A Mirror Descent Perspective of Smoothed Sign Descent

Shuyang Wang[1] **Diego Klabjan**[2]

[1]Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, Illinois, USA
[2]Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois, USA

Abstract

The optimization dynamics of gradient descent for overparameterized problems can be viewed as low-dimensional dual dynamics induced by a mirror map, providing a mirror descent perspective on the implicit regularization phenomenon. However, the dynamics of adaptive gradient descent methods that are widely used in practice remain less understood. Meanwhile, empirical evidence of performance gaps suggests fundamental differences in their underlying dynamics. In this work, we introduce the dual dynamics of smoothed sign descent with stability constant ε for regression problems, formulated using the mirror descent framework. Unlike prior methods, our approach applies to algorithms where update directions deviate from true gradients such as ADAM. We propose a mirror map that reveals the equivalent dual dynamics under some assumptions. By studying dual dynamics, we characterize the convergent solution as approximately minimizing a Bregman divergence style function closely related to the $l_{3/2}$ norm. Furthermore, we demonstrate the role of the stability constant ε in shaping the convergent solution. Our analyses offer new insights into the distinct properties of the smoothed sign descent algorithm, and show the potential of applying the mirror descent framework to study complex dynamics beyond gradient descent.

1 INTRODUCTION

Mirror descent (MD) is an optimization method that extends gradient descent (GD) beyond Euclidean geometries [Nemirovskij and Yudin, 1983]. Central to the MD framework is a mirror map that facilitates transformation between a primal space where iterates exist and a dual space where updates are performed. By defining an appropriate mirror map, MD can adapt to the geometry of the problem for efficient optimization. Since its introduction, MD has attracted considerable research interest in its regularization properties and has motivated development of efficient optimization algorithms [Beck and Teboulle, 2003, Radhakrishnan et al., 2020, Azizan et al., 2021, Gunasekar et al., 2021, Sun et al., 2022, 2023].

Recent studies reveal the power of adopting an MD perspective to interpret the optimization dynamics of GD for overparameterized problems [Woodworth et al., 2020, Li et al., 2022]. In an overparameterized setting, the number of parameters exceeds the number of examples, resulting in an underdetermined system and infinitely many solutions. This becomes an important setting for analyzing the behavior of optimization algorithms and characterizing the particular solutions they converge to among all solutions [Allen-Zhu et al., 2019, Oymak and Soltanolkotabi, 2019]. Given a parameterization of a problem, Woodworth et al. [2020], Li et al. [2022] formulate mirror maps that establish equivalence between GD dynamics and low-dimensional MD dynamics. The simplified dual dynamics lead to a characterization of the convergent solution among all solutions in terms of the Bregman divergence. Specifically, the convergent GD solution minimizes the Bregman divergence from the starting point. This method is further used to analyze the effects of the initialization shape [Azulay et al., 2021] and stochasticity [Pesme et al., 2021] on the convergent solution.

Such results have been shown on data where optimal solutions are easy to find, yet the underlying optimization dynamics are nontrivial. While the underlying dynamics for gradient descent have been examined, the dual dynamics for other popular gradient based methods remain less understood. The MD framework provides a powerful and elegant tool for analyzing high-dimensional optimization dynamics, however, the existence of such mirror maps is highly dependent on both the problem parameterization and the optimization algorithm. Existing analyses do not apply

to algorithms beyond (stochastic) GD. The challenges arise from both the formulation of a mirror map and the analysis of dual dynamics. For example, for adaptive gradient methods with coordinate-wise adaptive learning rates, the update directions deviate from the true gradients. The adaptivity alters the underlying dynamics, breaking the low-dimensional structure seen in GD and rendering existing approaches inapplicable. Our work addresses this limitation and proposes a method to apply the MD framework when update directions do not follow true gradients.

Among adaptive gradient descent methods, we examine a prototypical algorithm, smoothed sign descent, which can be viewed as a smoothed version of sign descent with a stability constant ε. Recent work shows a deep connection between smoothed sign descent and popular optimizers such as ADAM and RMSProp [Kunstner et al., 2023, Ma et al., 2022, Balles and Hennig, 2018, Bernstein et al., 2018]. While sign descent has been studied as a proxy to understand the dynamics of more complex adaptive gradient methods [Ma et al., 2023, Balles et al., 2020], studies [Wang et al., 2021, 2022] show that the stability constant plays a key role in determining the convergence direction for classification problems. This highlights the importance of studying smoothed sign descent and investigating the effect of the stability constant ε, which has been underexplored in literature. We study the dynamics of smoothed sign descent for a quadratically parameterized regression problem. Our results reveal dual dynamics that are distinct from those for GD, and explicitly characterize the relationship between the stability constant ε and the convergent solution.

In this work, we present an analysis of MD to interpret the optimization dynamics of smoothed sign descent. We identify an initial stage unique to smoothed sign descent, which allows us to formulate a mirror map for the main stage of the dynamics. Using the mirror map, we project the complex primal dynamics onto the dual space with a simplified structure. We further decompose the dual dynamics into a sign descent stage and a convergence stage. The dual dynamics interpretation enables us to connect the convergent solution to the approximate minimizer of a Bregman divergence style function closely related to the $l_{3/2}$-norm. Further analysis reveals the effect of the stability constant ε on reducing the deviation from the exact minimizer, corroborating the empirical findings on the sensitivity of the training and testing performance to the stability constant [De et al., 2018, Liu et al., 2020, Choi et al., 2019].

Our analysis introduces a three-stage decomposition of the complex dynamics, where each phase exhibits distinct characteristics. By carefully studying the behavior within each phase, we establish unique regularization properties of the convergent solution. However, to make the analysis of the underlying coupled nonlinear ODE system tractable, we adopt simplifying assumptions that may limit the direct applicability of our results to real-world settings.

Our contributions are as follows.

- We introduce the dual dynamics of smoothed sign descent for a quadratically parameterized regression problem using the MD framework.
- We show that after an initial stage, the dual dynamics begin a sign descent stage characterized by approximately linear growth with similar rates in all coordinates, and then transition into a convergence stage characterized by diminishing magnitude of gradients.
- We prove that the convergent solution approximately satisfies the KKT conditions for minimizing a Bregman divergence style function, in contrast to the already known exact Bregman divergence minimization property of GD dynamics. The convergent solution found by smoothed sign descent is the one that approximately minimizes the Bregman divergence style function from the starting point.
- We theoretically analyze the effect of the stability constant ε on bounding the deviation from the exact minimizer, emphasizing the benefit of tuning the stability constant.

In Section 2, we review previous research on the properties of MD and smoothed sign descent. In Section 3, we present our main results, including the formulation of dual dynamics and the characterization of convergent solutions. We conclude the paper in Section 4.

2 RELATED WORK

Recent works apply the MD framework to interpret dynamics of neural network training. The study [Woodworth et al., 2020] discovers the equivalent low-dimensional MD dynamics for the optimization dynamics of GD for overparameterized models, focusing on the effect of initialization scale. However, extending their methodology to more general cases remains a challenge. Li et al. [2022] identify a commutative property of neural network parameterization that enables the formulation of equivalent MD dynamics. Pesme et al. [2021] use a time-varying mirror map for stochastic GD and show the benefit of stochasticity for inducing sparsity of the convergent solution. Azulay et al. [2021] propose a warping technique to study the effect of the initialization shape on the equivalent MD dynamics of GD. We contribute to this line of research dealing with strict gradients by extending the framework beyond GD to a case where the adaptive learning rate breaks the gradient structure and showing distinct properties of the dual dynamics.

Research on regularization properties of MD algorithms dates back to the work [Beck and Teboulle, 2003], which reveals a local regularization effect in terms of Bregman divergence at each iteration. Recent study [Gunasekar et al., 2018] shows that MD converges to the solution that mini-

mizes the associated Bregman divergence from the starting point among all solutions. Subsequent works [Azizan and Hassibi, 2019, Azizan et al., 2021] extend this analysis to stochastic MD for nonlinear models and prove the Bregman divergence minimization property. Research so far primarily focuses on standard MD settings, where the dynamics follow the gradient directions in the dual space. In contrast, we study the case where the dual dynamics deviate from the gradients. We show that the convergent solution of smoothed sign descent satisfies the approximate KKT condition of minimizing a Bregman divergence style function by bounding the cumulative deviation.

The stability constant ε, designed to ensure numerical stability for algorithms such as ADAM and RMSProp, is typically set to a negligible value by default. Its impact on optimization dynamics is underexplored. De et al. [2018] experiment with different values of ε for ADAM and RMSProp and observe that training and testing performance is sensitive to ε. Studies [Nado et al., 2020, Liu et al., 2020, Choi et al., 2019] also provide empirical evidence supporting the benefit of tuning the stability constant ε. Yuan and Gao [2020] study the effect of modifying the location of ε in ADAM and propose an alternative optimizer to improve performance. We provide a theoretical justification for tuning the stability constant ε by explicitly showing its role in reducing the KKT error of the convergent solution. Carmon and Hinder [2022] introduce an algorithm for stochastic convex optimization, and show the role of ε in the regret bound, while our work reveals the role of ε in shaping the solution found by smoothed sign descent.

Our work also contributes an MD perspective to the ongoing discussion on the implicit regularization phenomenon in neural network training [Neyshabur et al., 2014, Zhang et al., 2021]. While many studies [Soudry et al., 2018, Arora et al., 2019, Lyu and Li, 2020] focus on GD, fewer have investigated adaptive gradient methods despite the performance gap observed in the paper [Wilson et al., 2017]. Notably, studies [Wang et al., 2021, 2022] find that ADAM achieves the same convergent direction as GD in classification problems, while we prove a distinct regularization property for smoothed sign descent compared to GD in regression problems. Recent study [Xie and Li, 2024] characterizes the convergent solution of AdamW as training time approaches infinity. In contrast, we characterize the entire dynamics of smoothed sign descent by formulating the equivalent dual dynamics which reveal an intrinsically simplified structure. We propose a three-stage decomposition of the dual dynamics that enables an in-depth analysis of the optimization dynamics. A related but different two-stage transition is observed empirically by [Ma et al., 2022] when optimizing a squared loss with Adam for fully connected neural networks, which exhibits an initial phase of fast convergence followed by oscillations, spikes, or a diverging pattern. While the initial phase is similar to our sign descent stage with sufficient update across all coordinates, we reveal a different behavior in the latter stages of convergence.

3 DUAL DYNAMICS OF SMOOTHED SIGN DESCENT

3.1 BACKGROUND

Let us consider the update rule of GD for minimizing a loss function $L(\boldsymbol{\beta})$ with step size $\eta > 0$:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta \nabla L(\boldsymbol{\beta}_t). \tag{1}$$

We suppose that the iterates $\boldsymbol{\beta}_t$ lie in the Euclidean space $\mathbb{R}^D$. Formally, the gradients $\nabla L(\boldsymbol{\beta}_t)$ lie in the dual space $\mathbb{R}^D$. In GD, we obtain the updated point by directly taking a linear combination of the iterate and the gradient as in (1). MD, however, formally distinguishes the primal and the dual spaces using a mirror map to transform between them. A mirror map $\nabla\Phi : \mathbb{R}^D \to \mathbb{R}^D$ is defined as the gradient of a potential function $\Phi : \mathbb{R}^D \to \mathbb{R}$, which is a differentiable and strictly convex function. The mirror map $\nabla\Phi$ maps the primal variable $\boldsymbol{\beta}$ to the dual variable denoted by $\boldsymbol{\phi} \in \mathbb{R}^D$. Each iteration of MD for minimizing $L(\boldsymbol{\beta})$ follows the following steps, where the step size $\eta > 0$:

$$\boldsymbol{\phi}_t = \nabla\Phi(\boldsymbol{\beta}_t) \tag{2}$$

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t - \eta \nabla L(\boldsymbol{\beta}_t) \tag{3}$$

$$\boldsymbol{\beta}_{t+1} = (\nabla\Phi)^{-1}(\boldsymbol{\phi}_{t+1}). \tag{4}$$

By plugging in (2), we can rewrite the MD update (3) in the dual space as:

$$\nabla\Phi(\boldsymbol{\beta}_{t+1}) = \nabla\Phi(\boldsymbol{\beta}_t) - \eta \nabla L(\boldsymbol{\beta}_t). \tag{5}$$

In the continuous-time limit when $\eta \to 0$, we get the **dual dynamics** of $\boldsymbol{\beta}(t)$:

$$\frac{d\nabla\Phi(\boldsymbol{\beta}(t))}{dt} = -\nabla L(\boldsymbol{\beta}(t)). \tag{6}$$

A key element of MD is the Bregman divergence that serves as the notion of measuring the distance between two points in the primal space.

Definition 3.1 (Bregman divergence)**.** *For $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^D$, the Bregman divergence associated with a potential function Φ from $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_2$ is defined as*

$$D_\Phi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \Phi(\boldsymbol{\beta}_1) - \Phi(\boldsymbol{\beta}_2) - \langle \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2, \nabla\Phi(\boldsymbol{\beta}_2)\rangle. \tag{7}$$

Bregman divergence generalizes squared Euclidean distance and captures different geometric structure of the space through the choice of Φ. When $\Phi(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2$, the associated Bregman divergence reduces to the squared Euclidean distance, the mirror map $\nabla\Phi$ becomes an identity map, and MD simplifies to GD.

3.2 PROBLEM SETUP

We suppose that there are N examples with $D > N$ features $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1,...,N}$, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^D, y^{(i)} \in \mathbb{R}$. Let us denote the data matrix by $X \in \mathbb{R}^{N\times D}$, where each row is $\boldsymbol{x}^{(i)}$, and denote the labels of the examples by $\boldsymbol{y} \in \mathbb{R}^N$. The Hadamard product is denoted by $\odot$. We consider a regression problem of minimizing the following loss function with respect to $\boldsymbol{w} := \begin{bmatrix}\boldsymbol{w}^+\\ \boldsymbol{w}^-\end{bmatrix} \in \mathbb{R}^{2D}$, where $\boldsymbol{w}^+, \boldsymbol{w}^- \in \mathbb{R}^D$:

$$L(\boldsymbol{w}) = \frac{1}{4}\left(X\left(\boldsymbol{w}^+ \odot \boldsymbol{w}^+ - \boldsymbol{w}^- \odot \boldsymbol{w}^-\right) - \boldsymbol{y}\right)^\top \left(X\left(\boldsymbol{w}^+ \odot \boldsymbol{w}^+ - \boldsymbol{w}^- \odot \boldsymbol{w}^-\right) - \boldsymbol{y}\right). \quad (8)$$

We let $\boldsymbol{\beta} := \boldsymbol{w}^+ \odot \boldsymbol{w}^+ - \boldsymbol{w}^- \odot \boldsymbol{w}^- \in \mathbb{R}^D$ denote the regression parameter, and $L(\boldsymbol{\beta}) = \frac{1}{4}\left(X\boldsymbol{\beta} - \boldsymbol{y}\right)^\top \left(X\boldsymbol{\beta} - \boldsymbol{y}\right)$ is the standard quadratic loss. This parameterization of $\boldsymbol{\beta}$ by $\boldsymbol{w}$ can also be viewed as a 2-layer diagonal linear neural network with weights $\boldsymbol{w} \in \mathbb{R}^{2D}$ (see Section 4 of the paper [Woodworth et al., 2020] for a detailed study of the model). Despite its simplicity, this setup has been used to prove numerous insightful results for neural networks training [Woodworth et al., 2020, Pesme et al., 2021, Nacson et al., 2022, Vivien et al., 2022].

