(X) = \bigcup_T V_T,$$

where T ranges over depth-J axis-aligned trees, viewed only through their induced partitions of X.

Let $A : \mathcal{H}_X \to \mathcal{H}_X$ be self-adjoint and positive definite. Define

$$\langle f, h \rangle_A := \langle f, Ah \rangle_X, \qquad \|f\|_A^2 := \langle f, f \rangle_A.$$

For each tree T, let $\Pi_{V_T}^A$ denote the A-orthogonal projection onto V_T. We define the primitive depth-J tree cosine by

$$\Theta_J(X; f \mid A) := \max_{g \in \mathcal{T}_J(X) \setminus \{0\}} \frac{\langle f, g \rangle_A}{\|f\|_A \|g\|_A}, \qquad f \in \mathcal{H}_X \setminus \{0\}.$$

Equivalently,

$$\Theta_J(X; f \mid A) = \max_T \frac{\|\Pi_{V_T}^A f\|_A}{\|f\|_A},$$

where T ranges over depth-J axis-aligned trees on X. Hence

$$\Theta_J(X; f \mid A)^2 = 1 - \inf_{g \in \mathcal{T}_J(X)} \frac{\|f - g\|_A^2}{\|f\|_A^2}.$$

When $A = I$ with respect to $\langle \cdot, \cdot \rangle_X$, we abbreviate

$$\Theta_J(X; f) := \Theta_J(X; f \mid I).$$

Along the realized boosting iterates, the depth-J weak Newton tree cosine from Sections 2–3 is the special case

$$\Theta_J^{(k)}(X, y, F_0) := \Theta_J(X; f_{k+1}^N \mid H_k).$$

When the realized iterates are fixed, we also write $\Theta_{k,J} := \Theta_J^{(k)}(X, y, F_0)$. All deterministic statements below therefore apply to weak Newton tree cosines by setting $A = H_k$ and $f = f_{k+1}^N$.

F.2 Deterministic empirical tree geometry

Proposition F.1 (Leaf-count obstruction and worst-case exactness). For every configuration X of n distinct points, exact representation of all empirical functions $f \in \mathcal{H}_X$ by depth-J axis-aligned trees is impossible when $2^J < n$. Conversely, exact representation of all empirical functions is possible for every such X once $J \geq n-1$. Equivalently, for every self-adjoint positive definite A on $\mathcal{H}_X$,

$$\Theta_J(X; f \mid A) = 1 \qquad \text{for every nonzero } f \in \mathcal{H}_X$$

whenever $J \geq n - 1$.

Proof. A depth-J binary tree has at most 2^J leaves. If $2^J < n$, then no depth-J tree can separate n distinct sample points into singleton leaves, so it cannot represent all empirical functions on X.

For the converse, argue by induction on n. The case $n = 1$ is trivial. If $n \geq 2$, choose two sample points and a coordinate on which they differ. A threshold between two consecutive values in that coordinate gives a nontrivial split into two subsets of sizes $n_1, n_2 \geq 1$ with $n_1 + n_2 = n$. By induction, the two children can be separated with depths at most $n_1 - 1$ and $n_2 - 1$, so the whole sample can be separated with depth at most

$$1 + \max\{n_1 - 1, n_2 - 1\} \leq n - 1.$$

Assigning arbitrary leaf values then realizes every empirical function. The equivalent statement for $\Theta_J(X; f \mid A)$ is immediate from the definition. □

Example F.2 (Parity obstruction). For the cube configuration

$$X = \{0, 1\}^d,$$

take $J = d - 1$ and define

$$f(x) := (-1)^{x_1 + \cdots + x_d}.$$

Then

$$\Theta_J(X; f) = 0.$$

Hence below the leaf-count regime there is, in general, no universal positive lower bound on the intrinsic tree cosine for empirical axis-aligned trees.

Proof. Every depth-$(d-1)$ leaf fixes at most $d - 1$ coordinates, so at least one coordinate remains free on each leaf. The parity function sums to zero over any such leaf, and therefore has zero inner product with every function that is constant on that leaf. Since a depth-$(d-1)$ tree is constant on each of its leaves, it follows that f is orthogonal to every such tree. □

This example should be read as a worst-case obstruction, not as a model of typical boosting iterations. In particular, for squared loss $H_k = I$, so the construction directly rules out assumption-free deterministic lower bounds for the iteration-wise quantity $\Theta_{k,J}$ below the leaf-count regime. For general Hessian metrics, it still shows why worst-case tree-angle analysis is too pessimistic and motivates the probabilistic proxy studied next.

F.3 Random model and dyadic lower bounds

The deterministic obstruction above shows that worst-case tree-angle bounds are too pessimistic. We therefore introduce a tractable random proxy for tree alignment. The proxy is not meant to assert that an actual normalized Newton direction is uniformly distributed on the empirical unit sphere. Rather, it is an uncoupled null model: it retains the sample and metric geometry, but replaces the Newton direction by a generic isotropic direction in that metric.

