

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 IMPLICIT BIAS OF PER-SAMPLE ADAM ON SEPARABLE DATA: DEPARTURE FROM THE FULL-BATCH REGIME

Anonymous authors

Paper under double-blind review

ABSTRACT

Adam (Kingma & Ba, 2015) is the de facto optimizer in deep learning, yet its theoretical understanding remains limited. Prior analyses show that Adam favors solutions aligned with ℓ_∞ -geometry, but these results are restricted to the full-batch regime. In this work, we study the implicit bias of incremental Adam (using one sample per step) for logistic regression on linearly separable data, and show that its bias can deviate from the full-batch behavior. As an extreme example, we construct datasets on which incremental Adam provably converges to the ℓ_2 -max-margin classifier, in contrast to the ℓ_∞ -max-margin bias of full-batch Adam. For general datasets, we characterize its bias using a proxy algorithm for the $\beta_2 \rightarrow 1$ limit. This proxy maximizes a *data-adaptive* Mahalanobis-norm margin, whose associated covariance matrix is determined by a *data-dependent* dual fixed-point formulation. We further present concrete datasets where this bias reduces to the standard ℓ_2 - and ℓ_∞ -max-margin classifiers. As a counterpoint, we prove that Signum (Bernstein et al., 2018) converges to the ℓ_∞ -max-margin classifier for any batch size. Overall, our results highlight that the implicit bias of Adam crucially depends on both the batching scheme and the dataset, while Signum remains invariant.

1 INTRODUCTION

The *implicit bias* of optimization algorithms plays a crucial role in training deep neural networks (Vardi, 2023). Even without explicit regularization, these algorithms steer learning toward solutions with specific structural properties. In over-parameterized models, where the training data can be perfectly classified and many global minima exist, the implicit bias dictates which solutions are selected. Understanding this phenomenon has become central to explaining why over-parameterized models often generalize well despite their ability to fit arbitrary labels (Zhang et al., 2017).

A canonical setting for studying implicit bias is linear classification on separable data with logistic loss. In this setup, achieving zero training loss requires the model’s weights to diverge to infinity, making the *direction of convergence*—which defines the decision boundary—the key object of study.

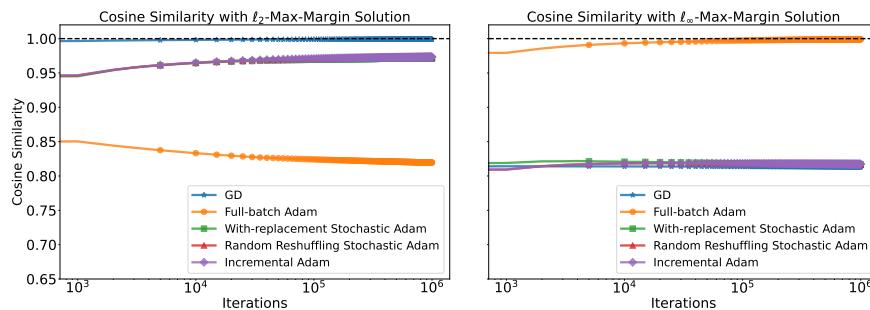


Figure 1: **Mini-batch Adam loses the ℓ_∞ -max-margin bias of full-batch Adam.** Cosine similarity between the weight vector and the ℓ_2 -max-margin (left) and ℓ_∞ -max-margin (right) solutions in a linear classification task on 10 data points drawn from the 50-dimensional standard Gaussian. Full-batch Adam with $(\beta_1, \beta_2) = (0.9, 0.95)$ converges to the ℓ_∞ -max-margin solution, whereas mini-batch variants with batch size 1 converge to the different direction (see Section 4 for the detailed characterization). See Appendix C for experimental details.

054 Seminal work by [Soudry et al. \(2018\)](#) establishes that gradient descent (GD) converges to the ℓ_2 -
 055 max-margin solution. This foundational result has inspired extensive research extending the analysis
 056 to neural networks, alternative optimizers, and other loss functions ([Gunasekar et al., 2018b](#); [Ji &](#)
 057 [Telgarsky, 2019, 2020](#); [Lyu & Li, 2020](#); [Chizat & Bach, 2020](#); [Yun et al., 2021](#)). In this work, we
 058 revisit the simplest setting—linear classification on separable data—to examine how the choice of
 059 optimizer shapes implicit bias.

060 Among modern optimization algorithms, Adam ([Kingma & Ba, 2015](#)) is one of the most widely used,
 061 making its implicit bias particularly important to understand. [Zhang et al. \(2024a\)](#) show that, unlike
 062 GD, full-batch Adam converges in direction to the ℓ_∞ -max-margin solution. This behavior is closely
 063 related to sign gradient descent (SignGD), which can be interpreted as normalized steepest descent
 064 in the ℓ_∞ -norm and is also known to converge to the ℓ_∞ -max-margin direction ([Gunasekar et al.,](#)
 065 [2018a](#); [Fan et al., 2025](#)). [Xie et al. \(2025\)](#) further attribute Adam’s empirical success in language
 066 model training to its ability to exploit the favorable ℓ_∞ -geometry of the loss landscape.

067 Yet, prior work on implicit bias in linear classification has almost exclusively focused on the full-batch
 068 setting. In contrast, modern training relies on stochastic mini-batches, a regime where theoretical
 069 understanding remains limited. Notably, [Nacson et al. \(2019\)](#) show that SGD preserves the same
 070 ℓ_2 -max-margin bias as GD, suggesting that mini-batching may not alter an optimizer’s implicit bias.
 071 But does this extend to adaptive methods such as Adam?

072 *Does Adam’s characteristic ℓ_∞ -bias persist under the mini-batch setting?*

073 Perhaps surprisingly, we find that the answer is *no*. Our experiments (Figure 1) illustrate that when
 074 trained on Gaussian data, full-batch Adam converges to the ℓ_∞ -max-margin direction, whereas
 075 mini-batch Adam variants with batch size 1 converge to **the different direction, which is even closer**
 076 **to the ℓ_2 -max-margin solution**. To explain this phenomenon, we develop a theoretical framework for
 077 analyzing the implicit bias of mini-batch Adam, focusing on the batch size 1 case as a representative
 078 contrast to the full-batch regime. To the best of our knowledge, this work provides the first theoretical
 079 evidence that Adam’s implicit bias is fundamentally altered in the mini-batch setting.

080 Our contributions are summarized as follows:

- 081 • We analyze *incremental Adam*, which processes one sample per step in a cyclic order. Despite
 082 its momentum-based updates, we show that its epoch-wise dynamics can be approximated by a
 083 recurrence depending only on the current iterate, which becomes a key tool in our analysis (see
 084 Section 2).
- 085 • We demonstrate a sharp contrast between full-batch and mini-batch Adam using a family of
 086 structured datasets, ***Scaled Rademacher (SR) data***. On **SR** data, we prove that incremental Adam
 087 converges to the ℓ_2 -max-margin solution, while full-batch Adam converges to the ℓ_∞ -max-margin
 088 solution (see Section 3).
- 089 • For general dataset, we introduce a *uniform-averaging proxy* that characterizes the limiting behavior
 090 of incremental Adam as $\beta_2 \rightarrow 1$. **We identify its convergence direction as the solution of a *data-***
 091 ***adaptive* margin-maximization problem**, induced by a Mahalanobis norm whose covariance matrix
 092 **determined by a *data-dependent* dual fixed-point equation**. We further present concrete datasets
 093 where this bias reduces to the standard ℓ_2 - and ℓ_∞ -max-margin classifiers (see Section 4).
- 094 • Finally, we prove that Signum (SignSGD with momentum; [Bernstein et al. \(2018\)](#)), unlike Adam,
 095 maintains its bias toward the ℓ_∞ -max-margin solution for *any* batch size when the momentum
 096 parameter is sufficiently close to 1 (see Section 5).

097 **2 HOW CAN WE APPROXIMATE WITHOUT-REPLACEMENT ADAM?**

100 **Notation.** For a vector \mathbf{v} , let $\mathbf{v}[k]$ denote its k -th entry, \mathbf{v}_t its value at time step t , and $\mathbf{v}_r^s \triangleq \mathbf{v}_{rN+s}$
 101 unless stated otherwise. For a matrix \mathbf{M} , let $\mathbf{M}[i, j]$ denote its (i, j) -th entry. We use Δ^{N-1} to
 102 denote the probability simplex in \mathbb{R}^N . Let $[N] = \{0, 1, \dots, N-1\}$ denote the set of the first N
 103 non-negative integers. For a PSD matrix \mathbf{M} , define the **Mahalanobis** norm as $\|\mathbf{x}\|_{\mathbf{M}} \triangleq \sqrt{\mathbf{x}^\top \mathbf{M} \mathbf{x}}$.
 104 For vectors, $\sqrt{\cdot}$, $(\cdot)^2$, and \div operations are applied entry-wise unless stated otherwise. Given two
 105 functions $f(t), g(t)$, we denote $f(t) = \mathcal{O}(g(t))$ if there exist $C, T > 0$ such that $t \geq T$ implies
 106 $|f(t)| \leq C|g(t)|$. For two vectors \mathbf{v} and \mathbf{w} , we denote $\mathbf{v} \propto \mathbf{w}$ if $\mathbf{v} = c \cdot \mathbf{w}$ for a *positive* scalar $c > 0$.
 107 Let $r = a \bmod b$ denote the remainder when dividing a by b , i.e., $0 \leq r < b$.

108

109

Algorithm 1 Det-Adam

110

111

Hyperparams: Learning rate schedule $\{\eta_t\}_{t=0}^{T-1}$,
momentum parameters $\beta_1, \beta_2 \in [0, 1]$

112

Input: Initial weight \mathbf{w}_0 , dataset $\{\mathbf{x}_i\}_{i \in [N]}$

113

- 1: Initialize momentum $\mathbf{m}_{-1} = \mathbf{v}_{-1} = \mathbf{0}$
- 2: **for** $t = 0, 1, 2, \dots, T-1$ **do**
- 3: $\mathbf{g}_t \leftarrow \nabla \mathcal{L}(\mathbf{w}_t)$
- 4: $\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t$
- 5: $\mathbf{v}_t \leftarrow \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2$
- 6: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}$
- 7: **end for**
- 8: **return** \mathbf{w}_T

114

115

116

117

118

119

120

121

122

123

Algorithms. We focus on incremental Adam (Inc-Adam), which processes mini-batch gradients sequentially from indices 0 to $N-1$ in each epoch. Studying Inc-Adam provides a tractable way to understand the implicit bias of mini-batch Adam: our experiments show that its iterates converge in directions closely aligned with mini-batch Adam of batch size 1 under both with-replacement and random-reshuffling sampling. Sharing the same mini-batch accumulation mechanism, Inc-Adam serves as a faithful surrogate for theoretical analysis. Pseudocodes for Inc-Adam and full-batch deterministic Adam (Det-Adam) are given in Algorithms 1 and 2.

124

125

126

127

128

129

130

131

132

133

134

135

136

137

Stability Constant ϵ . In practice, we often consider an additional ϵ term for numerical stability and update with $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t + \epsilon}}$. In fact, when investigating the asymptotic behavior of Adam, the stability constant significantly affects the converging direction, since $\mathbf{v}_t \rightarrow 0$ as $t \rightarrow \infty$ and ϵ dominates \mathbf{v}_t . Wang et al. (2021) investigate RMSprop and Adam with the stability constant, yielding their directional convergence to ℓ_2 -max-margin solution. More recent approaches, however, point out that analyzing Adam without the stability constant is more suitable for describing its intrinsic behavior (Xie & Li, 2024; Zhang et al., 2024a; Fan et al., 2025). We adopt this view and consider the version of Adam without ϵ .

138

139

140

Problem Settings. We primarily focus on binary linear classification tasks. To be specific, training data are given by $\{(\mathbf{x}_i, y_i)\}_{i \in [N]}$, where $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$. We aim to find a linear classifier \mathbf{w} which minimizes the loss

141

142

143

$$\mathcal{L}(\mathbf{w}) = \frac{1}{N} \sum_{i \in [N]} \ell(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) = \frac{1}{N} \sum_{i \in [N]} \mathcal{L}_i(\mathbf{w}),$$

144

145

146

147

where $\ell : \mathbb{R} \rightarrow \mathbb{R}$ is a surrogate loss for classification accuracy and $\mathcal{L}_i(\mathbf{w}) = \ell(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)$ denotes the loss value on the i -th data point. Without loss of generality, we assume $y_i = +1$, since we can newly define $\tilde{\mathbf{x}}_i = y_i \mathbf{x}_i$. In this paper, we consider two loss functions $\ell \in \{\ell_{\text{exp}}, \ell_{\text{log}}\}$, where $\ell_{\text{exp}}(z) = \exp(-z)$ denotes the exponential loss and $\ell_{\text{log}}(z) = \log(1 + e^{-z})$ denotes the logistic loss.

148

To investigate the implicit bias of Adam variants, we make the following assumptions.

149

150

Assumption 2.1 (Separable data). There exists $\mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{w}^\top \mathbf{x}_i > 0$, $\forall i \in [N]$.

151

Assumption 2.2. $\mathbf{x}_i[k] \neq 0$ for all $i \in [N], k \in [d]$.

152

Assumption 2.3 (Learning rate schedule). The sequence of learning rates, $\{\eta_t\}_{t=1}^\infty$, satisfies

153

154

(a) $\{\eta_t\}_{t=1}^\infty$ is decreasing in t , $\sum_{t=1}^\infty \eta_t = \infty$, and $\lim_{t \rightarrow \infty} \eta_t = 0$.

155

156

157

(b) For all $\beta \in (0, 1)$, $c_1 > 0$, there exist $t_1 \in \mathbb{N}_+$, $c_2 > 0$ such that $\sum_{\tau=0}^t \beta^\tau (e^{c_1 \sum_{\tau'=1}^\tau \eta_{t-\tau'}} - 1) \leq c_2 \eta_t$ for all $t \geq t_1$.

158

159

160

161

Assumption 2.1 guarantees linear separability of the data. Assumption 2.2 holds with probability 1 if the data is sampled from a continuous distribution. Assumption 2.3 originates from Zhang et al. (2024a) and it takes a crucial role to bound the error from the movement of weights. We note that a polynomial decaying learning rate schedule $\eta_t = (t+2)^{-a}$, $a \in (0, 1]$ satisfies Assumption 2.3, which is proved by Lemma C.1 in Zhang et al. (2024a).

Algorithm 2 Inc-Adam**Hyperparams:** Learning rate schedule $\{\eta_t\}_{t=0}^{T-1}$,
momentum parameters $\beta_1, \beta_2 \in [0, 1)$ **Input:** Initial weight \mathbf{w}_0 , dataset $\{\mathbf{x}_i\}_{i \in [N]}$ 1: Initialize momentum $\mathbf{m}_{-1} = \mathbf{v}_{-1} = \mathbf{0}$ 2: **for** $t = 0, 1, 2, \dots, T-1$ **do**3: $\mathbf{g}_t \leftarrow \nabla \mathcal{L}_i(\mathbf{w}_t)$, $i_t = t \bmod N$ 4: $\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t$ 5: $\mathbf{v}_t \leftarrow \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2$ 6: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}}$ 7: **end for**8: **return** \mathbf{w}_T

162 The dependence of the Adam update on the full gradient history makes its asymptotic analysis
 163 largely intractable. We address this challenge with the following propositions, which show that the
 164 *epoch-wise* updates of `Inc`-Adam and the updates of `Det`-Adam can be approximated by a function
 165 that depends only on the current iterate. [Detailed proofs are deferred to Appendix D](#).

166 **Proposition 2.4.** *Let $\{\mathbf{w}_t\}_{t=0}^\infty$ be the iterates of `Det`-Adam with $\beta_1 \leq \beta_2$. Then, under Assump-*
 167 *tions 2.2 and 2.3, if $\lim_{t \rightarrow \infty} \frac{\eta_t^{1/2} \mathcal{L}(\mathbf{w}_t)}{\|\nabla \mathcal{L}(\mathbf{w}_t)[k]\|} = 0$, then the update of k -th coordinate $\mathbf{w}_{t+1}[k] - \mathbf{w}_t[k]$*
 168 *can be represented by*

$$\mathbf{w}_{t+1}[k] - \mathbf{w}_t[k] = -\eta_t (\text{sign}(\nabla \mathcal{L}(\mathbf{w}_t)[k]) + \epsilon_t), \quad (1)$$

170 for some $\lim_{t \rightarrow \infty} \epsilon_t = 0$.

173 **Proposition 2.5.** *Let $\{\mathbf{w}_t\}_{t=0}^\infty$ be the iterates of `Inc`-Adam with $\beta_1 \leq \beta_2$. Then, under Assump-*
 174 *tions 2.2 and 2.3, the epoch-wise update $\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0$ can be represented by*

$$\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0 = -\eta_{rN} \left(C_{inc}(\beta_1, \beta_2) \sum_{i \in [N]} \frac{\sum_{j \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)^2}} + \epsilon_r \right), \quad (2)$$

180 where $\beta_1^{(i,j)} = \beta_1^{(i-j) \bmod N}$, $\beta_2^{(i,j)} = \beta_2^{(i-j) \bmod N}$, $C_{inc}(\beta_1, \beta_2) = \frac{1-\beta_1}{1-\beta_1^N} \sqrt{\frac{1-\beta_2^N}{1-\beta_2}}$ is a function of
 181 β_1, β_2 , and $\lim_{r \rightarrow \infty} \epsilon_r = \mathbf{0}$. If $\eta_t = (t+2)^{-a}$ for some $a \in (0, 1]$, then $\|\epsilon_r\|_\infty = \mathcal{O}(r^{-a/2})$.

184 **Discrepancy between `Det`-Adam and `Inc`-Adam.** Propositions 2.4 and 2.5 reveal a fundamental
 185 discrepancy between the behavior of `Det`-Adam and one of `Inc`-Adam. Proposition 2.4 demon-
 186 strates that `Det`-Adam can be approximated by SignGD, which has been reported by previous works
 187 ([Balles & Hennig, 2018](#); [Zou et al., 2023](#)). Note that the condition is not satisfied when $\nabla \mathcal{L}(\mathbf{w}_t)[k]$
 188 decays at a rate on the order of $\eta_t^{1/2} \mathcal{L}(\mathbf{w}_t)$, which often calls for a more detailed analysis (see [Zhang](#)
 189 [et al. \(2024a, Lemma 6.2\)](#)). Such an analysis establishes that `Det`-Adam asymptotically finds an ℓ_∞ -
 190 max-margin solution, a property that holds regardless of the choice of momentum hyperparameters
 191 satisfying $\beta_1 \leq \beta_2$ ([Zhang et al., 2024a](#)).

192 In stark contrast, our epoch-wise analysis illustrates that `Inc`-Adam’s updates more closely follow
 193 a weighted, preconditioned GD. This makes its behavior highly dependent on both the momentum
 194 parameters and the current iterate. The discrepancy originates from the use of mini-batch gradients;
 195 the preconditioner tracks the sum of squared mini-batch gradients, which diverges from the squared
 196 full-batch gradient. This discrepancy results in the highly complex dynamics of `Inc`-Adam, which
 197 are investigated in subsequent sections.

3 WARMUP: STRUCTURED DATA

200 **Eliminating Coordinate-Adaptivity.** To highlight the fundamental discrepancy between
 201 `Det`-Adam and `Inc`-Adam, we construct a scenario that completely nullifies the coordinate-wise
 202 adaptivity of `Inc`-Adam’s preconditioner by introducing the following family of structured datasets.

203 **Definition 3.1.** We define [Scaled Rademacher \(SR\)](#) data as a set of vectors $\{\mathbf{x}_i\}_{i \in [N]}$ which satisfy
 204 $|\mathbf{x}_i[k]| = |\mathbf{x}_i[l]|, \forall k, l \in [d]$, for each $i \in [N]$. We also assume that SR data satisfy Assumptions 2.1
 205 and 2.2, unless otherwise specified.

207 Applying Proposition 2.5 to the SR dataset, we obtain the following corollary.

208 **Corollary 3.2.** Consider `Inc`-Adam iterates $\{\mathbf{w}_t\}_{t=0}^\infty$ on SR data. Then, under Assumptions 2.2
 209 and 2.3, the epoch-wise update $\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0$ can be approximated by weighted normalized GD, i.e.,

$$\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0 = -\eta_{rN} \left(\sum_{i \in [N]} \frac{a_i(r)}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2} \nabla \mathcal{L}_i(\mathbf{w}_r^0) + \epsilon_r \right), \quad (3)$$

214 where $\lim_{r \rightarrow \infty} \epsilon_r = \mathbf{0}$ and $c_1 \leq a_i(r) \leq c_2$ for some positive constants c_1, c_2 only depending on
 215 $\beta_1, \beta_2, \{\mathbf{x}_i\}_{i \in [N]}$. If $\eta_t = (t+2)^{-a}$ for some $a \in (0, 1]$, then $\|\epsilon_r\|_\infty = \mathcal{O}(r^{-a/2})$.

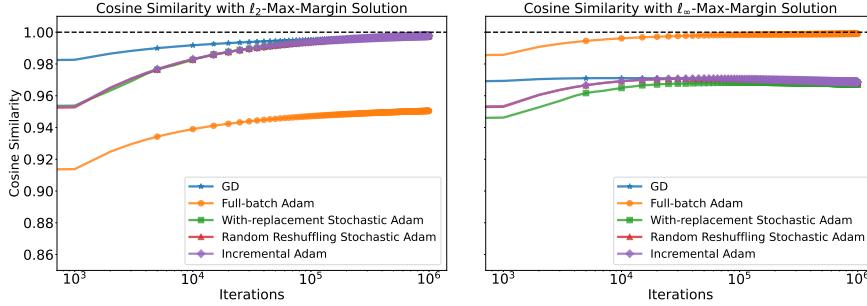


Figure 2: **Mini-batch Adam converges to the ℓ_2 -max-margin solution on the SR dataset.** We train on the dataset $\mathbf{x}_0 = (1, 1, 1, 1)$, $\mathbf{x}_1 = (2, 2, 2, -2)$, $\mathbf{x}_2 = (3, 3, -3, -3)$, and $\mathbf{x}_3 = (4, -4, 4, -4)$. Variants of mini-batch Adam with batch size 1 consistently converge to the ℓ_2 -max-margin direction, while full-batch Adam converges to the ℓ_∞ -max-margin direction.

Although using a structured dataset simplifies the denominator in Equation (2), the dynamics are still governed by weighted GD, which requires careful analysis. Prior work studies the implicit bias of weighted GD, particularly in the context of importance weighting (Xu et al., 2021; Zhai et al., 2023), but these analysis typically assume that the weights are constant or convergent. In our setting, the weight $a_i(r)$ varies with the epoch count r . We address this challenge and characterize the implicit bias of Inc-Adam on the SR data as follows.

Theorem 3.3. Consider Inc-Adam iterates $\{\mathbf{w}_t\}_{t=0}^\infty$ with $\beta_1 \leq \beta_2$ on SR data under Assumptions 2.1 to 2.3. If (a) $\mathcal{L}(\mathbf{w}_t) \rightarrow 0$ as $t \rightarrow \infty$ and (b) $\eta_t = (t+2)^{-a}$ for $a \in (2/3, 1]$, then it satisfies

$$\lim_{t \rightarrow \infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_2} = \hat{\mathbf{w}}_{\ell_2},$$

where $\hat{\mathbf{w}}_{\ell_2}$ denotes the (unique) ℓ_2 -max-margin solution of SR data $\{\mathbf{x}_i\}_{i \in [N]}$.

The analysis in Theorem 3.3 relies on Corollary 3.2, which ensures that the weights $a_i(r)$ are bounded by two positive constants, c_1 and c_2 . This condition is crucial to prevent any individual data from having a vanishing contribution, which could cause the Inc-Adam iterates to deviate from the ℓ_2 -max-margin direction. Furthermore, the controlled learning rate schedule is key to bounding the ϵ_r term in our analysis. The proof and further discussion are deferred to Appendix E. As shown in Figure 2, our experiments on SR data confirm that mini-batch Adam with batch size 1 converges in direction to the ℓ_2 -max-margin classifier, in contrast to the ℓ_∞ -bias of full-batch Adam.

Notably, Theorem 3.3 holds for any choice of momentum hyperparameters satisfying $\beta_1 \leq \beta_2$; see Figure 9 in Appendix B for empirical evidence. This invariance of the bias arises from the structure of SR data, which removes the coordinate adaptivity that momentum hyperparameters would normally affect. For general datasets, the invariance no longer holds; the adaptivity persists and varies with the choice of momentum hyperparameters, as discussed in Appendix A. In the next section, we introduce a proxy algorithm to study the regime where β_2 is close to 1 and characterize its implicit bias.

4 GENERALIZATION: ADAM PROXY

Uniform-Averaging Proxy. A key challenge in characterizing the limiting predictor of Inc-Adam for a general datasets is that its approximated update (Proposition 2.5) is difficult to analyze directly. To address this, we study a simpler *uniform-averaging* proxy, derived in Proposition 4.1 under the limit $\beta_2 \rightarrow 1$. This approximation is well-motivated, as β_2 is typically chosen close to 1 in practice.

Proposition 4.1. Let $\{\mathbf{w}_t\}_{t=0}^\infty$ be the iterates of Inc-Adam with $\beta_1 \leq \beta_2$. Then, under Assumptions 2.2 and 2.3, the epoch-wise update $\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0$ can be expressed as

$$\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0 = -\eta_{rN} \left(\sqrt{\frac{1 - \beta_2^N}{1 - \beta_2}} \frac{\nabla \mathcal{L}(\mathbf{w}_r^0)}{\sqrt{\sum_{i=1}^N \nabla \mathcal{L}_i(\mathbf{w}_r^0)^2}} + \epsilon_{\beta_2}(r) \right),$$

where $\limsup_{r \rightarrow \infty} \|\epsilon_{\beta_2}(r)\|_\infty \leq \epsilon(\beta_2)$ and $\lim_{\beta_2 \rightarrow 1} \epsilon(\beta_2) = 0$.

270 **Definition 4.2.** We define an update of AdamProxy as
 271

$$272 \quad \delta_t = \text{Prx}(\mathbf{w}_t) \triangleq \frac{\nabla \mathcal{L}(\mathbf{w}_t)}{\sqrt{\sum_{i=1}^N \nabla \mathcal{L}_i(\mathbf{w}_t)^2}}, \quad (4)$$

$$273$$

$$274 \quad \mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \delta_t.$$

$$275$$

276 **Proposition 4.3** (Loss convergence). *Under Assumptions 2.1 and 2.2, there exists a positive constant
 277 $\eta > 0$ depending only on the dataset $\{\mathbf{x}_i\}_{i \in [N]}$, such that if the learning rate schedule satisfies
 278 $\eta_t \leq \eta$ and $\sum_{t=0}^{\infty} \eta_t = \infty$, then AdamProxy iterates minimize the loss, i.e., $\lim_{t \rightarrow \infty} \mathcal{L}(\mathbf{w}_t) = 0$.*
 279

280 To characterize the convergence direction of AdamProxy, we further assume that the weights
 281 $\{\mathbf{w}_t\}_{t=0}^{\infty}$ and the updates $\{\delta_t\}_{t=0}^{\infty}$ converge in direction.
 282

283 **Assumption 4.4.** We assume that: (a) learning rates $\{\eta_t\}_{t=0}^{\infty}$ satisfy the conditions in Proposition 4.3,
 284 (b) $\exists \lim_{t \rightarrow \infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_2} \triangleq \hat{\mathbf{w}}$, and (c) $\exists \lim_{t \rightarrow \infty} \frac{\delta_t}{\|\delta_t\|_2} \triangleq \hat{\delta}$.
 285

286 **Lemma 4.5.** *Under Assumptions 2.1, 2.2 and 4.4, there exists $\mathbf{c} = (c_0, \dots, c_{N-1}) \in \Delta^{N-1}$ such
 287 that the limit direction $\hat{\mathbf{w}}$ of AdamProxy satisfies*

$$288 \quad \hat{\mathbf{w}} \propto \frac{\sum_{i \in [N]} c_i \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} c_i^2 \mathbf{x}_i^2}}, \quad (5)$$

$$289$$

$$290$$

291 and $c_i = 0$ for $i \notin S$, where $S = \arg \min_{i \in [N]} \hat{\mathbf{w}}^\top \mathbf{x}_i$ is the index set of support vectors of $\hat{\mathbf{w}}$.
 292

293 Prior research on the implicit bias of optimizers has predominantly focused on characterizing the
 294 convergence direction through the formulation of a corresponding optimization problem. For example,
 295 the solution to the ℓ_p -max-margin problem,
 296

$$297 \quad \max_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{w}\|_p^2 \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{x}_i - 1 \geq 0, \quad \forall i \in [N],$$

$$298$$

299 describes the implicit bias of the steepest descent algorithm with respect to the ℓ_p -norm in linear
 300 classification tasks (Gunasekar et al., 2018a). However, Equation (5) does not correspond to the KKT
 301 conditions of a conventional optimization problem. To address this, we introduce a novel framework
 302 to describe the convergence direction, based on a *parametric* optimization problem combined with
 303 *fixed-point analysis* between dual variables.
 304

305 **Definition 4.6.** Given $\mathbf{c} \in \Delta^{N-1}$, we define a parametric optimization problem $P_{\text{Adam}}(\mathbf{c})$ as
 306

$$307 \quad P_{\text{Adam}}(\mathbf{c}) : \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{w}\|_{\mathbf{M}(\mathbf{c})}^2 \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{x}_i - 1 \geq 0, \quad \forall i \in [N], \quad (6)$$

$$308$$

309 where $\mathbf{M}(\mathbf{c}) = \text{diag}(\sqrt{\sum_{j \in [N]} c_j^2 \mathbf{x}_j^2}) \in \mathbb{R}^{d \times d}$. We define $\mathbf{p}(\mathbf{c})$ as the set of global optimizers of
 310 $P_{\text{Adam}}(\mathbf{c})$ and $\mathbf{d}(\mathbf{c})$ as the set of corresponding dual solutions. Let $S(\mathbf{w}) = \{i \in [N] \mid \mathbf{w}^\top \mathbf{x}_i = 1\}$
 311 denote the index set for the support vectors for any $\mathbf{w} \in X(\mathbf{c})$.
 312

313 **Assumption 4.7** (Linear Independence Constraint Qualification). For any $\mathbf{c} \in \Delta^{N-1}$ and $\mathbf{w} \in \mathbf{p}(\mathbf{c})$,
 314 the set of support vectors $\{\mathbf{x}_i\}_{i \in S(\mathbf{w})}$ is linearly independent.
 315

316 Assumption 4.7 ensures the uniqueness of the dual solution for $P_{\text{Adam}}(\mathbf{c})$, which is essential for our
 317 framework. This assumption naturally holds in the overparameterized regime where the dataset
 318 $\{\mathbf{x}_i\}_{i \in [N]}$ consists of linearly independent vectors.
 319

320 **Theorem 4.8.** *Under Assumptions 2.1 and 4.7, $P_{\text{Adam}}(\mathbf{c})$ admits unique primal and dual solutions,
 321 so that $\mathbf{p}(\mathbf{c})$ and $\mathbf{d}(\mathbf{c})$ can be regarded as vector-valued functions. Moreover, under Assumptions 2.1,
 322 4.4 and 4.7, the following hold:*
 323

324 (a) $\mathbf{p} : \Delta^{N-1} \rightarrow \mathbb{R}^d$ is continuous.
 325

326 (b) $\mathbf{d} : \Delta^{N-1} \rightarrow \mathbb{R}_{\geq 0}^N \setminus \{\mathbf{0}\}$ is continuous. Consequently, the map $T(\mathbf{c}) \triangleq \frac{\mathbf{d}(\mathbf{c})}{\|\mathbf{d}(\mathbf{c})\|_1}$ is continuous.
 327

328 (c) The map $T : \Delta^{N-1} \rightarrow \Delta^{N-1}$ admits at least one fixed point.
 329

Algorithm 3 Fixed-Point Iteration

324
 325
 326 **Input:** Dataset $\{\mathbf{x}_i\}_{i \in [N]}$, initialization $\mathbf{c}_0 \in \Delta^{N-1}$, threshold $\epsilon_{\text{thr}} > 0$
 327 1: **repeat**
 328 2: Solve $P_{\text{Adam}}(\mathbf{c}_0) : \min \frac{1}{2} \|\mathbf{w}\|_{\mathbf{M}(\mathbf{c}_0)}^2$ subject to $\mathbf{w}^\top \mathbf{x}_i - 1 \geq 0, \forall i \in [N]$
 329 3: $\mathbf{w} \leftarrow \text{Primal}(P_{\text{Adam}})$
 330 4: $\mathbf{c}_1 \leftarrow \text{Dual}(P_{\text{Adam}})$
 331 5: $\delta \leftarrow \|\mathbf{c}_1 - \mathbf{c}_0\|_2$
 332 6: $\mathbf{c}_0 \leftarrow \mathbf{c}_1$
 333 7: **until** $\delta \leq \epsilon_{\text{thr}}$
 334 8: **return** \mathbf{w}
 335

336 (d) *There exists $\mathbf{c}^* \in \{\mathbf{c} \in \Delta^{N-1} : T(\mathbf{c}) = \mathbf{c}\}$ such that the convergence direction $\hat{\mathbf{w}}$ of*
 337 *AdamProxy is proportional to $\mathbf{p}(\mathbf{c}^*)$.*

338 Theorem 4.8 shows how the parametric optimization problem $P_{\text{Adam}}(\mathbf{c})$ captures the characterization
 339 from Lemma 4.5. The central idea is to treat the vector \mathbf{c} from Equation (5) in a dual role: as both
 340 the parameter of $P_{\text{Adam}}(\mathbf{c})$ and as its corresponding dual variable. The convergence direction is then
 341 identified at the point where these two roles coincide, leading naturally to the fixed-point formulation.
 342 **Detailed proofs are deferred to Appendix F.**

343 To computationally identify the convergence direction of AdamProxy based on Theorem 4.8, we
 344 introduce the fixed-point iteration described in Algorithm 3. Numerical experiments confirm that the
 345 resulting solution accurately predicts the limiting directions of both AdamProxy and Inc-Adam
 346 (see Example 4.10). However, the complexity of the mapping T makes it challenging to establish a
 347 formal convergence guarantee for Algorithm 3. A rigorous analysis is left for future work.