When GD is applied to minimize loss (8) with respect to $\boldsymbol{w}$, from the GD update rule with infinitesimal step size η we get

$$\frac{d\boldsymbol{w}^+(t)}{dt} = -\nabla_{\boldsymbol{w}^+} L(\boldsymbol{w}(t)), \quad (9)$$

$$\frac{d\boldsymbol{w}^-(t)}{dt} = -\nabla_{\boldsymbol{w}^-} L(\boldsymbol{w}(t)). \quad (10)$$

Using the chain rule, we get the optimization dynamics of $\boldsymbol{\beta}(t)$:

$$\frac{d\boldsymbol{\beta}(t)}{dt} = -2\boldsymbol{w}^+(t) \odot \nabla_{\boldsymbol{w}^+} L(\boldsymbol{w}(t)) + 2\boldsymbol{w}^-(t) \odot \nabla_{\boldsymbol{w}^-} L(\boldsymbol{w}(t)). \quad (11)$$

Previous work [Woodworth et al., 2020] shows that by defining a potential function:

$$\Psi_\alpha(\boldsymbol{\beta}) := \frac{1}{4}\left(\sum_{i=1}^{D} \beta_i \operatorname{arcsinh}\left(\frac{\beta_i}{2\alpha^2}\right) - \sqrt{\beta_i^2 + 4\alpha^4}\right), \quad (12)$$

where $\alpha > 0$ is the initialization scale, we can project the dynamics (11) onto the dual space using the mirror map $\nabla\Psi_\alpha$. Here the gradient is taken with respect to $\boldsymbol{\beta}$. By derivation in Appendix C, it follows that the dual dynamics are given by:

$$\frac{d\nabla\Psi_\alpha(\boldsymbol{\beta}(t))}{dt} = -\nabla_{\boldsymbol{\beta}} L(\boldsymbol{\beta}(t)). \quad (13)$$

Since (11) and (13) are equivalent, in the continuous-time limit, the evolution of $\boldsymbol{\beta}(t)$ using GD can be interpreted as following the MD algorithm (2)-(4) with mirror map $\nabla\Psi_\alpha$.

The dual dynamics (13) reveal an intrinsically low-dimensional structure of the dynamics of $\boldsymbol{\beta}(t)$ in the over-parameterized setting where $N < D$. Specifically, the gradients $\nabla_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$ in the right-hand side of (13) are confined in a subspace span$\{\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)}\}$, which has dimension of at most N. Furthermore, by analyzing the dual dynamics, previous work [Woodworth et al., 2020] proves that the convergent solution $\boldsymbol{\beta}^\infty := \lim_{t\to\infty} \boldsymbol{\beta}(t)$ satisfies the KKT conditions of the constrained optimization problem:

$$\boldsymbol{\beta}^\infty = \underset{\boldsymbol{\beta}\in\mathbb{R}^D \text{ s.t. } X\boldsymbol{\beta}=\boldsymbol{y}}{\operatorname{argmin}} D_{\Psi_\alpha}(\boldsymbol{\beta}, \boldsymbol{\beta}(0)). \quad (14)$$

In this work, we study the dynamics of smoothed sign descent for minimizing (8). For smoothed sign descent, the weights are updated according to

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \cdot \frac{\nabla_{\boldsymbol{w}} L(\boldsymbol{w}_t)}{|\nabla_{\boldsymbol{w}} L(\boldsymbol{w}_t)| + \varepsilon\mathbf{1}}, \quad (15)$$

where $\varepsilon > 0$ is the stability constant and the operations are taken element-wise. Smoothed sign descent can be viewed as an adaptive gradient method with coordinate-wise adaptive learning rate $\eta_{i,t} = \frac{\eta}{|[\nabla_{\boldsymbol{w}} L(\boldsymbol{w}_t)]_i| + \varepsilon}$ for each i. The magnitude of the gradient can differ vastly across all coordinates, and thus, the update in each coordinate is scaled differently. As a result, the update direction no longer follows the opposite of the true gradient, unlike normalized gradient descent, which preserves the direction by applying a uniform normalization scale to all coordinates.

We suppose that the weights are initialized by $\boldsymbol{w}(0) = \alpha\mathbf{1}$, $\alpha > 0$. In the continuous-time limit, the dynamics of the weights become

$$\frac{d\boldsymbol{w}(t)}{dt} = -\frac{\nabla_{\boldsymbol{w}} L(\boldsymbol{w}(t))}{|\nabla_{\boldsymbol{w}} L(\boldsymbol{w}(t))| + \varepsilon\mathbf{1}}. \quad (16)$$

This yields the dynamics of the regression parameter $\boldsymbol{\beta}(t)$ as follows, with $\boldsymbol{\beta}(0) = \mathbf{0}$:

$$\frac{d\boldsymbol{\beta}(t)}{dt} = -2\boldsymbol{w}^+(t) \odot \frac{\nabla_{\boldsymbol{w}^+} L(\boldsymbol{w}(t))}{|\nabla_{\boldsymbol{w}^+} L(\boldsymbol{w}(t))| + \varepsilon\mathbf{1}} + 2\boldsymbol{w}^-(t) \odot \frac{\nabla_{\boldsymbol{w}^-} L(\boldsymbol{w}(t))}{|\nabla_{\boldsymbol{w}^-} L(\boldsymbol{w}(t))| + \varepsilon\mathbf{1}}. \quad (17)$$

With coordinate-wise adaptive learning rate, the update direction deviates from the true gradients and the mirror map $\nabla\Psi_\alpha$ for GD no longer holds. It leads to two interesting questions:

1. Can we formulate a mirror map to show equivalent dual dynamics for (17)?
2. Can we use the dual dynamics to characterize the convergent solution among all solutions?

3.3 MAIN RESULTS

In this section, we present our answers to the two questions. Our results consist of three parts. In Propositions 3.6 and 3.7, we construct a mirror map and formulate the dual dynamics for smoothed sign descent. In Theorem 3.9 and Corollary 3.11, we prove a characterization of the convergent solution. In Corollaries 3.10 and 3.12, we further reveal the role of the stability constant in the convergent solution.

The weight dynamics (16) form a coupled system of nonlinear ODEs, with the stability constant ε adding another layer of complexity. By the Picard-Lindelof theorem, there exists a unique solution to (16). Solving this ODE system analytically is intractable. We make the following assumption to decouple the nonlinear ODE system into N autonomous systems. This decomposition allows us to analyze the interactions among an arbitrary number of dimensions within each system.

Assumption 3.2. *We assume that $y^{(n)}$ are non-zero, and that there exists a permutation of the columns of X such that $X^\top X$ is block-diagonal with N rank-1 blocks denoted by $B^{(n)} \in \mathbb{R}^{D_n \times D_n}$ for $n = 1, \dots, N$.*

It is easy to see that this condition is equivalent to requiring that each row of X has $D_n \geq 1$ non-zero elements denoted by $x_1^{(n)}, \dots, x_{D_n}^{(n)}$, where $\sum_{n=1}^{N} D_n = D$. While this assumption yields an easy optimization problem in the primal space, the dynamics of smoothed sign descent are very complex and intriguing.

We require the stability constant ε to be small relative to components of the initial gradient so that it does not overshadow the essential behavior of the dynamics as a smoothed version of sign descent. We notice that $\boldsymbol{w} = \mathbf{0}$ is a stationary point of the weight dynamics (16). Since the weights are initialized as $\boldsymbol{w}(0) = \alpha \mathbf{1}$ where $\alpha > 0$, we assume that α is chosen not so small to avoid being stuck near a stationary point. Moreover, we also avoid choosing a large initial value α that would dominate the value of the weights and overshadow the convergent behavior.

Assumption 3.3. *We assume that for each $n \in \{1, \dots, N\}$ and $i \in \{1, \dots, D_n\}$, the stability constant ε and the initialization scale α satisfy:*

$$0 \leq \varepsilon \leq \frac{1}{9} \frac{|x_i^{(n)}||y^{(n)}|^{\frac{3}{2}}}{\sqrt{2\sum_{k=1}^{D_n} |x_k^{(n)}|}}, \tag{18}$$

$$\frac{9\varepsilon}{4\left|x_i^{(n)} y^{(n)}\right|} \leq \alpha \leq \frac{1}{3}\sqrt{\frac{|y^{(n)}|}{2\sum_{k=1}^{D_n} |x_k^{(n)}|}}. \tag{19}$$

3.3.1 Three Stages

We begin by studying the sign and monotonicity of $\boldsymbol{w}^+(t)$ and $\boldsymbol{w}^-(t)$ by the following lemma assuming they satisfy (16). Proofs of the results in this section can be found in Appendix A.

Proposition 3.4. *For each coordinate $i \in \{1, \dots, D\}$,*

- *$w_i^+(t)$ and $w_i^-(t)$ are always non-negative,*
- *if $w_i^+(0)' > 0$, then $w_i^+(t)' \geq 0$ and $w_i^-(t)' \leq 0$ for all t,*
- *if $w_i^+(0)' \leq 0$, then $w_i^+(t)' \leq 0$ and $w_i^-(t)' \geq 0$ for all t.*

For each i, based on this proposition, either $w_i^+(t)$ or $w_i^-(t)$ is monotonically non-decreasing. We denote the dominating weight that is monotonically non-decreasing by u_i, and we denote the one that is non-increasing by v_i, i.e.,

$$u_i(t) := \begin{cases} w_i^+(t) & \text{if } w_i^+(0)' > 0, \\ w_i^-(t) & \text{else,} \end{cases}$$

$$v_i(t) := \begin{cases} w_i^-(t) & \text{if } w_i^+(0)' > 0, \\ w_i^+(t) & \text{else.} \end{cases}$$

A key identity in the derivation of the mirror map for GD is that $w_i^+(t)w_i^-(t) = \alpha^2$ holds throughout the dynamics. However, this quantity is not conserved when coordinate-wise adaptivity is applied. In fact, we can show that $w_i^+(t)w_i^-(t) < \alpha^2$ for $t > 0$. The adaptive learning rate ensures similar rate of change across all coordinates, and enables sufficient updates even when the gradient magnitude is relatively small. In particular, this allows the non-dominating weight $v_i(t)$ to diminish to negligible values early on. Based on this observation, we identify an initial stage of the dynamics where $v_i(t)$ decreases to and remains below a value on the order ε across all coordinates. The following lemma also shows that this initial stage lasts no longer than $t = 2\alpha$.

Proposition 3.5. *There exists $T_0 \in (0, 2\alpha]$ such that for all $t \geq T_0$, $v_i(t) \leq \frac{2\varepsilon}{|x_i^{(n)} y^{(n)}|}$ for all i.*

The proof hinges on bounding the value of $v_i(t)$ from above when the gradient component $[\nabla_{\boldsymbol{v}} L(\boldsymbol{w}(t))]_i$ reaches ε at $t = t_i$. Before t_i, the absolute value of the derivative $|v_i'(t)|$ is always greater than $\frac{1}{2}$, ensuring rapid decreasing of $v_i(t)$. Meanwhile, the non-negativity of $v_i(t)$ by Proposition 3.4 guarantees that the rapid decreasing stage lasts no longer than 2α. Based on the expression $[\nabla_{\boldsymbol{v}} L(\boldsymbol{w}(t))]_i = v_i(t)|x_i^{(n)} r^{(n)}(t)|$, we continue to bound the residual $|r^{(n)}(t)|$ from below using the maximal growth of $u_i(t)$ during this short time period. Finally, the lower bound of $|r^{(n)}(t_i)|$ leads to the upper bound of $v_i(t_i)$ at t_i. We complete the proof by letting T_0 be the largest t_i across all coordinate i.

During the initial stage, both $\boldsymbol{u}(t)$ and $\boldsymbol{v}(t)$ follow sign descent approximately, which allows us to approximate the

primal dynamics of $\boldsymbol{\beta}(t)$ by sign descent. After T_0, the dynamics of $\boldsymbol{\beta}(t)$ transition into the main stage, where $\boldsymbol{v}(t)$ remains small and the magnitude of $\boldsymbol{\beta}(t)$ is denominated by $\boldsymbol{u}(t)$. While the primal dynamics become complex, we formulate a mirror map so that the dual dynamics have a simplified structure that closely aligns with the sign of $\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))$.

Proposition 3.6 (Dual dynamics of smoothed sign descent)**.** *For $t > 0$, we define a potential function $\Phi_t(\boldsymbol{\beta}) = \frac{2}{3}\sum_{i=1}^{D}\left(|\beta_i| + v_{i,t}^2\right)^{\frac{3}{2}}$. The induced mirror map $\nabla\Phi_t : \mathbb{R}^D \to \mathbb{R}^D$ maps $\boldsymbol{\beta}(t)$ to the dual space. The dynamics in the dual space follow*

$$\frac{d\nabla\Phi_t(\boldsymbol{\beta}(t))}{dt} = -\operatorname{sgn}(\boldsymbol{\beta}(t)) \odot \frac{\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))}{|\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))| + \varepsilon \mathbf{1}}. \quad (20)$$

The potential function is time-varying with a time-dependent parameter $v_{i,t} := v_i(t)$. Pesme et al. [2021] also employ a time-varying potential function to construct a mirror map for the dynamics of stochastic GD. Radhakrishnan et al. [2020] conduct a thorough analysis of the convergence of MD with time-dependent mirrors. For $t \geq T_0$, since the non-dominating weights $v_i(t)$ diminish to small values by Proposition 3.5, the potential function has a close connection with the $l_{3/2}$-norm of $\boldsymbol{\beta}(t)$, in contrast with the potential function (12) for GD.

The dual dynamics (20) indeed reveal a greatly simplified structure compared to the primal dynamics (17). The right-hand side of the original dynamics (17) evolves in a complex way in the D-dimensional space as the weights are updated, while the right-hand side of the dual dynamics (20) reduces to two components, a sign vector and a vector approximating the sign of the gradient. The simplified structure enables us to understand the complex dynamics (17) by studying the evolution of the two sign vectors. However, the formulation of the dual dynamics (20) differs from standard MD dynamics (6) where the updates in the dual space align with the gradients exactly. The alignment has allowed previous work to show that the convergent solution satisfies the KKT conditions for Bregman divergence minimization as in (14). Therefore, further analysis of the dual dynamic (20) is required to understand the deviation from following the true gradients.