Let

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}([0, 1]^p),$$

and write $X = \{X_1, \ldots, X_n\}$, which is almost surely a set of n distinct points. A *metric model* Q consists of this sample law together with an arbitrary conditional law, given X, for a self-adjoint positive definite operator

$$A : \mathcal{H}_X \to \mathcal{H}_X.$$

For a realized pair (X, A), define the A-unit sphere

$$S_A(\mathcal{H}_X) := \{f \in \mathcal{H}_X : \|f\|_A = 1\},$$

and let

$$U_A \mid (X, A) \sim \text{Unif}(S_A(\mathcal{H}_X)),$$

where uniformity is understood with respect to the inner product $\langle \cdot, \cdot \rangle_A$. We define the random metric proxy depth-J tree cosine by

$$\widehat{\Theta}^{\text{prox},\mathsf{Q}}_{n,p,J} := \Theta_J(X; U_A \mid A).$$

The identity-metric static proxy is the special case where $A = I$ deterministically; in that case we write simply

$$\widehat{\Theta}^{\mathrm{prox}}_{n,p,J} := \widehat{\Theta}^{\mathrm{prox},\mathsf{Q}_I}_{n,p,J} = \Theta_J(X; U_I \mid I).$$

The intended comparison is with randomized iteration-level quantities generated by a chosen probabilistic model M for the data, labels, initialization, and fixed loss. Such a model induces coupled quantities

$$\widehat{\Theta}^{(k;\mathsf{M})}_{n,p,J} := \Theta_J(X; f^N_{k+1} \mid H_k),$$

where H_k and f^N_{k+1} are produced by the same optimization state. The corresponding uncoupled metric-matched proxy is

$$\widehat{\Theta}^{\mathrm{prox},(k;\mathsf{M})}_{n,p,J} := \Theta_J(X; U_{H_k} \mid H_k), \qquad U_{H_k} \mid (X, H_k) \sim \mathrm{Unif}(S_{H_k}(\mathcal{H}_X)).$$

The role of the proxy is to replace the coupled Newton direction by a generic direction in the same realized Hessian geometry. The first-iteration comparison motivating this interpretation is recorded next.

Remark F.3 (Proxy comparisons and asymptotic beta domination). At a fixed depth J, the basic model-dependent proxy comparison is the first-iteration expectation inequality

$$\mathbb{E}\left[\left(\widehat{\Theta}^{\mathrm{prox},(0;\mathsf{M})}_{n,p,J}\right)^2\right] \le \mathbb{E}\left[\left(\widehat{\Theta}^{(0;\mathsf{M})}_{n,p,J}\right)^2\right].$$

This should be read as a structured-signal hypothesis: it is expected when the first Newton direction contains feature-aligned, tree-readable signal, while unstructured null models should give approximate equality and parity-type targets need not satisfy it. This first-iteration comparison is conjectural, and Appendix G provides empirical evidence for precisely this interpretation of the proxy.

A stronger, genuinely dynamical conjecture is that, for non-adversarial boosting models M, the actual coupled weak Newton tree angle should asymptotically dominate the proxy beta law itself. More precisely, if $2^J < n$ and if the limiting coupled tree angle $\widehat{\Theta}^{(\infty;\mathsf{M})}_{n,p,J}$ exists in distribution, one may conjecture that

$$\left(\widehat{\Theta}^{(\infty;\mathsf{M})}_{n,p,J}\right)^2 \succeq_{\mathrm{st}} Z_J, \qquad Z_J \sim \mathrm{Beta}\left(2^{J-1}, \frac{n-2^J}{2}\right),$$

where $\succeq_{\mathrm{st}}$ denotes stochastic domination. Equivalently,

$$\mathbb{P}\left(\left(\widehat{\Theta}^{(\infty;\mathsf{M})}_{n,p,J}\right)^2 \ge t\right) \ge \mathbb{P}(Z_J \ge t), \qquad t \in [0,1].$$

This distributional statement is substantially stronger than anything proved or directly verified experimentally here. The consequence in expectation would be

$$\mathbb{E}\left[\left(\widehat{\Theta}^{(\infty;\mathsf{M})}_{n,p,J}\right)^2\right] \ge \frac{2^J}{n}.$$

More cautiously, at finite time this suggests the path-averaged expected-value heuristic

$$\liminf_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left(\widehat{\Theta}^{(k;\mathsf{M})}_{n,p,J}\right)^2\right] \ge \frac{\min(2^J, n)}{n}.$$

The experiments in Appendix E should be interpreted as evidence for this weaker along-trajectory scaling: the observed dependence of realized weak learner angles on tree depth follows the dyadic behavior predicted by the proxy theorem. Thus Appendix G supports the first-iteration proxy comparison, while Appendix E supports the weaker path-averaged expected-value consequence. Neither appendix should be read as establishing the full stochastic domination conjecture, which remains a stronger structural hypothesis about the late-time optimization dynamics. A rigorous treatment of this conjecture would require additional control of the limiting weak-Newton residual geometry; we leave this as an interesting direction for future work.

Remark F.4 (Distributional viewpoint). Let $\sigma_{X,A}$ denote the uniform probability measure on $S_A(\mathcal{H}_X)$. The metric proxy $\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}$ is obtained by first sampling (X, A) and then sampling a generic A-unit direction. Equivalently, its distribution function can be written as

$$\mathbb{P}(\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J} \le t) = \int \left(\int_{S_A(\mathcal{H}_X)} \mathbf{1}_{\{\Theta_J(X;u|A)\le t\}}\, d\sigma_{X,A}(u) \right) d\mathsf{Q}(X, A), \qquad 0 \le t \le 1.$$

Thus the proxy distribution averages the fixed-sample tree-angle observable first over generic directions in the realized metric and then over random sample–metric pairs.

Similarly, if M is a probabilistic model generating the iterates, then the coupled iteration-level tree cosine has distribution function

$$\mathbb{P}(\widehat{\Theta}^{(k;\mathsf{M})}_{n,p,J} \le t) = \int \mathbf{1}_{\{\Theta_J(X;f^N_{k+1}(X,y,F_0)|H_k(X,y,F_0))\le t\}}\, d\mathsf{M}(X, y, F_0).$$

The distinction is that $\widehat{\Theta}^{\mathrm{prox},(k;\mathsf{M})}_{n,p,J}$ preserves the realized Hessian geometry but randomizes the direction, while $\widehat{\Theta}^{(k;\mathsf{M})}_{n,p,J}$ uses the Newton direction coupled to that geometry.

We now prove the main stochastic domination and expected value bounds in the form of the subsequent two results; the combination of Lemma F.5 and Theorem F.6 is what entails Theorem 4.2.