348 **Data-dependent Limit Directions.** We illustrate how structural properties of the data shape the
 349 limit direction of AdamProxy through three case studies. These examples demonstrate that both
 350 AdamProxy and Inc-Adam converge to directions that are intrinsically data-dependent.

351 **Example 4.9** (Revisiting SR data). For SR data $\{\mathbf{x}_i\}_{i \in [N]}$, the matrix $\mathbf{M}(\mathbf{c})$ reduces to a scaled
 352 identity for every $\mathbf{c} \in \Delta^{N-1}$. Hence, the parametric optimization problem $P_{\text{Adam}}(\mathbf{c})$ narrows down
 353 to the standard SVM formulation

$$354 \min \frac{1}{2} \|\mathbf{w}\|_2^2 \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{x}_i - 1 \geq 0, \forall i \in [N].$$

355 Therefore, Theorem 4.8 implies that AdamProxy converges to the ℓ_2 -max-margin solution. This
 356 finding is consistent with Theorem 3.3, which establishes the directional convergence of Inc-Adam
 357 on SR data. Together, these results indicate that the structural property of SR data that eliminates
 358 coordinate adaptivity persists in the limit $\beta_2 \rightarrow 1$.

359 **Example 4.10** (Revisiting Gaussian data). We next validate the fixed-point characterization in
 360 Theorem 4.8 using the Gaussian dataset from Figure 1. The theoretical limit direction is given by
 361 the fixed point of T defined in Theorem 4.8, which we compute via the iteration in Algorithm 3. As
 362 shown in Figure 3, both AdamProxy and mini-batch Adam variants with batch size 1 converge to
 363 the predicted solution, confirming the fixed-point formulation and the effectiveness of Algorithm 3.
 364 Furthermore, this demonstrates that, depending on the dataset, the limit direction of mini-batch Adam
 365 may differ from both the conventional ℓ_2 - and ℓ_∞ -max-margin solutions.

366 **Example 4.11** (Shifted-diagonal data). Consider $N = d$ and $\{\mathbf{x}_i\}_{i \in [d]} \subseteq \mathbb{R}^d$ with $\mathbf{x}_i = x_i \mathbf{e}_i +$
 367 $\delta \sum_{j \neq i} \mathbf{e}_j$ for some $\delta > 0$ and $0 < x_0 < \dots < x_{d-1}$. Then, the ℓ_∞ -max-margin problem

$$368 \min \frac{1}{2} \|\mathbf{w}\|_\infty^2 \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{x}_i \geq 1, \forall i \in [N]$$

369 has the solution $\hat{\mathbf{w}}_\infty = (\frac{1}{x_0 + (d-1)\delta}, \dots, \frac{1}{x_0 + (d-1)\delta}) \in \mathbb{R}^d$. Notice that $\mathbf{c}^* = (1, 0, \dots, 0) \in \Delta^{d-1}$
 370 is a fixed point of T in Theorem 4.8, and $\hat{\mathbf{w}}_\infty = \mathbf{p}(\mathbf{c}^*)$; detailed calculations are deferred to
 371 Appendix F. Consequently, the ℓ_∞ -max-margin solution serves a candidate for the convergence
 372 direction of AdamProxy as predicted by Theorem 4.8. To verify this, we run AdamProxy and
 373 mini-batch Adam variants with batch size 1 on shifted-diagonal data given by $\mathbf{x}_0 = (1, \delta, \delta, \delta)$,

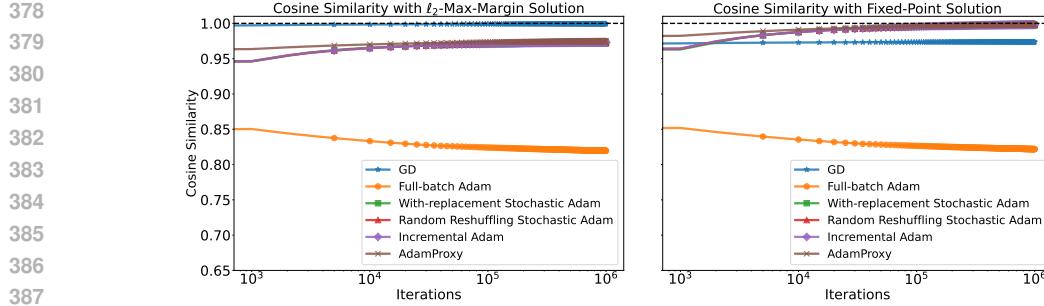


Figure 3: **Mini-batch Adam converges to the fixed-point solution on Gaussian data.** We train on the same Gaussian data as in Figure 1 and plot the cosine similarity of the weight vector with the ℓ_2 -max-margin solution (left) and the fixed-point solution (right). The results show that variants of mini-batch Adam with batch size 1 converge to the fixed-point solution obtained by Algorithm 3, consistent with our theoretical prediction (Theorem 4.8).

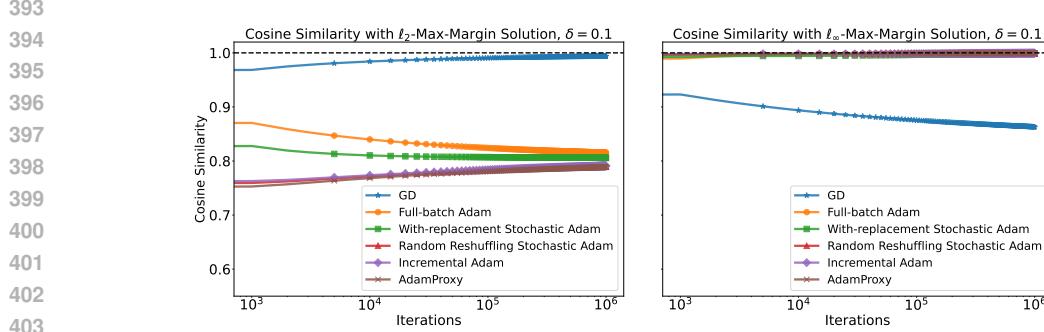


Figure 4: **Mini-batch Adam converges to the ℓ_∞ -max-margin solution on a shifted-diagonal dataset.** We train on the dataset $\mathbf{x}_0 = (1, \delta, \delta, \delta)$, $\mathbf{x}_1 = (\delta, 2, \delta, \delta)$, $\mathbf{x}_2 = (\delta, \delta, 4, \delta)$, and $\mathbf{x}_3 = (\delta, \delta, \delta, 8)$ with $\delta = 0.1$. Variants of mini-batch Adam with batch size 1 converge to the ℓ_∞ -max-margin direction.

$\mathbf{x}_1 = (\delta, 2, \delta, \delta)$, $\mathbf{x}_2 = (\delta, \delta, 4, \delta)$, and $\mathbf{x}_3 = (\delta, \delta, \delta, 8)$ with $\delta = 0.1$. As shown in Figure 4, all mini-batch Adam variants converge to the ℓ_∞ -max-margin solution, consistent with the theoretical prediction.

A key limitation of our analysis is that it assumes $\beta_2 \rightarrow 1$ and a batch size of 1. In Appendix A, we provide a preliminary analysis of how batch size and momentum hyperparameters affect the implicit bias of mini-batch Adam. In particular, Appendix A.2 explains why our fixed-point framework does not directly extend to finite β_2 .

5 SIGNUM CAN RETAIN ℓ_∞ -BIAS UNDER MINI-BATCH REGIME

In the previous section, we showed that Adam loses its ℓ_∞ -max-margin bias under mini-batch updates, drifting toward data-dependent solutions. This motivates the search for a *SignGD*-type algorithm that preserves ℓ_∞ -geometry even in the mini-batch regime. We prove that Signum (Bernstein et al., 2018) satisfies this property: with momentum close to 1, its iterates converge to the ℓ_∞ -max-margin direction for arbitrary mini-batch sizes.

Theorem 5.1. *Let $\delta > 0$. Then there exists $\epsilon > 0$ such that the iterates $\{\mathbf{w}_t\}_{t=0}^\infty$ of Inc-Signum (Algorithm 4) with batch size b and momentum $\beta \in (1 - \epsilon, 1)$, under Assumptions 2.1 and 2.3, satisfy*

$$\liminf_{t \rightarrow \infty} \frac{\min_{i \in [N]} \mathbf{x}_i^\top \mathbf{w}_t}{\|\mathbf{w}_t\|_\infty} \geq \gamma_\infty - \delta, \quad (7)$$

where

$$\gamma_\infty \triangleq \max_{\|\mathbf{w}\|_\infty \leq 1} \min_{i \in [N]} \mathbf{w}^\top \mathbf{x}_i, \quad D \triangleq \max_{i \in [N]} \|\mathbf{x}_i\|_1,$$

and

$$\epsilon = \frac{1}{2D \cdot \frac{N}{b} (\frac{N}{b} - 1)} \min\{\delta, \frac{\gamma_\infty}{2}\} \quad \text{if } b < N, \quad \epsilon = 1 \quad \text{if } b = N.$$

Theorem 5.1 demonstrates that, unlike Adam, Signum preserves ℓ_∞ -max-margin bias for any batch size, provided momentum is sufficiently close to 1. This generalizes the full-batch result of Fan et al. (2025). Moreover, the requirement $\beta \approx 1$ is not merely technical but *necessary* in the mini-batch setting to ensure convergence to the ℓ_∞ -max-margin solution; see Figure 10 in Appendix B for empirical evidence. As shown in Figure 5, our experiments on the Gaussian dataset from Figure 1 show that $\text{Inc-Signum}(\beta = 0.99)$ maintains ℓ_∞ -bias, regardless of the choice of batch size. Proofs and further discussion are deferred to Appendix G.

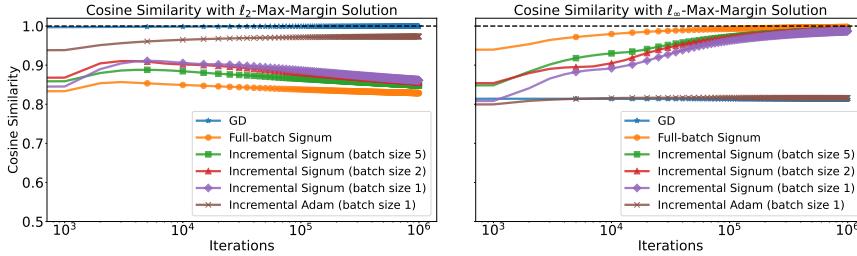


Figure 5: **Mini-batch Signum converges to the ℓ_∞ -max-margin solution.** We train on the same Gaussian data ($N = 10, d = 50$) as in Figure 1, using full-batch Signum and incremental Signum with $\beta = 0.99$, for batch sizes $b \in \{5, 2, 1\}$. Across all batch sizes, incremental Signum consistently converges to the ℓ_∞ -max-margin solution, in sharp contrast to incremental Adam.

6 RELATED WORK

Understanding Adam. Adam (Kingma & Ba, 2015) and its variant AdamW (Loshchilov & Hutter, 2019) are standard optimizers for large-scale models, particularly in domains like language modeling where SGD often falls short. A significant body of research seeks to explain this empirical success. One line focuses on convergence guarantees. The influential work of Reddi et al. (2018) demonstrates Adam’s failure to converge on certain convex problems, which motivates numerous studies establishing its convergence under various practical conditions (Défossez et al., 2022; Zhang et al., 2022; Li et al., 2023; Hong & Lin, 2024; Ahn & Cutkosky, 2024; Jin et al., 2025). Another line investigates why Adam outperforms SGD, attributing its success to robustness against heavy-tailed gradient noise (Zhang et al., 2020), better adaptation to ill-conditioned landscapes (Jiang et al., 2023; Pan & Li, 2023), and effectiveness in contexts of heavy-tailed class imbalance or gradient/Hessian heterogeneity (Kunstner et al., 2024; Zhang et al., 2024b; Tomihari & Sato, 2025). Ahn et al. (2024) further observe that this performance gap arises even in shallow linear Transformers. **Recent works investigate how the choice of momentum hyperparameters (Orvieto & Gower, 2025) and the rotation operation (Zhang et al., 2025) affect the performance of Adam.**

Implicit Bias and Connection to ℓ_∞ -Geometry. Recent work increasingly examines Adam’s implicit bias and its connection to ℓ_∞ -geometry. This link is motivated by Adam’s similarity to SignGD (Balles & Hennig, 2018; Bernstein et al., 2018), which performs normalized steepest descent under the ℓ_∞ -norm. Kunstner et al. (2023) show that the performance gap between Adam and SGD increases with batch size, while SignGD achieves performance similar to Adam in the full-batch regime, supporting this connection. Zhang et al. (2024a) prove that Adam without a stability constant converges to the ℓ_∞ -max-margin solution in separable linear classification, later extended to multi-class classification by Fan et al. (2025). Tsilivis et al. (2025) investigate implicit bias of steepest descent in homogeneous neural networks, supporting that SignGD describes a typical dynamics of Adam. Complementing these results, Xie & Li (2024) show that AdamW implicitly solves an ℓ_∞ -norm-constrained optimization problem, connecting its dynamics to the Frank-Wolfe algorithm. Exploiting this ℓ_∞ -geometry is argued to be a key factor in Adam’s advantage over SGD, particularly for language model training (Xie et al., 2025). Vasudeva et al. (2025) examine how Adam and GD show different implicit biases when training two-layer ReLU networks, describing Adam’s richer and more diverse decision boundary.

7 DISCUSSION AND FUTURE WORK

We studied the convergence directions of Adam and Signum for logistic regression on linearly separable data in the mini-batch regime. Unlike full-batch Adam, which always converges to the

486 ℓ_∞ -max-margin solution, mini-batch Adam exhibits data-dependent behavior, revealing a richer
 487 implicit bias, while Signum consistently preserves the ℓ_∞ -max-margin bias across all batch sizes.
 488

489 **Toward understanding the Adam–SGD gap.** Empirical evidence shows that Adam’s advan-
 490 tage over SGD is most pronounced in large-batch training, while the gap diminishes with smaller
 491 batches (Kunstner et al., 2023; Srećković et al., 2025; Marek et al., 2025). Our results suggest a
 492 possible explanation: the ℓ_∞ -adaptivity of Adam, proposed as the source of its advantage (Xie et al.,
 493 2025), may vanish in the mini-batch regime. An important direction for future work is to investigate
 494 whether this loss of ℓ_∞ -adaptivity extends beyond linear models and how it interacts with practical
 495 large-scale training.
 496

497 **Limitations.** Our analysis for general dataset relies on the asymptotic regime $\beta_2 \rightarrow 1$ and on
 498 incremental Adam as a tractable surrogate. Extending the framework to finite β_2 , larger batch sizes,
 499 and common sampling schemes (e.g., random reshuffling) would make the theory more complete.
 500 See Appendix A for further discussion. Relaxing technical assumptions and developing tools that
 501 apply under broader conditions also remain important directions.
 502

502 REPRODUCIBILITY STATEMENT

504 All assumptions and theorems for our theoretical results are stated in the main paper, with their
 505 complete proofs deferred to Appendices D to G. The primary experimental setups are described
 506 in the main paper and Appendix C, while details for supplementary experiments are provided in
 507 Appendices B and C.
 508

509 DECLARATION OF LLM USAGE

511 The authors utilized LLMs to improve the grammar and readability of this manuscript. The core
 512 conceptualization, analysis, and writing of the content were performed exclusively by the authors.
 513

514 REFERENCES

516 Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective
 517 for nonconvex optimization. In *The Thirty-eighth Annual Conference on Neural Information
 518 Processing Systems*, 2024. URL <https://openreview.net/forum?id=v416YLOQuU>.
 519 6

520 Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear
 521 attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth
 522 International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=0uI5415ry7>. 6

523 Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic
 524 gradients, 2018. URL <https://openreview.net/forum?id=S1EwLkW0W>. 2, 6

525 Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar.
 526 signSGD: Compressed optimisation for non-convex problems. In Jennifer Dy and Andreas
 527 Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80
 528 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 10–15 Jul 2018. URL
 529 <https://proceedings.mlr.press/v80/bernstein18a.html>. (document), 1, 5, 6

530 Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks
 531 trained with the logistic loss. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of
 532 Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning
 533 Research*, pp. 1305–1338. PMLR, 09–12 Jul 2020. URL <https://proceedings.mlr.press/v125/chizat20a.html>. 1

534 Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof
 535 of adam and adagrad. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL
 536 <https://openreview.net/forum?id=ZPQhzTSWA7>. 6

540 Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex
 541 optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016. [C](#)

542

543 Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on
 544 multiclass separable data, 2025. URL <https://arxiv.org/abs/2502.04664>. [1](#), [2](#), [5](#), [6](#),
 545 [G](#), [I.1](#)

546 Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in
 547 terms of optimization geometry. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th*
 548 *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning*
 549 *Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018a. URL <https://proceedings.mlr.press/v80/gunasekar18a.html>. [1](#), [4](#), [G](#)

550

551 Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias
 552 of gradient descent on linear convolutional networks. In S. Bengio, H. Wal-
 553 lach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Ad-
 554 vances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.,
 555 2018b. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/0e98aeeb54acf612b9eb4e48a269814c-Paper.pdf. [1](#)

556

557 Yusu Hong and Junhong Lin. On convergence of adam for stochastic optimization under relaxed
 558 assumptions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*,
 559 2024. URL <https://openreview.net/forum?id=x7usmidzxj>. [6](#)

560

561 Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In
 562 *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=HJflg30qKX>. [1](#)

563

564 Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In
 565 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neu-
 566 ral Information Processing Systems*, volume 33, pp. 17176–17186. Curran Associates, Inc.,
 567 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c76e4b2fa54f8506719a5c0dc14c2eb9-Paper.pdf. [1](#)

568

569 Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the
 570 regularization path for general losses. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings*
 571 *of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning*
 572 *Research*, pp. 2109–2136. PMLR, 09–12 Jul 2020. URL <https://proceedings.mlr.press/v125/ji20a.html>. [E.2](#), [E.2](#), [E.2](#), [E.3](#)

573

574 Kaiqi Jiang, Dhruv Malik, and Yuanzhi Li. How does adaptive optimization impact local neural net-
 575 work geometry? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.),
 576 *Advances in Neural Information Processing Systems*, volume 36, pp. 8305–8384. Curran Asso-
 577 ciates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1a5e6d0441a8e1eda9a50717b0870f94-Paper-Conference.pdf.
 578 [6](#)

579

580 Ruinan Jin, Xiao Li, Yaoliang Yu, and Baoxiang Wang. A comprehensive framework for analyzing
 581 the convergence of adam: Bridging the gap with sgd, 2025. URL <https://arxiv.org/abs/2410.04458>. [6](#)

582

583 Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd Inter-
 584 national Conference on Learning Representations*, 2015. URL <http://arxiv.org/abs/1412.6980>. [\(document\)](#), [1](#), [6](#)

584

585

586 Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not
 587 the main factor behind the gap between sgd and adam on transformers, but sign descent might
 588 be. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=a65YK0cqH8g>. [6](#), [7](#)

589

590

591 Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed
 592 class imbalance and why adam outperforms gradient descent on language models. In *The Thirty-
 593 eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=T56j6aV8Oc>. [6](#)

594 Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions.
 595 In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=yEewbkBNzi>. 6
 596
 597 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>. 6
 598
 599 Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks.
 600 In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SJeL1gBKPS>. 1
 601
 602 Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, and Micah Goldblum.
 603 Small batch size training for language models: When vanilla SGD works, and why gradient
 604 accumulation is wasteful. In *The Thirty-ninth Annual Conference on Neural Information Processing
 605 Systems*, 2025. URL <https://openreview.net/forum?id=52Ehpe0Lu5>. 7
 606
 607 Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable
 608 data: Exact convergence with a fixed learning rate. In Kamalika Chaudhuri and Masashi Sugiyama
 609 (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and
 610 Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3051–3059. PMLR,
 611 16–18 Apr 2019. URL <https://proceedings.mlr.press/v89/nacson19a.html>.
 612 1
 613
 614 Antonio Orvieto and Robert M. Gower. In search of adam’s secret sauce. In *The Thirty-ninth Annual
 615 Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=CH72XyZs4y>. 6
 616
 617 Yan Pan and Yuanzhi Li. Toward understanding why adam converges faster than sgd for transformers.
 618 In *arXiv preprint arXiv:2306.00204*, 2023. 6
 619
 620 Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In
 621 *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=ryQu7f-RZ>. 6
 622
 623 Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit
 624 bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57,
 625 2018. URL <http://jmlr.org/papers/v19/18-188.html>. 1, F3, F3
 626
 627 Teodora Srećković, Jonas Geiping, and Antonio Orvieto. Is your batch size the problem? revisiting
 628 the adam-sgd gap in language modeling. *arXiv preprint arXiv:2506.12543*, 2025. 7
 629
 630 Akiyoshi Tomihari and Issei Sato. Understanding why adam outperforms sgd: Gradient heterogeneity
 631 in transformers. *arXiv preprint arXiv:2502.00213*, 2025. 6
 632
 633 Nikolaos Tsilivis, Gal Vardi, and Julia Kempe. Flavors of margin: Implicit bias of steepest descent
 634 in homogeneous neural networks. In *The Thirteenth International Conference on Learning
 635 Representations*, 2025. URL <https://openreview.net/forum?id=BEpaPHD19r>. 6
 636
 637 Gal Vardi. On the implicit bias in deep-learning algorithms. *Commun. ACM*, 66(6):86–93, May 2023.
 638 ISSN 0001-0782. doi: 10.1145/3571070. URL <https://doi.org/10.1145/3571070>. 1
 639
 640 Bhavya Vasudeva, Jung Whan Lee, Vatsal Sharan, and Mahdi Soltanolkotabi. The rich and the
 641 simple: On the implicit bias of adam and SGD. In *The Thirty-ninth Annual Conference on
 642 Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=XLvHmzaHsx>. 6
 643
 644 Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization
 645 algorithms on homogeneous neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings
 646 of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine
 647 Learning Research*, pp. 10849–10858. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/wang21q.html>. 2

648 Shuo Xie and Zhiyuan Li. Implicit Bias of AdamW: ℓ_∞ -Norm Constrained Optimization. In
 649 *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=CmXkdl06JJ>. 2, 6
 650
 651 Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam Exploits ℓ_∞ -geometry of Loss Land-
 652 scape via Coordinate-wise Adaptivity. In *The Thirteenth International Conference on Learning*
 653 *Representations*, 2025. URL <https://openreview.net/forum?id=PUuD86UEK5>. 1,
 654 6, 7
 655
 656 Da Xu, Yuting Ye, and Chuanwei Ruan. Understanding the role of importance weighting for
 657 deep learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=_WnwtieRHxM. 3, E.2
 658
 659 Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training
 660 linear neural networks. In *International Conference on Learning Representations*, 2021. URL
 661 <https://openreview.net/forum?id=ZsZM-4iMQkH>. 1
 662
 663 Runtian Zhai, Chen Dan, J Zico Kolter, and Pradeep Kumar Ravikumar. Understanding why
 664 generalized reweighting does not improve over ERM. In *The Eleventh International Conference on*
 665 *Learning Representations*, 2023. URL https://openreview.net/forum?id=ashPce_W8F-. 3, E.2
 666
 667 Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. In
 668 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL
 669 <https://openreview.net/forum?id=xRQxan3WkM>. 1, 2, 2, 2, 6, D.1, G, I.1, (b), I.2,
 670 I.3, I.4, I.6
 671
 672 Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding
 673 deep learning requires rethinking generalization. In *International Conference on Learning*
 674 *Representations*, 2017. URL <https://openreview.net/forum?id=Sy8gdB9xx>. 1
 675
 676 Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, San-
 677 jiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In
 678 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neu-*
 679 *ral Information Processing Systems*, volume 33, pp. 15383–15393. Curran Associates, Inc.,
 680 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf. 6
 681
 682 Tianyue H. Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas,
 683 Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding adam re-
 684 quires better rotation dependent assumptions. In *The Thirty-ninth Annual Conference on Neural*
 685 *Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=KD4wgubh0>. 6
 686
 687 Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam
 688 can converge without any modification on update rules. In S. Koyejo, S. Mo-
 689 hamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural In-*
 690 *formation Processing Systems*, volume 35, pp. 28386–28399. Curran Associates, Inc.,
 691 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b6260ae5566442da053e5ab5d691067a-Paper-Conference.pdf. 6
 692
 693 Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why trans-
 694 formers need adam: A hessian perspective. In *The Thirty-eighth Annual Conference on Neural*
 695 *Information Processing Systems*, 2024b. URL <https://openreview.net/forum?id=X6rqEpbnj3>. 6
 696
 697 Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam
 698 in learning neural networks with proper regularization. In *The Eleventh International Confer-*
 699 *ence on Learning Representations*, 2023. URL <https://openreview.net/forum?id=iUYpN14qjTF>. 2, D, D.1, D
 700
 701

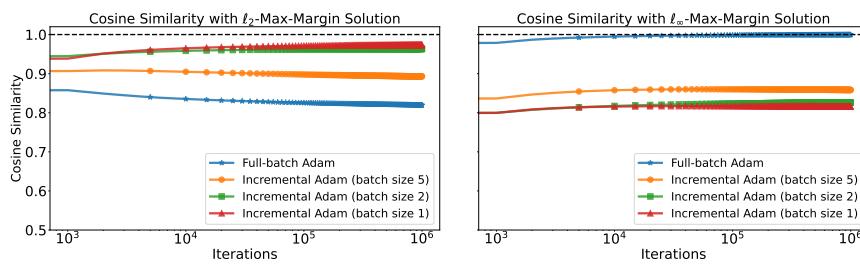
702 703 704 705 Appendix

706	A Further Discussion	15
707	A.1 Effect of Hyperparameters on Mini-batch Adam	15
708	A.2 Can We Directly Analyze Inc-Adam for General β_2 ?	15
710		
711	B Additional Experiments	17
712		
713	C Experimental Details	17
714		
715	D Missing Proofs in Section 2	20
716	D.1 Proof of Proposition 2.4	20
717	D.2 Proof of Proposition 2.5	21
718		
719		
720	E Missing Proofs in Section 3	26
721	E.1 Proof of Corollary 3.2	26
722	E.2 Proof of Theorem 3.3	27
723		
724		
725	F Missing Proofs in Section 4	31
726	F.1 Proof of Proposition 4.1	31
727	F.2 Proof of Proposition 4.3	31
728	F.3 Proof of Lemma 4.5	33
729	F.4 Proof of Theorem 4.8	37
730	F.5 Detailed Calculations of Example 4.11	38
731		
732		
733		
734		
735	G Missing Proofs in Section 5	39
736		
737	H Missing Proofs in Appendix A	47
738		
739	I Technical Lemmas	47
740	I.1 Proxy Function	47
741	I.2 Properties of Loss Functions	48
742	I.3 Auxiliary Results	49
743		
744		
745		
746		
747		
748		
749		
750		
751		
752		
753		
754		
755		

756 A FURTHER DISCUSSION
757758 A.1 EFFECT OF HYPERPARAMETERS ON MINI-BATCH ADAM
759

760 The scope of our analysis does not fully encompass the effects of batch sizes and momentum
761 hyperparameters on the limit direction of mini-batch Adam. To motivate further investigation, this
762 section presents preliminary empirical evidence that shows the sensitivity of the limit direction to
763 these choices.

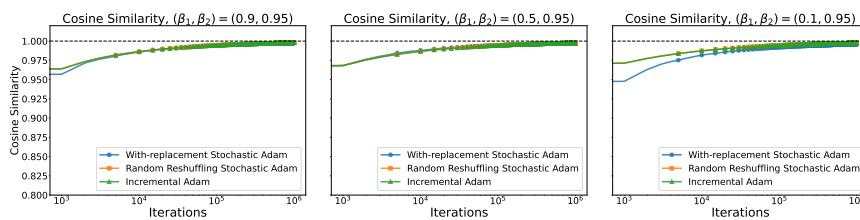
764 **Effect of Batch Size.** To investigate the effect of batch size on the limiting behavior of mini-batch
765 Adam, we run incremental Adam on the Gaussian data with $N = 10, d = 50$, varying batch sizes
766 among 1, 2, 5, and 10. Figure 6 shows that as the batch size increases, the cosine similarity between
767 the iterate and ℓ_∞ -max-margin solution increases. This result suggests that the choice of batch size
768 does affect the limiting behavior of mini-batch Adam, wherein larger batch sizes yield dynamics
769 that converge towards those of the full-batch regime. A formal characterization of this dependency
770 presents a compelling direction for future research.