Proposition 3.7. *There exists $T > T_0$ such that we can divide the dynamics into two stages:*

- *Sign descent stage: for $t \in [T_0, T)$, $|\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))|_i > \varepsilon$ for all i,*
- *Convergence stage: for $t \in [T, \infty)$, $\min_i |\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))|_i \leq \varepsilon$.*

At the beginning, the dual dynamics resemble sign descent when gradient components are relatively large compared to ε. The stability constant comes into effect when $|\nabla_{\boldsymbol{u}} L(\boldsymbol{w})|_i$ becomes small. In Proposition 3.7, we prove the transition between the two stages by studying the evolution of the magnitude of each gradient component. Importantly, Proposition 3.7 shows that once a gradient value reaches ε, it remains small for the duration of the dynamics. The dynamics then enter a convergence stage with diminishing magnitude of gradients. Eventually, the dynamics approximate the direction of $\nabla_{\boldsymbol{u}} L(\boldsymbol{w})$ as all gradient components approach zero (see Lemma A.3 in Appendix A).

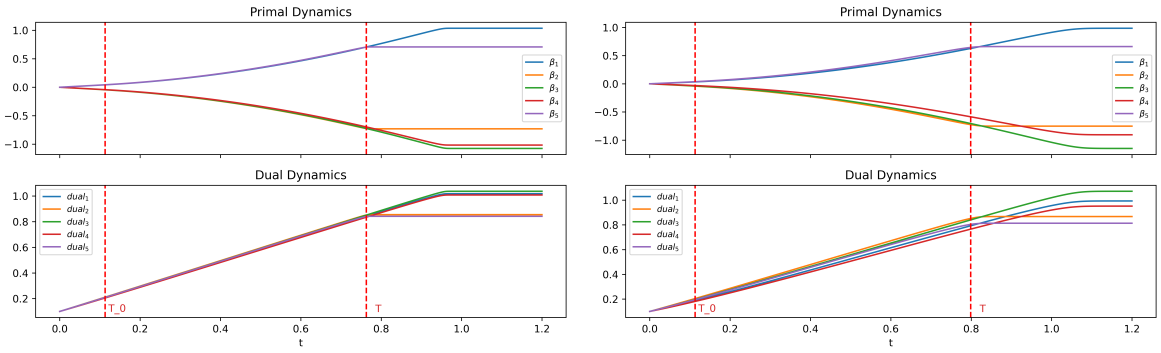
We illustrate the transition of the three stages in Figure 1. We randomly generate a dataset with $N = 2$ and $D = 5$ that satisfies Assumption 3.2 and set $\alpha = 0.1$. We simulate the dynamics (16) using the ODE solver in `SciPy` and visualize the evolution of primal and dual variables. In the experiments, T_0 is calculated as the value when $\max_i |\nabla_{\boldsymbol{v}} L(\boldsymbol{w}(t))|_i$ first becomes ε, while T is calculated as the value when $\min_i |\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))|_i$ first becomes ε. Based on smoothed sign descent (see (20)) and Proposition 3.7, we expect the change to be linear in $[T_0, T]$, and incoherent behavior in $[T, \infty)$. In the initial stage when $t < T_0$, we observe that the primal variable has linear change across all coordinates. During the sign descent stage when $T_0 \leq t < T$, the dual variable continues growing linearly with approximately uniform rate in all coordinates, while $\boldsymbol{\beta}(t)$ no longer changes linearly. After T, the dynamics enter the convergence stage, where the primal and dual variables gradually approach the convergent point. We also observe that the value of ε plays a key role in shaping the dynamics. For smaller ε, the dual variable follows the sign descent more closely and converges to values concentrated around two distinct points across all coordinates; while for larger ε, the dual variable shows greater dispersion across all coordinates. We quantify the relationship between the value of ε and the convergent solution in the following analysis.

3.3.2 Characterization of Convergent Solution by Bregman Divergence

The convergent solution of smoothed descent dynamics deviates from the exact KKT point of Bregman divergence minimization. However, we show that it satisfies the δ-KKT conditions for a Bregman divergence style function. In this section, we build on the results about stage transitions and conduct an in-depth analysis to quantify and bound the error δ. To emphasize the role of ε in bounding the error, we impose an additional assumption on the block-diagonal structure from Assumption 3.2.

Assumption 3.8. *We assume that each block $B^{(n)}$ of the block-diagonal matrix $X^\top X$ has size $D_n = 2$.*

The 2D block structure enables us to derive an explicit dependence of the bounds for δ on the stability constant ε, while keeping the overparameterization setting for smoothed



(a) $\varepsilon = 0.002$ (b) $\varepsilon = 0.01$

Figure 1: Evolution of primal variable $\boldsymbol{\beta}(t)$ and dual variable $\nabla\Phi_t(\boldsymbol{\beta}(t))$ in $\mathbb{R}^5$ of smoothed sign descent with different values of stability constant ε. The vertical line $t = T_0$ marks the transition from initial stage to the sign descent stage, and the line $t = T$ marks the transition to the convergence stage.

sign descent. By the spectral theorem, we can write $X^\top X = Q\Lambda Q^\top$ for an orthogonal matrix Q and a diagonal matrix Λ. The matrix Q is block-diagonal, where each block is expressed as a 2D rotation matrix parameterized by θ_n. We have

$$B^{(n)} = \begin{bmatrix} \cos\theta_n & -\sin\theta_n \\ \sin\theta_n & \cos\theta_n \end{bmatrix} \begin{bmatrix} \lambda_n & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos\theta_n & \sin\theta_n \\ -\sin\theta_n & \cos\theta_n \end{bmatrix}, \tag{21}$$

where $\lambda_n > 0$ and $\cos\theta_n, \sin\theta_n$ are non-zero by Assumption 3.2. Without loss of generality, we assume that $|\cos\theta_n| \geq |\sin\theta_n|$, which can be achieved by ordering the columns of X. We first present the result for $N = 1$ to illustrate the key findings and then generalize the results to $N > 1$. To this end, we show in Appendix A that there exists $\boldsymbol{v}^\infty = \lim_{t\to\infty} \boldsymbol{v}(t)$ and we let $\Phi_\infty(\boldsymbol{\beta}) = \frac{2}{3}\sum_{i=1}^D \left(|\beta_i| + (v_i^\infty)^2\right)^{\frac{3}{2}}$. By Proposition 3.5, we have $v_i^\infty = \mathcal{O}(\varepsilon)$ and $\Phi_\infty(\boldsymbol{\beta}) = \frac{2}{3}\sum_{i=1}^D |\beta_i|^{\frac{3}{2}} + \mathcal{O}(\varepsilon^2)$, i.e., approximately the $\frac{3}{2}$-th power of the $l_{3/2}$ norm.

We define a Bregman divergence style function E associated with the potential function Φ for smoothed sign descent by

$$E(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}}) := \Phi_\infty(\boldsymbol{\beta}) - \Phi_0(\bar{\boldsymbol{\beta}}) + \langle \nabla\Phi_0(\bar{\boldsymbol{\beta}}), \bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle. \tag{22}$$

We let $\boldsymbol{\beta}_0 := \boldsymbol{\beta}(0) = \mathbf{0}$ denote the starting point. Let us consider the constrained optimization problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^D \text{ s.t. } X\boldsymbol{\beta}=\boldsymbol{y}} E(\boldsymbol{\beta}, \boldsymbol{\beta}_0). \tag{23}$$

For $\delta \geq 0$, a solution $\boldsymbol{\beta}^*$ satisfies the δ-KKT conditions for (23) if $X\boldsymbol{\beta}^* = \boldsymbol{y}$ and there exists a scalar ν such that $\|\nabla_{\boldsymbol{\beta}} E(\boldsymbol{\beta}^*) - \nu\nabla_{\boldsymbol{\beta}}(X\boldsymbol{\beta}^*)\| \leq \delta$.

Theorem 3.9. *As $t \to \infty$, the regression parameter converges to an interpolating solution. We let $\boldsymbol{\beta}^\infty := \lim_{t\to\infty} \boldsymbol{\beta}(t)$, which exists by Lemma A.3 in Appendix A. We show that $\boldsymbol{\beta}^\infty$ satisfies the δ-KKT conditions for* (23) *with the error $\delta(\varepsilon)$ bounded by $\max\{|M_+|, |M_-|\}$, where*

$$M_+ = \left(|\cos\theta_1| - |\sin\theta_1|\right)\lambda_1^{-\frac{1}{4}}|y^{(1)}|^{\frac{1}{2}} + \mathcal{O}(\varepsilon),$$
$$M_- = \left(|\cos\theta_1| - |\sin\theta_1|\right)\left((2\lambda_1)^{-\frac{1}{4}}|y^{(1)}|^{\frac{1}{2}} - \alpha\right) + \mathcal{O}(\sqrt{\varepsilon}).$$

We present the main idea of the proof here and provide the full proof in Appendix B. The exact expressions for M_+, M_- can be found in the full proof. First, we observe the connection between the gradient of E and the integral of the dual dynamics (20) with respect to t. The dual dynamics structure enables us to calculate the deviation δ from satisfying the stationary condition using the dominating weights $\boldsymbol{u}^\infty$. Next, using an orthogonal projection, we reduce the problem to bounding the absolute value of $\Delta := |\cos\theta_1|\left(u_2^\infty - u_2(0)\right) - |\sin\theta_1|\left(u_1^\infty - u_1(0)\right)$. To bound Δ, we leverage the ratios between $u_1'(t)$ and $u_2'(t)$ in different stages of the dual dynamics, and focus on bounding the key quantity $u_2(T)$ at the transition between the two stages. During the sign descent stage, the leading terms of $u_1'(t)$ and $u_2'(t)$ are both 1 in the Taylor expansion at $\varepsilon = 0$, which guarantees a lower bound for $u_2(T)$. Being in the convergence stage, $u_1(t)$ dominates the growth, which allows us to derive an upper bound for u_2^∞. Finally, a lower bound for u_2^∞ leads to $\Delta \geq M_-$, while an upper bound leads to $\Delta \leq M_+$.

The derivation relies on the key quantity of $u_2(T)$ at the stage transition when the smallest gradient component reaches ε. The value of ε is crucial in determining the stage transition and it eventually affects the convergent solution. We further reveal the relationship between ε and the upper bound of δ in the following corollary. We provide the proof in Appendix B.

Corollary 3.10. *We let $\mathcal{I}_\varepsilon$ be the range of ε implied by*

Assumption 3.3. There exists a non-degenerate interval $\mathcal{I}' \subseteq \mathcal{I}_\varepsilon$ such that for all $\varepsilon \in \mathcal{I}'$,

$$\delta(\varepsilon) \leq \bar{M} - (|\cos\theta_1| - |\sin\theta_1|)\frac{\sqrt{2}\varepsilon}{4\lambda_1^{\frac{1}{2}}|y^{(1)}|}, \qquad (24)$$

where $\bar{M} := (|\cos\theta_1| - |\sin\theta_1|)\,\lambda_1^{-\frac{1}{4}}|y^{(1)}|^{\frac{1}{2}}$ is a quantity independent of ε.

The result highlights the role of ε in bounding the KKT error. Given a fixed dataset, while setting $\varepsilon = 0$ ensures a larger rate of change when the gradient magnitude becomes very small, the error is larger than that for smoothed sign descent with non-zero ε. Moreover, choosing a larger ε within a certain interval effectively shrinks the upper bound on the KKT error δ. It suggests that by using a proper value of ε, the dynamics can converge to a solution closer to the point with the E minimization property. Therefore, our result provides a theoretical ground for the benefit of tuning ε versus using a small default value for adaptive gradient methods.

Approximate KKT points are formally studied in [Andreani et al., 2011], which shows that when the KKT conditions are satisfied approximately, the point is close to solving the optimization problem. This concept is commonly used in practice such as in numerically solving an optimization problem. The iterative process is terminated after finding a solution satisfying approximate KKT conditions under a given tolerance of error δ, which can be justified by the approximate optimality of these points. Therefore, by showing a bound for the δ error, we establish the approximate optimality of the convergent solution.

To visualize the convergent solutions for different values of ε, we plot the trajectory of $\boldsymbol{\beta}(t)$ using randomly generated data with $N = 1$ and $D = 2$ in Figure 2. We note that as ε becomes larger, the convergent solution is closer to the solution with the minimal value of E to the initial point among all solutions. We also compute the value of E to the initial point for convergent solutions using different ε and plot the trend in Figure 3. The plot confirms that for larger ε, the convergent solutions have smaller values of $E(\boldsymbol{\beta}^\infty, \boldsymbol{\beta}_0)$.

Extension to $N > 1$. We generalize the results to the case when $N > 1$ in the following corollaries. The proofs can be found in Appendix B. We show that the convergent solution satisfies approximate KKT conditions of minimizing $E(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ among all solutions. Within a certain interval, a larger value of ε leads to a greater reduction of the KKT error. The implications for tuning the stability constant ε still hold.

Corollary 3.11. *For $N > 1$, let us suppose Assumption 3.8 is satisfied. As $t \to \infty$, the regression parameter converges to an interpolating solution $\boldsymbol{\beta}^\infty$ that satisfies the $\bar{\delta}$-KKT conditions for* (23) *with the error $\bar{\delta}(\varepsilon)$ bounded by*

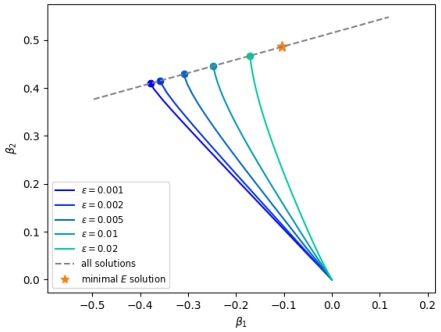


Figure 2: Trajectories of $\boldsymbol{\beta}(t)$ in $\mathbb{R}^2$ for different values of stability constant ε.

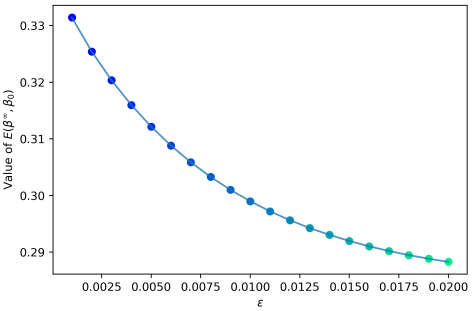


Figure 3: Bregman divergence style function value $E(\boldsymbol{\beta}^\infty, \boldsymbol{\beta}_0)$ of convergent solutions with different values of stability constant ε.