Lemma F.5 (Dyadic stochastic lower bound for the random tree cosine). Let $n \ge 2$, $p \ge 1$, and $J \ge 0$, and let Q be any metric model as above. If $2^J < n$, then there exists a random variable

$$Z_J \sim \mathrm{Beta}\left(2^{J-1}, \frac{n - 2^J}{2}\right),$$

where $\mathrm{Beta}(a, b)$ denotes the distribution on $[0, 1]$ with density proportional to $z^{a-1}(1 - z)^{b-1}$, such that

$$\left(\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}\right)^2 \ge Z_J \qquad \text{almost surely.}$$

If $2^J \ge n$, then

$$\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J} = 1 \qquad \text{almost surely.}$$

Proof. Because the sample law is continuous, the first-coordinate values are pairwise distinct almost surely. Condition on such a realization of (X, A) and relabel the points so that

$$x^{(1)}_1 < x^{(1)}_2 < \cdots < x^{(1)}_n.$$

Using only the first coordinate, recursively bisect the ordered list into consecutive blocks as evenly as possible. After depth J, this yields a depth-J axis-aligned tree, and let

$$V^{\mathrm{dyad}}_J(X) \subseteq \mathcal{H}_X$$

denote its associated leaf-constant subspace. By construction,

$$\dim V^{\mathrm{dyad}}_J(X) = \min(2^J, n).$$

First suppose $2^J < n$. Since $\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J} = \Theta_J(X; U_A \mid A)$ is defined using the best depth-J tree function in the A-metric, we have

$$\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J} \ge \max_{g \in V^{\mathrm{dyad}}_J(X)\setminus\{0\}} \frac{\langle U_A, g\rangle_A}{\|g\|_A} = \|\Pi^A_{V^{\mathrm{dyad}}_J(X)} U_A\|_A.$$

Hence

$$\left(\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}\right)^2 \ge \|\Pi^A_{V^{\mathrm{dyad}}_J(X)} U_A\|^2_A \qquad \text{almost surely.}$$

Choose any orthonormal basis of $(\mathcal{H}_X, \langle\cdot,\cdot\rangle_A)$ adapted to the decomposition

$$V^{\mathrm{dyad}}_J(X) \oplus_A \left(V^{\mathrm{dyad}}_J(X)\right)^{\perp_A}.$$

In this basis, the A-unit-sphere law is the standard spherical law on S^{n-1}, so

$$\|\Pi^A_{V_J^{\mathrm{dyad}}(X)} U_A\|_A^2 \overset{d}{=} \frac{G_1^2 + \cdots + G_{2^J}^2}{G_1^2 + \cdots + G_n^2},$$

where $G_1, \ldots, G_n \overset{\text{iid}}{\sim} N(0,1)$. Define

$$Z_J := \|\Pi^A_{V_J^{\mathrm{dyad}}(X)} U_A\|_A^2.$$

Then, conditionally on (X, A), and hence also unconditionally,

$$Z_J \sim \mathrm{Beta}\left(2^{J-1}, \frac{n-2^J}{2}\right),$$

which proves the stochastic domination.

If $2^J \geq n$, then the same first-coordinate construction isolates every sample point into a singleton leaf. Hence the resulting depth-J tree class equals $\mathcal{H}_X$, and therefore

$$\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J} = 1 \qquad \text{almost surely.}$$

□

Theorem F.6 (Dyadic expectation lower bounds). For every metric model Q, and for all $n \geq 2$, $p \geq 1$, and $J \geq 0$,

$$\mathbb{E}\left[\left(\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}\right)^2\right] \geq \frac{\min(2^J, n)}{n}.$$

Moreover, if $2^J < n$, then

$$\mathbb{E}[\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}] \geq \frac{\Gamma\left(\frac{2^J+1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma(2^{J-1})\,\Gamma\left(\frac{n+1}{2}\right)}.$$

In particular, if $2^J < n$ and $J, n \to \infty$, then

$$\frac{\Gamma\left(\frac{2^J+1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma(2^{J-1})\,\Gamma\left(\frac{n+1}{2}\right)} = \left(1 + O\left(2^{-J} + \frac{1}{n}\right)\right)\frac{2^{J/2}}{\sqrt{n}}.$$

If $2^J \geq n$, then

$$\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J} = 1 \qquad \text{almost surely,}$$

and consequently

$$\mathbb{E}[\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}] = \mathbb{E}\left[\left(\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}\right)^2\right] = 1.$$

Proof. If $2^J < n$, Lemma F.5 yields

$$\left(\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}\right)^2 \geq Z_J \qquad \text{almost surely,}$$

where

$$Z_J \sim \mathrm{Beta}\left(2^{J-1}, \frac{n-2^J}{2}\right).$$

Therefore

$$\mathbb{E}\left[\left(\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}\right)^2\right] \geq \mathbb{E}[Z_J] = \frac{2^{J-1}}{2^{J-1} + ((n-2^J)/2)} = \frac{2^J}{n} = \frac{\min(2^J, n)}{n}.$$

Also,

$$\mathbb{E}[\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}] = \mathbb{E}\left[\left((\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J})^2\right)^{1/2}\right] \geq \mathbb{E}[Z_J^{1/2}].$$

For $Z \sim \mathrm{Beta}(a, b)$,

$$\mathbb{E}[Z^{1/2}] = \frac{\Gamma(a+\frac{1}{2})\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+\frac{1}{2})}.$$

Applying this with

$$a = 2^{J-1}, \qquad b = \frac{n - 2^J}{2},$$

gives

$$\mathbb{E}[\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}] \geq \frac{\Gamma\left(\frac{2^J+1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma(2^{J-1})\,\Gamma\left(\frac{n+1}{2}\right)}.$$