771
772
773
774
775
776
777
778
779
780 Figure 6: **The choice of batch size influences the limit direction of mini-batch Adam.** We train
781 on the same Gaussian data ($N = 10, d = 50$) as in Figure 1 and plot the cosine similarity of the
782 weight vector with the ℓ_2 -max-margin solution (left) and the ℓ_∞ -max-margin solution (right), varying
783 batch sizes in $\{1, 2, 5, 10\}$. As the choice of batch size becomes closer to 10 (full-batch), the limit
784 direction aligns closer to ℓ_∞ -max-margin solution.

785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
17310
17311
17312
17313
17314
17315
17316
17317
17318
17319
17320
17321
17322
17323
17324
17325
17326
17327
17328
17329
17330
17331
17332
17333
17334
17335
17336
17337
17338
17339
17340
17341
17342
17343
17344
17345
17346
17347
17348
17349
17350
17351
17352
17353
17354
17355
17356
17357
17358
17359
17360
17361
17362
17363
17364
17365
17366
17367
17368
17369
17370
17371
17372
17373
17374
17375
17376
17377
17378
17379
17380
17381
17382
17383
17384
17385
17386
17387
17388
17389
17390
17391
17392
17393
17394
17395
17396
17397
17398
17399
17400
17401
17402
17403
17404
17405
17406
17407
17408
17409
17410
17411
17412
17413
17414
17415
17416
17417
17418
17419
17420
17421
17422
17423
17424
17425
17426
17427
17428
17429
17430
17431
17432
17433
17434
17435
17436
17437
17438
17439
17440
17441
17442
17443
17444
17445
17446
17447
17448
17449
17450
17451
17452
17453
17454
17455
17456
17457
17458
17459
17460
17461
17462
17463
17464
17465
17466
17467
17468
17469
17470
17471
17472
17473
17474
17475
17476
17477
17478
17479
17480
17481
17482
17483
17484
17485
17486
17487
17488
17489
17490
17491
17492
17493
17494
17495
17496
17497
17498
17499
17500
17501
17502
17503
17504
17505
17506
17507
17508
17509
17510
17511
17512
17513
17514
17515
17516
17517
17518
17519
17520
17521
17522
17523
17524
17525
17526
17527
17528
17529
17530
17531
17532
17533
17534
17535
17536
17537
17538
17539
17540
17541
17542
17543
17544
17545
17546
17547
17548
17549
17550
17551
17552
17553
17554
17555
17556
17557
17558
17559
17560
17561
17562
17563
17564
17565
17566
17567
17568
17569
17570
17571
17572
17573
17574
17575
17576
17577
17578
17579
17580
17581
17582
17583
17584
17585
17586
17587
17588
17589
17590
17591
17592
17593
17594
17595
17596
17597
17598
17599
17600
17601
17602
17603
17604
17605
17606
17607
17608
17609
17610
17611
17612
17613
17614
17615
17616
17617
17618
17619
17620
17621
17622
17623
17624
17625
17626
17627
17628
17629
17630
17631
17632
17633
17634
17635
17636
17637
17638
17639
17640
17641
17642
17643
17644
17645
17646
17647
17648
17649
17650
17651
17652
17653
17654
17655
17656
17657
17658
17659
17660
17661
17662
17663
17664
17665
17666
17667
17668
17669
17670
17671
17672
17673
17674
17675
17676
17677
17678
17679
17680
17681
17682
17683
17684
17685
17686
17687
17688
17689
17690
17691
17692
17693
17694
17695
17696
17697
17698
17699
17700
17701
17702
17703
17704
17705
17706
17707
17708
17709
17710
17711
17712
17713
17714
17715
17716
17717
17718
17719
17720
17721
17722
17723
17724
17725
17726
17727
17728
17729
17730
17731
17732
17733
17734
17735
17736
17737
17738
17739
17740
17741
17742
17743
17744
17745
17746
17747
17748
17749
17750
17751
17752
17753
17754
17755
17756
17757
17758
17759
17760
17761
17762
17763
17764
17765
17766
17767
17768
17769
17770
17771
17772
17773
17774
17775
17776
17777
17778
17779
17780
17781
17782
17783
17784
17785
17786
17787
17788
17789
17790
17791
17792
17793
17794
17795
17796
17797
17798
17799
17800
17801
17802
17803
17804
17805
17806
17807
17808
17809
17810
17811
17812
17813
17814
17815
17816
17817
17818
17819
17820
17821
17822
17823
17824
17825
17826
17827
17828
17829
17830
17831
17832
17833
17834
17835
17836
17837
17838
17839
17840
17841
17842
17843
17844
17845
17846
17847
17848
17849
17850
17851
17852
17853
17854
17855
17856
17857
17858
17859
17860
17861
17862
17863
17864
17865
17866
17867
17868
17869
17870
17871
17872
17873
17874
17875
17876
17877
17878
17879
17880
17881
17882
17883
17884
17885
17886
17887
17888
17889
17890
17891
17892
17893
17894
17895
17896
17897
17898
17899
17900
17901
17902
17903
17904
17905
17906
17907
17908
17909
17910
17911
17912
17913
17914
17915
17916
17917
17918
17919
17920
17921
17922
17923
17924
17925
17926
17927
17928
17929
17930
17931
17932
17933
17934
17935
17936
17937
17938
17939
17940
17941
17942
17943
17944
17945
17946
17947
17948
17949
17950
17951
17952
17953
17954
17955
17956
17957
17958
17959
17960
17961
17962
17963
17964
17965
17966
17967
17968
17969
17970
17971
17972
17973
17974
17975
17976
17977
17978
17979
17980
17981
17982
17983
17984
17985
17986
17987
17988
17989
17990
17991
17992
17993
17994
17995
17996
17997
17998
17999
18000
18001
18002
18003
18004
18005
18006
18007
18008
18009
18010
18011
18012
18013
18014
18015
18016
18

810



811

812

813

814

815

816

817

Figure 7: β_1 does not affect the convergence direction of mini-batch Adam for large β_2 . We train on the same Gaussian data as in Figure 1, varying $\beta_1 \in \{0.9, 0.5, 0.1\}$ with fixed $\beta_2 = 0.95$, and plot the cosine similarity between the weight vector and the fixed-point solution (Algorithm 3). All mini-batch Adam variants with batch size 1 consistently converge to the fixed-point solution.

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

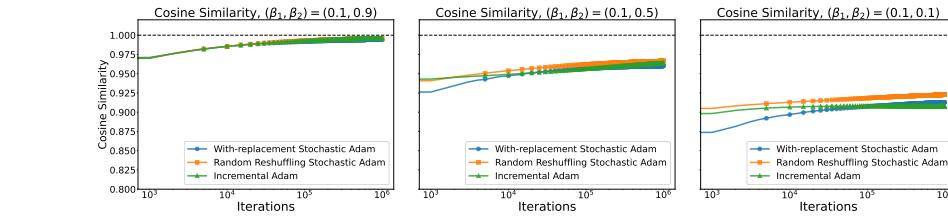
833

834

835

836

837



838

839

840

841

842

843

844

845

846

847

848

849

Figure 8: β_2 affects the convergence direction of mini-batch Adam. We train on the same Gaussian data as in Figure 1, varying $\beta_2 \in \{0.9, 0.5, 0.1\}$ with fixed $\beta_1 = 0.1$, and plot the cosine similarity between the weight vector and the fixed-point solution (Algorithm 3). Mini-batch Adam variants with batch size 1 deviate increasingly from the fixed-point solution as β_2 decreases.

850

851

852

853

854

855

856

857

858

859

860

861

862

863

Let $\{\mathbf{w}_t\}$ be the Inc-Adam iterates with $\beta_1 = 0$. For simplicity, we only consider the epoch-wise update and denote $\mathbf{w}_r = \mathbf{w}_r^0, \eta_r = C_{\text{inc}}(0, \beta_2)\eta_{rN}$ as an abuse of notation. By Proposition 2.5, \mathbf{w}_r can be written by

$$\delta_r \triangleq \underbrace{\sum_{i \in [N]} \frac{\nabla \mathcal{L}_i(\mathbf{w}_r)}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r)^2}}}_{\spadesuit} + \epsilon_r$$

$$\mathbf{w}_{r+1} - \mathbf{w}_r = -\eta_r \delta_r$$

for some $\epsilon_r \rightarrow 0$. Note that (\spadesuit) replaces AdamProxy in Section 4, incorporating the rich behavior induced by a general β_2 . Then, we provide a preliminary characterization of the limit direction of Inc-Adam as follows.

Lemma A.1. Suppose that (a) $\mathcal{L}(\mathbf{w}_r) \rightarrow 0$ and (b) $\mathbf{w}_r = \|\mathbf{w}_r\|_2 \hat{\mathbf{w}} + \rho(r)$ for some $\hat{\mathbf{w}}$ with $\exists \lim_{r \rightarrow \infty} \rho(r)$. Then, under Assumptions 2.1 and 2.2, there exists $\mathbf{c} = (c_0, \dots, c_{N-1}) \in \Delta^{N-1}$ such that the limit direction $\hat{\mathbf{w}}$ of Inc-Adam with $\beta_1 = 0$ satisfies

$$\hat{\mathbf{w}} \propto \sum_{i \in [N]} \frac{c_i \mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} c_j^2 \mathbf{x}_j^2}}, \quad (8)$$

and $c_i = 0$ for $i \notin S$, where $S = \arg \min_{i \in [N]} \hat{\mathbf{w}}^\top \mathbf{x}_i$ is the index set of support vectors of $\hat{\mathbf{w}}$.

We recall that the fixed-point formulation in Theorem 4.8 arises from constructing an optimization problem whose KKT conditions are given by Equation (5) fixing the c_i 's in the denominator; the convergence direction is then characterized when the dual solutions of the KKT conditions coincide with the c_i 's in the denominator. Therefore, to establish an analogous fixed-point type characterization, we should construct an optimization problem whose solution is given by $\mathbf{w}^* = \sum_{i \in [N]} \frac{d_i \mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} c_j^2 \mathbf{x}_j^2}}$ with dual variables $d_i \geq 0$ satisfying that $d_j = 0$ for $j \in S = \arg \min_{i \in [N]} \mathbf{w}^* \top \mathbf{x}_i$.

However, this cannot be formulated via KKT conditions of an optimization problem. The index set S indicates support vectors with respect to \mathbf{x}_i , while our dual variables are multiplied to

864 $\frac{\mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} c_j^2 \mathbf{x}_j^2}} = \hat{\mathbf{x}}_i(\mathbf{c})$. A notable direction for future work is to generalize the proposed
 865 methodology for arbitrary values of β_2 .
 866

867

868 **B ADDITIONAL EXPERIMENTS**
 869

870 **Supplementary Experiments in Section 3.** To investigate the universality of Theorem 3.3 with
 871 respect to the choice of the momentum hyperparameters, we run mini-batch Adam (with batch size
 872 1) on SR dataset $\mathbf{x}_0 = (1, 1, 1, 1)$, $\mathbf{x}_1 = (2, 2, 2, -2)$, $\mathbf{x}_2 = (3, 3, -3, -3)$, and $\mathbf{x}_3 = (4, -4, 4, -4)$,
 873 varying the momentum hyperparameters $(\beta_1, \beta_2) \in \{(0.1, 0.1), (0.5, 0.5), (0.9, 0.95)\}$. Figure 9
 874 demonstrates that its limiting behavior toward ℓ_2 -max-margin solution consistently holds on the
 875 broad choices of (β_1, β_2) .
 876

877 **Supplementary Experiments in Section 5.** Theorem 5.1 demonstrates that Inc-Signum main-
 878 tains its bias to ℓ_∞ -max-margin solution, while the momentum hyperparameter β should be close
 879 enough to 1 depending on the choice of batch size; the gap between β and 1 should decrease as
 880 batch size b decreases. To investigate this dependency, we run Inc-Signum on the same Gaus-
 881 sian data as in Figure 1, varying batch size $b \in \{1, 2, 5, 10\}$ and the momentum hyperparameter
 882 $\beta \in \{0.5, 0.9, 0.95, 0.99\}$. Figure 10 shows that to maintain the ℓ_∞ -bias, the choice of β should be
 883 closer to 1 as the batch size decreases.
 884

885 **C EXPERIMENTAL DETAILS**
 886

887 This section provides details for the experiments presented in the main text and appendix.

888 We generate synthetic separable data as follows:

- 890 • **Gaussian data (Figures 1, 3, 5, 6 to 8 and 10):** Samples are drawn from the standard Gaussian
 891 distribution $\mathcal{N}(0, I)$. We set the dimension $d = 50$ and sample $N = 10$ points, ensuring a positive
 892 margin so that the data is linearly separable.
- 893 • **Scaled Rademacher (SR) data (Figures 2 and 9):** We use $\mathbf{x}_0 = (1, 1, 1, 1)$, $\mathbf{x}_1 = (2, 2, 2, -2)$,
 894 $\mathbf{x}_2 = (3, 3, -3, -3)$, and $\mathbf{x}_3 = (4, -4, 4, -4)$.
- 895 • **Shifted-diagonal data (Figure 4):** We use $\mathbf{x}_0 = (1, \delta, \delta, \delta)$, $\mathbf{x}_1 = (\delta, 2, \delta, \delta)$, $\mathbf{x}_2 = (\delta, \delta, 4, \delta)$, and
 896 $\mathbf{x}_3 = (\delta, \delta, \delta, 8)$ with $\delta = 0.1$.
 897

898 We minimize the exponential loss using various algorithms. Momentum hyperparameters are
 899 $(\beta_1, \beta_2) = (0.9, 0.95)$ for Adam and $\beta = 0.99$ for Signum unless specified otherwise. For Adam
 900 and Signum variants, we use a learning rate schedule $\eta_t = \eta_0(t+2)^{-a}$ with $\eta_0 = 0.1$ and $a = 0.8$,
 901 following our theoretical analysis. Gradient descent uses a fixed learning rate $\eta_t = \eta_0 = 0.1$. Margins
 902 with respect to different norms are computed using CVXPY (Diamond & Boyd, 2016).

903 The fixed-point solution (Theorem 4.8) is obtained via fixed-point iteration (Algorithm 3) for Figures 3,
 904 7 and 8. We initialize $\mathbf{c}_0 = (1/N, \dots, 1/N) \in \Delta^{N-1}$, set the threshold $\epsilon_{\text{thr}} = 10^{-8}$, and converge
 905 to the fixed-point solution within 20 iterations in all settings.

906

907

908

909

910

911

912

913

914

915

916

917

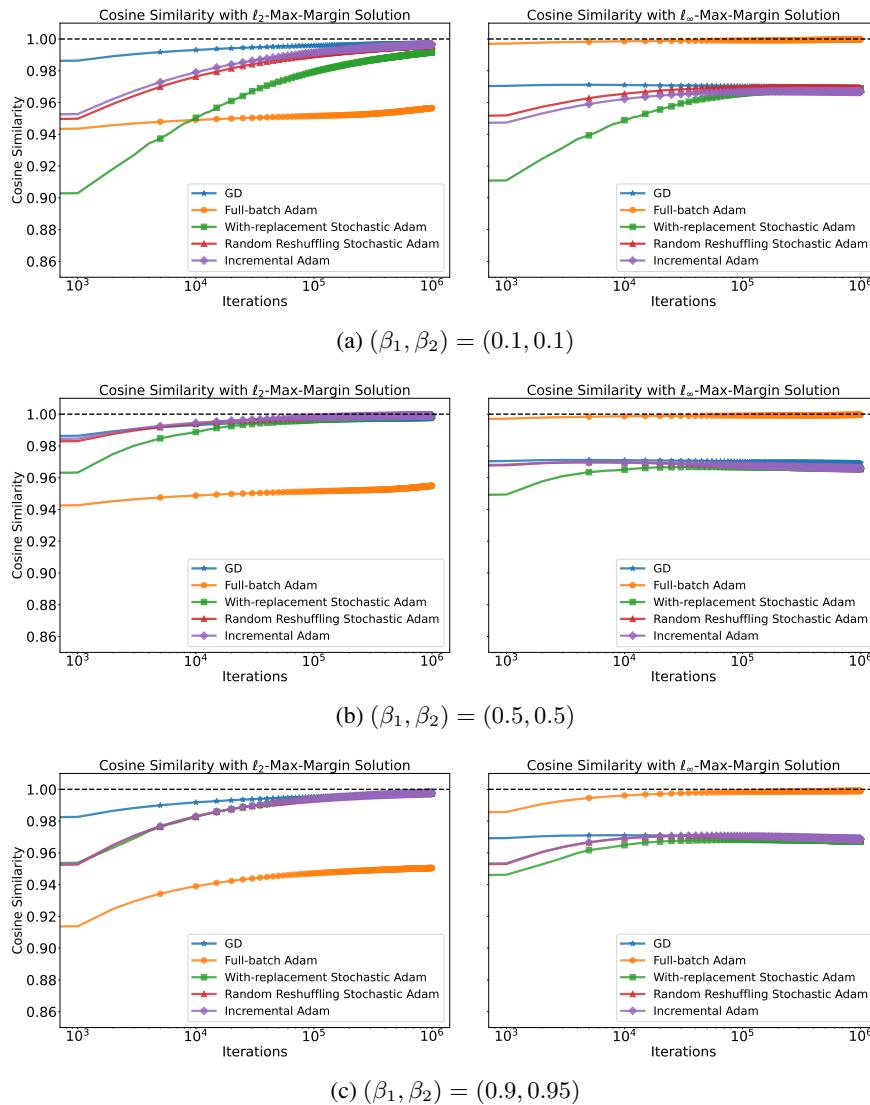


Figure 9: **Mini-batch Adam converges to the max ℓ_2 -margin solution for SR data.** We train on SR dataset $\mathbf{x}_0 = (1, 1, 1, 1)$, $\mathbf{x}_1 = (2, 2, 2, -2)$, $\mathbf{x}_2 = (3, 3, -3, -3)$, and $\mathbf{x}_3 = (4, -4, 4, -4)$, varying the momentum hyperparameters. In all tested configurations, the family of mini-batch Adam algorithms with batch size 1 converge to the ℓ_2 max-margin solution, which deviate significantly from the ℓ_∞ bias of full-batch Adam.

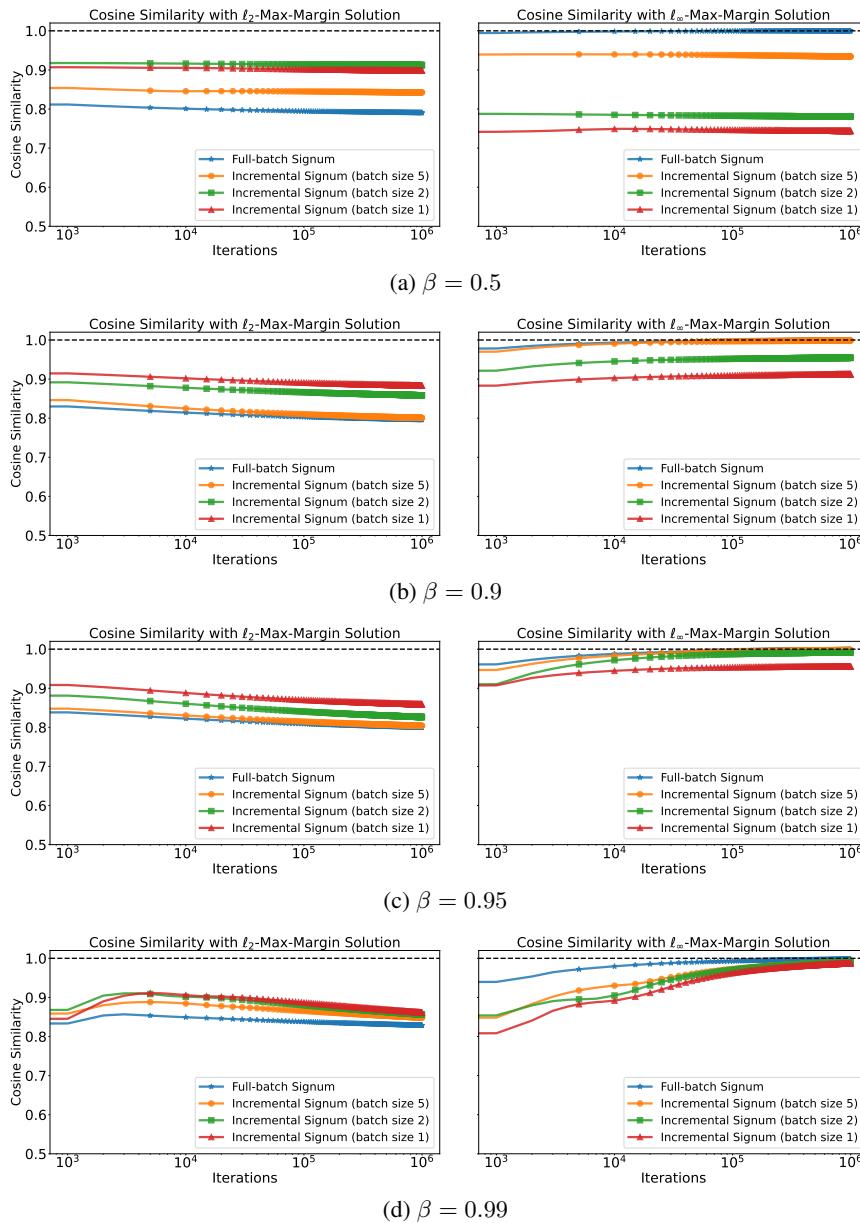
972
973
974
975
976

Figure 10: **Effect of Batch Size on Inc-Signum.** We run Inc-Signum on the same Gaussian data ($N = 10, d = 50$) as in Figure 1 and plot the cosine similarity of the weight vector with the ℓ_2 -max-margin solution (left) and the ℓ_∞ -max-margin solution (right), varying batch size $b \in \{1, 2, 5, 10\}$ and the momentum hyperparameter $\beta \in \{0.5, 0.9, 0.95, 0.99\}$. As the batch size decreases, we should choose β closer to 1 to maintain the limit direction toward ℓ_∞ -max-margin solution.

1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

1026 D MISSING PROOFS IN SECTION 2

1028 In this section, we provide the omitted proofs in Section 2, which describes asymptotic behaviors of
 1029 Det-Adam and Inc-Adam. We first introduce Lemma D.1 originated from Zou et al. (2023, Lemma
 1030 A.2), which gives a coordinate-wise upper bound of updates of both Det-Adam and Inc-Adam.
 1031 Then, we prove Propositions 2.4 and 2.5 by approximating two momentum terms.

1032 **Notation.** In this section, we introduce the proxy function $\mathcal{G} : \mathbb{R}^d \rightarrow \mathbb{R}$ defined as

$$1034 \mathcal{G}(\mathbf{w}) := -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i).$$

1035 **Lemma D.1** (Lemma A.2 in Zou et al. (2023)). *Assume $\beta_1^2 \leq \beta_2$ and let $\alpha = \sqrt{\frac{\beta_2(1-\beta_1)^2}{(1-\beta_2)(\beta_2-\beta_1^2)}}$.
 1036 Then, for both Det-Adam and Inc-Adam iterates, $\mathbf{m}_t[k] \leq \alpha \sqrt{\mathbf{v}_t[k]}$ for all $k \in [d]$.*

1037 *Proof.* Following the proof of Zou et al. (2023, Lemma A.2), we can easily show that the given
 1038 upper bound holds for both Det-Adam and Inc-Adam. We prove the case of Inc-Adam, while it
 1039 naturally extends to Det-Adam. By Cauchy-Schwartz inequality, we get

$$\begin{aligned} 1040 |\mathbf{m}_t[k]| &= \left| \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_{t-\tau})[k] \right| \\ 1041 &\leq \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) |\nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_{t-\tau})[k]| \\ 1042 &\leq \left(\sum_{\tau=0}^t \beta_2^\tau (1-\beta_2) |\nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_{t-\tau})[k]|^2 \right)^{1/2} \left(\sum_{\tau=0}^t \frac{\beta_1^{2\tau} (1-\beta_1)^2}{\beta_2^\tau (1-\beta_2)} \right)^{1/2} \quad (\text{CS inequality}) \\ 1043 &\leq \alpha \sqrt{\mathbf{v}_t[k]}. \end{aligned}$$

1044 The last inequality is from

$$\sum_{\tau=0}^t \frac{\beta_1^{2\tau} (1-\beta_1)^2}{\beta_2^\tau (1-\beta_2)} \leq \frac{(1-\beta_1)^2}{1-\beta_2} \sum_{\tau=0}^\infty \left(\frac{\beta_1^2}{\beta_2} \right)^\tau = \frac{\beta_2(1-\beta_1)^2}{(1-\beta_2)(\beta_2-\beta_1^2)} = \alpha^2,$$

1045 where the infinite sum is bounded from $\beta_1^2 \leq \beta_2$. □

1046 D.1 PROOF OF PROPOSITION 2.4

1047 **Proposition 2.4.** *Let $\{\mathbf{w}_t\}_{t=0}^\infty$ be the iterates of Det-Adam with $\beta_1 \leq \beta_2$. Then, under Assumptions 2.2 and 2.3, if $\lim_{t \rightarrow \infty} \frac{\eta_t^{1/2} \mathcal{L}(\mathbf{w}_t)}{|\nabla \mathcal{L}(\mathbf{w}_t)[k]|} = 0$, then the update of k -th coordinate $\mathbf{w}_{t+1}[k] - \mathbf{w}_t[k]$ can be represented by*

$$1048 \mathbf{w}_{t+1}[k] - \mathbf{w}_t[k] = -\eta_t (\text{sign}(\nabla \mathcal{L}(\mathbf{w}_t)[k]) + \epsilon_t), \quad (1)$$

1049 for some $\lim_{t \rightarrow \infty} \epsilon_t = 0$.

1050 *Proof.* We recall Lemma 6.1 in Zhang et al. (2024a), stating that

$$\begin{aligned} 1051 |\mathbf{m}_t[k] - (1-\beta_1^{t+1}) \nabla \mathcal{L}(\mathbf{w}_t)[k]| &\leq c_m \eta_t \mathcal{G}(\mathbf{w}_t), \\ 1052 \left| \sqrt{\mathbf{v}_t[k]} - \sqrt{1-\beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{w}_t)[k]| \right| &\leq c_v \sqrt{\eta_t} \mathcal{G}(\mathbf{w}_t) \end{aligned}$$

1053 for all $t > t_1$ and $k \in [d]$. Based on these results, we can rewrite $\mathbf{m}_t^s[k]$ and $\sqrt{\mathbf{v}_t^s[k]}$ as

$$\begin{aligned} 1054 \mathbf{m}_t[k] &= (1-\beta_1^{t+1}) \nabla \mathcal{L}(\mathbf{w}_t)[k] + \epsilon_m(t) \mathcal{G}(\mathbf{w}_t), \\ 1055 \sqrt{\mathbf{v}_t[k]} &= \sqrt{1-\beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{w}_t)[k]| + \epsilon_v(t) \mathcal{G}(\mathbf{w}_t), \end{aligned}$$

1080 where $\epsilon_{\mathbf{m}}(t) = \mathcal{O}(\eta_t)$, $\epsilon_{\mathbf{v}}(t) = \mathcal{O}(\sqrt{\eta_t})$. Note that $\frac{\mathcal{G}(\mathbf{w}_t)}{\mathcal{L}(\mathbf{w}_t)} \leq 1$ from Lemma I.1 and
 1081 $\left| \frac{a+\epsilon_1}{b+\epsilon_2} - \frac{a}{b} \right| \leq \left| \frac{\epsilon_1}{b+\epsilon_2} \right| + \left| \frac{a}{b} \cdot \frac{\epsilon_2}{b+\epsilon_2} \right| \leq \left| \frac{\epsilon_1}{b} \right| + \left| \frac{a}{b} \cdot \frac{\epsilon_2}{b} \right|$ for positive numbers $\epsilon_1, \epsilon_2, b$. Therefore,
 1082 if $\lim_{t \rightarrow \infty} \frac{\eta_t^{1/2} \mathcal{L}(\mathbf{w}_t)}{|\nabla \mathcal{L}(\mathbf{w}_t)[k]|} = 0$, then we get
 1083

$$\begin{aligned}
 & \left| \frac{\mathbf{m}_t[k]}{\sqrt{\mathbf{v}_t[k]}} - \frac{1 - \beta_1^{t+1}}{\sqrt{1 - \beta_2^{t+1}}} \operatorname{sign}(\nabla \mathcal{L}(\mathbf{w}_t)[k]) \right| \\
 & \leq \underbrace{\frac{\epsilon_{\mathbf{m}}(t) \mathcal{G}(\mathbf{w}_t)}{\sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{w}_t)[k]|}}_{\rightarrow 0} + \underbrace{\frac{1 - \beta_1^{t+1}}{\sqrt{1 - \beta_2^{t+1}}} \operatorname{sign}(\nabla \mathcal{L}(\mathbf{w}_t)[k]) \cdot \frac{\epsilon_{\mathbf{v}}(t) \mathcal{G}(\mathbf{w}_t)}{\sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{w}_t)[k]|}}_{\text{bounded}} \\
 & \rightarrow 0.
 \end{aligned}$$

1099 From $\beta_1^t, \beta_2^t \rightarrow 0$, we get $\mathbf{w}_{t+1}[k] - \mathbf{w}_t[k] = -\eta_t \frac{\mathbf{m}_t[k]}{\sqrt{\mathbf{v}_t[k]}} = \eta_t (\operatorname{sign}(\nabla \mathcal{L}(\mathbf{w}_t)[k]) + \epsilon_t)$ for some
 1100 $\lim_{t \rightarrow \infty} \epsilon_t = 0$. \square

1104 D.2 PROOF OF PROPOSITION 2.5

1107 To prove Proposition 2.5, we start by characterizing the first and second momentum terms $\mathbf{m}_t, \mathbf{v}_t$
 1108 in Inc-Adam, which track the exponential moving averages of the historical mini-batch gradients
 1109 and square gradients. As mentioned before, a key technical challenge of analyzing Adam is its
 1110 dependency in the full gradient history. The following lemma approximates momentum terms with
 1111 respect to a function of the *first* iterate in each epoch \mathbf{w}_r^0 , which is crucial for our *epoch-wise* analysis.