$\sum_{n=1}^{N} \max\left\{\left|M_+^{(n)}\right|, \left|M_-^{(n)}\right|\right\}$, *where*

$$M_+^{(n)} = (|\cos\theta_n| - |\sin\theta_n|)\,\lambda_n^{-\frac{1}{4}}|y^{(n)}|^{\frac{1}{2}} + \mathcal{O}(\varepsilon),$$
$$M_-^{(n)} = (|\cos\theta_n| - |\sin\theta_n|)\left((2\lambda_n)^{-\frac{1}{4}}|y^{(n)}|^{\frac{1}{2}} - \alpha\right) + \mathcal{O}(\sqrt{\varepsilon}).$$

Corollary 3.12. *There exists a non-degenerate interval $\mathcal{J} \subseteq \mathcal{I}_\varepsilon$ such that for all $\varepsilon \in \mathcal{J}$,*

$$\bar{\delta}(\varepsilon) \leq \sum_{n=1}^{N}(|\cos\theta_n| - |\sin\theta_n|)\left(\lambda_n^{-\frac{1}{4}}|y^{(n)}|^{\frac{1}{2}}\right) - \qquad (25)$$
$$\left(\sum_{n=1}^{N}(|\cos\theta_n| - |\sin\theta_n|)\frac{\sqrt{2}}{4\lambda_n^{\frac{1}{2}}|y^{(n)}|}\right)\varepsilon. \qquad (26)$$

In overparameterized regression problems, the dimension D is larger than the number of examples N, leading to infinitely many solutions. Our results establish that the solution found by smoothed sign descent approximately minimizes a measure of distance to the initial point related to the $l_{3/2}$-norm for quadratic parameterized models, and the error only scales with N.

Extension to Higher Order Models. Our analysis is generalizable to parameterizations with higher order $H \geq 2$ in weights, given by $\boldsymbol{\beta} = \boldsymbol{u}^H - \boldsymbol{v}^H$. Here s^H denotes applying Hadamard product H times on vector s. This parameterization can be interpreted as a diagonal linear neural network of depth H, as explained in [Woodworth et al., 2020]. The mirror map is induced by a potential function closely related to $l_{2-\frac{1}{H}}$-norm of $\boldsymbol{\beta}$, given by $\Phi_t^H(\boldsymbol{\beta}) := \sum_{i=1}^D \left(|\beta_i| + v_{i,t}^H\right)^{2-\frac{1}{H}}$, where $v_{i,t} = \mathcal{O}(\varepsilon)$. When the depth $H \to \infty$, the potential function approximates the squared l_2-norm.

4 CONCLUSION

In this work, we propose an MD perspective of the dynamics of smoothed sign descent for overparameterized regression problems. We extend existing results beyond GD to a case where update directions deviate from true gradients due to adaptivity, and formulate the equivalent dual dynamics with a simplified structure. We also study the role of the stability constant ε in bounding the deviation of the convergent solution from minimizing a Bregman divergence style function. The finding supports the benefit of tuning the stability constant ε. Future work may extend our analysis to widely used methods such as Adam, RMSProp, and AdaGrad. With additional approximations used for adapting the learning rates, further investigation is needed to understand the transition among the three stages and to analyze the impact of the stability constant on the convergent solution.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

Roberto Andreani, Gabriel Haeser, and José Màrio Martínez. On sequential optimality conditions for smooth constrained optimization. *Optimization*, 60(5):627–641, 2011.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. In *International Conference on Learning Representations*, 2019.

Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7717–7727, 2021.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477, 2021.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pages 404–413, 2018.

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569, 2018.

Yair Carmon and Oliver Hinder. Making SGD parameter-free. In *Conference on Learning Theory*, pages 2360–2389, 2022.

Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for RMSProp and Adam in non-convex optimization and an empirical comparison to Nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018.

Suriya Gunasekar, Blake Woodworth, and Nathan Srebro. Mirrorless mirror descent: A natural derivation of mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pages 2305–2313, 2021.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *International Conference on Learning Representations*, 2023.

Zhiyuan Li, Tianhao Wang, Jason D Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. *Advances in Neural Information Processing Systems*, 35:34626–34640, 2022.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Avery Ma, Yangchen Pan, and Amir massoud Farahmand. Understanding the robustness difference between stochastic gradient descent and adaptive gradient methods. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Chao Ma, Lei Wu, and E Weinan. A qualitative study of the dynamic behavior for adaptive gradient algorithms. In *Mathematical and Scientific Machine Learning*, pages 671–692, 2022.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pages 16270–16295, 2022.

Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960, 2019.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Linear convergence of generalized mirror descent with time-dependent mirrors. *arXiv preprint arXiv:2009.08574*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. *Advances in Neural Information Processing Systems*, 35:31089–31101, 2022.

Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24(393):1–58, 2023.

Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Conference on Learning Theory*, pages 2127–2159, 2022.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858, 2021.

Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does momentum change the implicit regularization on separable data? *Advances in Neural Information Processing Systems*, 35: 26764–26776, 2022.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673, 2020.

Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In *International Conference on Machine Learning*, 2024.

Wei Yuan and Kai-Xin Gao. EAdam optimizer: How ϵ impact Adam. *arXiv preprint arXiv:2011.02150*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A Mirror Descent Perspective of Smoothed Sign Descent (Supplementary Material)

Shuyang Wang[1] **Diego Klabjan**[2]

[1]Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, Illinois, USA
[2]Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois, USA

A PROOF OF RESULTS IN SECTION 3.3.1

Notation. Assumption 3.2 guarantees that the non-zero entries in X are non-overlapping across rows. Therefore, we can partition the index set $I = \{1, \dots, D\}$ into N disjoint subsets $I^{(1)}, \dots, I^{(N)}$ such that

$$I = \bigcup_{n=1}^{N} I^{(n)},\ I^{(n)} := \left\{ i \in [D] : x_i^{(n)} \neq 0 \right\}. \tag{27}$$

We define $\boldsymbol{w}^{+(n)}, \boldsymbol{w}^{-(n)}, \boldsymbol{\beta}^{(n)} \in \mathbb{R}^{D_n}$ as the subvectors of $\boldsymbol{w}^+$, $\boldsymbol{w}^-$ and $\boldsymbol{\beta}$ corresponding to the indices in $I^{(n)}$, respectively. Similarly, we define $\boldsymbol{g}^{+(n)}, \boldsymbol{g}^{-(n)}$ as subvectors of gradients $\nabla_{\boldsymbol{w}^+} L(\boldsymbol{w}), \nabla_{\boldsymbol{w}^-} L(\boldsymbol{w})$ corresponding to the indices in $I^{(n)}$. We let $\boldsymbol{w}^{(n)} := \begin{bmatrix} \boldsymbol{w}^{+(n)}, & \boldsymbol{w}^{-(n)} \end{bmatrix} \in \mathbb{R}^{2D_n}$ and $\boldsymbol{g}^{(n)} := \begin{bmatrix} \boldsymbol{g}^{+(n)}, & \boldsymbol{g}^{-(n)} \end{bmatrix} \in \mathbb{R}^{2D_n}$. The weight dynamics (16) can be decomposed into N autonomous ODE systems:

$$\frac{d\boldsymbol{w}^{(n)}(t)}{dt} = F^{(n)}\left(\boldsymbol{w}^{(n)}(t)\right) := -\frac{\boldsymbol{g}^{(n)}(t)}{\boldsymbol{g}^{(n)}(t) + \varepsilon \mathbf{1}}, \tag{28}$$

where $\boldsymbol{w}^{(n)}(0) = \alpha \mathbf{1}$ for each n. The residual for each n is defined by $r^{(n)}(t) := y^{(n)} - \sum_{i=1}^{D_n} x_i^{(n)} \beta_i^{(n)}(t)$. In this section, we prove the results for an arbitrary n. We omit the superscripts (n) when possible to simplify the notation.

A.1 PROOF OF PROPOSITION 3.4

Proof. For all $i = 1, \dots, D_n$, it is easy to see that $g_i^+(t) = -w_i^+(t) \cdot x_i \cdot r(t)$, $g_i^-(t) = w_i^-(t) \cdot x_i \cdot r(t)$. The dynamics follow $w_i^+(t)' = -\frac{g_i^+(t)}{|g_i^+(t)|+\varepsilon},\ w_i^-(t)' = -\frac{g_i^-(t)}{|g_i^-(t)|+\varepsilon}$.

First, we show that for all i, $w_i^+(t), w_i^-(t) \geq 0$ always hold. Suppose for contradiction that $w_i^+(t') < 0$ for some t'. Since $w_i^+(0) = w_i^-(0) = \alpha > 0$, by continuity of $w_i^+(t)$, there exists $t_0 \in (0, t')$ such that $w_i^+(t_0) = 0$ and $w_i^+(t_0)' < 0$. However, $w_i^+(t_0) = 0$ implies $g_i^+(t_0) = 0$ and $w_i^+(t_0)' = 0$. Therefore, $w_i^+(t)$ never changes sign and is always non-negative. Similarly, we can show that $w_i^-(t)$ is always non-negative.

Next, we show that for each i, if $w_i^+(0)' > 0$, then $w_i^+(t)' \geq 0, w_i^-(t) \leq 0$. Relation $w_i^+(0)' = \alpha x_i y > 0$ implies that $x_i y > 0$. Therefore, $x_i r(0) = x_i y > 0$. Let us suppose for contradiction that there exists $t' > 0$ such that $x_i r(t') < 0$. By continuity of $x_i r(t)$, there exists $t_0 \in (0, t')$ such that $x_i r(t_0) = 0$. Since $x_i \neq 0$ by assumption, we must have $r(t_0) = 0$. In turn, $x_j r(t_0) = 0$ and $g_j^+(t_0) = g_j^-(t_0) = 0$ for all $j = 1, \dots, D_n$. As a result, $F^{(n)}\left(\boldsymbol{w}^{(n)}(t_0)\right) = \mathbf{0}$ and $\boldsymbol{w}^{(n)}(t_0)$ is an equilibrium of the autonomous ODE system (28). It follows that for all $t \geq t_0$, $\boldsymbol{w}^{+(n)}(t) = \boldsymbol{w}^{+(n)}(t_0)$ and $\boldsymbol{w}^{-(n)}(t) = \boldsymbol{w}^{-(n)}(t_0)$. Therefore, we get $x_i r(t) = x_i r(t_0) = 0$ for all $t \geq t_0$. However, this contradicts that $x_i r(t') < 0$ and $t' > t_0$. Thus, we must have $x_i r(t) \geq 0$ for all $t \geq 0$. Since $w_i^+(t), w_i^-(t) \geq 0$, it follows that $g_i^+(t) \leq 0$ and $g_i^-(t) \geq 0$ for all t. We conclude $w_i^+(t)' \geq 0$ and $w_i^-(t)' \leq 0$ for all t.

If $w_i^+(0)' = \alpha x_i y \leq 0$, since x_i and y are non-zero by assumption, we must have $x_i y < 0$. Using similar arguments, it follows that $w_i^+(t)' \leq 0$ and $w_i^-(t)' \geq 0$ for all t. □

Lemma A.1. *Residual $r(t)$ never changes sign and its absolute value is always non-increasing.*

Proof. We have $r(0) = y \neq 0$ by assumption. When $r(0) > 0$, suppose for contradiction that there exists $t' > 0$ such that $r(t') < 0$. By continuity, we must have $r(t_0) = 0$ for some $t_0 \in (0, t')$. It follows that $g_i^+(t_0) = g_i^-(t_0) = 0$ for all $i = 1, \ldots, D_n$. In turn, $w_i^+(t_0)' = w_i^-(t_0)' = 0$ for all i and $\boldsymbol{w}^{(n)}(t_0)$ is an equilibrium of the autonomous ODE system (28). Therefore, $w_i^+(t) = w_i^+(t_0)$ and $w_i^-(t) = w_i^-(t_0)$ for all $t \geq t_0$. We conclude $r(t) = r(t_0) = 0$ for all $t \geq t_0$. This contradicts that $r(t') < 0$ for $t' > t_0$. As a result, $r(t) \geq 0$ for all t. Similarly, when $r(t) < 0$, it follows that $r(t) \leq 0$ for all t.

Next, we compute the derivative of $r(t)$ with respect to t as

$$\begin{aligned} r'(t) &= -2\sum_{i=1}^{D_n} x_i \left(w_i^+(t) \cdot w_i^+(t)' - w_i^-(t) \cdot w_i^-(t)' \right) \\ &= -2\sum_{i=1}^{D_n} x_i \left(w_i^+(t) \cdot \frac{w_i^+(t) x_i r(t)}{|w_i^+(t) x_i r(t)| + \varepsilon} - w_i^-(t) \cdot \frac{-w_i^-(t) x_i r(t)}{|w_i^-(t) x_i r(t)| + \varepsilon} \right) \\ &= -2\sum_{i=1}^{D_n} x_i^2 \left(\frac{(w_i^+(t))^2}{|w_i^+(t) x_i r(t)| + \varepsilon} + \frac{(w_i^-(t))^2}{|w_i^-(t) x_i r(t)| + \varepsilon} \right) r(t). \end{aligned}$$

Notice that $x_i^2 \left(\frac{(w_i^+(t))^2}{|w_i^+(t) x_i r(t)| + \varepsilon} + \frac{(w_i^-(t))^2}{|w_i^-(t) x_i r(t)| + \varepsilon} \right) \geq 0$. When $r(0) > 0$, we have shown that $r(t) \geq 0$ for all t. It follows that $r'(t) \leq 0$ for all t. Similarly, when $r(0) < 0$, we have $r(t) \leq 0$ and $r'(t) \geq 0$ for all t. Hence, the magnitude of the residual $r(t)$ is always non-increasing. □

Following the notation in Section 3.3.1, we let u_i denote the dominating weight, and let v_i represent the non-dominating weight. We repeat the definition here for clarity.

$$u_i(t) := \begin{cases} w_i^+(t) & \text{if } w_i^+(0)' > 0, \\ w_i^-(t) & \text{else} \end{cases}$$
$$v_i(t) := \begin{cases} w_i^-(t) & \text{if } w_i^+(0)' > 0, \\ w_i^+(t) & \text{else} \end{cases}$$

If $x_i y > 0$, then $\beta_i = u_i^2 - v_i^2$; if $x_i y < 0$, then $\beta_i = -u_i^2 + v_i^2$. Therefore, for all i,

$$\beta_i(t) = \operatorname{sgn}(x_i y) \left(u_i^2(t) - v_i^2(t) \right). \tag{29}$$

We let $f_i(t) := [\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))]_i$, $h_i(t) := [\nabla_{\boldsymbol{v}} L(\boldsymbol{w}(t))]_i$ denote the i-th component of the gradient with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively, which varies across all coordinates. By calculating the gradient, we derive the expressions

$$f_i(t) := -u_i(t)|x_i r(t)|, \; h_i(t) := v_i(t)|x_i r(t)|.$$

In turn, we have

$$\begin{aligned} u_i'(t) &= -\frac{f_i(t)}{|f_i(t)| + \varepsilon} = \frac{u_i(t)|x_i r(t)|}{u_i(t)|x_i r(t)| + \varepsilon}, \\ v_i'(t) &= -\frac{h_i(t)}{|h_i(t)| + \varepsilon} = -\frac{v_i(t)|x_i r(t)|}{v_i(t)|x_i r(t)| + \varepsilon}. \end{aligned}$$