If $2^J \geq n$, then Lemma F.5 gives $\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J} = 1$ almost surely. Hence both expectations equal 1, which is precisely $\min(2^J, n)/n$ for the squared-angle bound. It remains only to justify the final asymptotic statement. We use the standard gamma-ratio estimate

$$\frac{\Gamma(x+a)}{\Gamma(x+b)} = x^{a-b}\left(1 + O(x^{-1})\right), \qquad x \to \infty,$$

for fixed a, b. With $m = 2^J$, applying this first with $x = m/2$, $a = 1/2$, and $b = 0$, gives

$$\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} = \left(\frac{m}{2}\right)^{1/2}\left(1 + O(m^{-1})\right).$$

Applying the same estimate with $x = n/2$, $a = 0$, and $b = 1/2$, gives

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} = \left(\frac{n}{2}\right)^{-1/2}\left(1 + O(n^{-1})\right).$$

Multiplying these two estimates yields

$$\frac{\Gamma\left(\frac{m+1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n+1}{2}\right)} = \left(1 + O\left(m^{-1} + n^{-1}\right)\right)\sqrt{\frac{m}{n}}.$$

Substituting $m = 2^J$ gives the claimed

$$\left(1 + O\left(2^{-J} + \frac{1}{n}\right)\right)\frac{2^{J/2}}{\sqrt{n}}.$$

□

Remark F.7. The squared-angle estimate is the main quantity used in the paper, since the weak-AICN rates in Section 3 depend on squared cosine factors such as $\Theta^2_{k,J}$. Theorem F.6 therefore gives a lower benchmark for the uncoupled metric proxy $\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}$, uniformly over the metric model Q. In particular, choosing Q to be the marginal law of (X, H_k) induced by a probabilistic boosting model gives the corresponding benchmark for the metric-matched first-iteration or iteration-level proxy. The additional bound for $\mathbb{E}[\widehat{\Theta}^{\mathrm{prox},\mathsf{Q}}_{n,p,J}]$ is included only as a distributional consequence of the same Beta lower bound.

G Experimental Appendix: Proxy Tree-Angle Evaluation

This appendix tests the first-iteration proxy comparison from Remark F.3. For a probabilistic first-iteration model M and a tree-fitting rule A, the experimental analogue of the conjectural comparison is

$$\Delta_{J,\mathsf{A}}(\mathsf{M}) := \mathbb{E}_{\mathsf{M}}\Big[\Theta_{J,\mathsf{A}}(X; f_1^{\mathrm{N}} \mid H_0)^2 - \mathbb{E}_{U_{H_0}|X,H_0}\Theta_{J,\mathsf{A}}(X; U_{H_0} \mid H_0)^2\Big] \geq 0.$$

Here the second term is the uncoupled metric-isotropic baseline in the same Hessian geometry. The comparison is not meant to hold for arbitrary signals: we test it on *tree-readable* structured models, meaning signals generated from low-complexity coordinate functions which shallow axis-aligned trees can plausibly approximate. The Rademacher model serves as an unstructured control.

Data model and algorithmic tree cosines. Throughout, $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{Unif}([0,1]^p)$, $X = \{X_1, \dots, X_n\}$, $\mathcal{H}_X := \mathbb{R}^X$, and $\langle f, h\rangle_X := n^{-1}\sum_{x\in X} f(x)h(x)$. All empirical Hessians in these experiments are diagonal. For $A = \mathrm{diag}(a_i)$, write $\langle f, h\rangle_A := n^{-1}\sum_i a_i f(x_i)h(x_i)$ and $\|f\|_A^2 := \langle f, f\rangle_A$. Since exact maximization over all depth-J trees is not computationally feasible, we evaluate algorithmic tree cosines. For a fitting procedure A, let $\widehat{g}_{J,\mathsf{A}}(t; A)$ be its fitted depth-J tree approximation to $t \in \mathcal{H}_X$, computed in the A-metric, and define

$$\Theta_{J,\mathsf{A}}(X; t \mid A) := \frac{\langle t, \widehat{g}_{J,\mathsf{A}}(t; A)\rangle_A}{\|t\|_A \|\widehat{g}_{J,\mathsf{A}}(t; A)\|_A},$$

with value 0 if $\widehat{g}_{J,\mathsf{A}}(t; A) = 0$. For the weighted least-squares tree fits used below, equivalently $\Theta_{J,\mathsf{A}}(X; t \mid A)^2 = 1 - \|t - \widehat{g}_{J,\mathsf{A}}(t; A)\|_A^2 / \|t\|_A^2$. We use two fitting procedures: the dyadic construction from Algorithm 4, which is theorem-aligned, and a greedy CART regression tree fitted with sample weights a_i.

Coupled and uncoupled targets. For each sampled problem instance, compute $g_0 := \nabla L(F_0)$, $H_0 := \nabla^2 L(F_0)$, and $f_1^{\mathrm{N}} := -H_0^{-1}[g_0]$. We compare

$$\theta_{J,\mathsf{A}}^{2,\mathrm{cpl}} := \Theta_{J,\mathsf{A}}(X; f_1^{\mathrm{N}} \mid H_0)^2$$

with the metric-matched proxy

$$\theta_{J,\mathsf{A}}^{2,\mathrm{prox}} := \frac{1}{M_{\mathrm{prox}}} \sum_{m=1}^{M_{\mathrm{prox}}} \Theta_{J,\mathsf{A}}(X; U_{H_0}^{(m)} \mid H_0)^2, \qquad U_{H_0}^{(m)} \mid (X, H_0) \sim \mathrm{Unif}(S_{H_0}(\mathcal{H}_X)).$$

If $H_0 = \mathrm{diag}(h_i)$, we sample U_{H_0} by drawing $z_i \overset{\text{iid}}{\sim} N(0,1)$, setting $u_i = z_i/\sqrt{h_i}$, and normalizing so that $\|u\|_{H_0} = 1$.