1112 **Lemma D.2.** *Under Assumptions 2.2 and 2.3, there exists t_1 only depending on β_1, β_2 and the
 1113 dataset, such that*

$$\begin{aligned}
 & \left| \mathbf{m}_r^s[k] - \frac{1 - \beta_1}{1 - \beta_1^N} \sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k] \right| \leq \epsilon_{\mathbf{m}}(t) \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|, \\
 & \left| \mathbf{v}_r^s[k] - \frac{1 - \beta_2}{1 - \beta_2^N} \sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2 \right| \leq \epsilon_{\mathbf{v}}(t) \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|^2,
 \end{aligned}$$

1124 for all r, s satisfying $rN + s > t_1$ and $k \in [d]$, where

$$\begin{aligned}
 \epsilon_{\mathbf{m}}(t) & \triangleq (1 - \beta_1) e^{\alpha N D \eta_{rN}} c_2 \eta_t + (e^{\alpha N D \eta_{rN}} - 1) + \beta_1^{t+1}, \\
 \epsilon_{\mathbf{v}}(t) & \triangleq 3(1 - \beta_2) e^{2\alpha N D \eta_{rN}} c'_2 \eta_t + 3(e^{2\alpha N D \eta_{rN}} - 1) + \beta_2^{t+1},
 \end{aligned}$$

1132 $D = \max_{j \in [N]} \|\mathbf{x}_j\|_1$, and c_2, c'_2 are constants only depend on β_1, β_2 , and the dataset.
 1133

1134 *Proof.* Consider $t = rN + s$ and the gradient at time t is sampled from data with index s in r -th
 1135 epoch. Then we can decompose the error between $\mathbf{m}_r^s[k]$ and $\frac{1-\beta_1}{1-\beta_1^N} \sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]$ as
 1136

$$\begin{aligned}
 & |\mathbf{m}_r^s[k] - \frac{1-\beta_1}{1-\beta_1^N} \sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]| \\
 &= \left| \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_{t-\tau})[k] - \frac{1-\beta_1}{1-\beta_1^N} \sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k] \right| \\
 &\leq \underbrace{\left| \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_{t-\tau})[k] - \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_t)[k] \right|}_{(A): \text{error from movement of weights}} \\
 &\quad + \underbrace{\left| \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_t)[k] - \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k] \right|}_{(B): \text{error between } \mathbf{w}_t \text{ and } \mathbf{w}_r^0} \\
 &\quad + \underbrace{\left| \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k] - \frac{1-\beta_1}{1-\beta_1^N} \sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k] \right|}_{(C): \text{error from infinite-sum approximation}}
 \end{aligned}$$

1156 Note that

$$\begin{aligned}
 (A) &\leq \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) |\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_{i_{t-\tau}}) - \ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})| |\mathbf{x}_{i_{t-\tau}}[k]| \\
 &= \sum_{\tau=0}^t \beta_1^\tau (1-\beta_1) \left| \frac{\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_{i_{t-\tau}})}{\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})} - 1 \right| |\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})| |\mathbf{x}_{i_{t-\tau}}[k]| \\
 &\stackrel{(*)}{\leq} (1-\beta_1) \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_t)[k]| \sum_{\tau=0}^t \beta_1^\tau (e^{\alpha D \sum_{\tau'=1}^{\tau} \eta_{t-\tau'}} - 1) \\
 &\stackrel{(**)}{\leq} (1-\beta_1) c_2 \eta_t \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_t)[k]|, \\
 &\stackrel{(***)}{\leq} (1-\beta_1) e^{\alpha N D \eta_{rN}} c_2 \eta_t \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|
 \end{aligned}$$

1171 for some $c_2 > 0$ and $t > t_1$. Here, $(*)$ is from Lemma I.3 and

$$e^{|\mathbf{w}_t - \mathbf{w}_{t-\tau}|} - 1 \leq e^{|\mathbf{w}_t - \mathbf{w}_{t-\tau}|_\infty \|\mathbf{x}_{i_{t-\tau}}\|_1} - 1 \leq e^{\alpha D \sum_{\tau'=1}^{\tau} \eta_{t-\tau'}} - 1.$$

1175 Also, $(**)$ is from Assumption 2.3, and $(***)$ is from

$$\begin{aligned}
 \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_t)[k]| &\leq \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]| \cdot \max_{j \in [N]} \left| \frac{\nabla \mathcal{L}_j(\mathbf{w}_t)[k]}{\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]} \right| \\
 &= \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]| \cdot \max_{j \in [N]} \left| \frac{\ell'(\mathbf{w}_t^\top \mathbf{x}_j)}{\ell'(\mathbf{w}_r^0^\top \mathbf{x}_j)} \right| \\
 &\leq e^{\alpha N D \eta_{rN}} \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|,
 \end{aligned}$$

1184 where the last inequality is from Lemma I.3 and

$$\max_{j \in [N]} \left| \frac{\ell'(\mathbf{w}_t^\top \mathbf{x}_j)}{\ell'(\mathbf{w}_r^0^\top \mathbf{x}_j)} \right| \leq \max_{j \in [N]} e^{|\mathbf{w}_t - \mathbf{w}_r^0|^\top \mathbf{x}_j} \leq e^{\alpha N D \eta_{rN}}.$$

1188 Also, observe that
 1189

$$\begin{aligned}
 1190 \quad (B) &\leq \sum_{\tau=0}^t \beta_1^\tau (1 - \beta_1) |\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}}) - \ell'(\mathbf{w}_r^0{}^\top \mathbf{x}_{i_{t-\tau}})| |\mathbf{x}_{i_{t-\tau}}[k]| \\
 1191 \\
 1192 &= \sum_{\tau=0}^t \beta_1^\tau (1 - \beta_1) \left| \frac{\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})}{\ell'(\mathbf{w}_r^0{}^\top \mathbf{x}_{i_{t-\tau}})} - 1 \right| |\ell'(\mathbf{w}_r^0{}^\top \mathbf{x}_{i_{t-\tau}})| |\mathbf{x}_{i_{t-\tau}}[k]| \\
 1193 \\
 1194 &\stackrel{(*)}{\leq} (1 - \beta_1) \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]| (e^{\alpha N D \eta_{rN}} - 1) \sum_{\tau=0}^t \beta_1^\tau \\
 1195 \\
 1196 &\stackrel{(**)}{\leq} (e^{\alpha N D \eta_{rN}} - 1) \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|,
 \end{aligned}$$

1200 where $(*)$ is from Lemma I.3 and
 1201

$$\left| \frac{\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})}{\ell'(\mathbf{w}_r^0{}^\top \mathbf{x}_{i_{t-\tau}})} - 1 \right| \leq e^{|\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}}) - \ell'(\mathbf{w}_r^0{}^\top \mathbf{x}_{i_{t-\tau}})|} - 1 \leq e^{\|\mathbf{w}_t - \mathbf{w}_r^0\|_\infty \|\mathbf{x}_{i_{t-\tau}}\|_1} \leq e^{\alpha N D \eta_{rN}} - 1,$$

1202 and $(**)$ is from $\sum_{\tau=0}^t \beta_1^\tau \leq \frac{1}{1-\beta_1}$.
 1203

1204 Furthermore,

$$\begin{aligned}
 1205 \quad (C) &= \left| \sum_{\tau=0}^t \beta_1^\tau (1 - \beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k] - \sum_{\tau=t+1}^{\infty} \beta_1^\tau (1 - \beta_1) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k] \right| \\
 1206 \\
 1207 &\leq \sum_{\tau=t+1}^{\infty} \beta_1^\tau (1 - \beta_1) |\nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k]| \\
 1208 &\leq \beta_1^{t+1} \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|.
 \end{aligned}$$

1209 Therefore, we can conclude that
 1210

$$\begin{aligned}
 1211 \quad |\mathbf{m}_r^s[k] - \frac{1 - \beta_1}{1 - \beta_1^N} \sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]| \\
 1212 &\leq \underbrace{\left((1 - \beta_1) e^{\alpha N D \eta_{rN}} c_2 \eta_t + (e^{\alpha N D \eta_{rN}} - 1) + \beta_1^{t+1} \right)}_{\triangleq \epsilon_{\mathbf{m}}(t)} \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|.
 \end{aligned}$$

1213 Similarly,
 1214

$$\begin{aligned}
 1215 \quad |\mathbf{v}_r^s[k] - \frac{1 - \beta_2}{1 - \beta_2^N} \sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2| \\
 1216 &= \left| \sum_{\tau=0}^t \beta_2^\tau (1 - \beta_2) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_{t-\tau})[k]^2 - \frac{1 - \beta_2}{1 - \beta_2^N} \sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2 \right| \\
 1217 &\leq \underbrace{\left| \sum_{\tau=0}^t \beta_2^\tau (1 - \beta_2) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_{t-\tau})[k]^2 - \sum_{\tau=0}^t \beta_2^\tau (1 - \beta_2) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_t)[k]^2 \right|}_{(D): \text{error from movement of weights}} \\
 1218 &\quad + \underbrace{\left| \sum_{\tau=0}^t \beta_2^\tau (1 - \beta_2) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_t)[k]^2 - \sum_{\tau=0}^t \beta_2^\tau (1 - \beta_2) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k]^2 \right|}_{(E): \text{error between } \mathbf{w}_t \text{ and } \mathbf{w}_r^0} \\
 1219 &\quad + \underbrace{\left| \sum_{\tau=0}^t \beta_2^\tau (1 - \beta_2) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k]^2 - \frac{1 - \beta_2}{1 - \beta_2^N} \sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2 \right|}_{(F): \text{error from infinite-sum approximation}}.
 \end{aligned}$$

1242 Observe that

$$\begin{aligned}
 1244 \quad (D) &\leq \sum_{\tau=0}^t \beta_2^\tau (1-\beta_2) |\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_{i_{t-\tau}})^2 - \ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})^2| |\mathbf{x}_{i_{t-\tau}}[k]|^2 \\
 1245 \\
 1246 \quad &= \sum_{\tau=0}^t \beta_2^\tau (1-\beta_2) \left| \left(\frac{\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_{i_{t-\tau}})}{\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})} \right)^2 - 1 \right| |\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})|^2 |\mathbf{x}_{i_{t-\tau}}[k]|^2 \\
 1247 \\
 1248 \quad &\stackrel{(*)}{\leq} 3(1-\beta_2) \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_t)[k]|^2 \sum_{\tau=0}^t \beta_2^\tau (e^{2\alpha D \sum_{\tau'=1}^\tau \eta_{t-\tau'}} - 1) \\
 1249 \\
 1250 \quad &\stackrel{(**)}{\leq} 3(1-\beta_2) c'_2 \eta_t \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_t)[k]|^2, \\
 1251 \\
 1252 \quad &\stackrel{(***)}{\leq} 3(1-\beta_2) e^{2\alpha ND \eta_{rN}} c'_2 \eta_t \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|^2
 \end{aligned}$$

1253 for some $c'_2 > 0$ and $t > t'_1$. Here, $(*)$ is from Lemma I.4 and

$$\left| \left(\frac{\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_{i_{t-\tau}})}{\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})} \right)^2 - 1 \right| \leq 3(e^{2|(\mathbf{w}_t - \mathbf{w}_r^0)^\top \mathbf{x}_{i_{t-\tau}}|} - 1) \leq 3(e^{2\alpha D \sum_{\tau'=1}^\tau \eta_{t-\tau'}} - 1),$$

1254 $(**)$ is from Assumption 2.3, and $(***)$ can be derived similarly. Also, we get

$$\begin{aligned}
 1255 \quad (E) &\leq \sum_{\tau=0}^t \beta_2^\tau (1-\beta_2) |\ell'(\mathbf{w}_t^\top \mathbf{x}_{i_{t-\tau}})^2 - \ell'(\mathbf{w}_r^0 \top \mathbf{x}_{i_{t-\tau}})^2| |\mathbf{x}_{i_{t-\tau}}[k]|^2 \\
 1256 \quad &\leq 3(e^{2\alpha ND \eta_{rN}} - 1) \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|^2, \\
 1257 \quad (F) &= \left| \sum_{\tau=0}^t \beta_2^\tau (1-\beta_2) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k]^2 - \sum_{\tau=t+1}^\infty \beta_2^\tau (1-\beta_2) \nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k]^2 \right| \\
 1258 \quad &\leq \sum_{\tau=t+1}^\infty \beta_2^\tau (1-\beta_2) |\nabla \mathcal{L}_{i_{t-\tau}}(\mathbf{w}_r^0)[k]|^2 \\
 1259 \quad &\leq \beta_2^{t+1} \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|^2,
 \end{aligned}$$

1260 which can also be derived similarly to the previous part. Therefore, we can conclude that

$$\begin{aligned}
 1261 \quad &|\mathbf{v}_r^s[k] - \frac{1-\beta_2}{1-\beta_2^N} \sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2| \\
 1262 \quad &\leq \underbrace{(3(1-\beta_2) e^{2\alpha ND \eta_{rN}} c'_2 \eta_t + 3(e^{2\alpha ND \eta_{rN}} - 1) + \beta_2^{t+1})}_{\triangleq \epsilon_v(t)} \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|^2.
 \end{aligned}$$

1263 \square

1264 Notice that $\epsilon_m(t)$ and $\epsilon_v(t)$ defined in Lemma D.2 converge to 0 as $t \rightarrow \infty$, implying that each 1265 coordinate of two momentum terms can be effectively approximated by a weighted sum of mini-batch 1266 gradients and gradient squares, which emphasizes the discrepancy with Det-Adam and Inc-Adam. 1267 We also mention that the bound depends on $\max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|$, which converges to 0 as 1268 $\mathcal{L}(\mathbf{w}_r^0) \rightarrow 0$. Such approaches provide tight bounds, which enables the asymptotic analysis of 1269 Inc-Adam.

1270 **Proposition 2.5.** *Let $\{\mathbf{w}_t\}_{t=0}^\infty$ be the iterates of Inc-Adam with $\beta_1 \leq \beta_2$. Then, under Assumptions 1271 2.2 and 2.3, the epoch-wise update $\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0$ can be represented by*

$$\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0 = -\eta_{rN} \left(C_{inc}(\beta_1, \beta_2) \sum_{i \in [N]} \frac{\sum_{j \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)^2}} + \boldsymbol{\epsilon}_r \right), \quad (2)$$

1296 where $\beta_1^{(i,j)} = \beta_1^{(i-j) \bmod N}$, $\beta_2^{(i,j)} = \beta_2^{(i-j) \bmod N}$, $C_{inc}(\beta_1, \beta_2) = \frac{1-\beta_1}{1-\beta_2^N} \sqrt{\frac{1-\beta_2^N}{1-\beta_2}}$ is a function of
 1297 β_1, β_2 , and $\lim_{r \rightarrow \infty} \epsilon_r = \mathbf{0}$. If $\eta_t = (t+2)^{-a}$ for some $a \in (0, 1]$, then $\|\epsilon_r\|_\infty = \mathcal{O}(r^{-a/2})$.
 1298

1300 *Proof.* Since both $\mathbf{v}_r^s[k]$ and $\frac{1-\beta_2}{1-\beta_2^N} \sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2$ are positive and $|a^2 - b^2| = |a - b| |a + b| \geq |a - b|^2$ holds for two positive numbers a and b , Lemma D.2 implies that
 1301

$$1303 \left| \sqrt{\mathbf{v}_r^s[k]} - \sqrt{\frac{1-\beta_2}{1-\beta_2^N}} \sqrt{\sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2} \right| \leq \sqrt{\epsilon_{\mathbf{v}}(t)} \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|.$$

1306 Therefore, we can rewrite $\mathbf{m}_r^s[k]$ and $\sqrt{\mathbf{v}_r^s[k]}$ as
 1307

$$1308 \mathbf{m}_r^s[k] = \underbrace{\frac{1-\beta_1}{1-\beta_1^N} \sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]}_{(a)} + \underbrace{\epsilon'_{\mathbf{m}}(t) \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|}_{(\epsilon_1)},$$

$$1312 \sqrt{\mathbf{v}_r^s[k]} = \underbrace{\sqrt{\frac{1-\beta_2}{1-\beta_2^N}} \sqrt{\sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}}_{(b)} + \underbrace{\sqrt{\epsilon'_{\mathbf{v}}(t)} \max_{j \in [N]} |\nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]|}_{(\epsilon_2)},$$

1317 for some error terms $\epsilon'_{\mathbf{m}}(t), \epsilon'_{\mathbf{v}}(t)$ such that $|\epsilon'_{\mathbf{m}}(t)| \leq \epsilon_{\mathbf{m}}(t), |\epsilon'_{\mathbf{v}}(t)| \leq \epsilon_{\mathbf{v}}(t)$. Note that
 1318 $\left| \frac{a+\epsilon_1}{b+\epsilon_2} - \frac{a}{b} \right| \leq \left| \frac{\epsilon_1}{b+\epsilon_2} \right| + \left| \frac{a}{b} \cdot \frac{\epsilon_2}{b+\epsilon_2} \right| \leq \left| \frac{\epsilon_1}{b} \right| + \left| \frac{a}{b} \cdot \frac{\epsilon_2}{b} \right|$ for positive numbers $\epsilon_1, \epsilon_2, b$. Thus, we
 1319 can conclude that

$$1321 \left| \frac{\mathbf{m}_r^s[k]}{\sqrt{\mathbf{v}_r^s[k]}} - \frac{(a)}{(b)} \right| \leq \left| \frac{(\epsilon_1)}{(b)} \right| + \left| \frac{(a)}{(b)} \cdot \frac{(\epsilon_2)}{(b)} \right| \rightarrow 0, \quad (9)$$

1324 since

$$1325 \left| \frac{(\epsilon_1)}{(b)} \right| \leq \frac{1}{\sqrt{\frac{1-\beta_2}{1-\beta_2^N}} \sqrt{\beta_2^N}} \epsilon_{\mathbf{m}}(t) \rightarrow 0,$$

$$1328 \left| \frac{(a)}{(b)} \right| \leq \frac{\frac{1-\beta_1}{1-\beta_1^N}}{\sqrt{\frac{1-\beta_2}{1-\beta_2^N}}} \sqrt{N},$$

$$1331 \left| \frac{(\epsilon_2)}{(b)} \right| \leq \frac{1}{\sqrt{\frac{1-\beta_2}{1-\beta_2^N}} \sqrt{\beta_2^N}} \sqrt{\epsilon_{\mathbf{v}}(t)} \rightarrow 0.$$

1334 Now consider the epoch-wise update. From above results, we get
 1335

$$1336 \mathbf{w}_{r+1}^0[k] - \mathbf{w}_r^0[k] = - \sum_{s=0}^{N-1} \eta_s \frac{\mathbf{m}_r^s[k]}{\sqrt{\mathbf{v}_r^s[k]}}$$

$$1339 = - \sum_{s=0}^{N-1} \eta_{rN+s} \left(C_{inc}(\beta_1, \beta_2) \frac{\sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} + \epsilon_{rN+s}[k] \right), \quad (10)$$

1344 for some $\epsilon_t \rightarrow \mathbf{0}$. Since $\lim_{t \rightarrow \infty} \eta_t = 0$, the difference between η_{rN+s} for different $s \in [N]$
 1345 converges to 0, which proves the claim.

1346 Next, we consider the case $\eta_t = (t+2)^{-a}$ for some $a \in (0, 1]$. Then it is clear that
 1347

$$1348 \epsilon_{\mathbf{m}}(t) = (1-\beta_1) e^{\alpha N D \eta_{rN}} c_2 \eta_t + (e^{\alpha N D \eta_{rN}} - 1) + \beta_1^{t+1} = \mathcal{O}(t^{-a}),$$

$$1349 \epsilon_{\mathbf{v}}(t) = 3(1-\beta_2) e^{2\alpha N D \eta_{rN}} c_2' \eta_t + 3(e^{2\alpha N D \eta_{rN}} - 1) + \beta_2^{t+1} = \mathcal{O}(t^{-a}),$$

1350 where $D = \max_{j \in [N]} \|\mathbf{x}_j\|_1$. Therefore, from Equation (9), we get
 1351

$$1353 \quad \left| \frac{\mathbf{m}_r^s[k]}{\sqrt{\mathbf{v}_r^s[k]}} - C_{\text{inc}}(\beta_1, \beta_2) \frac{\sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} \right| = \mathcal{O}(t^{-a/2}),$$

1357 which implies $\epsilon_t[k] = \mathcal{O}(t^{-a/2})$ in Equation (10). Note that
 1358

$$1360 \quad \sum_{s=0}^{N-1} \eta_{rN+s} \left(\underbrace{C_{\text{inc}}(\beta_1, \beta_2) \frac{\sum_{j \in [N]} \beta_1^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{j \in [N]} \beta_2^{(s,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} + \epsilon_{rN+s}[k]}_{\triangleq p(s)} \right) \\ 1367 \quad = \eta_{rN} \sum_{s=0}^{N-1} \left(p(s) + \underbrace{\frac{\eta_{rN+s} - \eta_{rN}}{\eta_{rN}} p(s) + \frac{\eta_{rN+s}}{\eta_{rN}} \epsilon_{rN+s}[k]}_{\triangleq \epsilon'_{rN+s}[k]} \right).$$

1373 Furthermore,
 1374

$$1376 \quad \frac{\eta_{rN} - \eta_{(r+1)N}}{\eta_{rN}} = 1 - \left(1 + \frac{N}{rN+2} \right)^{-a} = \mathcal{O}(r^{-1}),$$

1380 from Lemma I.7. Since $p(s)$ is upper bounded by a constant from CS inequality, we get $\epsilon'_{rN+s}[k] =$
 1381 $\mathcal{O}(r^{-a/2})$, which ends the proof. \square
 1382

1384 E MISSING PROOFS IN SECTION 3

1387 In this section, we provide the omitted proofs in Section 3. We first introduce the proof of Corollary 3.2
 1388 describing how SR datasets eliminate coordinate-adaptivity of Inc-Adam. Then, we review previous
 1389 literature on the limit direction of weighted GD and prove Theorem 3.3.
 1390

1392 E.1 PROOF OF COROLLARY 3.2

1394 **Corollary 3.2.** *Consider Inc-Adam iterates $\{\mathbf{w}_t\}_{t=0}^\infty$ on SR data. Then, under Assumptions 2.2
 1395 and 2.3, the epoch-wise update $\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0$ can be approximated by weighted normalized GD, i.e.,*
 1396

$$1398 \quad \mathbf{w}_{r+1}^0 - \mathbf{w}_r^0 = -\eta_{rN} \left(\sum_{i \in [N]} \frac{a_i(r)}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2} \nabla \mathcal{L}_i(\mathbf{w}_r^0) + \epsilon_r \right), \quad (3)$$

1402 where $\lim_{r \rightarrow \infty} \epsilon_r = \mathbf{0}$ and $c_1 \leq a_i(r) \leq c_2$ for some positive constants c_1, c_2 only depending on
 1403 $\beta_1, \beta_2, \{\mathbf{x}_i\}_{i \in [N]}$. If $\eta_t = (t+2)^{-a}$ for some $a \in (0, 1]$, then $\|\epsilon_r\|_\infty = \mathcal{O}(r^{-a/2})$.

1404 *Proof.* Given SR data $\{\mathbf{x}_i\}_{i \in [N]}$, let $x_i = |\mathbf{x}_i[0]|$. Notice that

$$\begin{aligned}
 \sum_{i \in [N]} \frac{\sum_{j \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)^2}} &= \sum_{i \in [N]} \frac{\sum_{j \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)}{\sqrt{\sum_{l \in [N]} \beta_2^{(i,l)} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} \\
 &= \sum_{i \in [N]} \sum_{j \in [N]} \frac{\beta_1^{(i,j)}}{\sqrt{\sum_{l \in [N]} \beta_2^{(i,l)} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} \nabla \mathcal{L}_j(\mathbf{w}_r^0) \\
 &= \sum_{j \in [N]} \left(\sum_{i \in [N]} \frac{\beta_1^{(i,j)}}{\sqrt{\sum_{l \in [N]} \beta_2^{(i,l)} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} \right) \nabla \mathcal{L}_j(\mathbf{w}_r^0) \\
 &= \sum_{j \in [N]} \underbrace{\left(\sum_{i \in [N]} \frac{\beta_1^{(i,j)} \|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2}{\sqrt{\sum_{l \in [N]} \beta_2^{(i,l)} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} \right)}_{a_j(r)} \frac{\nabla \mathcal{L}_j(\mathbf{w}_r^0)}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2}.
 \end{aligned}$$

1421 Therefore, it is enough to show that $a_j(r)$ is bounded. Note that

$$\begin{aligned}
 a_j(r) &\leq \frac{N}{\sqrt{\beta_2^{N-1}}} \frac{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2}{\sqrt{\sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} = \frac{1}{\sqrt{\beta_2^{N-1}}} \frac{\|\sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)| \mathbf{x}_l\|_2}{\sqrt{\sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} \\
 &\leq \frac{\sqrt{d}}{\sqrt{\beta_2^{N-1}}} \frac{\sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)| x_l}{\sqrt{\sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} \leq \frac{\sqrt{dN}}{\sqrt{\beta_2^{N-1}}}.
 \end{aligned}$$

1429 To find lower bound of $a_j(r)$, we use Assumption 2.1. Take $\mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{v}\|_2 = 1$ and
1430 $\mathbf{v}^\top \mathbf{x}_i > 0, \forall i \in [N]$. Let $\gamma \triangleq \min_{i \in [N]} \mathbf{v}^\top \mathbf{x}_i > 0$. Note that

$$(-\mathbf{v})^\top \nabla \mathcal{L}(\mathbf{w}_r^0) = \frac{1}{N} \sum_{l \in [N]} (-\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)) \cdot \mathbf{v}^\top \mathbf{x}_l \geq \frac{\gamma}{N} \sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|,$$

1435 and by CS inequality,

$$\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2 = \|-\mathbf{v}\|_2 \|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2 \geq \langle -\mathbf{v}, \nabla \mathcal{L}(\mathbf{w}_r^0) \rangle \geq \frac{\gamma}{N} \sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|. \quad (11)$$

1439 Therefore, we can conclude that

$$\begin{aligned}
 a_j(r) &\geq N \beta_1^{N-1} \frac{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2}{\sqrt{\sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} \stackrel{(*)}{\geq} \gamma \beta_1^{N-1} \frac{\sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|}{\sqrt{\sum_{l \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_l \rangle)|^2 x_l^2}} \\
 &\geq \frac{\gamma \beta_1^{N-1}}{\max_{l \in [N]} x_l}
 \end{aligned}$$

1447 where $(*)$ is from Equation (11). Now we can take $c_1 = \frac{\gamma \beta_1^{N-1}}{\max_{l \in [N]} x_l}$ and $c_2 = \frac{\sqrt{dN}}{\sqrt{\beta_2^{N-1}}}$ only depending
1448 on $\beta_1, \beta_2, \{\mathbf{x}_i\}$. \square

1450 E.2 PROOF OF THEOREM 3.3

1452 **Related Work.** We now turn to the proof of Theorem 3.3, building upon the foundational work of
1453 [Ji et al. \(2020\)](#), who characterized the convergence direction of GD via its regularization path. Sub-
1454 sequent research has extended this characterization to weighted GD, which optimizes the weighted
1455 empirical risk $\mathcal{L}_{\mathbf{q}(t)}(\mathbf{w}) = \sum_{i \in [N]} q_i(t) \ell(\mathbf{w}^\top \mathbf{x}_i)$. [Xu et al. \(2021\)](#) proved that weighted GD con-
1456 verges to ℓ_2 -max-margin direction on the same linear classification task when the weights are fixed
1457 during training. This condition was later relaxed by [Zhai et al. \(2023\)](#), who demonstrated that the
same convergence guarantee holds provided the weights converge to a limit, i.e., $\exists \lim_{t \rightarrow \infty} \mathbf{q}(t) = \hat{\mathbf{q}}$.

Our setting, however, introduces distinct technical challenges. First, the weights are bounded but not guaranteed to converge. The most relevant existing result is Theorem 7 in [Zhai et al. \(2023\)](#), which establishes the same limit direction but requires the stronger combined assumptions of lower-bounded weights, loss convergence, and directional convergence of the iterates. A further complication in our analysis is an additional error term, ϵ_r in [Corollary 3.2](#), which must be carefully controlled. Our fine-grained analysis overcomes these issues by extending the methodology of [Ji et al. \(2020\)](#), enabling us to manage the error term under the sole, weaker assumption of loss convergence.

Definition E.1. Given $\mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$, we define \mathbf{a} -weighted loss as $\mathcal{L}^{\mathbf{a}}(\mathbf{w}) \triangleq \sum_{i \in [N]} a_i \mathcal{L}_i(\mathbf{w})$. We denote the regularized solution as $\bar{\mathbf{w}}^{\mathbf{a}}(B) \triangleq \arg \min_{\|\mathbf{w}\|_2 \leq B} \mathcal{L}^{\mathbf{a}}(\mathbf{w})$.

By introducing \mathbf{a} -weighted loss, we can regard weighted GD as vanilla GD with respect to weighted loss. To follow the line of [Ji et al. \(2020\)](#), we show that the regularization path converges in direction to ℓ_2 -max-margin solution, regardless of the choice of the weight vector \mathbf{a} if it is bounded by two positive constants, and such convergence is uniform; we can take sufficiently large B to be close the ℓ_2 solution for any $\mathbf{a} \in [c_1, c_2]^N$.

Lemma E.2 (Adaptation of [Proposition 10 in Ji et al. \(2020\)](#)). *Let $\hat{\mathbf{u}} = \arg \max_{\|\mathbf{v}\|_2 \leq 1} \min_{i \in [N]} \langle \mathbf{v}, \mathbf{x}_i \rangle$ be the (unique) ℓ_2 -max-margin solution and c_1, c_2 be two positive constants. Then, for any $\mathbf{a} \in [c_1, c_2]^N$,*

$$\lim_{B \rightarrow \infty} \frac{\bar{\mathbf{w}}^{\mathbf{a}}(B)}{B} = \hat{\mathbf{u}}.$$

Furthermore, given $\epsilon > 0$, there exists $M(c_1, c_2, \epsilon, N) > 0$ only depending on c_1, c_2, ϵ, N such that $B > M$ implies $\|\frac{\bar{\mathbf{w}}^{\mathbf{a}}(B)}{B} - \hat{\mathbf{u}}\| < \epsilon$ for any $\mathbf{a} \in [c_1, c_2]^N$.

Proof. We first have to show the uniqueness of ℓ_2 -max-margin solution. This proof was introduced by [Ji et al. \(2020\)](#), but we provide it for completeness. Suppose that there exist two distinct unit vectors \mathbf{u}_1 and \mathbf{u}_2 such that both of them achieve the max-margin $\hat{\gamma}$. Take $\mathbf{u}_3 = \frac{\mathbf{u}_1 + \mathbf{u}_2}{2}$ as a middle point of \mathbf{u}_1 and \mathbf{u}_2 . Then we get

$$\mathbf{u}_3^\top \mathbf{x}_i = \frac{1}{2}(\mathbf{u}_1^\top \mathbf{x}_i + \mathbf{u}_2^\top \mathbf{x}_i) \geq \hat{\gamma},$$

for all $i \in [N]$, which implies that $\min_{i \in [N]} \mathbf{u}_3^\top \mathbf{x}_i \geq \hat{\gamma}$. Since $\mathbf{u}_1 \neq \mathbf{u}_2$, we get $\|\mathbf{u}_3\| < 1$, implying that $\frac{\mathbf{u}_3}{\|\mathbf{u}_3\|}$ achieves a larger margin than $\hat{\gamma}$. This makes a contradiction.

Now we prove the main claim. Let $\hat{\gamma} = \min_{i \in [N]} \langle \hat{\mathbf{u}}, \mathbf{x}_i \rangle$ be the margin of $\hat{\mathbf{u}}$. Then, it satisfies

$$c_1 \ell\left(\min_{i \in [N]} \langle \bar{\mathbf{w}}^{\mathbf{a}}(B), \mathbf{x}_i \rangle\right) \leq \mathcal{L}^{\mathbf{a}}(\bar{\mathbf{w}}^{\mathbf{a}}(B)) \leq \mathcal{L}^{\mathbf{a}}(B\hat{\mathbf{u}}) \leq Nc_2 \ell(B\hat{\gamma}). \quad (12)$$

For $\ell = \ell_{\text{exp}}$, we get $\min_{i \in [N]} \langle \bar{\mathbf{w}}^{\mathbf{a}}(B), \mathbf{x}_i \rangle \geq B\hat{\gamma} - \log \frac{Nc_2}{c_1}$, which implies

$$\min_{i \in [N]} \langle \frac{\bar{\mathbf{w}}^{\mathbf{a}}(B)}{B}, \mathbf{x}_i \rangle \geq \hat{\gamma} - \frac{1}{B} \log \frac{Nc_2}{c_1}. \quad (13)$$

Since ℓ_2 -max-margin solution is unique, $\frac{\bar{\mathbf{w}}^{\mathbf{a}}(B)}{B}$ converges to $\hat{\mathbf{u}}$. Note that the lower bound in [Equation \(13\)](#) does not depend on $\mathbf{a} \in [c_1, c_2]^N$. Therefore, the choice of M in [Lemma E.2](#) only depends on c_1, c_2, ϵ, N .