The residual can be written as

$$r(t) = y - \sum_{k=1}^{D_n} x_k \beta_i(t) \tag{30}$$

$$= y - \sum_{k=1}^{D_n} \operatorname{sgn}(y)\operatorname{sgn}(x_k) \cdot x_k \left(u_k^2(t) - v_k^2(t)\right) \tag{31}$$

$$= \operatorname{sgn}(y) \left(|y| - \sum_{k=1}^{D_n} |x_k| \left(u_k^2(t) - v_k^2(t)\right) \right). \tag{32}$$

By Lemma A.1, $r(t)$ never changes sign. Since $r(0) = \operatorname{sgn}(y)|y|$, then for all t,

$$|r(t)| = |y| - \sum_{k=1}^{D_n} |x_k| \left(u_k^2(t) - v_k^2(t)\right). \tag{33}$$

A.2 PROOF OF PROPOSITION 3.5

Proof. First, we show the existence of $t_i > 0$ such that $h_i(t_i) = \varepsilon$ for each i. By Assumption 3.3, $h_i(0) = \alpha|x_i y| \geq 2\varepsilon$. Let us suppose for contradiction that $h_i(t) > \varepsilon$ for all t. Then $v_i'(t) < -\frac{\varepsilon}{\varepsilon+\varepsilon} = -\frac{1}{2}$ for all t, and for $t > 2\alpha$, $v_i(t) < \alpha - \frac{1}{2}t < 0$. However, by Proposition 3.4, $v_i(t)$ is always non-negative. It yields $h_i(t_i') < \varepsilon$ for some t_i'. Since $h_i(0) \geq 2\varepsilon$, by continuity of $h_i(t)$, there exists t_i such that $h_i(t_i) = \varepsilon$. It follows that $t_i \leq 2\alpha$. Because $h_2(t) \geq \varepsilon$ for $t \leq t_i$, $v_2'(t) \leq -\frac{1}{2}$. If $t_i > 2\alpha$, then $v_2(t_i) < \alpha - \frac{1}{2}t_i < 0$, which is a contradiction. We conclude $t_i \in (0, 2\alpha]$.

Next, we show that $|r(t_i)|$ is lower bounded. Using (33), we get

$$\begin{aligned} |r(t)| &= |y| - \sum_{i=1}^{D_n} |x_i| \left(u_i^2(t) - v_i^2(t)\right) \\ &\geq |y| - \sum_{i=1}^{D_n} |x_i| u_i^2(t). \end{aligned}$$

For all i, $u_i'(t) \leq 1$ always holds. It follows that $u_i(t) \leq \alpha + t$ for all t. Since $|r(t)|$ is non-increasing by Lemma A.1, and using $t_i \leq 2\alpha$, it follows that

$$|r(t_i)| \geq |r(2\alpha)| \geq |y| - \sum_{i=1}^{D_n} |x_i| \left(\alpha + 2\alpha\right)^2 = |y| - 9\alpha^2 \sum_{i=1}^{D_n} |x_i|.$$

By Assumption 3.3, $9\alpha^2 \leq \frac{|y|}{2\sum_{i=1}^{D_n} |x_i|}$, and thus

$$|r(t_i)| \geq |r(2\alpha)| \geq |y| - \frac{|y|}{2} = \frac{|y|}{2}. \tag{34}$$

Since $h_i(t_i) = v_i(t_i)|x_i r(t_i)| = \varepsilon$, we get $v_i(t_i) \leq \frac{2\varepsilon}{|x_i y|}$. Function $v_i(t)$ is non-increasing by Proposition 3.4, so for all $t \geq t_i$ we have $v_i(t) \leq \frac{2\varepsilon}{|x_i y|}$. The argument holds for all $i = 1, \ldots, D_n$ and for all n. We complete the proof by letting $T_0 := \max\{t_i\}$. □

A.3 PROOF OF PROPOSITION 3.6

Proof. Let us define the potential function by $\Phi_t(\boldsymbol{\beta}(t)) := \frac{2}{3}\sum_{i=1}^{D} \left(|\beta_i(t)| + v_{i,t}^2\right)^{\frac{3}{2}}$ for all t, where $v_{i.t} := v_i(t)$ is a parameter for the time-varying potential. We get the dual variable using the mirror map

$$\nabla \Phi_t(\boldsymbol{\beta}(t)) = \operatorname{sgn}(\boldsymbol{\beta}(t)) \odot \left(|\boldsymbol{\beta}(t)| + \boldsymbol{v}_t^2\right)^{\frac{1}{2}},$$

where operations are taken element-wise. The Hessian $\nabla^2\Phi_t(\boldsymbol{\beta}(t))$ is a diagonal matrix with diagonal elements $\frac{\operatorname{sgn}(\beta_i(t))}{2\left(|\beta_i(t)|+v_{i,t}^2\right)^{\frac{1}{2}}}$. Using the chain rule, we compute the dual dynamics

$$\begin{aligned}\frac{d\nabla\Phi_t(\boldsymbol{\beta}(t))}{dt} =&\langle\nabla^2\Phi_t(\boldsymbol{\beta}(t)), \frac{d\boldsymbol{\beta}(t)}{dt}\rangle + \langle\nabla_{\boldsymbol{v}}\nabla\Phi_t(\boldsymbol{\beta}(t)), \frac{d\boldsymbol{v}(t)}{dt}\rangle\\ =&\operatorname{sgn}(\boldsymbol{\beta}(t)) \odot \left(|\boldsymbol{\beta}(t)| + \boldsymbol{v}_t^2\right)^{-\frac{1}{2}} \odot \left(\boldsymbol{u}(t) \odot \frac{d\boldsymbol{u}(t)}{dt} - \boldsymbol{v}(t) \odot \frac{d\boldsymbol{v}(t)}{dt}\right)\\ &+ \operatorname{sgn}(\boldsymbol{\beta}(t)) \odot \left(|\boldsymbol{\beta}(t)| + \boldsymbol{v}_t^2\right)^{-\frac{1}{2}} \odot \boldsymbol{v}(t) \odot \frac{d\boldsymbol{v}(t)}{dt}\\ =&\operatorname{sgn}(\boldsymbol{\beta}(t)) \odot \frac{d\boldsymbol{u}(t)}{dt}\\ =&-\operatorname{sgn}(\boldsymbol{\beta}(t)) \odot \frac{\nabla_{\boldsymbol{u}}L(\boldsymbol{w}(t))}{|\nabla_{\boldsymbol{u}}L(\boldsymbol{w}(t))| + \varepsilon\mathbf{1}}.\end{aligned}$$

□

Lemma A.2. *For all $i, j \in \{1, \ldots, D_n\}$, $|x_i| \geq |x_j|$ implies $u_i(t) \geq u_j(t)$ for all $t \geq 0$.*

Proof. If $|x_i| = |x_j|$, since $u_i(0) = u_j(0) = \alpha$, then $u_i'(t) = u_j'(t)$ and $u_i(t) = u_j(t)$. Suppose $|x_i| > |x_j|$. Let $\bar{u}(t) := u_i(t) - u_j(t)$. Then we have $\bar{u}(0) = \alpha - \alpha = 0$.

First, we show that there exists a small neighborhood $\mathcal{B}$ such that $\bar{u}(t) > 0$ for $t \in \mathcal{B}$. Because $u_i(t), u_j(t)$ are differentiable everywhere, $\bar{u}(t)$ is differentiable for all $t \geq 0$. Inequality $|x_i| > |x_j|$ implies $u_i'(0) = \frac{\alpha|x_i y|}{\alpha|x_i y|+\varepsilon} > \frac{\alpha|x_j y|}{\alpha|x_j y|+\varepsilon} = u_j'(0)$. As a result, $\bar{u}'(0) > 0$. Using differentiability of $\bar{u}(t)$ at $t = 0$, we get

$$\lim_{\tau\to 0^+} \frac{\bar{u}(\tau) - \bar{u}(0)}{\tau - 0} = \lim_{\tau\to 0^+} \frac{\bar{u}(\tau)}{\tau} = \bar{u}'(0). \tag{35}$$

Let $\epsilon_\tau := \frac{\bar{u}'(0)}{3} > 0$. By definition of limit in (35), there exists $\delta_\tau > 0$ such that for all $\tau \in (0, \delta_\tau)$, $\left|\frac{\bar{u}(\tau)}{\tau} - \bar{u}'(0)\right| < \epsilon_\tau$. Therefore, $\frac{\bar{u}(\tau)}{\tau} - \bar{u}'(0) > -\epsilon_\tau = -\frac{\bar{u}'(0)}{3}$. It follows that $\bar{u}(\tau) > \frac{2\tau}{3}\bar{u}'(0) > 0$ for all $\tau \in (0, \delta_\tau)$.

Next, we show that $\bar{u}(t) \geq 0$ for all $t > 0$. Suppose for contradiction that there exists $t > 0$ such that $\bar{u}(t) < 0$. Let $t_0 := \inf\{t : t > 0,\ \bar{u}(t) < 0\}$. We have $\bar{u}(t_0) \leq 0$ and $\bar{u}(t) \geq 0$ for $t \in (0, t_0)$ by definition. We must have $t_0 \geq \delta_\tau > \frac{\delta_\tau}{2} > 0$ as we have shown that $\bar{u}(t) > 0$ for $t \in (0, \delta_\tau)$. Since $\bar{u}(t)$ is differentiable, by the Mean Value Theorem, there exists $t_1 \in (\frac{\delta_\tau}{2}, t_0)$ such that $\bar{u}'(t_1) = \frac{\bar{u}(t_0)-\bar{u}(\frac{\delta_\tau}{2})}{t_0-\frac{\delta_\tau}{2}}$. Since $\bar{u}(t_0) \leq 0$ and $\bar{u}(\frac{\delta_\varepsilon}{2}) > 0$, we have $\bar{u}'(t_1) < 0$. Therefore,

$$\begin{aligned}\bar{u}'(t_1) &= u_i'(t_1) - u_j'(t_1)\\ &= \frac{u_i(t_1)|x_i r(t_1)|}{u_i(t_1)|x_i r(t_1)| + \varepsilon} - \frac{u_j(t_1)|x_j r(t_1)|}{u_j(t_1)|x_j r(t_1)| + \varepsilon}\\ &< 0.\end{aligned}$$

The inequality implies that $u_i(t_1)|x_i r(t_1)| < u_j(t_1)|x_j r(t_1)|$. By assumption, we have $|x_i| > |x_j|$, and thus we must have $u_i(t_1) < u_j(t_1)$, i.e., $\bar{u}(t_1) < 0$. However, $t_1 < t_0$ and this contradicts that $\bar{u}(t) \geq 0$ for all $t \in (0, t_0)$. Thus, $u_i(t) \geq u_j(t)$ always holds. □

A.4 PROOF OF PROPOSITION 3.7

Proof. First, we show that for all i, there exists T_i such that $|\nabla_{\boldsymbol{u}}L(\boldsymbol{w}(T_i))|_i = f_i(T_i) = \varepsilon$. Suppose for contradiction that $f_i(t) > \varepsilon$ for all t. We have $u_i'(t) > \frac{1}{2}$ and $u_i(t) \geq \alpha + \frac{t}{2}$. Without loss of generality, we assume $r(0) = y > 0$. For

$t > 2\sqrt{\frac{|y|}{|x_i|}}$, the residual is negative due to

$$\begin{aligned} r(t) &= y - |x_i|\left(u_i^2(t) - v_i^2(t)\right) - \sum_{k\neq i}^{D_n} |x_k|\left(u_i^2(t) - v_i^2(t)\right) \\ &\leq y - |x_i|\left(u_i^2(t) - v_i^2(t)\right) \\ &\leq y - |x_i|\left(\left(\alpha + \sqrt{\frac{|y|}{|x_i|}}\right)^2 - \alpha^2\right) \\ &< y - |x_i|\frac{|y|}{|x_i|} \\ &= 0. \end{aligned}$$

However, this contradicts that $r(t)$ never flips sign by Lemma A.1. Hence, there exists T_i' such that $f_i(T_i') \leq \varepsilon$. By continuity, there exists $t \in (0, T_i']$ such that $f_i(t) = \varepsilon$. Let $T_i := \min\{t : 0 \leq t \leq T_i',\ f_i(t) = \varepsilon\}$. Therefore,

$$f_i(T_i) = \varepsilon, \text{ and } f_i(t) > \varepsilon \text{ for } t < T_i. \tag{36}$$

Next, we show that $T_i > 2\alpha \geq T_0$. In (34) we have proved that $|r(2\alpha)| \geq \frac{|y|}{2}$. Since $u_i(t) \geq \alpha$ and $|r(t)| \geq |r(2\alpha)|$ for $t \leq 2\alpha$, then $f_i(t) = u_i(t)|x_i r(t)| \geq \alpha\frac{|x_i y|}{2}$. By Assumption 3.3, we have $\alpha > \frac{2\varepsilon}{|x_i y|}$. As a result, $f_i(t) > \varepsilon$ for all $t \leq 2\alpha$. Therefore, we must have $T_i' > 2\alpha \geq T_0$.