Label and loss models. We use $n = 64$, $p = 16$, and $J = 0, \dots, 6$. For $v = (v_1, \dots, v_n)$, write $\bar{v} := n^{-1}\sum_i v_i$ and $\mathrm{std}(v)_i := (v_i - \bar{v})/(n^{-1}\sum_j (v_j - \bar{v})^2)^{1/2}$. Define

$$c := \frac{(1, 2, \dots, p)^\top}{\|(1, 2, \dots, p)\|_2}, \qquad s_{\mathrm{lin}} := \mathrm{std}\big((c^\top x_i)_{i=1}^n\big),$$

and

$$s_{\mathrm{sm}} := \mathrm{std}\big((\sin(2\pi x_i^{(1)}) + 0.75\cos(2\pi x_i^{(2)}) + 0.50\sin(4\pi x_i^{(3)}))_{i=1}^n\big).$$

For squared loss, $\ell_{\mathrm{sq}}(z, y) = \frac{1}{2}(z - y)^2$, we set $F_0 = 0$, hence $H_0 = I$ and $f_1^{\mathrm{N}} = y$, and use $y_i \overset{\text{iid}}{\sim} \mathrm{Unif}\{-1, 1\}$, $y_i = (s_{\mathrm{lin}})_i$, and $y_i = (s_{\mathrm{sm}})_i$. For binary logistic loss, $\ell_{\log}(z, y) = \log(1 + \exp z) - yz$, $y \in \{0, 1\}$, with $\mathrm{sig}(z) = (1 + e^{-z})^{-1}$, we use $y_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(1/2)$, $y_i \sim \mathrm{Bernoulli}(\mathrm{sig}(2(s_{\mathrm{lin}})_i))$, and $y_i \sim \mathrm{Bernoulli}(\mathrm{sig}(2(s_{\mathrm{sm}})_i))$, independently conditional on X. The logistic warm start is $F_0(x_i) = 1.5\ \mathrm{std}(0.75 s_{\mathrm{lin}} + 0.75 s_{\mathrm{sm}} + 0.15\xi)_i$, with $\xi_i \overset{\text{iid}}{\sim} N(0,1)$, so $h_i = \mathrm{sig}(F_0(x_i))(1 - \mathrm{sig}(F_0(x_i)))$ is nonconstant.

Dyadic constructive benchmark. The proof of Lemma F.5 uses an explicit depth-J tree obtained by sorting the sample by the first coordinate and recursively bisecting into consecutive blocks. In the metric setting, the leaf values are weighted averages.

Reported experiments and uncertainty. We report three figures. For each loss, label model, algorithm, and depth, we estimate $\mathbb{E}[\theta_{J,\mathsf{A}}^{2,\mathrm{cpl}}]$, $\mathbb{E}[\theta_{J,\mathsf{A}}^{2,\mathrm{prox}}]$, and $\Delta_{J,\mathsf{A}} := \mathbb{E}[\theta_{J,\mathsf{A}}^{2,\mathrm{cpl}} - \theta_{J,\mathsf{A}}^{2,\mathrm{prox}}]$. All curves

Algorithm 4 Dyadic constructive tree benchmark

Input: Sample $X = \{x_1, \dots, x_n\} \subset [0,1]^p$, target $t \in \mathcal{H}_X$, metric weights $a_i > 0$, depth J
Output: Dyadic fit g_J^{dyad} and squared cosine $(\theta_J^{\mathrm{dyad}})^2$
Relabel the sample so that $x_1^{(1)} < x_2^{(1)} < \cdots < x_n^{(1)}$
Initialize $\mathcal{B}_0 \leftarrow \big\{\{1, \dots, n\}\big\}$
for $j = 1, 2, \dots, J$ **do**
  Initialize $\mathcal{B}_j \leftarrow \emptyset$
  for each block $B \in \mathcal{B}_{j-1}$ **do**
    if $|B| = 1$ **then**
      Add B to $\mathcal{B}_j$
    else
      Split B into two consecutive sub-blocks $B_{\mathrm{L}}, B_{\mathrm{R}}$ as evenly as possible
      Add B_{L} and B_{R} to $\mathcal{B}_j$
    end if
  end for
end for
Let $\mathcal{B}_J$ be the final partition
for each block $B \in \mathcal{B}_J$ **do**
  For every $i \in B$, set $g_J^{\mathrm{dyad}}(x_i) \leftarrow \big(\sum_{k \in B} a_k t(x_k)\big) / \big(\sum_{k \in B} a_k\big)$
end for
Set $(\theta_J^{\mathrm{dyad}})^2 \leftarrow \big(\sum_i a_i t(x_i) g_J^{\mathrm{dyad}}(x_i)\big)^2 / \big((\sum_i a_i t(x_i)^2)(\sum_i a_i g_J^{\mathrm{dyad}}(x_i)^2)\big)$, with value 0 if $g_J^{\mathrm{dyad}} = 0$.