For $\ell = \ell_{\log}$, [Equation \(12\)](#) implies that $\ell(\min_{i \in [N]} \langle \bar{\mathbf{w}}^{\mathbf{a}}(B), \mathbf{x}_i \rangle) \leq \frac{Nc_2}{c_1} \ell(B\hat{\gamma})$. Notice that $\frac{Nc_2}{c_1} > 1$ and $\min_{i \in [N]} \langle \bar{\mathbf{w}}^{\mathbf{a}}(B), \mathbf{x}_i \rangle > 0, B\hat{\gamma} > 0$ hold for sufficiently large B from [Lemma I.2](#). From [Lemma I.5](#), we get

$$\min_{i \in [N]} \langle \frac{\bar{\mathbf{w}}^{\mathbf{a}}(B)}{B}, \mathbf{x}_i \rangle \geq \hat{\gamma} - \frac{1}{B} \log(2^{\frac{Nc_2}{c_1}} - 1).$$

Following the proof of the previous part, we can easily show that the statement also holds in this case. \square

Lemma E.3 (Adaptation of [Lemma 9 in Ji et al. \(2020\)](#)). *Let $\alpha, c_1, c_2 > 0$ be given. Then, there exists $\rho(\alpha) > 0$ such that $\|\mathbf{w}\|_2 > \rho(\alpha) \Rightarrow \mathcal{L}^{\mathbf{a}}((1 + \alpha)\|\mathbf{w}\|_2 \hat{\mathbf{u}}) \leq \mathcal{L}^{\mathbf{a}}(\mathbf{w})$ for any $\mathbf{a} \in [c_1, c_2]^N$.*

1512 *Proof.* Let $\hat{\mathbf{u}}$ be the ℓ_2 -max-margin solution and $\hat{\gamma} = \max_{i \in [N]} \langle \hat{\mathbf{u}}, \mathbf{x}_i \rangle$ be its margin. From the
 1513 uniform convergence in Lemma E.2, we can choose $\rho(\alpha)$ large enough so that
 1514

$$1515 \quad \|\mathbf{w}\|_2 > \rho(\alpha) \Rightarrow \left\| \frac{\bar{\mathbf{w}}^{\mathbf{a}}(\|\mathbf{w}\|_2)}{\|\mathbf{w}\|_2} - \hat{\mathbf{u}} \right\|_2 \leq \alpha \hat{\gamma},$$

1517 for any $\mathbf{a} \in [c_1, c_2]^N$. For $1 \leq i \leq n$, we get
 1518

$$1519 \quad \langle \bar{\mathbf{w}}^{\mathbf{a}}(\|\mathbf{w}\|_2), \mathbf{x}_i \rangle = \langle \bar{\mathbf{w}}^{\mathbf{a}}(\|\mathbf{w}\|_2) - \|\mathbf{w}\|_2 \hat{\mathbf{u}}, \mathbf{x}_i \rangle + \langle \|\mathbf{w}\|_2 \hat{\mathbf{u}}, \mathbf{x}_i \rangle$$

$$1520 \quad \leq \alpha \hat{\gamma} \|\mathbf{w}\|_2 + \langle \|\mathbf{w}\|_2 \hat{\mathbf{u}}, \mathbf{x}_i \rangle$$

$$1521 \quad \leq (1 + \alpha) \|\mathbf{w}\|_2 \langle \hat{\mathbf{u}}, \mathbf{x}_i \rangle.$$

1522 This implies that
 1523

$$1524 \quad \mathcal{L}^{\mathbf{a}}((1 + \alpha) \|\mathbf{w}\|_2 \hat{\mathbf{u}}) \leq \mathcal{L}^{\mathbf{a}}(\bar{\mathbf{w}}^{\mathbf{a}}(\|\mathbf{w}\|_2)) \leq \mathcal{L}^{\mathbf{a}}(\mathbf{w}),$$

1525 for any $\mathbf{a} \in [c_1, c_2]^N$. □
 1526

1527 **Theorem 3.3.** Consider Inc-Adam iterates $\{\mathbf{w}_t\}_{t=0}^{\infty}$ with $\beta_1 \leq \beta_2$ on SR data under Assumptions 2.1 to 2.3. If (a) $\mathcal{L}(\mathbf{w}_t) \rightarrow 0$ as $t \rightarrow \infty$ and (b) $\eta_t = (t + 2)^{-a}$ for $a \in (2/3, 1]$, then it
 1528 satisfies
 1529

$$1530 \quad \lim_{t \rightarrow \infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_2} = \hat{\mathbf{w}}_{\ell_2},$$

1532 where $\hat{\mathbf{w}}_{\ell_2}$ denotes the (unique) ℓ_2 -max-margin solution of SR data $\{\mathbf{x}_i\}_{i \in [N]}$.
 1533

1534 *Proof.* From Corollary 3.2, we can rewrite the update as
 1535

$$1536 \quad \mathbf{w}_{r+1}^0 - \mathbf{w}_r^0 = - \frac{\eta_{rN}}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2} \sum_{i \in [N]} a_i(r) \nabla \mathcal{L}_i(\mathbf{w}_r^0) - \eta_{rN} \boldsymbol{\epsilon}_r$$

$$1538 \quad = - \frac{\eta_{rN}}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2} \nabla \mathcal{L}^{\mathbf{a}(r)}(\mathbf{w}_r^0) - \eta_{rN} \boldsymbol{\epsilon}_r,$$

1540 where $c_1 \leq a_i(r) \leq c_2$ for some positive constants c_1, c_2 and $\lim_{r \rightarrow \infty} \boldsymbol{\epsilon}_r = \mathbf{0}$.
 1541

1542 First, we show that $\lim_{r \rightarrow \infty} \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^0\|_2} = \hat{\mathbf{w}}_{\ell_2}$. Let $\epsilon > 0$ be given. Then, we can take $\alpha = \frac{\epsilon}{1-\epsilon}$ so
 1543 that $\frac{1}{1+\alpha} = 1 - \epsilon$. Since $\|\mathbf{w}_t\|_2 \rightarrow \infty$, we can choose r_0 such that $t \geq r_0 N \implies \|\mathbf{w}_t\|_2 >$
 1544 $\max\{\rho(\alpha), 1\}$, where $\rho(\alpha)$ is given by Lemma E.3. Then for any $r \geq r_0$, we get
 1545

$$1546 \quad \langle \nabla \mathcal{L}^{\mathbf{a}}(\mathbf{w}_r^0), \mathbf{w}_r^0 - (1 + \alpha) \|\mathbf{w}_r^0\|_2 \hat{\mathbf{u}} \rangle \geq \mathcal{L}^{\mathbf{a}}(\mathbf{w}_r^0) - \mathcal{L}^{\mathbf{a}}((1 + \alpha) \|\mathbf{w}_r^0\|_2 \hat{\mathbf{u}}) \geq 0,$$

1547 which implies
 1548

$$1549 \quad \langle \nabla \mathcal{L}^{\mathbf{a}}(\mathbf{w}_r^0), \mathbf{w}_r^0 \rangle \geq (1 + \alpha) \|\mathbf{w}_r^0\|_2 \langle \nabla \mathcal{L}^{\mathbf{a}}(\mathbf{w}_r^0), \hat{\mathbf{u}} \rangle.$$

1550 Therefore, we get
 1551

$$1552 \quad \langle \mathbf{w}_{r+1}^0 - \mathbf{w}_r^0, \hat{\mathbf{u}} \rangle$$

$$1553 \quad = \left\langle -\frac{\eta_{rN}}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2} \nabla \mathcal{L}^{\mathbf{a}(r)}(\mathbf{w}_r^0), \hat{\mathbf{u}} \right\rangle + \langle -\eta_{rN} \boldsymbol{\epsilon}_r, \hat{\mathbf{u}} \rangle$$

$$1554 \quad \geq \frac{1}{(1 + \alpha) \|\mathbf{w}_r^0\|_2} \left\langle -\frac{\eta_{rN}}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2} \nabla \mathcal{L}^{\mathbf{a}(r)}(\mathbf{w}_r^0), \mathbf{w}_r^0 \right\rangle + \langle -\eta_{rN} \boldsymbol{\epsilon}_r, \hat{\mathbf{u}} \rangle$$

$$1555 \quad = \frac{1}{(1 + \alpha) \|\mathbf{w}_r^0\|_2} \langle \mathbf{w}_{r+1}^0 - \mathbf{w}_r^0, \mathbf{w}_r^0 \rangle + \frac{1}{(1 + \alpha) \|\mathbf{w}_r^0\|_2} \langle \eta_{rN} \boldsymbol{c}, \mathbf{w}_r^0 \rangle + \langle -\eta_{rN} \boldsymbol{\epsilon}_r, \hat{\mathbf{u}} \rangle$$

$$1556 \quad = \frac{1}{(1 + \alpha) \|\mathbf{w}_r^0\|_2} \left(\frac{1}{2} \|\mathbf{w}_{r+1}^0\|_2^2 - \frac{1}{2} \|\mathbf{w}_r^0\|_2^2 - \frac{1}{2} \|\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0\|_2^2 \right) + \langle -\eta_{rN} \boldsymbol{\epsilon}_r, \hat{\mathbf{u}} - \frac{\mathbf{w}_r^0}{(1 + \alpha) \|\mathbf{w}_r^0\|_2} \rangle$$

$$1557 \quad \geq \frac{1}{(1 + \alpha) \|\mathbf{w}_r^0\|_2} \left(\frac{1}{2} \|\mathbf{w}_{r+1}^0\|_2^2 - \frac{1}{2} \|\mathbf{w}_r^0\|_2^2 - \frac{1}{2} \|\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0\|_2^2 \right) - 2\eta_{rN} \|\boldsymbol{\epsilon}_r\|_2,$$

1559 where the last inequality is from $\langle \eta_{rN} \boldsymbol{\epsilon}_r, \hat{\mathbf{u}} - \frac{\mathbf{w}_r^0}{(1 + \alpha) \|\mathbf{w}_r^0\|_2} \rangle \leq \eta_{rN} \|\boldsymbol{\epsilon}_r\|_2 \left\| \hat{\mathbf{u}} - \frac{\mathbf{w}_r^0}{(1 + \alpha) \|\mathbf{w}_r^0\|_2} \right\|_2 \leq$
 1560 $2\eta_{rN} \|\boldsymbol{\epsilon}_r\|_2$.
 1561

1566 Note that

$$1567 \frac{\frac{1}{2}\|\mathbf{w}_{r+1}^0\|_2^2 - \frac{1}{2}\|\mathbf{w}_r^0\|_2^2}{\|\mathbf{w}_r^0\|_2} \geq \|\mathbf{w}_{r+1}^0\|_2 - \|\mathbf{w}_r^0\|_2.$$

1570 Furthermore,

$$1572 \frac{\|\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0\|_2^2}{2(1+\alpha)\|\mathbf{w}_r^0\|_2} \leq \frac{\|\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0\|_2^2}{2} \leq \frac{1}{2} \left(\eta_{rN}^2 \frac{\|\nabla \mathcal{L}^{\mathbf{a}(r)}(\mathbf{w}_r^0)\|_2^2}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2^2} + \eta_{rN} \|\boldsymbol{\epsilon}_r\|_2^2 \right) \\ 1574 \leq c_3 r^{-2a},$$

1576 for some $c_3 > 0$ and sufficiently large r , since $\eta_{rN} = \mathcal{O}(r^{-a})$, $\|\boldsymbol{\epsilon}_r\| = \mathcal{O}(r^{-a/2})$, and $\frac{\|\nabla \mathcal{L}^{\mathbf{a}(r)}(\mathbf{w}_r^0)\|_2^2}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2^2}$ is upper bounded from

$$1579 \frac{\|\nabla \mathcal{L}^{\mathbf{a}(r)}(\mathbf{w}_r^0)\|_2^2}{\|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2^2} \stackrel{(*)}{\leq} \frac{\left(c_2 \sqrt{d} \max_{i \in [N]} x_i \sum_{i \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_i \rangle)| \right)^2}{\left(\frac{\gamma}{N} \sum_{i \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_i \rangle)| \right)^2} = \frac{c_2^2 d N^2 (\max_{i \in [N]} x_i)^2}{\gamma^2},$$

1583 with $\gamma = \min_{i \in [N]} \langle \hat{\mathbf{w}}_{\ell_2}, \mathbf{x}_i \rangle > 0$. Note that $(*)$ is from

$$1585 \|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2^2 = \|\hat{\mathbf{w}}_{\ell_2}\|_2^2 \|\nabla \mathcal{L}(\mathbf{w}_r^0)\|_2^2 \geq \langle \hat{\mathbf{w}}_{\ell_2}, \frac{1}{N} \sum_{i \in [N]} \ell'(\langle \mathbf{w}_r^0, \mathbf{x}_i \rangle) \mathbf{x}_i \rangle^2 \geq \left(\frac{\gamma}{N} \sum_{i \in [N]} |\ell'(\langle \mathbf{w}_r^0, \mathbf{x}_i \rangle)| \right)^2.$$

1588 Therefore, we get

$$1590 \langle \mathbf{w}_r^0 - \mathbf{w}_{r_0}^0, \hat{\mathbf{u}} \rangle \geq \frac{\|\mathbf{w}_r^0\|_2 - \|\mathbf{w}_{r_0}^0\|_2}{1+\alpha} - \sum_{s=r_0}^r c_3 s^{-2a} - 2 \sum_{s=r_0}^r \eta_{sN} \|\boldsymbol{\epsilon}_s\|_2 \\ 1592 \geq (1-\epsilon)(\|\mathbf{w}_r^0\|_2 - \|\mathbf{w}_{r_0}^0\|_2) - \underbrace{\left(\sum_{s=r_0}^{\infty} c_3 s^{-2a} + \sum_{s=r_0}^{\infty} c_4 s^{-\frac{3}{2}a} \right)}_{=c_5 < \infty},$$

1598 since $\|\boldsymbol{\epsilon}_r\| = \mathcal{O}(r^{-a/2})$ and $a \in (2/3, 1]$. As a result, we can conclude that

$$1600 \langle \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^0\|_2}, \hat{\mathbf{u}} \rangle \geq \frac{(1-\epsilon)(\|\mathbf{w}_r^0\|_2 - \|\mathbf{w}_{r_0}^0\|_2) + \langle \mathbf{w}_{r_0}^0, \hat{\mathbf{u}} \rangle + c_5}{\|\mathbf{w}_r^0\|_2},$$

1603 which implies

$$1605 \liminf_{r \rightarrow \infty} \langle \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^0\|_2}, \hat{\mathbf{u}} \rangle \geq 1 - \epsilon.$$

1607 Since we choose $\epsilon > 0$ arbitrarily, we get $\lim_{r \rightarrow \infty} \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^0\|_2} = \hat{\mathbf{w}}_{\ell_2}$.

1609 Second, we claim that $\lim_{t \rightarrow \infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_2} = \hat{\mathbf{w}}_{\ell_2}$. It suffices to show that $\lim_{r \rightarrow \infty} \left\| \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^0\|_2} - \frac{\mathbf{w}_r^s}{\|\mathbf{w}_r^s\|_2} \right\|_2 = 0$ for all $s \in [N]$. Note that

$$1612 \left\| \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^0\|_2} - \frac{\mathbf{w}_r^s}{\|\mathbf{w}_r^s\|_2} \right\|_2 \leq \left\| \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^0\|_2} - \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^s\|_2} \right\|_2 + \left\| \frac{\mathbf{w}_r^0}{\|\mathbf{w}_r^s\|_2} - \frac{\mathbf{w}_r^s}{\|\mathbf{w}_r^s\|_2} \right\|_2 \\ 1614 \leq \frac{\|\mathbf{w}_r^s\|_2 - \|\mathbf{w}_r^0\|_2}{\|\mathbf{w}_r^s\|_2} + \frac{\|\mathbf{w}_r^s - \mathbf{w}_r^0\|_2}{\|\mathbf{w}_r^s\|_2} \\ 1616 \leq 2 \frac{\|\mathbf{w}_r^s - \mathbf{w}_r^0\|_2}{\|\mathbf{w}_r^s\|_2} \rightarrow 0,$$

1619 which ends the proof. \square

1620 **F MISSING PROOFS IN SECTION 4**

1621 **F.1 PROOF OF PROPOSITION 4.1**

1622 **Proposition 4.1.** *Let $\{\mathbf{w}_t\}_{t=0}^\infty$ be the iterates of Inc-Adam with $\beta_1 \leq \beta_2$. Then, under Assumptions 2.2 and 2.3, the epoch-wise update $\mathbf{w}_{r+1}^0 - \mathbf{w}_r^0$ can be expressed as*

$$1623 \mathbf{w}_{r+1}^0 - \mathbf{w}_r^0 = -\eta_{rN} \left(\sqrt{\frac{1-\beta_2^N}{1-\beta_2}} \frac{\nabla \mathcal{L}(\mathbf{w}_r^0)}{\sqrt{\sum_{i=1}^N \nabla \mathcal{L}_i(\mathbf{w}_r^0)^2}} + \epsilon_{\beta_2}(r) \right),$$

1630 where $\limsup_{r \rightarrow \infty} \|\epsilon_{\beta_2}(r)\|_\infty \leq \epsilon(\beta_2)$ and $\lim_{\beta_2 \rightarrow 1} \epsilon(\beta_2) = 0$.

1632 *Proof.* Note that

$$1633 \sum_{i \in [N]} \frac{\sum_{j \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{j \in [N]} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} = \frac{\sum_{j \in [N]} \left(\sum_{i \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k] \right)}{\sqrt{\sum_{j \in [N]} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} \\ 1638 = \frac{1-\beta_1^N}{1-\beta_1} \frac{\nabla \mathcal{L}(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{i=1}^N \nabla \mathcal{L}_i(\mathbf{w}_r^0)[k]^2}}.$$

1641 Furthermore,

$$1642 \left| \frac{\sum_{j \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} - \frac{\sum_{j \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{j \in [N]} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} \right| \\ 1647 \leq \left| \frac{\sum_{j \in [N]} \beta_1^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} \right| \left| 1 - \frac{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}}{\sqrt{\sum_{j \in [N]} \nabla \mathcal{L}_j(\mathbf{w}_r^0)[k]^2}} \right| \\ 1653 \leq \sqrt{\sum_{j \in [N]} \frac{\beta_1^{(i,j)}^2}{\beta_2^{(i,j)}}} \left(1 - \sqrt{\beta_2^{N-1}} \right) \leq \underbrace{\sqrt{\sum_{j \in [N]} \frac{1}{\beta_2^{(i,j)}}} \left(1 - \sqrt{\beta_2^{N-1}} \right)}_{\triangleq \epsilon(\beta_2)},$$

1655 where $\lim_{\beta_2 \rightarrow 1} \epsilon(\beta_2) = 0$. Substituting to Equation (2), we get

$$1656 \mathbf{w}_{r+1}^0[k] - \mathbf{w}_r^0[k] = -\eta_{rN} \left(C_{\text{inc}}(\beta_1, \beta_2) \frac{1-\beta_1^N}{1-\beta_1} \frac{\nabla \mathcal{L}(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{i=1}^N \nabla \mathcal{L}_i(\mathbf{w}_r^0)[k]^2}} + \epsilon_{\beta_2}(r)[k] \right) \\ 1662 = -\eta_{rN} \left(C_{\text{proxy}}(\beta_2) \frac{\nabla \mathcal{L}(\mathbf{w}_r^0)[k]}{\sqrt{\sum_{i=1}^N \nabla \mathcal{L}_i(\mathbf{w}_r^0)[k]^2}} + \epsilon_{\beta_2}(r)[k] \right),$$

1664 where $C_{\text{proxy}}(\beta_2) = \sqrt{\frac{1-\beta_2^N}{1-\beta_2}}$, $\limsup_{r \rightarrow \infty} \|\epsilon_{\beta_2}(r)\|_\infty \leq N\epsilon(\beta_2)$, and $\lim_{\beta_2 \rightarrow 1} \epsilon(\beta_2) = 0$. \square

1667 **F.2 PROOF OF PROPOSITION 4.3**

1669 To prove Proposition 4.3, we begin with identifying AdamProxy as normalized steepest descent
1670 with respect to an energy norm, where the inducing matrix depends on the current iterate and the
1671 dataset. The following lemma shows that the matrix is always non-degenerate; the energy norm is
1672 bounded above and below with respect to ℓ_2 -norm multiplied by two constants only depending on the
1673 dataset. This result takes a crucial role to make the convergence guarantee of AdamProxy.

1674 **Lemma F.1.** *Consider AdamProxy iterates $\{\mathbf{w}_t\}$ under Assumptions 2.1 and 2.2. Then, it satisfies*

1674 (a) $\text{Prx}(\mathbf{w}) = \arg \min_{\|\mathbf{v}\|_{\mathbf{P}(\mathbf{w})}=1} \langle \nabla \mathcal{L}(\mathbf{w}), \mathbf{v} \rangle$, where $\tilde{\mathbf{P}}(\mathbf{w}) = \text{diag} \left(\sqrt{\sum_{i \in [N]} \nabla \mathcal{L}_i(\mathbf{w})^2} \right)$ and $\mathbf{P}(\mathbf{w}) =$
 1675 $\frac{1}{\|\nabla \mathcal{L}(\mathbf{w})\|_{\tilde{\mathbf{P}}^{-1}(\mathbf{w})}^2} \tilde{\mathbf{P}}(\mathbf{w})$.
 1676
 1677

1678 (b) There exist positive constants c_1, c_2 depending only on the dataset $\{\mathbf{x}_i\}_{i \in [N]}$ such that $c_1 \|\mathbf{v}\|_2 \leq$
 1679 $\|\mathbf{v}\|_{\mathbf{P}(\mathbf{w})} \leq c_2 \|\mathbf{v}\|_2$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$.
 1680

1681 *Proof.* (a) Note that $\text{Prx}(\mathbf{w}) = -\tilde{\mathbf{P}}(\mathbf{w})^{-1} \nabla \mathcal{L}(\mathbf{w}) = \arg \min_{\mathbf{v}} \langle \nabla \mathcal{L}(\mathbf{w}), \mathbf{v} \rangle + \frac{1}{2} \|\mathbf{v}\|_{\tilde{\mathbf{P}}(\mathbf{w})}^2$. There-
 1682 fore, normalizing by $\|\nabla \mathcal{L}(\mathbf{w})\|_{\tilde{\mathbf{P}}^{-1}(\mathbf{w})}^2$, we get $\text{Prx}(\mathbf{w}) = \arg \min_{\|\mathbf{v}\|_{\mathbf{P}(\mathbf{w})}=1} \langle \nabla \mathcal{L}(\mathbf{w}), \mathbf{v} \rangle$
 1683

1684 (b) It is enough to show that every element of $\mathbf{P}(\mathbf{w})$ is bounded for some $c_1, c_2 > 0$. For simplicity,
 1685 we denote $|\ell'(\mathbf{w}^\top \mathbf{x}_i)| = r_i$, $\min_{i \in [N], j \in [d]} |\mathbf{x}_i[j]| = B_1 > 0$ and $\max_{i \in [N], j \in [d]} |\mathbf{x}_i[j]| =$
 1686 $B_2 > 0$.
 1687

1688 Note that

$$\begin{aligned} \mathbf{P}(\mathbf{w})[k, k] &= \sqrt{\sum_{i \in [N]} r_i^2 \mathbf{x}_i[k]^2} \times \frac{1}{\sum_{j \in [d]} \frac{\nabla \mathcal{L}(\mathbf{w})[j]^2}{\sqrt{\sum_{i \in [N]} r_i^2 \mathbf{x}_i[j]^2}}} \\ &\geq B_1 \sqrt{\sum_{i \in [N]} r_i^2} \times \frac{1}{\sum_{j \in [d]} \frac{(\sum_{i \in [N]} r_i B_2)^2}{\sqrt{\sum_{i \in [N]} r_i^2 B_2^2}}} \\ &= \frac{B_1^2}{B_2^2} \cdot \frac{1}{d} \frac{\sum_{i \in [N]} r_i^2}{(\sum_{i \in [N]} r_i)^2} \geq \frac{1}{Nd} \cdot \frac{B_1^2}{B_2^2}. \end{aligned}$$

1689 Let $\mathbf{v} \in \mathbb{R}^d$ s.t. $\|\mathbf{v}\|_2 = 1$ and $\mathbf{v}^\top \mathbf{x}_i > 0, \forall i \in [N]$ (since $\{\mathbf{x}_i\}$ is linearly separable). Let
 1690 $\min_{i \in [N]} \mathbf{v}^\top \mathbf{x}_i = \gamma > 0$. Then, we get $\mathbf{v}^\top \nabla \mathcal{L}(\mathbf{w}) = \sum_{i \in [N]} r_i \mathbf{v}^\top \mathbf{x}_i \geq \gamma \sum_{i \in [N]} r_i$, which
 1691 implies $\|\mathbf{v}\|_{\tilde{\mathbf{P}}(\mathbf{w})}^2 \|\nabla \mathcal{L}(\mathbf{w})\|_{\tilde{\mathbf{P}}(\mathbf{w})^{-1}}^2 \geq \langle \mathbf{v}, \nabla \mathcal{L}(\mathbf{w}) \rangle^2 \geq \gamma^2 \left(\sum_{i \in [N]} r_i \right)^2$
 1692

1693 Note that $\|\mathbf{v}\|_{\tilde{\mathbf{P}}(\mathbf{w})}^2 = \sum_{j \in [d]} \left(\sum_{i \in [N]} r_i^2 |\mathbf{x}_i[j]|^2 \cdot \mathbf{v}[j]^2 \right) \leq dB_2 \sqrt{\sum_{i \in [N]} r_i^2}$. To wrap up, we
 1694 get
 1695

$$\|\nabla \mathcal{L}(\mathbf{w})\|_{\tilde{\mathbf{P}}(\mathbf{w})^{-1}}^2 \geq \frac{\gamma^2}{dB_2} \frac{\left(\sum_{i \in [N]} r_i \right)^2}{\sqrt{\sum_{i \in [N]} r_i^2}},$$

1696 and therefore,
 1697

$$\mathbf{P}(\mathbf{w})[k, k] = \frac{\sqrt{\sum_{i \in [N]} r_i^2 \mathbf{x}_i[k]^2}}{\|\nabla \mathcal{L}(\mathbf{w})\|_{\tilde{\mathbf{P}}(\mathbf{w})^{-1}}^2} \leq \sqrt{\sum_{i \in [N]} r_i^2 \mathbf{x}_i[k]^2} \frac{dB_2}{\gamma^2} \frac{\sqrt{\sum_{i \in [N]} r_i^2}}{(\sum_{i \in [N]} r_i)^2} \leq \frac{dB_2^2}{\gamma^2}.$$

1698 As a result, we can conclude that
 1699

$$\frac{B_1^2}{dB_2^2 N} \|\mathbf{v}\| \leq \|\mathbf{v}\|_{\mathbf{P}(\mathbf{w})} \leq \frac{dB_2^2}{\gamma^2} \|\mathbf{v}\|, \quad \forall \mathbf{v}, \mathbf{w} \in \mathbb{R}^d,$$

1700 and take $c_1 = \frac{B_1^2}{dB_2^2 N}$ and $c_2 = \frac{dB_2^2}{\gamma^2}$.
 1701

□

1702 **Proposition 4.3** (Loss convergence). *Under Assumptions 2.1 and 2.2, there exists a positive constant
 1703 $\eta > 0$ depending only on the dataset $\{\mathbf{x}_i\}_{i \in [N]}$, such that if the learning rate schedule satisfies
 1704 $\eta_t \leq \eta$ and $\sum_{t=0}^{\infty} \eta_t = \infty$, then AdamProxy iterates minimize the loss, i.e., $\lim_{t \rightarrow \infty} \mathcal{L}(\mathbf{w}_t) = 0$.*

1728 *Proof.* First, we start with the descent lemma for AdamProxy, following the standard techniques in
 1729 the analysis of normalized steepest descent.