We need to show that the derivative of $f_i(t)$ is always non-positive for $t \geq T_i$. Using the expression for $|r(t)|$ in (33), we get

$$\begin{aligned} f_i'(t) &= |x_i|\left(u_i'(t)|r(t)| + u_i(t)|r(t)|'\right) \\ &= |x_i|\left(u_i'(t)|r(t)| + u_i(t)\left(-2\sum_{k=1}^{D_n}|x_k|u_k(t)u_k'(t) + 2\sum_{k=1}^{D_n}|x_k|v_k(t)v_k'(t)\right)\right). \end{aligned}$$

Since $v_k'(t) \leq 0$ for all k, we get

$$f_i'(t) \leq |x_i|\left(u_i'(t)|r(t)| - 2u_i(t)\sum_{k=1}^{D_n}|x_k|u_k(t)u_k'(t)\right). \tag{37}$$

Next, we want to find a lower bound for $2u_i(T_i)\sum_{k=1}^{D_n}|x_k|u_k(T_i)u_k'(T_i)$. We denote the index set by $\mathcal{I} := \{1, \ldots, D_n\}$ that we partition as $\mathcal{I} = \mathcal{I}_i^+ \cup \mathcal{I}_i^-$, where $\mathcal{I}_i^+ := \{k : |x_k| \geq |x_i|\}$ and $\mathcal{I}_i^- := \{k : |x_k| < |x_i|\}$. For $k \in \mathcal{I}_i^-$, since $|x_k| < |x_i|$, we have $u_k(t) \leq u_i(t)$ by Lemma A.2. As a result, $u_k(t)|x_k r(t)| \leq u_i(t)|x_i r(t)|$. In turn, for all $k \in \mathcal{I}_i^-$ and for all t,

$$\begin{aligned} u_i(t)u_k'(t) &= u_i(t)\frac{u_k(t)|x_k r(t)|}{u_k(t)|x_k r(t)| + \varepsilon} && (38)\\ &\geq u_i(t)\frac{u_k(t)|x_k r(t)|}{u_i(t)|x_i r(t)| + \varepsilon} && (39)\\ &\geq \frac{u_k(t)|x_k|}{|x_i|}\cdot\frac{u_i(t)|x_i r(t)|}{u_i(t)|x_i r(t)| + \varepsilon} && (40)\\ &= \frac{|x_k|}{|x_i|}u_k(t)u_i'(t). && (41) \end{aligned}$$

For $k \in \mathcal{I}_i^+$, similarly, we have $u_k(t) \geq u_i(t)$ for all t. We also have $f_k(t) = u_k(t)|x_k r(t)| \geq u_i(t)|x_i r(t)| = f_i(t)$. In turn, $u_k'(t) \geq u_i'(t)$. Therefore, for all $k \in \mathcal{I}_i^+$ and for all t, we have

$$u_i(t)u_k'(t) \geq u_i(t)u_i'(t). \tag{42}$$

Using (41) and (42), we get

$$\begin{aligned}
2u_i(t)\sum_{k=1}^{D_n}|x_k|u_k(t)u_k'(t) &= 2\sum_{k\in\mathcal{I}_i^+}|x_k|u_k(t)u_i(t)u_k'(t) + 2\sum_{k\in\mathcal{I}_i^-}|x_k|u_k(t)u_i(t)u_k'(t)\\
&\geq 2\sum_{k\in\mathcal{I}_i^+}|x_k|u_k(t)u_i(t)u_i'(t) + 2\sum_{k\in\mathcal{I}_i^-}\frac{|x_k|}{|x_i|}\cdot|x_k|u_k^2(t)u_i'(t)\\
&= 2u_i'(t)\left(\sum_{k\in\mathcal{I}_i^+}|x_k|u_i(t)u_k(t) + \sum_{k\in\mathcal{I}_i^-}\frac{|x_k|}{|x_i|}|x_k|u_k^2(t)\right).
\end{aligned}$$

Because $u_i(t)$ is non-decreasing for all i, it follows that for $t \geq T_i$,

$$2u_i'(t)\left(\sum_{k\in\mathcal{I}_i^+}|x_k|u_i(t)u_k(t) + \sum_{k\in\mathcal{I}_i^-}\frac{|x_k|}{|x_i|}|x_k|u_k^2(t)\right) \geq 2u_i'(t)\left(\sum_{k\in\mathcal{I}_i^+}|x_k|u_i(T_i)u_k(T_i) + \sum_{k\in\mathcal{I}_i^-}\frac{|x_k|}{|x_i|}|x_k|u_k^2(T_i)\right).$$

Moreover, for $t \in [0, T_i]$, $u_i'(t) \geq \frac{1}{2}$ and $u_k'(t) \leq 1$. As a result, $u_i'(t) \geq \frac{1}{2}u_k'(t)$. In turn, we have

$$u_i(T_i) \geq \alpha + \frac{1}{2}\left(u_k(T_i) - \alpha\right) > \frac{1}{2}u_k(T_i). \tag{43}$$

We also know that $\frac{|x_k|}{|x_i|} \geq \frac{\min_j\{|x_j|\}}{|x_i|}$ for all $k \in \mathcal{I}_i^-$, and $1 \geq \frac{\min_j\{|x_j|\}}{|x_i|}$. Using (43), we get

$$\begin{aligned}
2u_i(t)\sum_{k=1}^{D_n}|x_k|u_k(T_i)u_k'(T_i) &\geq u_i'(t)\left(\sum_{k\in\mathcal{I}_i^+}2|x_k|\frac{1}{2}u_k(T_i)u_k(T_i) + \sum_{k\in\mathcal{I}_i^-}\frac{2|x_k|}{|x_i|}|x_k|u_k^2(T_i)\right) &(44)\\
&= u_i'(t)\left(\sum_{k\in\mathcal{I}_i^+}|x_k|u_k^2(T_i) + \sum_{k\in\mathcal{I}_i^-}\frac{2|x_k|}{|x_i|}|x_k|u_k^2(T_i)\right) &(45)\\
&\geq u_i'(t)\left(\sum_{k\in\mathcal{I}_i^+}\frac{\min_j\{|x_j|\}}{|x_i|}|x_k|u_k^2(T_i) + \sum_{k\in\mathcal{I}_i^-}\frac{\min_j\{|x_j|\}}{|x_i|}|x_k|u_k^2(T_i)\right) &(46)\\
&= u_i'(t)\frac{\min_j\{|x_j|\}}{|x_i|}\sum_{k=1}^{D_n}|x_k|u_k^2(T_i). &(47)
\end{aligned}$$

Next, we consider (37) by using (47). Since $|r(t)|$ is non-increasing, for all $t \geq T_i$, we have

$$\begin{aligned}
f_i'(t) &\leq |x_i|\left(u_i'(t)|r(t)| - u_i'(t)\frac{\min_j\{|x_j|\}}{|x_i|}\sum_{k=1}^{D_n}|x_k|u_k^2(T_i)\right) &(48)\\
&\leq |x_i|\left(u_i'(t)|r(T_i)| - u_i'(t)\frac{\min_j\{|x_j|\}}{|x_i|}\sum_{k=1}^{D_n}|x_k|u_k^2(T_i)\right) &(49)\\
&\leq |x_i|u_i'(t)\left(|r(T_i)| - \frac{\min_j\{|x_j|\}}{|x_i|}\sum_{k=1}^{D_n}|x_k|u_k^2(T_i)\right). &(50)
\end{aligned}$$

At $t = T_i$, we know that $u_i(T_i) \geq \alpha$ and $f_i(T_i) = u_i(T_i)|x_i r(T_i)| = \varepsilon$. By Assumption 3.3, we have $\alpha > \frac{2\varepsilon}{|x_j y|}$ for all j. We must have

$$|r(T_i)| = \frac{f_i(t)}{u_i(t)|x_i|} \leq \frac{\varepsilon}{\alpha|x_i|} < \frac{\varepsilon}{|x_i|}\cdot\frac{\min_j\{|x_j|\}|y|}{2\varepsilon} = \frac{1}{2}\frac{\min_j\{|x_j|\}}{|x_i|}|y|, \tag{51}$$

which implies

$$|y| - \sum_{k=1}^{D_n}|x_k|\left(u_k^2(T_i) - v_k^2(T_i)\right) < \frac{1}{2}\frac{\min_j\{|x_j|\}}{|x_i|}|y|.$$

Thus,

$$\sum_{k=1}^{D_n} |x_k| u_k^2(T_i) \geq \sum_{k=1}^{D_n} |x_k| \left(u_k^2(T_i) - v_k^2(T_i)\right) > \left(1 - \frac{1}{2}\frac{\min_j\{|x_j|\}}{|x_i|}\right)|y| \geq \frac{1}{2}|y|. \tag{52}$$

By using (51) and (52) in (50), we get

$$\begin{aligned} f_i'(t) &\leq |x_i| u_i'(t) \left(\frac{1}{2}\frac{\min_j\{|x_j|\}}{|x_i|}|y| - \frac{\min_j\{|x_j|\}}{|x_i|}\sum_{k=1}^{D_n}|x_k|u_k^2(T_i)\right) \\ &\leq |x_i| u_i'(t) \left(\frac{1}{2}\frac{\min_j\{|x_j|\}}{|x_i|}|y| - \frac{\min_j\{|x_j|\}}{|x_i|}\frac{1}{2}|y|\right) \\ &\leq 0. \end{aligned}$$

Hence, for all $t \geq T_i$, $f_i'(t)$ is non-increasing. We conclude that for each i, there exists $T_i > T_0$ such that $f_i'(t) > \varepsilon$ for $t < T_i$, and $f_i'(t) \leq \varepsilon$ for $t \geq T_i$. □

Lemma A.3 (Convergence)**.** *As $t \to \infty$, for every n we have*

$$\begin{aligned} &\lim_{t\to\infty} r^{(n)}(t) = 0, \\ &\lim_{t\to\infty} \nabla_{\boldsymbol{w}} L(\boldsymbol{w}(t)) = \mathbf{0}, \\ &\boldsymbol{u}^\infty := \lim_{t\to\infty}(\boldsymbol{u}(t)) \textit{ with } u_i^\infty < \infty \ \forall i, \\ &\boldsymbol{v}^\infty := \lim_{t\to\infty}(\boldsymbol{v}(t)) \textit{ with } v_i^\infty < \infty \ \forall i. \end{aligned}$$

Proof. Without loss of generality, we assume $r(0) = y > 0$. By Lemma 33, $r(t)$ is bounded below by 0 and monotonically non-increasing in t. Therefore, $r(t)$ converges as $t \to \infty$ by calculus. Let $R_0 := \lim_{t\to\infty} r(t) \geq 0$. We want to show that $R_0 = 0$. Suppose for contradiction that $R_0 > 0$. We have $r(t) \geq R_0 > 0$ for all $t \geq 0$.

We first show that $u_k'(t)$ is bounded below by a positive number for all k. Since $u_k(t) \geq \alpha$ and $r(t) \geq R_0$ for all t, we have $f_k(t) = u_k(t)|x_k|r(t) \geq \alpha|x_k|R_0 > 0$. Therefore, for all $t \geq 0$,

$$u_k'(t) = \frac{f_k(t)}{f_k(t) + \varepsilon} \geq \frac{\alpha|x_k|R_0}{\alpha|x_k|R_0 + \varepsilon} > 0.$$

As a result, $u_k(t) \geq \alpha + t \cdot \frac{\alpha|x_k|R_0}{\alpha|x_k|R_0+\varepsilon}$. Recall that

$$\begin{aligned} r(t) &= y - \sum_{k=1}^{D_n} |x_k| \left(u_k^2(t) - v_k^2(t)\right) \\ &\leq y - \sum_{k=1}^{D_n} |x_k| \left(u_k^2(t) - \alpha^2\right). \end{aligned}$$

As $t \to \infty$, $u_k^2(t) \to \infty$, and the summation $\sum_{k=1}^{D_n} |x_k| u_k^2(t)$ is unbounded. We conclude that $r(t) < 0$ for sufficiently large t. This contradicts that $r(t) \geq 0$ for all t by Lemma A.1. Thus, we must have $R_0 = \lim_{t\to\infty} r(t) = 0$.

The argument holds for all n, so $\lim_{t\to\infty} r^{(n)}(t) = 0$ for all $n = 1, \ldots, N$. As a result, we have $\lim_{t\to\infty}[\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))]_i = 0$ and $\lim_{t\to\infty}[\nabla_{\boldsymbol{v}} L(\boldsymbol{w}(t))]_i = 0$ for all i. It follows that $\lim_{t\to\infty} \nabla_{\boldsymbol{w}} L(\boldsymbol{w}(t)) = \mathbf{0}$.

Next, we show that the weights converge as $t \to \infty$. Without loss of generality, we suppose $r(0) = y > 0$. Because $r(t)$ never changes sign by Lemma A.1, we have $0 \leq r(t) \leq y - \sum_{k=1}^{D_n} |x_k| \left(u_k^2(t) - \alpha^2\right)$. As a result, $u_k(t)$ is upper bounded. Since $u_k(t)$ is non-decreasing, we have $u_k^\infty := \lim_{t\to\infty} u_k(t) < \infty$ by calculus. Using a similar argument for $v_k(t)$ which is non-increasing, $v_k^\infty := \lim_{t\to\infty} v_k(t) < \infty$. The proof holds for all k and all n. Therefore, $\boldsymbol{u}^\infty := \lim_{t\to\infty} \boldsymbol{u}(t)$ exists with $u_i^\infty < \infty$ for all i, and $\boldsymbol{v}^\infty := \lim_{t\to\infty} \boldsymbol{v}(t)$ exists with $v_i^\infty < \infty$ for all i. □

B PROOF OF RESULTS IN SECTION 3.3.2

Using Assumption 3.8, we parameterize the dynamics using θ_1 and λ_1 with $|\cos\theta_1| \geq |\sin\theta_1| > 0$ and $\lambda_1 > 0$. We let $y := y^{(1)}$, $\theta := \theta_1$, $\lambda := \lambda_1$ and $\tilde{y} := \frac{y^{(1)}}{\sqrt{\lambda_1}}$ to simplify the notation in the proofs. We have

$$
\begin{aligned}
|r(t)| &= |\tilde{y}| - |\cos\theta| \left(u_1^2(t) - v_1^2(t)\right) - |\sin\theta| \left(u_2^2(t) - v_2^2(t)\right), \\
f_1(t) &= \lambda u_1(t) |\cos\theta r(t)|, \\
f_2(t) &= \lambda u_2(t) |\sin\theta r(t)|, \\
u_1'(t) &= \frac{f_1(t)}{f_1(t)+\varepsilon},\ u_2'(t) := \frac{f_2(t)}{f_2(t)+\varepsilon}.
\end{aligned}
$$

Lemma B.1. *We have $u_1'(t) \geq u_2'(t)$ for $t \in [0, T)$, and $u_1'(t) \geq \frac{2|\cot\theta|}{1+|\cot\theta|} u_2'(t)$ for $t \in [T, \infty)$. Quantity T is the stage transition time as in Proposition 3.7.*

Proof. First, we show that for all $t \geq 0$,

$$
u_1'(t) \geq \frac{|\cot\theta| \left(f_2(t)+\varepsilon\right)}{|\cot\theta| f_2(t) + \varepsilon} u_2'(t). \tag{53}
$$

Since $|\cos\theta| \geq |\sin\theta| > 0$ by Assumption 3.8, we have $u_1(t) \geq u_2(t)$ by Lemma A.2. As a result,

$$
\begin{aligned}
f_1(t) &= \lambda u_1(t) |\cos\theta r(t)| \\
&= |\cot\theta| \lambda u_1(t) |\sin\theta r(t)| \\
&\geq |\cot\theta| \lambda u_2(t) |\sin\theta r(t)| \\
&= |\cot\theta| f_2(t).
\end{aligned}
$$

Therefore,

$$
u_1'(t) = \frac{f_1(t)}{f_1(t)+\varepsilon} = 1 - \frac{\varepsilon}{f_1(t)+\varepsilon} \geq 1 - \frac{\varepsilon}{|\cot\theta| f_2(t) + \varepsilon} = \frac{|\cot\theta| f_2(t)}{|\cot\theta| f_2(t) + \varepsilon}. \tag{54}
$$

When $u_2'(t) = 0$, (53) holds since $u_1'(t)$ is always non-negative. When $u_2'(t) \neq 0$, using (54), we have that (53) holds:

$$
\begin{aligned}
\frac{u_1'(t)}{u_2'(t)} &= \frac{f_1(t)}{f_1(t)+\varepsilon} \cdot \frac{f_2(t)+\varepsilon}{f_2(t)} \\
&\geq \frac{|\cot\theta| f_2(t)}{|\cot\theta| f_2(t) + \varepsilon} \cdot \frac{f_2(t)+\varepsilon}{f_2(t)} \\
&= \frac{|\cot\theta| (f_2(t)+\varepsilon)}{|\cot\theta| f_2(t) + \varepsilon}.
\end{aligned}
$$

By Proposition 3.7, there exist stage transition times T_1, T_2 for $f_1(t)$ and $f_2(t)$, respectively. We know that $f_1(t) \leq \varepsilon$ for $t \geq T_1$. Since $|\cos\theta| \geq |\sin\theta|$, $f_1(t) \geq f_2(t) > \varepsilon$ for $t \in [0, T_2)$. As a result, we must have $T_1 \geq T_2$. By definition, $T := \min\{T_1, T_2\} = T_2$. For all t, $|\cos\theta| \geq |\sin\theta|$ implies $u_1(t) \geq u_2(t)$ and $f_1(t) \geq f_2(t)$. Therefore, we conclude $u_1'(t) \geq u_2'(t)$ for $t \in [0, T)$.