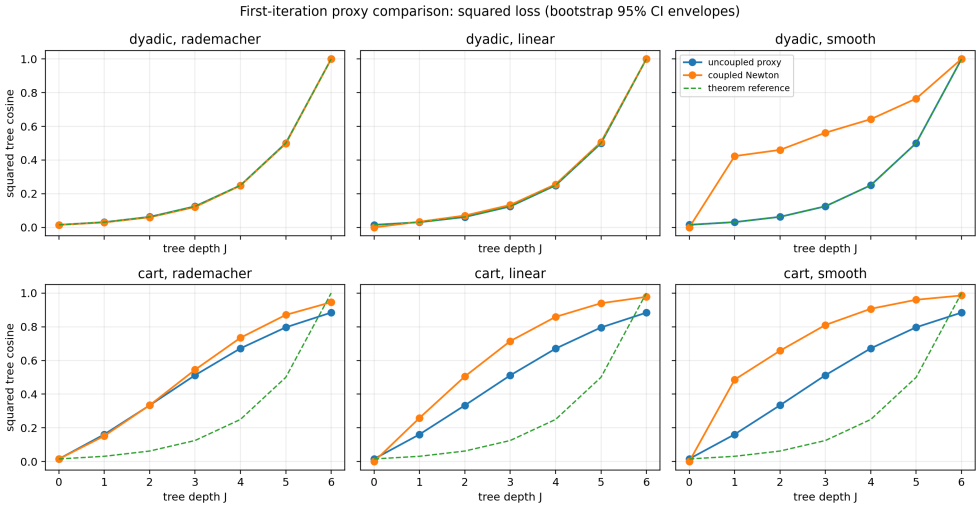


Figure 7: Squared-loss first-iteration proxy comparison for $n = 64$, $p = 16$, and $J = 0, \dots, 6$. Rows correspond to dyadic and CART fitting; columns correspond to Rademacher, linear, and smooth label models. Each panel compares the coupled target $\Theta_{J,\mathsf{A}}(X; f_1^{\mathrm{N}} \mid I)^2$ with the uncoupled proxy $\Theta_{J,\mathsf{A}}(X; U_I \mid I)^2$. Shaded bands are bootstrap 95% confidence intervals for the mean.

use 2000 independent problem instances and $M_{\mathrm{prox}} = 16$ proxy directions per instance. Shaded bands are nonparametric bootstrap 95% confidence intervals for the mean, computed by resampling outer problem instances. The curve $\min(2^J, n)/n$ is included only as theorem reference from Theorem F.6, not as an algorithmic lower bound for CART.

Squared-loss proxy comparison. Since $F_0 = 0$, the squared-loss Newton direction is the label function itself and $H_0 = I$. Thus this experiment compares structured labels to an uncoupled identity-isotropic proxy.

The Rademacher model is an unstructured control. Positive separation for the linear and smooth models indicates that tree-readable structure improves alignment relative to an isotropic proxy.

Logistic-loss metric-matched proxy comparison. For logistic loss, the warm start gives nonconstant H_0. The uncoupled baseline is therefore sampled from the H_0-unit sphere.

This tests the proxy comparison in nontrivial Hessian geometry. The adaptive CART fits show a clear positive gap for the structured label models, while the unstructured control remains close to the proxy.

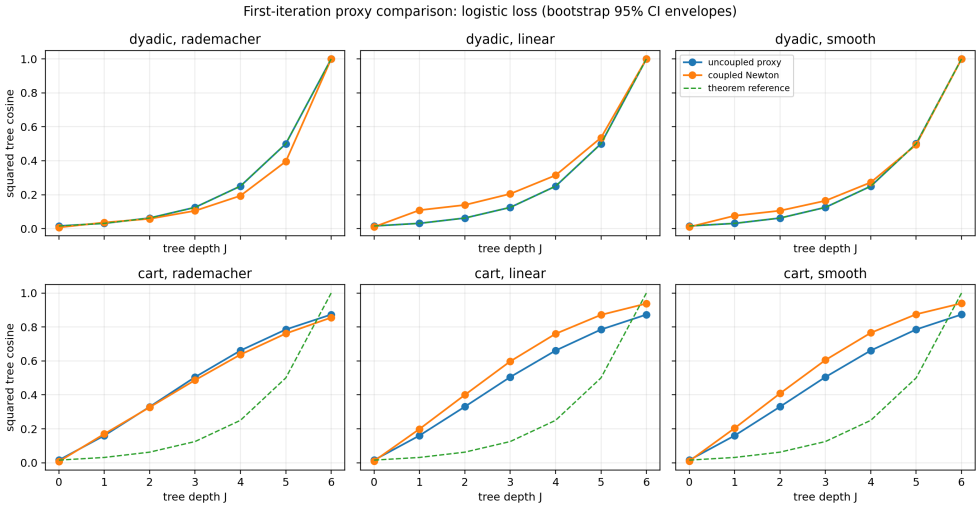


Figure 8: Binary logistic-loss first-iteration proxy comparison for $n = 64$, $p = 16$, and $J = 0, \dots, 6$. Rows correspond to dyadic and CART fitting; columns correspond to Rademacher, linear, and smooth label models. Each panel compares the coupled Newton direction $\Theta_{J,\mathrm{A}}(X; f_1^{\mathrm{N}} \mid H_0)^2$ with the uncoupled metric-matched proxy $\Theta_{J,\mathrm{A}}(X; U_{H_0} \mid H_0)^2$. Shaded bands are bootstrap 95% confidence intervals for the mean.