1730 Let $D = \sup_{\mathbf{w} \in \mathbb{R}^d} \max_{i \in [N]} \|\mathbf{x}_i\|_{\mathbf{P}^{-1}(\mathbf{w})}$. Notice that $D \leq c_2 \max_{i \in [N]} \|\mathbf{x}_i\|_2 < \infty$ by Lemma F.1.
 1731 Also, we define

$$1733 \quad \gamma_{\mathbf{w}} = \max_{\|\mathbf{v}\|_{\mathbf{P}(\mathbf{w})} \leq 1} \min_{i \in [N]} \mathbf{v}^\top \mathbf{x}_i$$

1735 be the $\|\cdot\|_{\mathbf{P}(\mathbf{w})}$ -max-margin. Also notice that $\bar{\gamma} \triangleq \sup_{\mathbf{w} \in \mathbb{R}^d} \gamma_{\mathbf{w}} < \infty$, since

$$1737 \quad \max_{\|\mathbf{v}\|_{\mathbf{P}(\mathbf{w})} \leq 1} \min_{i \in [N]} \mathbf{v}^\top \mathbf{x}_i \leq \max_{\|\mathbf{v}\|_2 \leq \frac{1}{c_1}} \min_{i \in [N]} \mathbf{v}^\top \mathbf{x}_i$$

1739 for any $\mathbf{w} \in \mathbb{R}^d$ by Lemma F.1. Then, we get

$$\begin{aligned} 1740 \quad \mathcal{L}(\mathbf{w}_{t+1}) &= \mathcal{L}(\mathbf{w}_t) + \eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \text{Prx}(\mathbf{w}_t) \rangle + \frac{\eta_t^2}{2} \text{Prx}(\mathbf{w}_t)^\top \nabla^2 \mathcal{L}(\mathbf{w}_t + \beta(\mathbf{w}_{t+1} - \mathbf{w}_t)) \text{Prx}(\mathbf{w}_t) \\ 1741 \quad &\stackrel{(*)}{\leq} \mathcal{L}(\mathbf{w}_t) - \eta_t \|\nabla \mathcal{L}(\mathbf{w}_t)\|_{\mathbf{P}^{-1}(\mathbf{w}_t)} + \frac{\eta_t^2 D^2}{2} \sup\{\mathcal{G}(\mathbf{w}_t), \mathcal{G}(\mathbf{w}_{t+1})\} \\ 1742 \quad &\stackrel{(**)}{\leq} \mathcal{L}(\mathbf{w}_t) - \eta_t \|\nabla \mathcal{L}(\mathbf{w}_t)\|_{\mathbf{P}^{-1}(\mathbf{w}_t)} + \frac{\eta_t^2 D^2 e^{\eta_0 D}}{2} \mathcal{G}(\mathbf{w}_t) \\ 1743 \quad &\stackrel{(***)}{\leq} \mathcal{L}(\mathbf{w}_t) - \left(\eta_t - \frac{\eta_t^2 D^2 e^{\eta_0 D}}{2} \gamma_{\mathbf{w}_t} \right) \|\nabla \mathcal{L}(\mathbf{w}_t)\|_{\mathbf{P}^{-1}(\mathbf{w}_t)} \\ 1744 \quad &\leq \mathcal{L}(\mathbf{w}_t) - \frac{\eta_t}{2} \|\nabla \mathcal{L}(\mathbf{w}_t)\|_{\mathbf{P}^{-1}(\mathbf{w}_t)}, \end{aligned}$$

1749 for $\eta_t \leq \frac{1}{\bar{\gamma} D^2 e^{\eta_0 D}} \triangleq \eta$. Note that $(*)$ is from
 1750

$$\begin{aligned} 1753 \quad \text{Prx}(\mathbf{w}_t)^\top \nabla^2 \mathcal{L}(\mathbf{w}) \text{Prx}(\mathbf{w}_t) &= \frac{1}{N} \sum_{i \in [N]} \ell''(\mathbf{w})(\text{Prx}(\mathbf{w}_t)^\top \mathbf{x}_i)^2 \\ 1754 \quad &\leq \frac{1}{N} \sum_{i \in [N]} \ell''(\mathbf{w}) \|\text{Prx}(\mathbf{w}_t)\|_\infty^2 \|\mathbf{x}_i\|_1^2 \leq D^2 \mathcal{G}(\mathbf{w}), \end{aligned}$$

1755 where the last inequality is from Lemma I.1, and $(**)$, $(***)$ are also from Lemma I.1. Telescoping
 1756 this inequality, we get

$$1761 \quad \frac{1}{2} \sum_{t=t_0}^T \eta_t \|\nabla \mathcal{L}(\mathbf{w}_t)\|_{\mathbf{P}^{-1}(\mathbf{w}_t)} \leq \mathcal{L}(\mathbf{w}_{t_0}) - \mathcal{L}(\mathbf{w}_T) \leq \mathcal{L}(\mathbf{w}_{t_0}),$$

1764 which implies $\sum_{t=t_0}^\infty \eta_t \|\nabla \mathcal{L}(\mathbf{w}_t)\|_{\mathbf{P}^{-1}(\mathbf{w}_t)} < \infty$. Since $\sum_{t=t_0}^T \eta_t = \infty$, we get
 1765 $\liminf_{t \rightarrow \infty} \|\nabla \mathcal{L}(\mathbf{w}_t)\|_{\mathbf{P}^{-1}(\mathbf{w}_t)} = 0$. From Lemma F.1, we get $\liminf_{t \rightarrow \infty} \|\nabla \mathcal{L}(\mathbf{w}_t)\|_2 = 0$, also
 1766 implying $\liminf_{t \rightarrow \infty} \mathcal{L}(\mathbf{w}_t) = 0$. Since $\mathcal{L}(\mathbf{w}_t)$ is monotonically decreasing, we get $\mathcal{L}(\mathbf{w}_t) \rightarrow 0$. \square

1768 F.3 PROOF OF LEMMA 4.5

1769 **Intuition.** Before we provide a rigorous proof of Lemma 4.5, we first demonstrate its intuitive
 1770 explanation motivated by Soudry et al. (2018). For simplicity, assume $\ell = \ell_{\text{exp}}$ and let $\mathbf{w}_t =$
 1771 $g(t)\hat{\mathbf{w}} + \boldsymbol{\rho}(t)$ where $g(t) = \|\mathbf{w}_t\|_2 \rightarrow \infty$, $\boldsymbol{\rho}(t) \in \mathbb{R}^d$, and $\frac{1}{g(t)}\boldsymbol{\rho}(t) \rightarrow 0$. Then, the mini-batch
 1772 gradient can be represented by

$$1774 \quad \nabla \mathcal{L}_i(\mathbf{w}) = -\exp(-\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i = -\exp(-g(t)\hat{\mathbf{w}}^\top \mathbf{x}_i) \exp(-\boldsymbol{\rho}(t)^\top \mathbf{x}_i) \mathbf{x}_i.$$

1775 As $g(t) \rightarrow \infty$, the coefficient exponentially decays to 0. It implies that only terms with the smallest
 1776 $\hat{\mathbf{w}}^\top \mathbf{x}_i$ will contribute to the update of AdamProxy. Therefore, the limit direction $\hat{\mathbf{w}}$ will be described
 1777 by $\frac{\sum_{i \in [N]} c_i \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} c_i^2 \mathbf{x}_i^2}}$ where c_i is the contribution of the i -th sample to the update and it vanishes for
 1778 $i \notin S$ where $S = \arg \min_{i \in [N]} \hat{\mathbf{w}}^\top \mathbf{x}_i$.

1781 Building upon this intuition, we first establish the following technical lemma, characterizing limit
 1782 points of a sequence in a form of AdamProxy.

1782 **Lemma F.2.** Let $(\mathbf{a}(t))_{t \geq 0}$ be a sequence of real vectors in $\mathbb{R}_{>0}^N$ and $\{\mathbf{x}_i\}_{i \in S} \subseteq \mathbb{R}^d$ be the dataset
 1783 with nonzero entries for an index set $S \subseteq [N]$. Suppose that $\mathbf{b}_t = \frac{\sum_{i \in S} a_i(t) \mathbf{x}_i}{\sqrt{\sum_{i \in S} a_i(t)^2 \mathbf{x}_i^2}}$ satisfies $\|\mathbf{b}_t\|_2 \geq$
 1784 $C > 0$ for all $t \geq 0$. Then every limit point of $\frac{\mathbf{b}_t}{\|\mathbf{b}_t\|_2}$ is positively proportional to $\frac{\sum_{i \in [N]} c_i \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} c_i^2 \mathbf{x}_i^2}}$ for
 1785 some $\mathbf{c} \in \Delta^{N-1}$ satisfying $c_i = 0$ for $i \notin S$.
 1786

1787 *Proof.* Define a function $F : \Delta^{|S|-1} \rightarrow \mathbb{R}^d$ as

$$1788 F(\mathbf{d}) = \frac{\sum_{i \in S} d_i \mathbf{x}_i}{\sqrt{\sum_{i \in S} d_i^2 \mathbf{x}_i^2}}.$$

1790 Since $\{\mathbf{x}_i\}_{i \in S}$ has nonzero entries, F is continuous. Let $A = \{\mathbf{d} \in \Delta^{|S|-1} : \|F(\mathbf{d})\|_2 \geq C\}$.
 1791 Since F is continuous, A is a closed subset of $\Delta^{|S|-1}$. Furthermore, since $\|\delta_t\|_2 \geq C$ for all $t \geq 0$,
 1792 $\{\mathbf{a}(t)\}_{t \geq 0} \subseteq A$.
 1793

1794 Now let $\hat{\delta}$ be a limit point of $\frac{\delta_t}{\|\delta_t\|_2}$. Define a function $G : A \subseteq \Delta^{|S|-1} \rightarrow \mathbb{R}^d$ as
 1795

$$1796 G(\mathbf{d}) = \frac{1}{\left\| \frac{\sum_{i \in S} d_i \mathbf{x}_i}{\sqrt{\sum_{i \in S} d_i^2 \mathbf{x}_i^2}} \right\|_2} \cdot \frac{\sum_{i \in S} d_i \mathbf{x}_i}{\sqrt{\sum_{i \in S} d_i^2 \mathbf{x}_i^2}}.$$

1803 Notice that G is continuous on A and $\hat{\delta} = \lim_{t \rightarrow \infty} G(\mathbf{a}(t))$. Since A is bounded and closed, Bolzano-
 1804 Weierstrass Theorem tells us that there exists a subsequence $\mathbf{a}(t_n)$ such that $\exists \lim_{n \rightarrow \infty} \mathbf{a}(t_n) = \mathbf{c} \in$
 1805 A . Therefore, we get

$$1806 \hat{\delta} = \lim_{n \rightarrow \infty} G(\mathbf{a}(t_n)) = G(\lim_{n \rightarrow \infty} \mathbf{a}(t_n)) = G(\mathbf{c}).$$

1807 Hence, the limit point $\hat{\delta}$ is proportional to $\frac{\sum_{i \in S} c_i \mathbf{x}_i}{\sqrt{\sum_{i \in S} c_i^2 \mathbf{x}_i^2}}$. Then we regard $\mathbf{c} \in \Delta^{N-1}$ by taking $c_i = 0$
 1808 for $i \notin S$. \square

1809 **Lemma 4.5.** Under Assumptions 2.1, 2.2 and 4.4, there exists $\mathbf{c} = (c_0, \dots, c_{N-1}) \in \Delta^{N-1}$ such
 1810 that the limit direction $\hat{\mathbf{w}}$ of AdamProxy satisfies

$$1811 \hat{\mathbf{w}} \propto \frac{\sum_{i \in [N]} c_i \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} c_i^2 \mathbf{x}_i^2}}, \quad (5)$$

1812 and $c_i = 0$ for $i \notin S$, where $S = \arg \min_{i \in [N]} \hat{\mathbf{w}}^\top \mathbf{x}_i$ is the index set of support vectors of $\hat{\mathbf{w}}$.
 1813

1814 *Proof.* We start with the case of $\ell = \ell_{\text{exp}}$. First step is to characterize $\hat{\delta}$, the limit direction of δ_t . To
 1815 begin with, we introduce some new notations.
 1816

- 1817 · From Assumption 4.4, let $\mathbf{w}_t = g(t)\hat{\mathbf{w}} + \boldsymbol{\rho}(t)$ where $g(t) = \|\mathbf{w}_t\|_2 \rightarrow \infty$, $\boldsymbol{\rho}(t) \in \mathbb{R}^d$, and
 $\frac{1}{g(t)} \boldsymbol{\rho}(t) \rightarrow \mathbf{0}$.
- 1818 · Let $\gamma = \min_i \langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle$, $\bar{\gamma}_i = \langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle$, $\bar{\gamma} = \min_{i \notin S} \langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle$. Then it satisfies $S = \{i \in [N] : \langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle = \gamma\}$. Here, note that $\bar{\gamma} > \gamma > 0$.
- 1819 · Let $\boldsymbol{\alpha}(t) \in \mathbb{R}^N$ be $\alpha_i(t) = \exp(-\boldsymbol{\rho}(t)^\top \mathbf{x}_i)$.
- 1820 · Let $B_0 = \max_i \|\mathbf{x}_i\|_2$, $B_1 = \min_{i \in [N], j \in [d]} |\mathbf{x}_i[j]| > 0$, and $B_2 = \max_{i \in [N], j \in [d]} |\mathbf{x}_i[j]|$.

1821 Since $\|\boldsymbol{\rho}(t)\|/g(t) \rightarrow 0$ and $\gamma, \bar{\gamma} > 0$, there exist $t_{\epsilon_1}, t_{\epsilon_2} > 0$ such that

$$1822 \boldsymbol{\rho}(t)^\top \mathbf{x}_i \leq \|\boldsymbol{\rho}(t)\|_2 B_0 \leq \epsilon_1 \gamma g(t), \quad \forall t > t_{\epsilon_1}, \forall i \in [N],$$

$$1823 \boldsymbol{\rho}(t)^\top \mathbf{x}_i \geq -\|\boldsymbol{\rho}(t)\|_2 B_0 \geq -\epsilon_2 \bar{\gamma} g(t), \quad \forall t > t_{\epsilon_2}, \forall i \in [N],$$

1836 for all $\epsilon_1, \epsilon_2 > 0$. Then, we can decompose dominant and residual terms in the update rule.

$$\begin{aligned} 1838 \quad \delta_t &= \frac{\sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2}} + \frac{\sum_{i \in S^c} \exp(-\bar{\gamma}_i g(t)) \exp(-\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2}} \\ 1839 \quad &\triangleq \mathbf{d}(t) + \mathbf{r}(t). \end{aligned}$$

1840 To investigate the limit direction of δ_t , we first show that $\mathbf{d}(t)$ dominates $\mathbf{r}(t)$, i.e., $\lim_{t \rightarrow \infty} \frac{\|\mathbf{r}(t)\|_2}{\|\mathbf{d}(t)\|_2} = 0$. Let $\mathbf{M}_t = \text{diag} \left(\sqrt{\sum_{i \in [N]} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2} \right)$. Notice that

$$1847 \quad \|\mathbf{M}_t \hat{\mathbf{w}}\|_2 \|\mathbf{d}(t)\|_2 \geq \langle \mathbf{M}_t \hat{\mathbf{w}}, \mathbf{d}(t) \rangle = \gamma \sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i).$$

1849 Since the diagonals of \mathbf{M}_t are upper bounded by $B_2 \sqrt{\sum_{i \in [N]} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i)}$, we
1850 get
1851

$$1852 \quad \|\mathbf{d}(t)\|_2 \geq \frac{\gamma \sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i)}{B_2 \sqrt{\sum_{i \in [N]} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i)}}.$$

1855 Also, notice that
1856

$$1857 \quad \|\mathbf{r}(t)\|_2 \leq \frac{B_2 \sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i)}{B_1 \sqrt{\sum_{i \in [N]} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i)}}.$$

1860 From the following inequalities

$$\begin{aligned} 1863 \quad \sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i) &\geq \exp(-\gamma g(t)) \exp(-\epsilon_1 \gamma g(t)) \\ 1865 \quad &= \exp(-(1 + \epsilon_1) \gamma g(t)), \end{aligned}$$

$$\begin{aligned} 1867 \quad \sum_{i \in S^c} \exp(-\bar{\gamma}_i g(t)) \exp(-\rho(t)^\top \mathbf{x}_i) &\leq N \exp(-\bar{\gamma} g(t)) \exp(\epsilon_2 \bar{\gamma} g(t)) \\ 1869 \quad &= N \exp(-(1 - \epsilon_2) \bar{\gamma} g(t)), \end{aligned}$$

1871 we conclude that

$$\begin{aligned} 1873 \quad \frac{\|\mathbf{r}(t)\|_2}{\|\mathbf{d}(t)\|_2} &= \frac{B_2^2}{\gamma B_1} \frac{\sum_{i \in S^c} \exp(-\bar{\gamma}_i g(t)) \exp(-\rho(t)^\top \mathbf{x}_i)}{\sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i)} \\ 1875 \quad &\leq \frac{N B_2^2}{\gamma B_1} \exp\left(-\frac{1}{2}(\bar{\gamma} - \gamma) g(t)\right) \rightarrow 0. \end{aligned}$$

1878 Next, we claim that every limit point of $\frac{\mathbf{d}(t)}{\|\mathbf{d}(t)\|_2}$ is positively proportional to $\frac{\sum_{i \in [N]} c_i \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} c_i^2 \mathbf{x}_i^2}}$ for some
1879 $\mathbf{c} = (c_0, \dots, c_{N-1}) \in \Delta^{N-1}$ satisfying $c_i = 0$ for $i \notin S$. Notice that

$$\begin{aligned} 1881 \quad \mathbf{d}(t)[k] &= \frac{\sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i[k]}{\sqrt{\sum_{i \in [N]} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2[k]}} \\ 1883 \quad &= \frac{\sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i[k]}{\sqrt{\sum_{i \in S} \exp(-2\gamma g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2[k] + \sum_{i \in S^c} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2[k]}} \\ 1885 \quad &= \frac{\sum_{i \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i[k]}{\sqrt{\sum_{i \in S} \exp(-2\gamma g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2[k]}} \frac{1}{\sqrt{1 + \frac{\sum_{i \in S^c} \exp(-2\bar{\gamma}_i g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2[k]}{\sum_{i \in S} \exp(-2\gamma g(t)) \exp(-2\rho(t)^\top \mathbf{x}_i) \mathbf{x}_i^2[k]}}}. \end{aligned}$$

1890 Let $\mathbf{b}_t = \frac{\sum_{i \in S} \exp(-\gamma g(t)) \exp(-\boldsymbol{\rho}(t)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{i \in S} \exp(-2\gamma g(t)) \exp(-2\boldsymbol{\rho}(t)^\top \mathbf{x}_i) \mathbf{x}_i^2}} = \frac{\sum_{i \in S} \exp(-\boldsymbol{\rho}(t)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{i \in S} \exp(-2\boldsymbol{\rho}(t)^\top \mathbf{x}_i) \mathbf{x}_i^2}}.$ Since
 1891
 1892
 1893
 1894
 1895 every limit point of $\frac{\mathbf{d}(t)}{\|\mathbf{d}(t)\|_2}$ is represented by a limit point of $\frac{\mathbf{b}_t}{\|\mathbf{b}_t\|_2}.$ Notice that \mathbf{b}_t is an update
 1896 of AdamProxy under the dataset $\{\mathbf{x}_i\}_{i \in S},$ which implies $\|\mathbf{b}_t\|_2$ is lower bounded by a positive
 1897 constant from Lemma F.1. Therefore, Lemma F.2 proves the claim.
 1898

1899 Hence, we can characterize $\hat{\boldsymbol{\delta}}$ as

$$\begin{aligned} 1900 \hat{\boldsymbol{\delta}} &= \lim_{t \rightarrow \infty} \frac{\boldsymbol{\delta}_t}{\|\boldsymbol{\delta}_t\|_2} = \lim_{t \rightarrow \infty} \frac{\mathbf{d}(t) + \mathbf{r}(t)}{\|\mathbf{d}(t) + \mathbf{r}(t)\|_2} \\ 1901 &= \lim_{t \rightarrow \infty} \frac{\mathbf{d}(t)}{\|\mathbf{d}(t) + \mathbf{r}(t)\|_2} + \lim_{t \rightarrow \infty} \frac{\mathbf{r}(t)}{\|\mathbf{d}(t) + \mathbf{r}(t)\|_2} \\ 1902 &= \lim_{t \rightarrow \infty} \frac{\mathbf{d}(t)}{\|\mathbf{d}(t)\|_2} \propto \frac{\sum_{i \in [N]} c_i \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} c_i^2 \mathbf{x}_i^2}}, \\ 1903 & \\ 1904 & \\ 1905 & \\ 1906 & \\ 1907 \end{aligned}$$

1908 for some $\mathbf{c} \in \Delta^{N-1}$ satisfying $c_i = 0$ for $i \notin S.$

1909 Second step is to connect the limiting behavior of $\boldsymbol{\delta}_t$ to the limit direction $\hat{\mathbf{w}}$ using Stolz-Cesaro
 1910 theorem. From the first step, we can represent

$$1911 \boldsymbol{\delta}_t = h(t) \hat{\boldsymbol{\delta}} + \boldsymbol{\sigma}(t),$$

1912 where $h(t) = \|\boldsymbol{\delta}_t\|_2$ and $\frac{1}{h(t)} \boldsymbol{\sigma}(t) \rightarrow 0.$ Notice that $\mathbf{w}_t - \mathbf{w}_0 = \sum_{s=0}^{t-1} \eta_s h(s) (\hat{\boldsymbol{\delta}} + \frac{1}{h(s)} \boldsymbol{\sigma}(t)).$ Since
 1913 $\hat{\boldsymbol{\delta}} + \frac{1}{h(s)} \boldsymbol{\sigma}(t)$ is bounded, we get $\sum_{s=0}^{t-1} \eta_s h(s) \rightarrow \infty.$ Then we take
 1914

$$\begin{aligned} 1915 \mathbf{a}_t &= \mathbf{w}_t - \mathbf{w}_0 = \sum_{s=0}^{t-1} \eta_s h(s) (\hat{\boldsymbol{\delta}} + \frac{1}{h(s)} \boldsymbol{\sigma}(t)) \\ 1916 & \\ 1917 & \\ 1918 & \\ 1919 b_t &= \sum_{s=0}^{t-1} \eta_s h(s). \\ 1920 & \\ 1921 & \end{aligned}$$

1922 Then, $\{b_t\}_{t=1}^\infty$ is strictly monotone and diverging. Also, $\lim_{t \rightarrow \infty} \frac{\mathbf{a}_{t+1} - \mathbf{a}_t}{b_{t+1} - b_t} = \hat{\boldsymbol{\delta}}.$ Then, by Stolz-
 1923 Cesaro theorem, we get

$$1924 \lim_{t \rightarrow \infty} \frac{\mathbf{a}_t}{b_t} = \hat{\boldsymbol{\delta}}.$$

1925 This implies $\mathbf{w}_t = b_t \hat{\boldsymbol{\delta}} + \boldsymbol{\tau}(t)$ where $\frac{\boldsymbol{\tau}(t)}{b_t} \rightarrow 0.$ Also notice that $\mathbf{w}_t = g(t) \hat{\mathbf{w}} + \boldsymbol{\rho}(t).$ Dividing by
 1926 $g(t),$ we get

$$1927 \hat{\mathbf{w}} = \lim_{t \rightarrow \infty} \frac{g(t) \hat{\mathbf{w}} + \boldsymbol{\rho}(t)}{g(t)} = \lim_{t \rightarrow \infty} \frac{b_t}{g(t)} \left(\hat{\boldsymbol{\delta}} + \frac{\boldsymbol{\tau}(t)}{b_t} \right).$$

1928 Since ℓ_2 norm is continuous, we get

$$1929 1 = \|\hat{\mathbf{w}}\|_2 = \lim_{t \rightarrow \infty} \frac{b_t}{g(t)} \left\| \hat{\boldsymbol{\delta}} + \frac{\boldsymbol{\tau}(t)}{b_t} \right\|_2 = \lim_{t \rightarrow \infty} \frac{b_t}{g(t)},$$

1930 which implies $\hat{\mathbf{w}} = \hat{\boldsymbol{\delta}}.$

1931 Then we move on to the case of $\ell = \ell_{\log}.$ This kind of extension is possible since the logistic loss has
 1932 a similar tail behavior of the exponential loss, following the line of Soudry et al. (2018). We adopt
 1933 the same notation with previous part, and we decompose dominant and residual terms as follows:

$$\begin{aligned} 1934 \boldsymbol{\delta}_t &= \frac{\sum_{i \in S} |\ell'(\gamma g(t) + \boldsymbol{\rho}(t)^\top \mathbf{x}_i)| \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} |\ell'(\gamma_i g(t) + \boldsymbol{\rho}(t)^\top \mathbf{x}_i)|^2 \mathbf{x}_i^2}} + \frac{\sum_{i \in S} |\ell'(\bar{\gamma}_i g(t) + \boldsymbol{\rho}(t)^\top \mathbf{x}_i)| \mathbf{x}_i}{\sqrt{\sum_{i \in [N]} |\ell'(\bar{\gamma}_i g(t) + \boldsymbol{\rho}(t)^\top \mathbf{x}_i)|^2 \mathbf{x}_i^2}} \\ 1935 & \\ 1936 & \triangleq \mathbf{d}(t) + \mathbf{r}(t). \end{aligned}$$

1944 Notice that $\lim_{z \rightarrow \infty} \frac{|\ell_{\log}(z)|}{|\ell'_{\exp}(z)|} = \lim_{z \rightarrow \infty} \frac{1}{1+e^{-z}} = 1$. Therefore, the limit behavior of $\mathbf{d}(t)$ and $\mathbf{r}(t)$ is
 1945 identical to the previous $\ell = \ell_{\exp}$ case. This implies the same proof also holds for the logistic loss,
 1946 which ends the proof. \square
 1947

1948 **F.4 PROOF OF THEOREM 4.8**
 1949

1950 **Theorem 4.8.** *Under Assumptions 2.1 and 4.7, $P_{\text{Adam}}(\mathbf{c})$ admits unique primal and dual solutions,
 1951 so that $\mathbf{p}(\mathbf{c})$ and $\mathbf{d}(\mathbf{c})$ can be regarded as vector-valued functions. Moreover, under Assumptions 2.1,
 1952 2.2, 4.4 and 4.7, the following hold:*

1953 (a) $\mathbf{p} : \Delta^{N-1} \rightarrow \mathbb{R}^d$ is continuous.
 1954 (b) $\mathbf{d} : \Delta^{N-1} \rightarrow \mathbb{R}_{\geq 0}^N \setminus \{\mathbf{0}\}$ is continuous. Consequently, the map $T(\mathbf{c}) \triangleq \frac{\mathbf{d}(\mathbf{c})}{\|\mathbf{d}(\mathbf{c})\|_1}$ is continuous.
 1955 (c) The map $T : \Delta^{N-1} \rightarrow \Delta^{N-1}$ admits at least one fixed point.
 1956 (d) There exists $\mathbf{c}^* \in \{\mathbf{c} \in \Delta^{N-1} : T(\mathbf{c}) = \mathbf{c}\}$ such that the convergence direction $\hat{\mathbf{w}}$ of
 1957 AdamProxy is proportional to $\mathbf{p}(\mathbf{c}^*)$.
 1958

1959 *Proof.* We first show that $P_{\text{Adam}}(\mathbf{c})$ has a unique solution and $\mathbf{p}(\mathbf{c})$ can be identified as a vector-valued
 1960 function. Since $\mathbf{M}(\mathbf{c})$ is positive definite for every $\mathbf{c} \in \Delta^{N-1}$, $\frac{1}{2}\|\mathbf{w}\|_{\mathbf{M}(\mathbf{c})}$ is strictly convex. Since
 1961 the feasible set is convex, there exists a unique optimal solution of $P_{\text{Adam}}(\mathbf{c})$ and we can redefine
 1962 $\mathbf{p}(\mathbf{c})$ as a vector-valued function.

1963 Since the inequality constraints are linear, $P_{\text{Adam}}(\mathbf{c})$ satisfies Slater's condition, which implies that
 1964 there exists a dual solution. From Assumption 4.7, such dual solution is unique.
 1965

1966 (a) Let $f(\mathbf{w}, \mathbf{c}) = \frac{1}{2}\|\mathbf{w}\|_{\mathbf{M}(\mathbf{c})}$ be the objective function of $P_{\text{Adam}}(\mathbf{c})$ and $F = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w}^\top \mathbf{x}_i - 1 \geq 0, \forall i \in [N]\}$ be the feasible set. It is clear that such f is continuous on \mathbf{w} and \mathbf{c} . Let
 1967 $\bar{\mathbf{c}} \in \Delta^{N-1}$ and assume \mathbf{p} is not continuous on $\bar{\mathbf{c}}$. Then there exists $\{\mathbf{c}_k\} \subset \Delta^{N-1}$ such that
 1968 $\lim_{k \rightarrow \infty} \mathbf{c}_k = \bar{\mathbf{c}}$ but $\|\mathbf{p}(\mathbf{c}_k) - \mathbf{p}(\bar{\mathbf{c}})\|_2 \geq \epsilon$ for some $\epsilon > 0$. We denote $\mathbf{w}_k = \mathbf{p}(\mathbf{c}_k)$ and
 1969 $\bar{\mathbf{w}} = \mathbf{p}(\bar{\mathbf{c}})$.
 1970

1971 First, construct $\{\mathbf{u}_k\} \subset F$ such that $\lim_{k \rightarrow \infty} \mathbf{u}_k = \bar{\mathbf{w}}$. Then we get a natural relationship
 1972 between \mathbf{w}_k and \mathbf{u}_k as
 1973

$$\frac{1}{2}\mathbf{w}_k^\top \mathbf{M}(\mathbf{c}_k) \mathbf{w}_k \leq \frac{1}{2}\mathbf{u}_k^\top \mathbf{M}(\mathbf{c}_k) \mathbf{u}_k.$$

1974 Second, consider the case when $\{\mathbf{w}_k\}$ is bounded. Then we can take a subsequence $\mathbf{w}_{k_n} \rightarrow \mathbf{w}_0$.
 1975 Since $\{\mathbf{w}_{k_n}\} \subset F$ and F is closed, we get $\mathbf{w}_0 \in F$. Also, since f is continuous, $f(\mathbf{w}_{k_n}, \mathbf{c}_{k_n}) \rightarrow f(\mathbf{w}_0, \bar{\mathbf{c}})$. Therefore,
 1976

$$f(\mathbf{w}_{k_n}, \mathbf{c}_{k_n}) \leq f(\bar{\mathbf{w}}, \mathbf{c}_{k_n}) \xrightarrow{n \rightarrow \infty} f(\mathbf{w}_0, \bar{\mathbf{c}}) \leq f(\bar{\mathbf{w}}, \bar{\mathbf{c}}),$$

1977 which implies $\mathbf{w}_0 = \bar{\mathbf{w}}$. This makes a contradiction to $\|\mathbf{p}(\mathbf{c}_k) - \mathbf{p}(\bar{\mathbf{c}})\|_2 = \|\mathbf{w}_k - \bar{\mathbf{w}}\|_2 \geq \epsilon$.
 1978

1979 Lastly, consider the case when $\{\mathbf{w}_k\}$ is not bounded. By taking a subsequence, we can assume
 1980 that $\|\mathbf{w}_k\|_2 \rightarrow \infty$ without loss of generality. Define $\mathbf{v}_k = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2}$. Since \mathbf{v}_k is bounded, we can
 1981 take a convergent subsequence and consider $\lim_{k \rightarrow \infty} \mathbf{v}_k = \bar{\mathbf{v}}$ without loss of generality. Then,
 1982

$$\frac{1}{2}\mathbf{w}_k^\top \mathbf{M}(\mathbf{c}_k) \mathbf{w}_k \leq \frac{1}{2}\mathbf{u}_k^\top \mathbf{M}(\mathbf{c}_k) \mathbf{u}_k \Rightarrow \frac{1}{2}\mathbf{v}_k^\top \mathbf{M}(\mathbf{c}_k) \mathbf{v}_k \leq \frac{1}{2} \left(\frac{\mathbf{u}_k}{\|\mathbf{w}_k\|_2} \right)^\top \mathbf{M}(\mathbf{c}_k) \left(\frac{\mathbf{u}_k}{\|\mathbf{w}_k\|_2} \right).$$

1983 Since f is continuous and $\{\mathbf{u}_k\}$ is bounded, we get
 1984

$$\begin{aligned} \frac{1}{2}\bar{\mathbf{v}}^\top \mathbf{M}(\bar{\mathbf{c}}) \bar{\mathbf{v}} &= f(\bar{\mathbf{v}}, \bar{\mathbf{c}}) = \lim_{k \rightarrow \infty} f(\mathbf{v}_k, \mathbf{c}_k) = \lim_{k \rightarrow \infty} \frac{1}{2}\mathbf{v}_k^\top \mathbf{M}(\mathbf{c}_k) \mathbf{v}_k \\ &\leq \limsup_{k \rightarrow \infty} \frac{1}{2} \left(\frac{\mathbf{u}_k}{\|\mathbf{w}_k\|_2} \right)^\top \mathbf{M}(\mathbf{c}_k) \left(\frac{\mathbf{u}_k}{\|\mathbf{w}_k\|_2} \right) = 0. \end{aligned}$$

1985 Note that $\mathbf{M}(\bar{\mathbf{c}})$ is positive definite and $\frac{1}{2}\bar{\mathbf{v}}^\top \mathbf{M}(\bar{\mathbf{c}}) \bar{\mathbf{v}} = 0$ implies $\bar{\mathbf{v}} = 0$, which makes a
 1986 contradiction.
 1987

1998 (b) Let $\mathbf{c}_0 \in \Delta^{N-1}$ be given and take $\mathbf{w}^* = \mathbf{p}(\mathbf{c}_0)$. From KKT conditions of $P_{\text{Adam}}(\mathbf{c}_0)$, the dual
 1999 solution $\mathbf{d}(\mathbf{c}_0)$ is given by
 2000

2001
$$\mathbf{M}(\mathbf{c}_0)\mathbf{w}^* = \sum_{i \in S(\mathbf{w}^*)} d_i(\mathbf{c}_0) \mathbf{x}_i$$

 2002
 2003

2004 and such $d_i(\mathbf{c}_0) \geq 0$ is uniquely determined since $\{\mathbf{x}_i\}_{i \in S(\mathbf{w}^*)}$ is a set of linearly independent
 2005 vectors by Assumption 4.7.