For $t \in [T, \infty)$, we establish $f_2(t) \leq \varepsilon$. Notice that $\frac{|\cot\theta|(f_2+\varepsilon)}{|\cot\theta| f_2+\varepsilon} = 1 + \frac{(|\cot\theta|-1)\varepsilon}{|\cot\theta| f_2+\varepsilon}$. Since $|\cot\theta| \geq 1$, the ratio $\frac{|\cot\theta|(f_2+\varepsilon)}{|\cot\theta| f_2+\varepsilon}$ is non-increasing in $f_2 \geq 0$. Using $f_2(t) \leq \varepsilon$, we get

$$
\frac{|\cot\theta|(f_2(t)+\varepsilon)}{|\cot\theta| f_2(t) + \varepsilon} \geq \frac{|\cot\theta|(\varepsilon+\varepsilon)}{|\cot\theta|\varepsilon + \varepsilon} = \frac{2|\cot\theta|}{1+|\cot\theta|}.
$$

Using (53), we conclude that for $t \in [T, \infty)$,

$$
u_1'(t) \geq \frac{|\cot\theta| \left(f_2(t)+\varepsilon\right)}{|\cot\theta| f_2(t) + \varepsilon} u_2'(t) \geq \frac{2|\cot\theta|}{1+|\cot\theta|} u_2'(t).
$$

□

Let us consider the cubic equation $x\left(A-Bx^2\right)=\epsilon$, where $A>0, B>0, x>0$. We assume that $\epsilon \geq 0$ is small. The largest solution x^* is approximately

$$x^*=\sqrt{\frac{A}{B}}-\frac{1}{2A}\epsilon-\frac{3}{8}B^{\frac{1}{2}}A^{-\frac{5}{2}}\epsilon^2+\mathcal{O}\left(\epsilon^3\right).$$

This can be established by using elementary perturbation theory.

Lemma B.2. *We have*

$$\Delta:=|\cos\theta|\left(u_2^\infty-u_2(0)\right)-|\sin\theta|\left(u_1^\infty-u_1(0)\right)\leq M_+,$$

where $M_+:=(|\cos\theta|-|\sin\theta|)\left(\lambda^{-\frac{1}{4}}|y|^{\frac{1}{2}}-\frac{\sqrt{2}\varepsilon}{4\lambda^{\frac{1}{2}}|y|}\right)$.

Proof. We complete the proof in three steps.

Step 1. We show an upper bound for $u_2(T)$.

Let us define $p(U):=\lambda|\sin\theta|U\left(|\tilde{y}|+(|\cos\theta|+|\sin\theta|)\alpha^2-(|\cos\theta|+|\sin\theta|)U^2\right)$, which is a cubic function of $U\in\mathbb{R}$. Let $\hat{U}$ be the largest solution to $p(U)=\varepsilon$. Let us define $f_+(t):=(p\circ u_2)(t)$. We want to show that $f_+(t)\geq f_2(t)$ for $t\in[0,T]$. Indeed, since $u_1(t)\geq u_2(t), v_1(t), v_2(t)\leq\alpha$ always hold, we have

$$\begin{aligned}
f_+(t)&=\lambda|\sin\theta|u_2(t)\left(|\tilde{y}|+(|\cos\theta|+|\sin\theta|)\alpha^2-(|\cos\theta|+|\sin\theta|)u_2^2(t)\right)\\
&\geq\lambda|\sin\theta|u_2(t)\left(|\tilde{y}|+|\cos\theta|v_1^2(t)+|\sin\theta|v_2^2(t)-|\cos\theta|u_1^2(t)-|\sin\theta|u_2^2(t)\right)\\
&=f_2(t).
\end{aligned}$$

We know that $f_2(T)=\varepsilon$, so $f_+(t)=p(u_2(T))\geq\varepsilon$. Meanwhile, $p(\hat{U})=\varepsilon$. We want to show that $u_2(T)\leq\hat{U}$. Suppose for contradiction that $u_2(T)>\hat{U}$. By studying the behavior of the cubic function $p(U)$, we observe that $p(U)<0$ for sufficiently large U. Since $p(u_2(T))\geq\varepsilon$, by continuity, there exists $U'\geq u_2(T)$ such that $p(U')=\varepsilon$. However, $U'\geq u_2(T)>\hat{U}$, which contradicts that $\hat{U}$ is the largest solution to $p(U)=\varepsilon$. Thus, $u_2(T)\leq\hat{U}$. By using the expansion of the cubic root $\hat{U}$ in ε, it is easy to show that $\hat{U}<\tilde{u}_2:=\sqrt{\frac{|\tilde{y}|}{|\cos\theta|+|\sin\theta|}+\alpha^2}-\frac{\varepsilon}{2\lambda|\sin\theta\tilde{y}|}$ under Assumption 3.3. As a result,

$$u_2(T)\leq\hat{U}<\tilde{u}_2:=\sqrt{\frac{|\tilde{y}|}{|\cos\theta|+|\sin\theta|}+\alpha^2}-\frac{\varepsilon}{2\lambda|\sin\theta\tilde{y}|}. \tag{55}$$

Step 2. We show that $u_1(T)-u_1(0)\geq u_2(T)-u_2(0)$ and $u_1^\infty-u_1(T)\geq\frac{2|\cot\theta|}{1+|\cot\theta|}\left(u_2^\infty-u_2(T)\right)$.

For all t, we know that $u_1'(t)\geq u_2'(t)$. By integrating both sides with respect to t from 0 to T, we get

$$u_1(T)-u_1(0)\geq u_2(T)-u_2(0). \tag{56}$$

For $t\geq T$, by Lemma B.1 we have $u_1'(t)\geq\frac{2|\cot\theta|}{1+|\cot\theta|}u_2'(t)$. Again by integrating both sides, we get

$$u_1^\infty-u_1(T)\geq\frac{2|\cot\theta|}{1+|\cot\theta|}\left(u_2^\infty-u_2(T)\right). \tag{57}$$

Step 3. We derive an upper bound for Δ.

We can write $\Delta=\Delta_1+\Delta_2$, where

$$\begin{aligned}
\Delta_1&:=|\cos\theta|(u_2(T)-u_2(0))-|\sin\theta|(u_1(T)-u_1(0)),\\
\Delta_2&:=|\cos\theta|(u_2^\infty-u_2(T))-|\sin\theta|(u_1^\infty-u_1(T)).
\end{aligned}$$

Using (56) and (57) from Step 2, we get

$$\Delta_1\leq(|\cos\theta|-|\sin\theta|)(u_2(T)-u_2(0)), \tag{58}$$

$$\Delta_2\leq\left(|\cos\theta|-|\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|}\right)\left(u_2^\infty-u_2(T)\right). \tag{59}$$

Adding (59) and (58), we get

$$\begin{aligned}\Delta &\leq \left(|\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|}\right)(u_2^\infty - u_2(T)) + (|\cos\theta| - |\sin\theta|)\,(u_2(T) - u_2(0)) \\ &= \left(|\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|}\right)(u_2^\infty - \tilde{u}_2) \\ &\quad + \left(|\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|}\right)(\tilde{u}_2 - u_2(T)) + (|\cos\theta| - |\sin\theta|)(u_2(T) - \tilde{u}_2) \\ &\quad + (|\cos\theta| - |\sin\theta|)(\tilde{u}_2 - u_2(0)).\end{aligned}$$

We have shown that $\tilde{u}_2 \geq u_2(T)$ in (55), and $|\cos\theta| \geq |\sin\theta|$ implies $|\cos\theta| - |\sin\theta| \geq |\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|} \geq 0$. As a result,

$$\begin{aligned}&\left(|\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|}\right)(\tilde{u}_2 - u_2(T)) \leq (|\cos\theta| - |\sin\theta|)\,(\tilde{u}_2 - u_2(T)) \\ &\left(|\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|}\right)(\tilde{u}_2 - u_2(T)) + (|\cos\theta| - |\sin\theta|)(u_2(T) - \tilde{u}_2) \leq 0.\end{aligned}$$

Therefore,

$$\Delta \leq \left(|\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|}\right)(u_2^\infty - \tilde{u}_2) + (|\cos\theta| - |\sin\theta|)(\tilde{u}_2 - u_2(0)). \tag{60}$$

Moreover, by Lemma A.3, we know that the residual converges to zero. It follows that $\lim_{t\to\infty} r(t) = 0$, and

$$|\tilde{y}| = |\cos\theta|\left((u_1^\infty)^2 - (v_1^\infty)^2\right) + |\sin\theta|\left((u_2^\infty)^2 - (v_2^\infty)^2\right).$$

Because $u_1(t) \geq u_2(t)$ and $v_1(t), v_2(t) \leq \alpha$ always hold, it follows that $u_2^\infty \leq \sqrt{\frac{|\tilde{y}|}{|\cos\theta|+|\sin\theta|} + \alpha^2}$. Using $\tilde{u}_2$ from (55), we get $u_2^\infty - \tilde{u}_2 \leq \frac{\varepsilon}{2\lambda|\sin\theta\tilde{y}|}$. Continuing with (60) and using $u_2(0) = \alpha$, we get

$$\begin{aligned}\Delta &\leq \left(|\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|}\right)\frac{\varepsilon}{2\lambda|\sin\theta\tilde{y}|} \\ &\quad + (|\cos\theta| - |\sin\theta|)\left(\sqrt{\frac{|\tilde{y}|}{|\cos\theta|+|\sin\theta|} + \alpha^2} - \frac{\varepsilon}{2\lambda|\sin\theta\tilde{y}|} - u_2(0)\right) \\ &= (|\cos\theta| - |\sin\theta|)\left(\sqrt{\frac{|\tilde{y}|}{|\cos\theta|+|\sin\theta|} + \alpha^2} - u_2(0)\right) \\ &\quad + \left(|\cos\theta| - |\sin\theta|\frac{2|\cot\theta|}{1+|\cot\theta|} - |\cos\theta| + |\sin\theta|\right)\frac{\varepsilon}{2\lambda|\sin\theta\tilde{y}|} \\ &= (|\cos\theta| - |\sin\theta|)\left(\sqrt{\frac{|\tilde{y}|}{|\cos\theta|+|\sin\theta|} + \alpha^2} - \alpha\right) - \left(\frac{|\cos\theta| - |\sin\theta|}{|\cos\theta|+|\sin\theta|}\right)\frac{\varepsilon}{2\lambda|\tilde{y}|} \\ &\leq (|\cos\theta| - |\sin\theta|)\sqrt{\frac{|\tilde{y}|}{|\cos\theta|+|\sin\theta|}} - \left(\frac{|\cos\theta| - |\sin\theta|}{|\cos\theta|+|\sin\theta|}\right)\frac{\varepsilon}{2\lambda|\tilde{y}|} \\ &\leq (|\cos\theta| - |\sin\theta|)\sqrt{|\tilde{y}|} - (|\cos\theta| - |\sin\theta|)\frac{\sqrt{2}\varepsilon}{4\lambda|\tilde{y}|} \\ &= (|\cos\theta| - |\sin\theta|)\left(|y|^{\frac{1}{2}}\lambda^{-\frac{1}{4}} - \frac{\sqrt{2}\varepsilon}{4\lambda^{\frac{1}{2}}|y|}\right) \\ &= M_+.\end{aligned}$$

We conclude $\Delta \leq M_+$. □

Lemma B.3. *We have*

$$\Delta := |\cos\theta|\,(u_2^\infty - u_2(0)) - |\sin\theta|\,(u_1^\infty - u_1(0)) \geq M_-,$$

where $M_- := (|\cos\theta| - |\sin\theta|)\left((2\lambda)^{-\frac{1}{4}}\,|y|^{\frac{1}{2}} - \alpha\right) - 2\sqrt{\frac{2\varepsilon}{\lambda^{\frac{3}{4}}|\sin\theta||y|^{\frac{1}{2}}}} - \frac{3\sqrt{2}\varepsilon}{\lambda^{\frac{1}{2}}|\sin\theta y|}\ln\left(\frac{\lambda^{\frac{1}{4}}|\sin\theta||y|^{\frac{3}{2}}}{\sqrt{2}\varepsilon}\right)$.