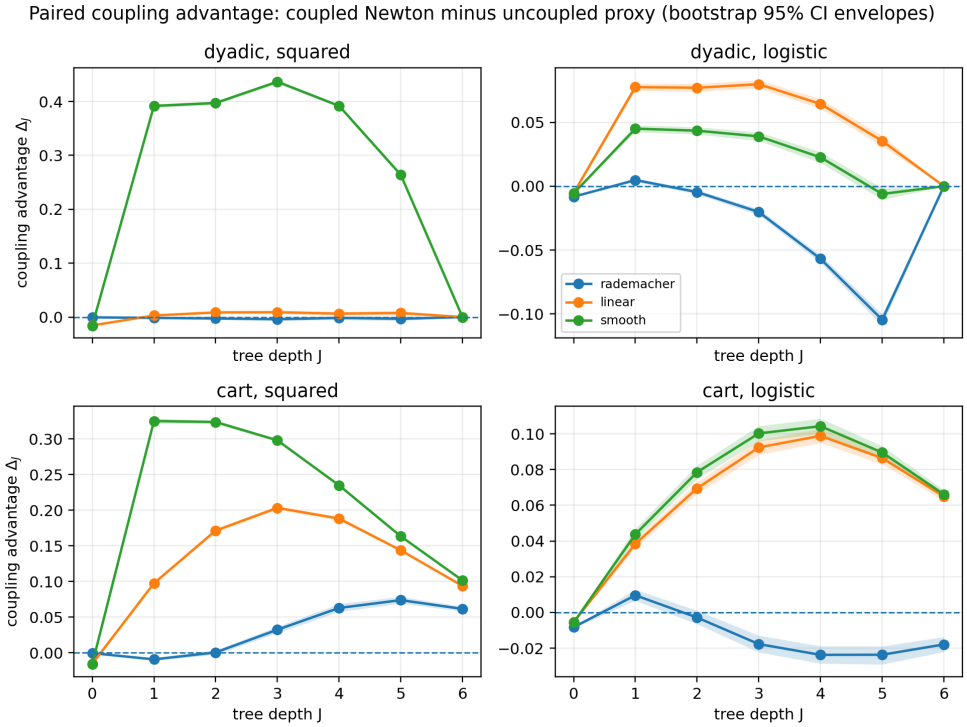


Figure 9: Coupling advantage $\Delta_{J,\mathrm{A}}$ as a function of tree depth. Rows correspond to dyadic and CART fitting; columns correspond to squared and logistic losses. Shaded bands are bootstrap 95% confidence intervals for the mean paired difference. Positive values indicate that the coupled first Newton direction is more tree-alignable than an uncoupled metric-isotropic direction in the same geometry.

Coupling advantage. Finally, we plot the paired difference $\Delta_{J,\mathrm{A}} = \mathbb{E}[\theta_{J,\mathrm{A}}^{2,\mathrm{cpl}} - \theta_{J,\mathrm{A}}^{2,\mathrm{prox}}]$. This is the finite-sample algorithmic version of the proxy comparison above; positive values indicate that the coupled first Newton direction is more tree-alignable than the uncoupled metric-isotropic baseline.

Taken together, Figs. 7 to 9 support this first-iteration proxy comparison in the intended tree-readable regime. The Rademacher controls remain close to the uncoupled proxy, while the linear and smooth models show a clear positive coupling advantage, especially for the adaptive CART fits. The dyadic method is theorem-aligned but rigid, and therefore mainly benefits when the signal aligns with its fixed first-coordinate splitting rule. Thus the experiments support the proxy comparison as a model-dependent structured-signal phenomenon, not as a universal statement over all targets, all fitting rules, or later optimization iterations.

H Alternative Decision Tree Angle Bound

Theorem H.1. Consider the decision tree specialization in Section 4.1 and Appendix F.1. Let $X = \{x_1, \dots, x_n\}$ consist of pairwise distinct training inputs. Fix a positive definite Hessian H_k and a depth-J tree T^* with exactly 2^J non-empty leaves, where $2^J < n$. Suppose $f_{k+1}^{\mathrm{N}} = \mathfrak{t} + \epsilon$, where $\mathfrak{t} \in V_{T^*}$ is a fixed tree-aligned signal and ϵ has independent $\mathcal{N}(0, \sigma^2)$ coordinates in an H_k-orthonormal basis, with $\sigma > 0$. Then the weak learner cosine angle Θ_k satisfies

$$\mathbb{E}\left[\Theta_k^2\right] \geq \mathbb{E}\left[\frac{\|\mathfrak{t} + \epsilon_V\|_{H_k}^2}{\|\mathfrak{t} + \epsilon_V\|_{H_k}^2 + B}\right] \geq \mathbb{E}\left[\frac{A}{A+B}\right] = \frac{2^J}{n}, \tag{79}$$

where $A \sim \sigma^2 \chi_m^2$ and $B \sim \sigma^2 \chi_{n-m}^2$ are independent scaled chi-squared variables with $m = \dim V_{T^*} = 2^J$, and $\epsilon_V = \Pi_{T^*}^{H_k} \epsilon$ is the projected noise. Moreover,

$$\mathbb{E}[\Theta_k] \geq \frac{\Gamma\left(\frac{2^J+1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma(2^{J-1})\,\Gamma\left(\frac{n+1}{2}\right)}. \tag{80}$$

Additionally, if $\mathfrak{t} = 0$, then Θ_k^2 dominates a random variable following a Beta distribution:

$$\Theta_k^2 \geq Z, \quad \text{where} \quad Z \sim \mathrm{Beta}\left(2^{J-1}, \frac{n - 2^J}{2}\right). \tag{81}$$

Proof. All analysis and projections take place in the Hessian-induced norm $\|\cdot\|_{H_k}$. Following the discussion in Appendix F.1, Θ_k can be expressed by projecting the exact Newton direction to the tree subspace that maximizes the projection of the exact Newton step, yielding:

$$\Theta_k^2 = \max_{T \in \mathcal{T}_J(X)} \frac{\|\Pi_T^{H_k} f_{k+1}^N\|_{H_k}^2}{\|f_{k+1}^N\|_{H_k}^2} \geq \frac{\|\Pi_{T^*}^{H_k} f_{k+1}^N\|_{H_k}^2}{\|f_{k+1}^N\|_{H_k}^2}, \tag{82}$$

where T^* is the fixed tree from the hypothesis, and $\Pi_T^{H_k}$ denotes orthogonal projection in the H_k-inner product. By assumption, T^* has exactly 2^J non-empty leaves, so $\dim V_{T^*} = 2^J$.