2006 Now we claim that $\mathbf{d}(\mathbf{c})$ is continuous at $\mathbf{c} = \mathbf{c}_0$. Notice that $\min_{i \notin S(\mathbf{w}^*)} \mathbf{w}^{*\top} \mathbf{x}_i > 1$. Since
 2007 \mathbf{p} is continuous at \mathbf{c}_0 , there exists $\delta > 0$ such that $\mathbf{p}(\mathbf{c})^\top \mathbf{x}_i - 1 > 0$ for $i \notin S(\mathbf{w}^*)$ and
 2008 $\mathbf{c} \in \Delta^{N-1} \cap B_\delta(\mathbf{c}_0)$. Therefore, $S(\mathbf{p}(\mathbf{c})) \subseteq S(\mathbf{w}^*)$ on $\mathbf{c} \in \Delta^{N-1} \cap B_\delta(\mathbf{c}_0)$.

2009 Let \mathbf{X} be a matrix whose columns are the support vectors of \mathbf{w}^* . On $\mathbf{c} \in \Delta^{N-1} \cap B_\delta(\mathbf{c}_0)$, KKT
 2010 conditions tells us that
 2011

2012
$$\begin{aligned} \mathbf{M}(\mathbf{c})\mathbf{p}(\mathbf{c}) &= \sum_{i \in S(\mathbf{p}(\mathbf{c}))} d_i(\mathbf{c}) \mathbf{x}_i \stackrel{(*)}{=} \sum_{i \in S(\mathbf{w}^*)} d_i(\mathbf{c}) \mathbf{x}_i = \mathbf{X}\mathbf{d}(\mathbf{c}) \\ &\stackrel{(**)}{\Leftrightarrow} \mathbf{d}(\mathbf{c}) = (\mathbf{X}^\top|_{\text{im } \mathbf{X}^\top})^{-1} \mathbf{M}(\mathbf{c})\mathbf{p}(\mathbf{c}), \end{aligned}$$

 2013
 2014
 2015
 2016

2017 where $(*)$ is from $S(\mathbf{p}(\mathbf{c})) \subseteq S(\mathbf{w}^*)$ and $(**)$ is from the linear independence of columns of \mathbf{X} .
 2018 Notice that $\mathbf{M}(\mathbf{c})$ and $\mathbf{w}^*(\mathbf{c})$ are continuous on $\mathbf{c} = \mathbf{c}_0$, which implies that $\mathbf{d}(\mathbf{c})$ is continuous
 2019 on $\mathbf{c} = \mathbf{c}_0$.
 2020

2021 Since at least one of the dual solutions is strictly positive, \mathbf{d} is a continuous map from Δ^{N-1} to
 2022 $\mathbb{R}_{\geq 0}^N \setminus \{\mathbf{0}\}$. This implies that T is continuous, since $\mathbf{d} \mapsto \frac{\mathbf{d}}{\sum_{i \in [N]} d_i}$ is continuous on $\mathbb{R}_{\geq 0}^N \setminus \{\mathbf{0}\}$.
 2023

2024 (c) Since Δ^{N-1} is a nonempty convex compact subset of \mathbb{R}^N , there exists a fixed point of T by
 2025 Brouwer fixed-point theorem.

2026 (d) From Lemma 4.5, there exists $\mathbf{c}^* \in \Delta^{N-1}$ such that $\hat{\mathbf{w}} \propto \frac{\sum_{i=1}^N c_i^* \mathbf{x}_i}{\sqrt{\sum_{i=1}^N c_i^{*2} \mathbf{x}_i^2}}$ with $c_i^* = 0$ for $i \notin S'$
 2027 where $S' = \arg \min_{i \in [N]} \hat{\mathbf{w}}^\top \mathbf{x}_i$. Then we take $\hat{\mathbf{w}} = \frac{\sum_{i \in S} k c_i^* \mathbf{x}_i}{\sqrt{\sum_{i \in S} c_i^{*2} \mathbf{x}_i^2}}$ for some $k > 0$. We claim
 2028 that such \mathbf{c}^* becomes a fixed point of T and $\hat{\mathbf{w}} \propto \mathbf{p}(\mathbf{c}^*)$.
 2029

2030 Consider the optimization problem $P_{\text{Adam}}(\mathbf{c}^*)$ and its unique primal solution $\mathbf{w}^* = \mathbf{p}(\mathbf{c}^*)$.
 2031 Notice that $\min_{i \in [N]} \hat{\mathbf{w}}^\top \mathbf{x}_i = \gamma > 0$ since AdamProxy minimizes the loss. Therefore,
 2032 $\mathbf{w}^* = \frac{1}{\gamma} \hat{\mathbf{w}}$ and $d_i(\mathbf{c}^*) = \frac{k c_i^*}{\gamma}$ satisfy the following KKT conditions
 2033

2034
$$\begin{aligned} \mathbf{M}(\mathbf{c}^*)\mathbf{w}^* &= \sum_{i \in S^*} d_i \mathbf{x}_i, d_i \geq 0, \\ &\mathbf{w}^{*\top} \mathbf{x}_i - 1 \geq 0, \forall i \in [N], \end{aligned}$$

 2035
 2036
 2037

2038 where $S^* = \{i \in [N] : \mathbf{w}^{*\top} \mathbf{x}_i - 1 = 0\}$ is the index set of support vectors of \mathbf{w}^* . This implies
 2039 that $T(\mathbf{c}^*) = \mathbf{c}^*$ and $\hat{\mathbf{w}} = \gamma \mathbf{w}^* \propto \mathbf{w}^* = \mathbf{p}(\mathbf{c}^*)$, which proves the claim.
 2040

□

2044 F.5 DETAILED CALCULATIONS OF EXAMPLE 4.11

2045 Consider $N = d$ and $\{\mathbf{x}_i\}_{i \in [d]} \subseteq \mathbb{R}^d$ where $\mathbf{x}_i = x_i \mathbf{e}_i + \delta \sum_{j \neq i} \mathbf{e}_j$ for some $0 < \delta$ and $0 < x_0 < \dots < x_{d-1}$. ℓ_∞ -max-margin problem is given by
 2046

2047
$$\min \|\mathbf{w}\|_\infty \text{ subject to } \mathbf{w}^\top \mathbf{x}_i \geq 1, \forall i \in [N].$$

2052 (For the convenience of calculation, we use the objective $\|\mathbf{w}\|_\infty$ rather than $\frac{1}{2}\|\mathbf{w}\|_\infty^2$.) Its KKT
 2053 conditions are given by
 2054

$$\begin{aligned} 2055 \quad \partial\|\mathbf{w}\|_\infty &\ni \sum_{i \in [N]} \lambda_i \mathbf{x}_i, \\ 2056 \quad \sum_{i \in [N]} \lambda_i (\mathbf{w}^\top \mathbf{x}_i - 1) &= 0, \\ 2057 \quad \lambda_i \geq 0, \mathbf{w}^\top \mathbf{x}_i - 1 \geq 0, \forall i \in [N]. \\ 2058 \end{aligned}$$

2061 Note that $\mathbf{w}^* = (\frac{1}{x_0 + (d-1)\delta}, \dots, \frac{1}{x_0 + (d-1)\delta}) \in \mathbb{R}^d$ and $\lambda^* = (\frac{1}{x_0 + (d-1)\delta}, 0, \dots, 0) \in \mathbb{R}^d$ satisfy
 2062 the KKT conditions since
 2063

$$\begin{aligned} 2064 \quad \partial\|\mathbf{w}\|_\infty \Big|_{\mathbf{w}=\mathbf{w}^*} &= \Delta^{d-1} \ni \frac{1}{x_0 + (d-1)\delta} \mathbf{x}_0 = \sum_{i \in [N]} \lambda_i^* \mathbf{x}_i, \\ 2065 \quad \sum_{i \in [N]} \lambda_i^* (\mathbf{w}^{*\top} \mathbf{x}_i - 1) &= \lambda_1^* \left(\frac{x_0 + (d-1)\delta}{x_0 + (d-1)\delta} - 1 \right) = 0, \\ 2066 \quad \lambda_i^* \geq 0, \mathbf{w}^{*\top} \mathbf{x}_i - 1 \geq 0, \forall i \in [N]. \\ 2067 \end{aligned}$$

2068 Now we show that $\mathbf{c}^* = (1, 0, \dots, 0) \in \Delta^{d-1}$ is a fixed point of T in Theorem 4.8 and $\mathbf{w}^* = \mathbf{p}(\mathbf{c}^*)$.
 2069 Note that for $k = \frac{1}{x_0 + (d-1)\delta} > 0$, it satisfies
 2070

$$\begin{aligned} 2071 \quad \mathbf{M}(\mathbf{c}^*) \mathbf{w}^* &= \text{diag}(x_0, \delta, \dots, \delta) \mathbf{w}^* = k \mathbf{x}_0 = k \sum_{i \in [N]} c_i^* \mathbf{x}_i \\ 2072 \quad \sum_{i \in [N]} c_i^* (\mathbf{w}^{*\top} \mathbf{x}_i - 1) &= 0, \\ 2073 \quad c_i^* \geq 0, \mathbf{w}^{*\top} \mathbf{x}_i - 1 \geq 0, \forall i \in [N], \\ 2074 \end{aligned}$$

2075 which implies $T(\mathbf{c}^*) = \mathbf{c}^*$ and $\mathbf{w}^* = \mathbf{p}(\mathbf{c}^*)$.
 2076

2077 G MISSING PROOFS IN SECTION 5

2078 **Algorithm 4** Inc-Signum

2079 **Hyperparams:** Learning rate schedule $\{\eta_t\}_{t=0}^{T-1}$, momentum parameter $\beta \in [0, 1)$, batch size b
 2080 **Input:** Initial weight \mathbf{w}_0 , dataset $\{\mathbf{x}_i\}_{i \in [N]}$

2081 1: Initialize momentum $\mathbf{m}_{-1} = \mathbf{0}$
 2082 2: **for** $t = 0, 1, 2, \dots, T-1$ **do**
 2083 3: $\mathcal{B}_t \leftarrow \{(t \cdot b + i) \pmod N\}_{i=0}^{b-1}$
 2084 4: $\mathbf{g}_t \leftarrow \nabla \mathcal{L}_{\mathcal{B}_t}(\mathbf{w}_t) = \frac{1}{b} \sum_{i \in \mathcal{B}_t} \ell'(\mathbf{w}_t^\top \mathbf{x}_i) \mathbf{x}_i$
 2085 5: $\mathbf{m}_t \leftarrow \beta \mathbf{m}_{t-1} + (1 - \beta) \mathbf{g}_t$
 2086 6: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \text{sign}(\mathbf{m}_t)$
 2087 7: **end for**
 2088 8: **return** \mathbf{w}_T

2089
 2090
 2091
 2092
 2093
 2094
 2095
 2096
 2097
 2098
 2099
 2100 **Related Work.** Our proof of Theorem 5.1 builds on standard techniques from the analysis of the
 2101 implicit bias of normalized steepest descent on linearly separable data (Gunasekar et al., 2018a;
 2102 Zhang et al., 2024a; Fan et al., 2025). The most closely related result is due to Fan et al. (2025),
 2103 who showed that full-batch Signum converges in direction to the maximum ℓ_∞ -margin solution.
 2104 Theorem 5.1 extends this result to the mini-batch setting, establishing that the mini-batch variant
 2105 of Inc-Signum (Algorithm 4) also converges in direction to the maximum ℓ_∞ -margin solution,
 2106 provided the momentum parameter is chosen sufficiently close to 1.

Technical Contribution. The key technical contribution enabling the mini-batch analysis is Lemma G.2. Importantly, requiring momentum parameter β close to 1 is not merely a technical convenience but intrinsic to the mini-batch setting ($b < N$), as formalized in Lemma G.2 and supported empirically in Figure 10 of Appendix B.

Implicit Bias of SignSGD. We note that as an extreme case, Inc-Signum with $\beta = 0$ and batch size 1 (i.e., SignSGD) has a simple implicit bias: its iterates converge in direction to $\sum_{i \in [N]} \text{sign}(\mathbf{x}_i)$, which corresponds to neither the ℓ_2 - nor the ℓ_∞ -max-margin solution.

Notation. We introduce additional notation to analyze Inc-Signum (Algorithm 4) with arbitrary mini-batch size b . Let $\mathcal{B}_t \subseteq [N]$ denote the set of indices in the mini-batch sampled at iteration t . The corresponding mini-batch loss $\mathcal{L}_{\mathcal{B}_t}(\mathbf{w})$ is defined as

$$\mathcal{L}_{\mathcal{B}_t}(\mathbf{w}) \triangleq \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \ell(\mathbf{w}^\top \mathbf{x}_i).$$

We define the maximum normalized ℓ_∞ -margin as

$$\gamma_\infty \triangleq \max_{\|\mathbf{w}\|_\infty \leq 1} \min_{i \in [N]} \mathbf{w}^\top \mathbf{x}_i > 0,$$

and again introduce the proxy $\mathcal{G} : \mathbb{R}^d \rightarrow \mathbb{R}$ defined as

$$\mathcal{G}(\mathbf{w}) \triangleq -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i).$$

As before, we consider ℓ to be either the logistic loss $\ell_{\log}(z) = \log(1 + \exp(-z))$ or the exponential loss $\ell_{\exp}(z) = \exp(-z)$. Finally, let D be an upper bound on the ℓ_1 -norm of the data, i.e., $\|\mathbf{x}_i\|_1 \leq D$ for all $i \in [N]$.

Lemma G.1 (Descent inequality). Inc-Signum iterates $\{\mathbf{w}_t\}$ satisfy

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t) - \eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \Delta_t \rangle + C_H \eta_t^2 \mathcal{G}(\mathbf{w}_t), \quad \Delta_t := \text{sign}(\mathbf{m}_t),$$

where $C_H = \frac{1}{2} D^2 e^{\eta_0 D}$.

Proof. By Taylor's theorem,

$$\mathcal{L}(\mathbf{w}_{t+1}) = \mathcal{L}(\mathbf{w}_t - \eta_t \Delta_t) = \mathcal{L}(\mathbf{w}_t) - \eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \Delta_t \rangle + \frac{1}{2} \eta_t^2 \Delta_t^\top \nabla^2 \mathcal{L}(\mathbf{w}_t - \zeta \eta_t \Delta_t) \Delta_t,$$

for some $\zeta \in (0, 1)$. Note that for any $\mathbf{w} \in \mathbb{R}^d$,

$$\Delta_t^\top \nabla^2 \mathcal{L}(\mathbf{w}) \Delta_t = \frac{1}{N} \sum_{i \in [N]} \ell''(\mathbf{w}^\top \mathbf{x}_i) (\Delta_t^\top \mathbf{x}_i)^2 \leq \frac{1}{N} \sum_{i \in [N]} \ell''(\mathbf{w}^\top \mathbf{x}_i) \|\Delta_t\|_\infty^2 \|\mathbf{x}_i\|_1^2 \leq D^2 \mathcal{G}(\mathbf{w}),$$

where we used $\mathcal{G}(\mathbf{w}) \geq \frac{1}{N} \sum_{i \in [N]} \ell''(\mathbf{w}^\top \mathbf{x}_i)$ from Lemma I.1. Then,

$$\begin{aligned} \mathcal{L}(\mathbf{w}_{t+1}) &\leq \mathcal{L}(\mathbf{w}_t) - \eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \Delta_t \rangle + \frac{1}{2} \eta_t^2 \Delta_t^\top \nabla^2 \mathcal{L}(\mathbf{w}_t - \zeta \eta_t \Delta_t) \Delta_t \\ &\leq \mathcal{L}(\mathbf{w}_t) - \eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \Delta_t \rangle + \frac{1}{2} \eta_t^2 D^2 \mathcal{G}(\mathbf{w}_t - \zeta \eta_t \Delta_t) \\ &\leq \mathcal{L}(\mathbf{w}_t) - \eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \Delta_t \rangle + \frac{1}{2} \eta_t^2 D^2 e^{\eta_0 D} \mathcal{G}(\mathbf{w}), \end{aligned}$$

where we used $\mathcal{G}(\mathbf{w}') \leq e^{D\|\mathbf{w}' - \mathbf{w}\|_\infty} \mathcal{G}(\mathbf{w})$ for all \mathbf{w}, \mathbf{w}' from Lemma I.1. Finally, choosing $C_H := \frac{1}{2} D^2 e^{\eta_0 D}$, we obtain the desired inequality. \square

Lemma G.2 (EMA misalignment). We denote $\mathbf{e}_t := \mathbf{m}_t - \nabla \mathcal{L}(\mathbf{w}_t)$. Suppose that $\beta \in (\frac{N-b}{N}, 1)$. Then, there exists $t_0 \in \mathbb{N}$ such that for all $t \geq t_0$,

$$\|\mathbf{e}_t\|_1 = \|\mathbf{m}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|_1 \leq [(1 - \beta) D \frac{N}{b} (\frac{N}{b} - 1) + C_1 \eta_t + C_2 \beta^t] \mathcal{G}(\mathbf{w}_t)$$

where $C_1, C_2 > 0$ are constants determined by β, N, b , and D .

2160 *Proof.* The momentum \mathbf{m}_t can be written as:
 2161

$$2162 \quad \mathbf{m}_t = (1 - \beta) \sum_{\tau=0}^t \beta^\tau \mathbf{g}_{t-\tau} = (1 - \beta) \sum_{\tau=0}^t \beta^\tau \nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_{t-\tau}),$$

2164 and the full-batch gradient $\nabla \mathcal{L}(\mathbf{w}_t)$ can be written as:
 2165

$$2166 \quad \nabla \mathcal{L}(\mathbf{w}_t) = \beta^{t+1} \nabla L(\mathbf{w}_t) + (1 - \beta) \sum_{\tau=0}^t \beta^\tau \nabla \mathcal{L}(\mathbf{w}_t),$$

2168 Consequently, the misalignment $\mathbf{e}_t = \mathbf{m}_t - \nabla \mathcal{L}(\mathbf{w}_t)$ can be decomposed as:
 2169

$$2170 \quad \mathbf{e}_t = (1 - \beta) \sum_{\tau=0}^t \beta^\tau (\nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_{t-\tau}) - \nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_t)) \\ 2171 \\ 2172 \\ 2173 \\ 2174 \\ 2175 \\ 2176$$

$$+ (1 - \beta) \sum_{\tau=0}^t \beta^\tau (\nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_t)) \\ - \beta^{t+1} \nabla \mathcal{L}(\mathbf{w}_t),$$

2177 and thus

$$2178 \quad \|\mathbf{e}_t\|_1 = \underbrace{\left\| (1 - \beta) \sum_{\tau=0}^t \beta^\tau (\nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_{t-\tau}) - \nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_t)) \right\|_1}_{\triangleq (A)} \\ 2179 \\ 2180 \\ 2181 \\ 2182 \\ 2183 \\ 2184 \\ 2185 \\ 2186 \\ 2187 \\ 2188$$

$$+ \underbrace{\left\| (1 - \beta) \sum_{\tau=0}^t \beta^\tau (\nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_t)) \right\|_1}_{\triangleq (B)} \\ + \underbrace{\left\| \beta^{t+1} \nabla \mathcal{L}(\mathbf{w}_t) \right\|_1}_{\triangleq (C)}.$$

2189 We upper bound each term separately.

2190 First, the term (A) represents the misalignment by the weight movement, which can be bounded as:
 2191

$$2192 \quad (A) = \left\| (1 - \beta) \sum_{\tau=0}^t \beta^\tau (\nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_{t-\tau}) - \nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_t)) \right\|_1 \\ 2193 \\ 2194 \\ 2195 \\ 2196 \\ 2197 \\ 2198 \\ 2199 \\ 2200 \\ 2201 \\ 2202 \\ 2203 \\ 2204 \\ 2205 \\ 2206 \\ 2207 \\ 2208 \\ 2209$$

$$\leq (1 - \beta) \sum_{\tau=0}^t \beta^\tau \|\nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_{t-\tau}) - \nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_t)\|_1 \\ = (1 - \beta) \sum_{\tau=0}^t \beta^\tau \left\| \frac{1}{b} \sum_{i \in \mathcal{B}_{t-\tau}} (\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_i) - \ell'(\mathbf{w}_t^\top \mathbf{x}_i)) \mathbf{x}_i \right\|_1 \\ \leq (1 - \beta) \sum_{\tau=0}^t \beta^\tau \frac{D}{b} \sum_{i \in \mathcal{B}_{t-\tau}} |\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_i) - \ell'(\mathbf{w}_t^\top \mathbf{x}_i)| \\ \leq \frac{(1 - \beta)D}{b} \sum_{\tau=0}^t \beta^\tau \sum_{i \in \mathcal{B}_{t-\tau}} |\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_i)| \left| \frac{\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_i)}{\ell'(\mathbf{w}_t^\top \mathbf{x}_i)} - 1 \right| \\ \leq \frac{(1 - \beta)DN}{b} \mathcal{G}(\mathbf{w}_t) \sum_{\tau=0}^t \beta^\tau \sum_{i \in \mathcal{B}_{t-\tau}} \left| \frac{\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_i)}{\ell'(\mathbf{w}_t^\top \mathbf{x}_i)} - 1 \right|,$$

2210 where we used $N\mathcal{G}(\mathbf{w}) = -\sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i) = \sum_{i \in [N]} |\ell'(\mathbf{w}^\top \mathbf{x}_i)| \geq \max_{i \in [N]} |\ell'(\mathbf{w}^\top \mathbf{x}_i)|$ in
 2211 the last inequality. For all $i \in [N]$,

$$2212 \quad \left| \frac{\ell'(\mathbf{w}_{t-\tau}^\top \mathbf{x}_i)}{\ell'(\mathbf{w}_t^\top \mathbf{x}_i)} - 1 \right| \leq e^{|\mathbf{w}_t - \mathbf{w}_{t-\tau}|^\top \mathbf{x}_i} - 1 \leq e^{\|\mathbf{w}_t - \mathbf{w}_{t-\tau}\|_\infty \|\mathbf{x}_i\|_1} - 1 \leq e^{D \sum_{\tau'=1}^{\tau} \eta_{t-\tau'}} - 1.$$

2214 By Assumption 2.3, there exists $t_0 \in \mathbb{N}$ and constant $c_1 > 0$ determined by β and D such that
 2215 $\sum_{\tau=0}^t \beta^\tau (e^{D \sum_{\tau'=1}^\tau \eta_{t-\tau'}} - 1) \leq c_1 \eta_t$ for all $t \geq t_0$. Then, for all $t \geq t_0$, we have
 2216

$$\begin{aligned}
 2217 \text{(A)} &\leq \frac{(1-\beta)DN}{b} \mathcal{G}(\mathbf{w}_t) \sum_{\tau=0}^t \beta^\tau b (e^{D \sum_{\tau'=1}^\tau \eta_{t-\tau'}} - 1) \\
 2218 \\
 2219 &= (1-\beta)DN \mathcal{G}(\mathbf{w}_t) \sum_{\tau=0}^t \beta^\tau e^{D \sum_{\tau'=1}^\tau \eta_{t-\tau'}} - 1 \\
 2220 \\
 2221 &\leq (1-\beta)DN c_1 \eta_t \mathcal{G}(\mathbf{w}_t).
 \end{aligned}$$

2222
 2223 Second, the term (B) represents the misalignment by mini-batch updates. Denote the number of
 2224 mini-batches in a single epoch as $m := \frac{N}{b}$. Since $\mathcal{B}_t = \{(t \cdot b + i) \pmod{N}\}_{i=0}^{b-1}$, note that $\mathcal{B}_i = \mathcal{B}_j$
 2225 if and only if $i \equiv j \pmod{m}$. Now, the term (B) can be upper bounded as
 2226

$$\begin{aligned}
 2227 \text{(B)} &= \left\| (1-\beta) \sum_{\tau=0}^t \beta^\tau (\nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_t)) \right\|_1 \\
 2228 \\
 2229 &= \left\| (1-\beta) \sum_{\tau=0}^t \beta^\tau \left[\nabla \mathcal{L}_{\mathcal{B}_{t-\tau}}(\mathbf{w}_t) - \frac{1}{m} \sum_{j=1}^m \nabla \mathcal{L}_{\mathcal{B}_j}(\mathbf{w}_t) \right] \right\|_1 \\
 2230 \\
 2231 &= \left\| (1-\beta) \sum_{j=1}^m \left(\sum_{\tau \leq t: (t-\tau) \equiv j \pmod{m}} \beta^\tau - \frac{1}{m} \sum_{\tau=0}^t \beta^\tau \right) \nabla \mathcal{L}_{\mathcal{B}_j}(\mathbf{w}_t) \right\|_1 \\
 2232 \\
 2233 &\leq (1-\beta) m \cdot \max_{j \in [m]} \left| \sum_{\tau \leq t: (t-\tau) \equiv j \pmod{m}} \beta^\tau - \frac{1}{m} \sum_{\tau=0}^t \beta^\tau \right| \cdot \max_{j \in [m]} \|\nabla \mathcal{L}_{\mathcal{B}_j}(\mathbf{w}_t)\|_1 \\
 2234 \\
 2235 &\leq (1-\beta) D m^2 \mathcal{G}(\mathbf{w}_t) \cdot \max_{j \in [m]} \left| \sum_{\tau \leq t: (t-\tau) \equiv j \pmod{m}} \beta^\tau - \frac{1}{m} \sum_{\tau=0}^t \beta^\tau \right|,
 \end{aligned}$$

2236 where the last inequality holds since
 2237

$$\max_{j \in [m]} \|\nabla \mathcal{L}_{\mathcal{B}_j}(\mathbf{w})\|_1 = \frac{1}{b} \max_{j \in [m]} \left\| \sum_{i \in \mathcal{B}_j} \ell'(\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i \right\|_1 \leq \frac{1}{b} \sum_{i=1}^N |\ell'(\mathbf{w}^\top \mathbf{x}_i)| \cdot D = \frac{DN}{b} \mathcal{G}(\mathbf{w}) = Dm \mathcal{G}(\mathbf{w}),$$

2238 for all $\mathbf{w} \in \mathbb{R}^d$.
 2239

2268 It remains to upper bound $\max_{j \in [m]} \left| \sum_{\tau \leq t: (t-\tau) \equiv j \pmod{m}} \beta^\tau - \frac{1}{m} \sum_{\tau=0}^t \beta^\tau \right|$. Fix arbitrary $j \in$
 2269 $[m]$. Note that

2271

2272

2273

2274

2275

2276

2277

2278

2279

2280

2281

2282

2283

$$\begin{aligned}
 & (1 - \beta) \left(\sum_{\tau \leq t: (t-\tau) \equiv j \pmod{m}} \beta^\tau - \frac{1}{m} \sum_{\tau=0}^t \beta^\tau \right) \\
 & \leq (1 - \beta) \sum_{k=0}^{\lfloor \frac{t}{m} \rfloor} \beta^{mk} - (1 - \beta) \frac{1}{m} \sum_{\tau=0}^t \beta^\tau \\
 & = (1 - \beta) \sum_{k=0}^{\lfloor \frac{t}{m} \rfloor} \beta^{mk} - (1 - \beta) \sum_{k=0}^{\lfloor \frac{t}{m} \rfloor - 1} \left(\frac{1}{m} \beta^{mk} \sum_{\tau=0}^{m-1} \beta^\tau \right) - (1 - \beta) \frac{1}{m} \sum_{\tau=m(\lfloor \frac{t}{m} \rfloor - 1) + 1}^t \beta^\tau \\
 & \leq (1 - \beta) \beta^m \lfloor \frac{t}{m} \rfloor + \sum_{k=0}^{\lfloor \frac{t}{m} \rfloor - 1} \beta^{mk} \left[(1 - \beta) - \frac{1}{m} (1 - \beta^m) \right] \\
 & \stackrel{(*)}{\leq} (1 - \beta) \beta^{t-m} + \sum_{k=0}^{\lfloor \frac{t}{m} \rfloor - 1} \beta^{mk} \frac{(m-1)(1-\beta)^2}{2} \\
 & \leq (1 - \beta) \beta^{t-m} + \frac{1}{1 - \beta^m} \cdot \frac{(m-1)(1-\beta)^2}{2} \\
 & \stackrel{(**)}{\leq} (1 - \beta) \beta^{t-m} + \frac{2}{m(1-\beta)} \cdot \frac{(m-1)(1-\beta)^2}{2} \\
 & = (1 - \beta) \beta^{t-m} + \frac{m-1}{m} (1 - \beta),
 \end{aligned}$$

2308

2309

2310

2311

2312

2313

2314

2315

2316

2317

2318

2319

2320

2321 where the inequalities $(*)$ and $(**)$ hold since $(1 - \epsilon)^m \leq 1 - m\epsilon + \frac{m(m-1)}{2} \epsilon^2 \leq 1 - \frac{m}{2} \epsilon$ for all
 $0 \leq \epsilon \leq \frac{1}{m-1}$ and choose $\epsilon = 1 - \beta$.

2322 Similarly, we have
 2323
 2324
$$(1 - \beta) \left(\frac{1}{m} \sum_{\tau=0}^t \beta^\tau - \sum_{\tau \leq t: (t-\tau) \equiv j \pmod{m}} \beta^\tau \right)$$

 2325
 2326
 2327
$$\leq (1 - \beta) \frac{1}{m} \sum_{\tau=0}^t \beta^\tau - (1 - \beta) \sum_{k=0}^{\lfloor \frac{t+1}{m} \rfloor - 1} \beta^{m(k+1)-1}$$

 2328
 2329
 2330
$$= (1 - \beta) \sum_{k=0}^{\lfloor \frac{t+1}{m} \rfloor - 1} \left(\frac{1}{m} \beta^{mk} \sum_{\tau=0}^{m-1} \beta^\tau \right) + (1 - \beta) \frac{1}{m} \sum_{\tau=m \lfloor \frac{t+1}{m} \rfloor}^t \beta^\tau - (1 - \beta) \sum_{k=0}^{\lfloor \frac{t+1}{m} \rfloor - 1} \beta^{m(k+1)-1}$$

 2331
 2332
 2333
 2334
 2335
$$\leq (1 - \beta) \frac{1}{m} \sum_{\tau=t-m+2}^t \beta^\tau + \sum_{k=0}^{\lfloor \frac{t+1}{m} \rfloor - 1} \beta^{mk} \left[\frac{1}{m} (1 - \beta^m) - (1 - \beta) \beta^{m-1} \right]$$

 2336
 2337
 2338
 2339
$$= \frac{1}{m} \beta^{t-m+2} (1 - \beta^{m-1}) + \sum_{k=0}^{\lfloor \frac{t+1}{m} \rfloor - 1} \beta^{mk} \left[\frac{1}{m} (1 - \beta^m) - (1 - \beta) \beta^{m-1} \right]$$

 2340
 2341
 2342
$$\leq \frac{1}{m} \beta^{t-m+2} (1 - \beta^{m-1}) + \sum_{k=0}^{\lfloor \frac{t+1}{m} \rfloor - 1} \beta^{mk} \frac{(m-1)(1-\beta)^2}{2}$$

 2343
 2344
 2345
$$\leq \frac{1}{m} \beta^{t-m+2} (1 - \beta^{m-1}) + \frac{1}{1 - \beta^m} \cdot \frac{(m-1)(1-\beta)^2}{2}$$

 2346
 2347
 2348
$$\leq (1 - \beta) \beta^{t-m} + \frac{m-1}{m} (1 - \beta).$$

2349 Combining the bounds, we get
 2350
 2351

$$(B) \leq (1 - \beta) D m (\beta^{t-m} m + m - 1) \mathcal{G}(\mathbf{w}_t).$$

2352 Finally,
 2353

$$(C) = \|\beta^{t+1} \nabla \mathcal{L}(\mathbf{w}_t)\|_1 \leq \beta^{t+1} D \mathcal{G}(\mathbf{w}_t).$$

2354 Therefore, we conclude
 2355

$$\|\mathbf{e}\|_1 \leq [(1 - \beta) D m (m - 1) + C_1 \eta_t + C_2 \beta^t] \mathcal{G}(\mathbf{w}_t)$$

2356 where $C_1, C_2 > 0$ are constants determined by β, m , and D . \square
 2357

2358 **Corollary G.3.** Suppose that $\beta \in (\frac{N-b}{N}, 1)$. Then, there exists $t_0 \in \mathbb{N}$ such that for all $t \geq t_0$,
 2359 Inc-Signum iterates $\{\mathbf{w}_t\}$ satisfy

2360
$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t) - \eta_t (\gamma_\infty - 2(1 - \beta) D \frac{N}{b} (\frac{N}{b} - 1) - (2C_1 + C_H) \eta_t - 2C_2 \beta^t) \mathcal{G}(\mathbf{w}_t),$$

 2361 where $C_H, C_1, C_2 > 0$ are constants in Lemmas G.1 and G.2.
 2362

2363 *Proof.* By Lemma I.1, we get
 2364

$$\begin{aligned} \langle \nabla \mathcal{L}(\mathbf{w}_t), \Delta_t \rangle &= \langle \mathbf{m}_t, \Delta_t \rangle - \langle \mathbf{e}_t, \Delta_t \rangle \\ &\geq \|\mathbf{m}_t\|_1 - \|\mathbf{e}_t\|_1 \|\Delta_t\|_\infty \\ &\geq (\|\nabla \mathcal{L}(\mathbf{w}_t)\|_1 - \|\mathbf{e}_t\|_1) - \|\mathbf{e}_t\|_1 \\ &= \|\nabla \mathcal{L}(\mathbf{w}_t)\|_1 - 2\|\mathbf{e}_t\|_1 \\ &\geq \gamma_\infty \mathcal{G}(\mathbf{w}_t) - 2\|\mathbf{e}_t\|_1. \end{aligned}$$

2370 Now using Lemma G.1 and Lemma G.2, we conclude
 2371

$$\begin{aligned} \mathcal{L}(\mathbf{w}_{t+1}) &\leq \mathcal{L}(\mathbf{w}_t) - \eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \Delta_t \rangle + C_H \eta_t^2 \mathcal{G}(\mathbf{w}_t) \\ &\leq \mathcal{L}(\mathbf{w}_t) - \eta_t (\gamma_\infty \mathcal{G}(\mathbf{w}_t) - 2\|\mathbf{e}_t\|_1) + C_H \eta_t^2 \mathcal{G}(\mathbf{w}_t) \\ &\leq \mathcal{L}(\mathbf{w}_t) - \eta_t (\gamma_\infty - 2(1 - \beta) D \frac{N}{b} (\frac{N}{b} - 1) - (2C_1 + C_H) \eta_t - 2C_2 \beta^t) \mathcal{G}(\mathbf{w}_t), \end{aligned}$$

2372 which ends the proof. \square
 2373
 2374
 2375

2376 **Proposition G.4** (Loss convergence). *Suppose that $\beta \in (1 - \frac{\gamma_\infty}{4C_0}, 1)$ if $b < N$ and $\beta \in (0, 1)$ if
2377 $b = N$, where $C_0 := D \frac{N}{b} (\frac{N}{b} - 1)$. Then, $\mathcal{L}(\mathbf{w}_t) \rightarrow 0$ as $t \rightarrow \infty$.*

2379 *Proof.* Note that $\beta \in (\frac{N-b}{N}, 1)$ since $\gamma_\infty = \max_{\|\mathbf{w}\|_\infty \leq 1} \min_{i \in [N]} \mathbf{w}^\top \mathbf{x}_i \leq D$. By Corollary G.3,
2380 there exists $t_0 \in \mathbb{N}$ such that for all $t \geq t_0$,
2381

$$2382 \eta_t(\gamma_\infty - 2C_0(1 - \beta) - (2C_1 + C_H)\eta_t - 2C_2\beta^t)\mathcal{G}(\mathbf{w}_t) \leq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_{t+1}).$$

2383 Since $\eta_t, \beta^t \rightarrow 0$ as $t \rightarrow \infty$, there exists $t_1 \geq t_0$ such that for all $t \geq t_1$,

$$2385 (2C_1 + C_H)\eta_t + 2C_2\beta^t < \frac{\gamma_\infty}{4}.$$

2387 Then,

$$2389 \frac{\gamma_\infty}{4} \sum_{t=t_1}^{\infty} \eta_t \mathcal{G}(\mathbf{w}_t) \leq \sum_{t=t_1}^{\infty} \eta_t (\gamma_\infty - 2C_0(1 - \beta) - (2C_1 + C_H)\eta_t - 2C_2\beta^t) \mathcal{G}(\mathbf{w}_t) \leq \sum_{t=t_1}^{\infty} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_{t+1}) < \infty.$$

2391 Thus, $\sum_{t=t_0}^{\infty} \eta_t \mathcal{G}(\mathbf{w}_t) < \infty$ and since $\sum_{t=t_0}^{\infty} \eta_t = \infty$, this implies $\mathcal{G}(\mathbf{w}_t) \rightarrow 0$ and therefore
2392 $\mathcal{L}(\mathbf{w}_t) \rightarrow 0$ as $t \rightarrow \infty$. \square
2393

2394 **Proposition G.5** (Unnormalized margin lower bound). *Suppose that $\beta \in (1 - \frac{\gamma_\infty}{4C_0}, 1)$ if $b < N$ and
2395 $\beta \in (0, 1)$ if $b = N$, where $C_0 := D \frac{N}{b} (\frac{N}{b} - 1)$. Then, there exists $t_s \in \mathbb{N}$ such that for all $t \geq t_s$,*

$$2397 \min_{i \in [N]} \mathbf{w}^\top \mathbf{x}_i \leq (\gamma_\infty - 2C_0(1 - \beta)) \sum_{\tau=t_s}^{t-1} \eta_\tau \frac{\mathcal{G}(\mathbf{w}_\tau)}{\mathcal{L}(\mathbf{w}_\tau)} - (2C_1 + C_H) \sum_{\tau=t_s}^{t-1} \eta_\tau^2 - \frac{2C_2\eta_0}{1 - \beta},$$

2400 where $C_0 := D \frac{N}{b} (\frac{N}{b} - 1)$ and $C_H, C_1, C_2 > 0$ are constants in Lemmas G.1 and G.2.
2401

2402 *Proof.* By Proposition G.4, there exists time step $t_s \in \mathbb{N}$ such that $\mathcal{L}(\mathbf{w}_t) \leq \frac{\log 2}{N}$ for all $t \geq t_s$.
2403 Then, $\ell(\mathbf{w}_t^\top \mathbf{x}_i) \leq \frac{1}{N} \mathcal{L}(\mathbf{w}_t) \leq \log 2 < 1$, and thus $\min_{i \in [N]} \mathbf{w}_t^\top \mathbf{x}_i \geq 0$ for all $t \geq t_s$. Then, for all
2404 $t \geq t_s$,

$$2406 \exp(-\min_{i \in [N]} \mathbf{w}_t^\top \mathbf{x}_i) = \max_{i \in [N]} \exp(-\mathbf{w}_t^\top \mathbf{x}_i) \leq \frac{1}{\log 2} \max_{i \in [N]} \log(1 + \exp(-\mathbf{w}_t^\top \mathbf{x}_i)) \leq \frac{N \mathcal{L}(\mathbf{w}_t)}{\log 2},$$

2408 for logistic loss, and $\exp(-\min_{i \in [N]} \mathbf{w}_t^\top \mathbf{x}_i) \leq N \mathcal{L}(\mathbf{w}_t) \leq \frac{N \mathcal{L}(\mathbf{w}_t)}{\log 2}$ for exponential loss.
2409

2410 Using Corollary G.3 and $\mathcal{G}(\mathbf{w}) \leq \mathcal{L}(\mathbf{w})$ from Lemma I.1, we get
2411

$$2412 \mathcal{L}(\mathbf{w}_t) \leq \mathcal{L}(\mathbf{w}_{t-1}) \left(1 - (\gamma_\infty - 2C_0(1 - \beta))\eta_{t-1} \frac{\mathcal{G}(\mathbf{w}_{t-1})}{\mathcal{L}(\mathbf{w}_{t-1})} + (2C_1 + C_H)\eta_{t-1}^2 + 2C_2\beta^{t-1}\eta_{t-1} \right)$$

$$2414 \leq \mathcal{L}(\mathbf{w}_{t-1}) \exp \left(-(\gamma_\infty - 2C_0(1 - \beta))\eta_{t-1} \frac{\mathcal{G}(\mathbf{w}_{t-1})}{\mathcal{L}(\mathbf{w}_{t-1})} + (2C_1 + C_H)\eta_{t-1}^2 + 2C_2\beta^{t-1}\eta_{t-1} \right)$$

$$2417 \leq \mathcal{L}(\mathbf{w}_{t_s}) \exp \left(-(\gamma_\infty - 2C_0(1 - \beta)) \sum_{\tau=t_s}^{t-1} \eta_\tau \frac{\mathcal{G}(\mathbf{w}_\tau)}{\mathcal{L}(\mathbf{w}_\tau)} + (2C_1 + C_H) \sum_{\tau=t_s}^{t-1} \eta_\tau^2 + 2C_2 \sum_{\tau=t_s}^{t-1} \beta^\tau \eta_\tau \right)$$

$$2420 \leq \frac{\log 2}{N} \exp \left(-(\gamma_\infty - 2C_0(1 - \beta)) \sum_{\tau=t_s}^{t-1} \eta_\tau \frac{\mathcal{G}(\mathbf{w}_\tau)}{\mathcal{L}(\mathbf{w}_\tau)} + (2C_1 + C_H) \sum_{\tau=t_s}^{t-1} \eta_\tau^2 + \frac{2C_2\eta_0}{1 - \beta} \right).$$

2423 Thus, we get

$$2424 \exp(-\min_{i \in [N]} \mathbf{w}_t^\top \mathbf{x}_i) \leq \frac{N \mathcal{L}(\mathbf{w}_t)}{\log 2}$$

$$2427 \leq \exp \left(-(\gamma_\infty - 2C_0(1 - \beta)) \sum_{\tau=t_s}^{t-1} \eta_\tau \frac{\mathcal{G}(\mathbf{w}_\tau)}{\mathcal{L}(\mathbf{w}_\tau)} + (2C_1 + C_H) \sum_{\tau=t_s}^{t-1} \eta_\tau^2 + \frac{2C_2\eta_0}{1 - \beta} \right),$$

2429 which gives the desired inequality. \square

2430 **Theorem 5.1.** Let $\delta > 0$. Then there exists $\epsilon > 0$ such that the iterates $\{\mathbf{w}_t\}_{t=0}^\infty$ of Inc-Signum
2431 (Algorithm 4) with batch size b and momentum $\beta \in (1 - \epsilon, 1)$, under Assumptions 2.1 and 2.3, satisfy
2432

$$2433 \quad \liminf_{t \rightarrow \infty} \frac{\min_{i \in [N]} \mathbf{x}_i^\top \mathbf{w}_t}{\|\mathbf{w}_t\|_\infty} \geq \gamma_\infty - \delta, \quad (7)$$

2435 where

$$2437 \quad \gamma_\infty \triangleq \max_{\|\mathbf{w}\|_\infty \leq 1} \min_{i \in [N]} \mathbf{w}^\top \mathbf{x}_i, \quad D \triangleq \max_{i \in [N]} \|\mathbf{x}_i\|_1,$$

2439 and

$$2440 \quad \epsilon = \frac{1}{2D \cdot \frac{N}{b} (\frac{N}{b} - 1)} \min\{\delta, \frac{\gamma_\infty}{2}\} \quad \text{if } b < N, \quad \epsilon = 1 \quad \text{if } b = N.$$

2442 *Proof.* Let $C_0 := D \frac{N}{b} (\frac{N}{b} - 1)$ so that $\epsilon := \min\{\frac{\delta}{2C_0}, \frac{\gamma_\infty}{4C_0}\}$ if $b < N$ and $\epsilon := 1$ if $b = N$. Note
2443 that $C_0 = 0$ if $b = N$. Suppose that $\beta \in (1 - \epsilon, 1)$.

2444 Let t_0 be a time step that satisfy Corollary G.3. By Proposition G.4, there exists $t^* \geq t_0$ such that
2445 $(2C_1 + C_H)\eta_t + 2C_2\beta^t < \frac{\gamma_\infty}{8}$ and $\mathcal{L}(\mathbf{w}_t) \leq \frac{\log 2}{N}$ for all $t \geq t^*$. Then, for each $t \geq t^*$, we get
2446 $\frac{\mathcal{G}(\mathbf{w}_t)}{\mathcal{L}(\mathbf{w}_t)} \geq 1 - \frac{N\mathcal{L}(\mathbf{w}_t)}{2} \geq \frac{1}{2}$. By Corollary G.3, for all $t \geq t^*$,

$$2447 \quad \begin{aligned} \mathcal{L}(\mathbf{w}_t) &\leq \mathcal{L}(\mathbf{w}_{t-1}) \left(1 - (\gamma_\infty - 2C_0(1 - \beta))\eta_{t-1} \frac{\mathcal{G}(\mathbf{w}_{t-1})}{\mathcal{L}(\mathbf{w}_{t-1})} + (2C_1 + C_H)\eta_{t-1}^2 + 2C_2\beta^{t-1}\eta_{t-1} \right) \\ 2448 &\leq \mathcal{L}(\mathbf{w}_{t-1}) \left(1 - \frac{1}{4}\gamma_\infty\eta_{t-1} + \frac{1}{8}\gamma_\infty\eta_{t-1} \right) \\ 2449 &\leq \mathcal{L}(\mathbf{w}_{t-1}) \exp\left(-\frac{1}{8}\gamma_\infty\eta_{t-1}\right) \\ 2450 &\leq \mathcal{L}(\mathbf{w}_{t^*}) \exp\left(-\frac{\gamma_\infty}{8} \sum_{\tau=t^*}^{t-1} \eta_\tau\right) \\ 2451 &\leq \frac{\log 2}{N} \exp\left(-\frac{\gamma_\infty}{8} \sum_{\tau=t^*}^{t-1} \eta_\tau\right). \end{aligned}$$

2452 Consequently, by Lemma I.1, we have

$$2453 \quad \frac{\mathcal{G}(\mathbf{w}_t)}{\mathcal{L}(\mathbf{w}_t)} \geq 1 - \frac{N\mathcal{L}(\mathbf{w}_t)}{2} \geq 1 - \exp\left(-\frac{\gamma_\infty}{8} \sum_{\tau=t^*}^{t-1} \eta_\tau\right),$$

2454 for all $t \geq t^*$.

2455 Finally, using Proposition G.5, we get

$$2456 \quad \begin{aligned} \gamma_\infty - 2C_0(1 - \beta) - \frac{\min_{i \in [N]} \mathbf{w}_t^\top \mathbf{x}_i}{\|\mathbf{w}_t\|_\infty} \\ 2457 &\leq \frac{(\gamma_\infty - 2C_0(1 - \beta)) \left(\|\mathbf{w}_0\| + \sum_{\tau=0}^{t^*-1} \eta_\tau + \sum_{\tau=t^*}^t \eta_\tau e^{-\frac{\gamma_\infty}{8} \sum_{\tau=t^*}^{t-1} \eta_\tau} \right) + (2C_1 + C_H) \sum_{\tau=t^*}^{t-1} \eta_\tau^2 + \frac{2C_2\eta_0}{1-\beta}}{\|\mathbf{w}_0\| + \sum_{\tau=0}^{t-1} \eta_\tau} \\ 2458 &= \mathcal{O}\left(\frac{\sum_{\tau=0}^{t^*-1} \eta_\tau + \sum_{\tau=t^*}^t \eta_\tau e^{-\frac{\gamma_\infty}{8} \sum_{\tau=t^*}^{t-1} \eta_\tau} + \sum_{\tau=t^*}^{t-1} \eta_\tau^2}{\sum_{\tau=0}^{t-1} \eta_\tau}\right) \end{aligned}$$

2459 Therefore, we conclude

$$2460 \quad \liminf_{t \rightarrow \infty} \frac{\min_{i \in [N]} \mathbf{w}_t^\top \mathbf{x}_i}{\|\mathbf{w}_t\|_\infty} \geq \gamma_\infty - 2C_0(1 - \beta) \geq \gamma - \delta.$$

2461 \square

2484 H MISSING PROOFS IN APPENDIX A
2485

2486 **Lemma A.1.** Suppose that (a) $\mathcal{L}(\mathbf{w}_r) \rightarrow 0$ and (b) $\mathbf{w}_r = \|\mathbf{w}_r\|_2 \hat{\mathbf{w}} + \boldsymbol{\rho}(r)$ for some $\hat{\mathbf{w}}$ with
2487 $\exists \lim_{r \rightarrow \infty} \boldsymbol{\rho}(r)$. Then, under Assumptions 2.1 and 2.2, there exists $\mathbf{c} = (c_0, \dots, c_{N-1}) \in \Delta^{N-1}$ such
2488 that the limit direction $\hat{\mathbf{w}}$ of Inc-Adam with $\beta_1 = 0$ satisfies

$$2489 \hat{\mathbf{w}} \propto \sum_{i \in [N]} \frac{c_i \mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} c_j^2 \mathbf{x}_j^2}}, \quad (8)$$

2492 and $c_i = 0$ for $i \notin S$, where $S = \arg \min_{i \in [N]} \hat{\mathbf{w}}^\top \mathbf{x}_i$ is the index set of support vectors of $\hat{\mathbf{w}}$.
2493

2494 *Proof.* We start with the case of $\ell = \ell_{\text{exp}}$. First step is to characterize $\hat{\delta}$, the limit of δ_r . Notice that (b)
2495 is a strictly stronger assumption than Assumption 4.4 and it simplifies the analysis, while maintaining
2496 the intuition that the terms of support vectors dominate the update direction. Let $\lim_{r \rightarrow \infty} \boldsymbol{\rho}(r) = \hat{\boldsymbol{\rho}}$.
2497 We recall previous notations as $\gamma = \min_i \langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle$, $\bar{\gamma}_i = \langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle$, $\bar{\gamma} = \min_{i \notin S} \langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle$. Then it
2498 satisfies $S = \{i \in [N] : \langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle = \gamma\}$ and $\bar{\gamma} > \gamma > 0$. We can decompose dominant and residual
2499 terms in the update rule as follows.
2500

$$\begin{aligned} 2501 \delta_r &= \sum_{i \in S} \frac{\exp(-\gamma g(r)) \exp(-\boldsymbol{\rho}(r)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \exp(-2\bar{\gamma}_j g(r)) \exp(-2\boldsymbol{\rho}(r)^\top \mathbf{x}_j) \mathbf{x}_j^2}} \\ 2502 &\quad + \sum_{i \in S^c} \frac{\exp(-\bar{\gamma}_i g(r)) \exp(-\boldsymbol{\rho}(r)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \exp(-2\bar{\gamma}_j g(r)) \exp(-2\boldsymbol{\rho}(r)^\top \mathbf{x}_j) \mathbf{x}_j^2}} + \epsilon_r \\ 2503 &= \sum_{i \in S} \frac{\exp(-\boldsymbol{\rho}(r)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \exp(-2(\bar{\gamma}_j - \gamma) g(r)) \exp(-2\boldsymbol{\rho}(r)^\top \mathbf{x}_j) \mathbf{x}_j^2}} \\ 2504 &\quad + \sum_{i \in S^c} \frac{\exp(-(\bar{\gamma}_j - \gamma) g(r)) \exp(-\boldsymbol{\rho}(r)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} \exp(-2(\bar{\gamma}_j - \gamma) g(r)) \exp(-2\boldsymbol{\rho}(r)^\top \mathbf{x}_j) \mathbf{x}_j^2}} + \epsilon_r \\ 2505 &\triangleq \mathbf{d}(r) + \mathbf{r}(r) + \epsilon_r. \\ 2506 \end{aligned}$$

2515 Since $\bar{\gamma}_j > \gamma$ and $g(r) \rightarrow \infty$, $\mathbf{r}(r)$ converges to 0. Therefore, we get

$$\begin{aligned} 2516 \hat{\delta} &\triangleq \lim_{r \rightarrow \infty} \delta_r = \lim_{r \rightarrow \infty} \mathbf{d}(r) = \lim_{r \rightarrow \infty} \sum_{i \in S} \frac{\exp(-\boldsymbol{\rho}(r)^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{j \in S} \beta_2^{(i,j)} \exp(-2\boldsymbol{\rho}(r)^\top \mathbf{x}_j) \mathbf{x}_j^2}} \\ 2517 &= \sum_{i \in S} \frac{\exp(-\hat{\boldsymbol{\rho}}^\top \mathbf{x}_i) \mathbf{x}_i}{\sqrt{\sum_{j \in S} \beta_2^{(i,j)} \exp(-2\hat{\boldsymbol{\rho}}^\top \mathbf{x}_j) \mathbf{x}_j^2}} \\ 2518 &= \sum_{i \in [N]} \frac{c_i \mathbf{x}_i}{\sqrt{\sum_{j \in [N]} \beta_2^{(i,j)} c_j^2 \mathbf{x}_j^2}}, \\ 2519 \end{aligned}$$

2520 for some $\mathbf{c} \in \Delta^{N-1}$ satisfying $c_i = 0$ for $i \notin S$. Using the same technique based on Stolz-Cesaro
2521 theorem, we can also deduce that $\hat{\mathbf{w}} = \hat{\delta}$. Since we can extend this result to $\ell = \ell_{\text{log}}$ following the
2522 proof of Lemma 4.5, the statement is proved. \square
2523

2524 I TECHNICAL LEMMAS
25252526 I.1 PROXY FUNCTION
2527

2528 **Lemma I.1** (Proxy function). The proxy function \mathcal{G} satisfy the following properties: for any given
2529 weights $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ and any norm $\|\cdot\|$,

2530 (a) $\gamma_{\|\cdot\|} \mathcal{G}(\mathbf{w}) \leq \|\nabla \mathcal{L}(\mathbf{w})\|_* \leq D \mathcal{G}(\mathbf{w})$, where $D = \max_{i \in [N]} \|\mathbf{x}_i\|_*$ and $\gamma_{\|\cdot\|} =$
2531 $\max_{\|\mathbf{w}\| \leq 1} \min_{i \in [N]} \mathbf{w}^\top \mathbf{x}_i$ is the $\|\cdot\|$ -normalized max margin,
2532

2538 (b) $1 - \frac{N\mathcal{L}(\mathbf{w})}{2} \leq \frac{\mathcal{G}(\mathbf{w})}{\mathcal{L}(\mathbf{w})} \leq 1$,
 2539
 2540 (c) $\mathcal{G}(\mathbf{w}) \geq \frac{1}{N} \sum_{i \in [N]} \ell''(\mathbf{w}^\top \mathbf{x}_i)$,
 2541
 2542 (d) $\mathcal{G}(\mathbf{w}') \leq e^{B\|\mathbf{w}' - \mathbf{w}\|} \mathcal{G}(\mathbf{w})$, where $D = \max_{i \in [N]} \|\mathbf{x}_i\|_*$.
 2543

2544 *Proof.* This lemma (or a similar variant) is proved in [Zhang et al. \(2024a\)](#) and [Fan et al. \(2025\)](#).
 2545 Below, we provide a proof for completeness.
 2546

2547 (a) First, by duality we get

$$\begin{aligned} 2549 \|\nabla \mathcal{L}(\mathbf{w})\|_* &= \max_{\|\mathbf{g}\| \leq 1} \langle \mathbf{g}, -\nabla \mathcal{L}(\mathbf{w}) \rangle \geq \max_{\|\mathbf{g}\| \leq 1} -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i) \mathbf{g}^\top \mathbf{x}_i \\ 2550 &\geq \mathcal{G}(\mathbf{w}) \max_{\|\mathbf{g}\| \leq 1} \min_{i \in [N]} \mathbf{g}^\top \mathbf{x}_i \\ 2551 &= \gamma_{\|\cdot\|} \mathcal{G}(\mathbf{w}). \\ 2552 \\ 2553 \end{aligned}$$

2554 Second, we can obtain the lower bound as
 2555

$$\begin{aligned} 2556 \|\nabla \mathcal{L}(\mathbf{w})\|_* &= \left\| -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i \right\|_* \leq -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i) \|\mathbf{x}_i\|_* \leq D \mathcal{G}(\mathbf{w}). \\ 2557 \\ 2558 \end{aligned}$$

2559 (b) For exponential loss, $\frac{\mathcal{G}(\mathbf{w})}{\mathcal{L}(\mathbf{w})} = 1$. For logistic loss, the lower bound $\frac{\mathcal{G}(\mathbf{w})}{\mathcal{L}(\mathbf{w})} \geq 1 - \frac{N\mathcal{L}(\mathbf{w})}{2}$ follows
 2560 from [Zhang et al. \(2024a, Lemma C.7\)](#). The upper bound follows from the elementary inequality
 2561 $-\ell'_{\log}(z) = \frac{\exp(-z)}{1+\exp(-z)} \leq \log(1 + \exp(-z)) = \ell_{\log}(z)$ for all $z \in \mathbb{R}$.
 2562

2563 (c) For exponential loss, the equality holds. For logistic loss, the elementary inequality $-\ell'_{\log}(z) = \frac{\exp(-z)}{1+\exp(-z)} \geq \frac{\exp(-z)}{(1+\exp(-z))^2} = \ell''_{\log}(z)$ for all $z \in \mathbb{R}$, which results in
 2564

$$\begin{aligned} 2565 \mathcal{G}(\mathbf{w}) &= -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i) \geq \frac{1}{N} \sum_{i \in [N]} \ell''(\mathbf{w}^\top \mathbf{x}_i). \\ 2566 \\ 2567 \end{aligned}$$

2568 (d) First, for exponential loss, $-\ell'_{\exp}(z') = -\exp(z - z') \ell'_{\exp}(z) \leq -\exp(|z' - z|) \ell'_{\exp}(z)$, and
 2569 for logistic loss, $-\ell'_{\log}(z') = \frac{\exp(z)+1}{\exp(z')+1} \ell'_{\log}(z) \leq -\exp(|z' - z|) \ell'_{\log}(z)$ hold for any $z, z' \in \mathbb{R}$.
 2570 By duality, we get
 2571

$$\begin{aligned} 2572 \mathcal{G}(\mathbf{w}') &= -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}'^\top \mathbf{x}_i) = -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i + (\mathbf{w}' - \mathbf{w})^\top \mathbf{x}_i) \\ 2573 &\leq -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i) \exp(|(\mathbf{w}' - \mathbf{w})^\top \mathbf{x}_i|) \\ 2574 &\leq -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i) \exp(\|\mathbf{w}' - \mathbf{w}\| \|\mathbf{x}_i\|_*) \\ 2575 &\leq -\frac{1}{N} \sum_{i \in [N]} \ell'(\mathbf{w}^\top \mathbf{x}_i) \exp(D \|\mathbf{w}' - \mathbf{w}\|) \\ 2576 &= e^{D\|\mathbf{w}' - \mathbf{w}\|} \mathcal{G}(\mathbf{w}). \\ 2577 \\ 2578 \end{aligned}$$

2579 \square
 2580
 2581

1.2 PROPERTIES OF LOSS FUNCTIONS

2582 **Lemma I.2** (Lemma C.4 in [Zhang et al. \(2024a\)](#)). *For $\ell \in \{\ell_{\exp}, \ell_{\log}\}$, either $\mathcal{G}(\mathbf{w}) < \frac{1}{2n}$ or
 2583 $\mathcal{L}(\mathbf{w}) < \frac{\log 2}{n}$ implies $\mathbf{w}^\top \mathbf{x}_i > 0$ for all $i \in [N]$.*

2592
2593**Lemma I.3** (Lemma C.5 in Zhang et al. (2024a)). *For $\ell \in \{\ell_{\exp}, \ell_{\log}\}$ and any $z_1, z_2 \in \mathbb{R}$, we have*2594
2595
2596

$$\left| \frac{\ell'(z_1)}{\ell'(z_2)} - 1 \right| \leq e^{|z_1 - z_2|} - 1.$$

2597
2598**Lemma I.4** (Lemma C.6 in Zhang et al. (2024a)). *For $\ell \in \{\ell_{\exp}, \ell_{\log}\}$ and any $z_1, z_2, z_3, z_4 \in \mathbb{R}$, we have*2599
2600
2601

$$\left| \frac{\ell'(z_1)\ell'(z_3)}{\ell'(z_2)\ell'(z_4)} - 1 \right| \leq (e^{|z_1 - z_2|} - 1) + (e^{|z_3 - z_4|} - 1) + (e^{|z_1 + z_3 - z_2 - z_4|} - 1).$$

2602
2603**Lemma I.5.** *For $a > 1$ and $z_1, z_2 > 0$, if $\ell_{\log}(z_1) \leq a\ell_{\log}(z_2)$, then $z_1 \geq z_2 - \log(2^a - 1)$.*

2604

Proof. Note that2605
2606

$$\log(1 + e^{-z_1}) \leq a \log(1 + e^{-z_2}) \implies e^{-z_1} \leq (1 + e^{-z_2})^a - 1,$$

2607
2608
2609
2610and define $f(x) = \frac{(1+x)^a - 1}{x}$. Since f is an increasing function on the interval $(0, 1)$, we get $\sup_{x \in (0, 1)} f(x) = f(1) = 2^a - 1$. This implies $(1+x)^a - 1 \leq (2^a - 1)x$ for $x \in (0, 1)$. Since $z_1, z_2 > 0$, it satisfies $e^{-z_1}, e^{-z_2} \in (0, 1)$. Therefore, we get

2611

$$e^{-z_1} \leq (1 + e^{-z_2})^a - 1 \leq (2^a - 1)e^{-z_2}.$$

2612
2613By taking the natural logarithm of both sides, we get the desired inequality. \square 2614
2615

I.3 AUXILIARY RESULTS

2616
2617**Lemma I.6** (Lemma C.1 in Zhang et al. (2024a)). *The learning rate $\eta_t = (t+2)^{-a}$ with $a \in (0, 1]$ satisfies Assumption 2.3.*2618
2619**Lemma I.7** (Bernoulli's Inequality). (a) *If $r \geq 1$ and $x \geq -1$, then $(1+x)^r \geq 1 + rx$.*

2620

(b) *If $0 \leq r \leq 1$ and $x \geq -1$, then $(1+x)^r \leq 1 + rx$.*2621
2622
2623**Lemma I.8** (Stolz-Cesaro Theorem). *Let $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$ be the two sequences of real numbers. Assume that $(b_n)_{n \geq 1}$ is strictly monotone and divergent sequence and the following limit exists:*2624
2625

$$\lim_{n \rightarrow \infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n} = l.$$

2626
2627
2628
2629*Then it satisfies that*

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = l.$$

2630
2631**Lemma I.9** (Brouwer Fixed-point Theorem). *Every continuous function from a nonempty convex compact subset of \mathbb{R}^d to itself has a fixed point.*2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645