We decompose the noise ϵ into components within the subspace V_{T^*} and its H_k-orthogonal complement $V_{T^*}^{\perp}$:

$$\epsilon = \epsilon_V + \epsilon_\perp, \quad \text{where} \quad \epsilon_V = \Pi_{T^*}^{H_k} \epsilon, \quad \epsilon_\perp = \Pi_{V_{T^*}^{\perp}}^{H_k} \epsilon. \tag{83}$$

Because ϵ has independent $\mathcal{N}(0, \sigma^2)$ coordinates in an H_k-orthonormal basis, the orthogonal projections ϵ_V and $\epsilon_\perp$ are independent. Letting $m = \dim V_{T^*} = 2^J$, their squared H_k-norms follow scaled chi-squared distributions:

$$A := \|\epsilon_V\|_{H_k}^2 \sim \sigma^2 \chi_m^2, \quad B := \|\epsilon_\perp\|_{H_k}^2 \sim \sigma^2 \chi_{n-m}^2. \tag{84}$$

Since the signal $\mathfrak{t}$ lies entirely in V_{T^*}, we have $\Pi_{T^*}^{H_k} f_{k+1}^N = \mathfrak{t} + \epsilon_V$. Consequently, the total squared norm decomposes as $\|f_{k+1}^N\|_{H_k}^2 = \|\mathfrak{t} + \epsilon_V\|_{H_k}^2 + \|\epsilon_\perp\|_{H_k}^2$. Substituting this into our lower bound for Θ_k^2 yields

$$\Theta_k^2 \geq \frac{\|\mathfrak{t} + \epsilon_V\|_{H_k}^2}{\|\mathfrak{t} + \epsilon_V\|_{H_k}^2 + \|\epsilon_\perp\|_{H_k}^2} = \frac{\|\mathfrak{t} + \epsilon_V\|_{H_k}^2}{\|\mathfrak{t} + \epsilon_V\|_{H_k}^2 + B}. \tag{85}$$

To bound the expectation of this ratio, observe that since $B > 0$ almost surely, the scalar function $b(a) = \frac{a}{a+B}$ is increasing for $a \geq 0$. For any realization of the signal $\mathfrak{t}$, the term $\|\mathfrak{t} + \epsilon_V\|_{H_k}^2$ follows a scaled non-central chi-squared distribution $\sigma^2 \chi_m^2(\lambda)$ with non-centrality parameter $\lambda = \|\mathfrak{t}\|_{H_k}^2 / \sigma^2$. A well-known property of the non-central chi-squared distribution is that it stochastically dominates the central chi-squared distribution. Because $b(a)$ is an increasing function, this stochastic dominance preserves the inequality in expectation:

$$\mathbb{E}\left[\Theta_k^2\right] \geq \mathbb{E}\left[\frac{\|\mathfrak{t} + \epsilon_V\|_{H_k}^2}{\|\mathfrak{t} + \epsilon_V\|_{H_k}^2 + B}\right] \geq \mathbb{E}\left[\frac{A}{A+B}\right]. \tag{86}$$

Since A and B are independent scaled chi-squared variables, the ratio $Z := \frac{A}{A+B}$ follows a Beta distribution: $Z \sim \mathrm{Beta}\left(\frac{m}{2}, \frac{n-m}{2}\right) = \mathrm{Beta}\left(2^{J-1}, \frac{n-2^J}{2}\right)$. The expected value of Z is $\frac{m/2}{m/2+(n-m)/2} = \frac{2^J}{n}$, completing the proof of the squared-angle expectation bound.

To bound $\mathbb{E}[\Theta_k]$, we apply the same stochastic-domination argument to the square root of the ratio. The variable $\|\mathfrak{t} + \epsilon_V\|^2_{H_k}$ is independent of B, and, conditional on B, the function $a \mapsto \sqrt{a/(a+B)}$ is increasing. Therefore,

$$\mathbb{E}[\Theta_k] \geq \mathbb{E}\left[\sqrt{\frac{\|\mathfrak{t} + \epsilon_V\|^2_{H_k}}{\|\mathfrak{t} + \epsilon_V\|^2_{H_k} + B}}\right] \tag{87}$$

$$\geq \mathbb{E}\left[\sqrt{\frac{A}{A+B}}\right] = \mathbb{E}[\sqrt{Z}]. \tag{88}$$

Evaluating this expectation using the Beta density gives

$$\mathbb{E}[\sqrt{Z}] = \frac{\Gamma\left(\frac{m+1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n+1}{2}\right)}. \tag{89}$$

Substituting $m = 2^J$ yields the stated bound.

Finally, for the special case where there is no signal ($\mathfrak{t} = 0$), the ratio instead directly evaluates to the Beta-distributed random variable Z:

$$\Theta_k^2 \geq \frac{A}{A+B} =: Z \sim \mathrm{Beta}\left(2^{J-1}, \frac{n-2^J}{2}\right). \tag{90}$$

□

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: theoretical claims made in the abstract and introduction are explicitly backed by corresponding theorems and rigorous proofs provided in the main text and appendix.

   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: The paper includes a dedicated "Limitations" section at the end of the main paper body.

   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 theorems include clearly stated assumptions, with complete formal proofs provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All 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: Full experimental details, including hyperparameters and dataset configurations, are extensively documented in the Appendices, and code is provided in the supplementary material to reproduce all results.

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: Our code is publicly 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: All necessary training details, hyperparameters, and evaluation setups are explicitly detailed in the respective experimental appendix sections.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The tree-angle benchmark experiments in Appendix G include 95% confidence intervals. The remaining experiments do not require confidence intervals since we are optimizing deterministic functions with deterministic optimization schemes.

   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: The paper explicitly states that all experiments were run 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: The research focuses on theoretical convex optimization and gradient boosting theory, raising no ethical concerns.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The research focuses on theoretical convex optimization and gradient boosting theory, raising no foreseeable negative societal impacts.

    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 high-risk models or datasets.

    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: Standard open-source datasets (UCI Wine, HIGGS) and libraries (PyTorch, JAX) are properly cited.

    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 any new datasets or similar assets beyond the provided supplementary code for the experiments.

    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: The core methodology and development of this research do not involve LLMs.

    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.