Proof. We begin by exhibiting a lower bounding function for $f_2(t)$ for $t \in [0,T]$. Let $\bar{v} := |\cos\theta|\,(v_1^\infty)^2 + |\sin\theta|\,(v_2^\infty)^2$. Since $v_1(t), v_2(t)$ are non-increasing and non-negative, we have

$$0 \leq \bar{v} \leq |\cos\theta| v_1^2(t) + |\sin\theta| v_2^2(t) \leq (|\cos\theta| + |\sin\theta|)\,\alpha^2. \tag{61}$$

Let us define

$$f_-(t) := \frac{1}{2}\lambda|\sin\theta|\,(\alpha+t)\left(|\tilde{y}| + \bar{v} - (|\cos\theta| + |\sin\theta|)\,(\alpha+t)^2\right).$$

For $t \leq T$, since $f_2(t) \geq \varepsilon$ and $u_2'(t) \geq \frac{1}{2}$, we have $u_2(t) \geq \alpha + \frac{1}{2}t > \frac{1}{2}(\alpha+t)$. Moreover, $u_1(t) \leq \alpha + t$ and $u_2(t) \leq \alpha + t$ always hold. Therefore, for $t \in [0,T]$, we establish that

$$\begin{aligned}
f_-(t) &= \frac{1}{2}\lambda|\sin\theta|\,(\alpha+t)\left(|\tilde{y}| + \bar{v} - (|\cos\theta| + |\sin\theta|)\,(\alpha+t)^2\right)\\
&< \lambda|\sin\theta| u_2(t)\left(|\tilde{y}| + \bar{v} - |\cos\theta| u_1^2(t) - |\sin\theta| u_2^2(t)\right)\\
&\leq \lambda|\sin\theta| u_2(t)\left(|\tilde{y}| - |\cos\theta|\left(u_1^2(t) - v_2^2(t)\right) - |\sin\theta|\left(u_2^2(t) - v_2^2(t)\right)\right)\\
&= f_2(t).
\end{aligned}$$

As a result, we get

$$f_-(T) < f_2(T) = \varepsilon. \tag{62}$$

Assumption 3.3 guarantees $\sqrt{\frac{|\tilde{y}|+\bar{v}}{3(|\cos\theta|+|\sin\theta|)}} - \alpha > 0$. The derivative of the cubic function $f_-(t)$ shows that $f_-(t)$ is increasing on $\left[0,\ \sqrt{\frac{|\tilde{y}|+\bar{v}}{3(|\cos\theta|+|\sin\theta|)}} - \alpha\right)$ and decreasing for $t > \sqrt{\frac{|\tilde{y}|+\bar{v}}{3(|\cos\theta|+|\sin\theta|)}} - \alpha$. Because $f_-(0) > \varepsilon$ by Assumption 3.3 and $f_-(T) < \varepsilon$ by (62), it follows that there exists a unique $T' \in \left(\sqrt{\frac{|\tilde{y}|+\bar{v}}{3(|\cos\theta|+|\sin\theta|)}} - \alpha,\ T\right)$ such that $f_-(T') = \varepsilon$, and $f_-(t) < \varepsilon$ for $t > T'$.

Next, we show a lower bound for $\alpha + T'$. Since we already have $\alpha + T' > \sqrt{\frac{|\tilde{y}|+\bar{v}}{3(|\cos\theta|+|\sin\theta|)}}$, then

$$\begin{aligned}
\varepsilon = f_-(T') &= \frac{1}{2}\lambda|\sin\theta|\,(\alpha+T')\left(|\tilde{y}| + \bar{v} - (|\cos\theta| + |\sin\theta|)\,(\alpha+T')^2\right)\\
&\geq \frac{1}{2}\lambda|\sin\theta|\sqrt{\frac{|\tilde{y}|+\bar{v}}{3\,(|\cos\theta|+|\sin\theta|)}}\left(|\tilde{y}| + \bar{v} - (|\cos\theta| + |\sin\theta|)\,(\alpha+T')^2\right).
\end{aligned}$$

Therefore, we have

$$\frac{2\varepsilon}{\lambda|\sin\theta|}\sqrt{\frac{3(|\cos\theta|+|\sin\theta|)}{|\tilde{y}|+\bar{v}}} \geq |\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)\left(\alpha+T'\right)^2 \tag{63}$$

$$(|\cos\theta|+|\sin\theta|)\left(\alpha+T'\right)^2 \geq |\tilde{y}|+\bar{v}-\frac{2\varepsilon}{\lambda|\sin\theta|}\sqrt{\frac{3(|\cos\theta|+|\sin\theta|)}{|\tilde{y}|+\bar{v}}} \tag{64}$$

$$(\alpha+T')^2 \geq \frac{|\tilde{y}|+\bar{v}}{|\cos\theta|+|\sin\theta|}-\frac{2\varepsilon}{\lambda|\sin\theta|}\sqrt{\frac{3}{(|\tilde{y}|+\bar{v})(|\cos\theta|+|\sin\theta|)}} \tag{65}$$

$$(\alpha+T')^2 \geq \frac{|\tilde{y}|+\bar{v}}{|\cos\theta|+|\sin\theta|}-\frac{2\sqrt{3}\varepsilon}{\lambda|\sin\theta|(|\tilde{y}|+\bar{v})^{\frac{1}{2}}} \tag{66}$$

$$(\alpha+T')^2 \geq \frac{|\tilde{y}|+\bar{v}}{|\cos\theta|+|\sin\theta|}-\frac{4\varepsilon}{\lambda|\sin\theta|(|\tilde{y}|+\bar{v})^{\frac{1}{2}}} \tag{67}$$

$$\alpha+T' \geq \left(\frac{|\tilde{y}|+\bar{v}}{|\cos\theta|+|\sin\theta|}-\frac{4\varepsilon}{\lambda|\sin\theta|(|\tilde{y}|+\bar{v})^{\frac{1}{2}}}\right)^{\frac{1}{2}} \tag{68}$$

$$\alpha+T' \geq \sqrt{\frac{|\tilde{y}|+\bar{v}}{|\cos\theta|+|\sin\theta|}}-2\sqrt{\frac{\varepsilon}{\lambda|\sin\theta|(|\tilde{y}|+\bar{v})^{\frac{1}{2}}}}. \tag{69}$$

Next, we want to find a lower bound for $u_2(T)$. Because $u_2(t)$ is non-decreasing, $T > T'$ implies $u_2(T) \geq u_2(T')$. For all $t \in [0, T']$, $f_-(t) \leq f_2(t)$ holds, and therefore

$$\begin{aligned} u_2'(t) &= \frac{f_2(t)}{f_2(t)+\varepsilon} \\ &\geq \frac{f_-(t)}{f_-(t)+\varepsilon} \\ &= 1-\frac{2\varepsilon}{\lambda|\sin\theta|\,(\alpha+t)\left(|\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)\,(\alpha+t)^2\right)}. \end{aligned}$$

This lower bounding function is explicit in t, which makes it possible to obtain a lower bound for $u_2(T')$ by integrating it with respect to t from 0 to T', which yields

$$u_2(T')-u_2(0) \geq \int_0^{T'} 1-\frac{2\varepsilon}{\lambda|\sin\theta|\,(\alpha+t)\left(|\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)\,(\alpha+t)^2\right)}\,dt \tag{70}$$

$$u_2(T') \geq \alpha+T'-\frac{2\varepsilon}{\lambda|\sin\theta|}\int_0^{T'}\frac{1}{(\alpha+t)\left(|\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)\,(\alpha+t)^2\right)}\,dt. \tag{71}$$

Let $\tau := \alpha + t$. We compute the integral

$$\begin{aligned} J := \int_\alpha^{\alpha+T'}\frac{1}{\tau(|\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)\tau^2)}\,d\tau &= \frac{1}{2(|\tilde{y}|+\bar{v})}\ln\frac{\tau^2}{|\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)\tau^2}\Bigg|_\alpha^{\alpha+T'} \\ &= \frac{1}{2(|\tilde{y}|+\bar{v})}\ln\frac{(\alpha+T')^2(|\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)\alpha^2)}{\alpha^2(|\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)(\alpha+T')^2)}. \end{aligned}$$

Since $f_-(T') = \varepsilon$, we get

$$|\tilde{y}|+\bar{v}-(|\cos\theta|+|\sin\theta|)\left(\alpha+T'\right)^2 = \frac{2\varepsilon}{\lambda|\sin\theta|(\alpha+T')}.$$

Moreover, $(\alpha + T')^2 \leq |\tilde{y}| + \bar{v}$. Using $\bar{v} \leq (|\cos\theta| + |\sin\theta|)\,\alpha^2$ from (61) and Assumption 3.3, we get

$$\begin{aligned}
J &\leq \frac{1}{2(|\tilde{y}| + \bar{v})} \ln \frac{(\alpha + T')^3}{2\varepsilon/(\lambda|\sin\theta|)} \frac{|\tilde{y}|}{\alpha^2} \\
&\leq \frac{1}{2(|\tilde{y}| + \bar{v})} \ln \frac{(|\tilde{y}| + \bar{v})^{\frac{3}{2}}}{2\varepsilon/(\lambda|\sin\theta|)} \frac{|\tilde{y}|}{(2\varepsilon/(\lambda|\sin\theta\tilde{y}|))^2} \\
&= \frac{1}{2(|\tilde{y}| + \bar{v})} \ln \left((|\tilde{y}| + \bar{v})^{\frac{3}{2}} \left(\frac{\lambda|\sin\theta\tilde{y}|}{2\varepsilon} \right)^3 \right) \\
&= \frac{3}{2(|\tilde{y}| + \bar{v})} \ln \left((|\tilde{y}| + \bar{v})^{\frac{1}{2}} \left(\frac{\lambda|\sin\theta\tilde{y}|}{2\varepsilon} \right) \right).
\end{aligned}$$

Using Assumption 3.3, we obtain that

$$|\tilde{y}| + \bar{v} \leq |\tilde{y}| + (|\cos\theta| + |\sin\theta|)\,\alpha^2 \leq |\tilde{y}| + \frac{1}{18}|\tilde{y}| = \frac{19}{18}|\tilde{y}|.$$

In turn we have

$$\begin{aligned}
J &\leq \frac{3}{2(|\tilde{y}| + \bar{v})} \ln \left((|\tilde{y}| + \bar{v})^{\frac{1}{2}} \left(\frac{\lambda|\sin\theta\tilde{y}|}{2\varepsilon} \right) \right) \\
&\leq \frac{3}{2(|\tilde{y}| + \bar{v})} \ln \left(\left(\frac{19}{18}|\tilde{y}| \right)^{\frac{1}{2}} \left(\frac{\lambda|\sin\theta\tilde{y}|}{2\varepsilon} \right) \right) \\
&\leq \frac{3}{2(|\tilde{y}| + \bar{v})} \ln \left(\sqrt{2}|\tilde{y}|^{\frac{1}{2}} \left(\frac{\lambda|\sin\theta\tilde{y}|}{2\varepsilon} \right) \right) \\
&= \frac{3}{2(|\tilde{y}| + \bar{v})} \ln \left(\frac{\lambda|\sin\theta||\tilde{y}|^{\frac{3}{2}}}{\sqrt{2}\varepsilon} \right).
\end{aligned}$$

Assumption 3.3 implies that $\varepsilon \leq \frac{\lambda|\sin\theta||\tilde{y}|^{\frac{3}{2}}}{9\sqrt{2(|\cos\theta|+|\sin\theta|)}}$. Therefore, $\ln\left(\frac{\lambda|\sin\theta||\tilde{y}|^{\frac{3}{2}}}{\sqrt{2}\varepsilon}\right)$ is guaranteed to be positive, and we get

$$J \leq \frac{3}{2|\tilde{y}|} \ln \left(\frac{\lambda|\sin\theta||\tilde{y}|^{\frac{3}{2}}}{\sqrt{2}\varepsilon} \right). \tag{72}$$

Combining (69), (71) and (72), we get

$$\begin{aligned}
u_2(T') &\geq \sqrt{\frac{|\tilde{y}| + \bar{v}}{|\cos\theta| + |\sin\theta|}} - 2\sqrt{\frac{\varepsilon}{\lambda|\sin\theta|(|\tilde{y}| + \bar{v})^{\frac{1}{2}}}} - \frac{3\varepsilon}{\lambda|\sin\theta\tilde{y}|} \ln \left(\frac{\lambda|\sin\theta||\tilde{y}|^{\frac{3}{2}}}{\sqrt{2}\varepsilon} \right) \\
&\geq \sqrt{\frac{|\tilde{y}| + \bar{v}}{|\cos\theta| + |\sin\theta|}} - 2\sqrt{\frac{\varepsilon}{\lambda|\sin\theta||\tilde{y}|^{\frac{1}{2}}}} - \frac{3\varepsilon}{\lambda|\sin\theta\tilde{y}|} \ln \left(\frac{\lambda|\sin\theta||\tilde{y}|^{\frac{3}{2}}}{\sqrt{2}\varepsilon} \right).
\end{aligned}$$

Let $P := \sqrt{\frac{|\tilde{y}|+\bar{v}}{|\cos\theta|+|\sin\theta|}}$, $Q := 2\sqrt{\frac{\varepsilon}{\lambda|\sin\theta||\tilde{y}|^{\frac{1}{2}}}} + \frac{3\varepsilon}{\lambda|\sin\theta\tilde{y}|} \ln\left(\frac{\lambda|\sin\theta||\tilde{y}|^{\frac{3}{2}}}{\sqrt{2}\varepsilon}\right)$, and $P, Q > 0$. Therefore, we have $u_2(T') \geq P - Q$. Using Lemma A.3, we obtain that

$$\begin{aligned}
|\cos\theta|\,(P+Q)^2 + |\sin\theta|\,(P-Q)^2 &= (|\cos\theta| + |\sin\theta|)P^2 + 2PQ(|\cos\theta| - |\sin\theta|) + (|\cos\theta| + |\sin\theta|)Q^2 \\
&\geq (|\cos\theta| + |\sin\theta|)P^2 \\
&= |\tilde{y}| + \bar{v} \\
&= |\tilde{y}| + |\cos\theta|\,(v_1^\infty)^2 + |\sin\theta|\,(v_2^\infty)^2 \\
&= |\cos\theta|\,(u_1^\infty)^2 + |\sin\theta|\,(u_2^\infty)^2.
\end{aligned}$$

Since $u_2(t)$ is non-decreasing, we get $u_2^\infty \geq u_2(T') \geq P - Q$. As a result, we must have $u_1^\infty \leq P + Q$. We derive

$$|\cos\theta| u_2^\infty - |\sin\theta| u_1^\infty \geq (|\cos\theta| - |\sin\theta|)\, P - (|\cos\theta| + |\sin\theta|)\, Q \tag{73}$$

$$= (|\cos\theta| - |\sin\theta|) \sqrt{\frac{|\tilde{y}| + \bar{v}}{|\cos\theta| + |\sin\theta|}} - (|\cos\theta| + |\sin\theta|)\, Q \tag{74}$$

$$\geq (|\cos\theta| - |\sin\theta|) \sqrt{\frac{|\tilde{y}|}{|\cos\theta| + |\sin\theta|}} - \sqrt{2} Q \tag{75}$$

$$\geq (|\cos\theta| - |\sin\theta|) \sqrt{\frac{|\tilde{y}|}{\sqrt{2}}} - \sqrt{2} Q. \tag{76}$$

Additionally, we have $u_1(0) = u_2(0) = \alpha$, which implies

$$-|\cos\theta| u_2(0) + |\sin\theta| u_1(0) = -\alpha(|\cos\theta| - |\sin\theta|). \tag{77}$$

Adding (76) and (77), and substituting in Q, we get

$$\begin{aligned}
\Delta &= |\cos\theta|(u_2^\infty - u_2(0)) - |\sin\theta|(u_1^\infty - u_1(0)) \\
&= |\cos\theta| u_2^\infty - |\sin\theta| u_1^\infty - \alpha(|\cos\theta| - |\sin\theta|) \\
&\geq (|\cos\theta| - |\sin\theta|) \left(\sqrt{\frac{|\tilde{y}|}{\sqrt{2}}} - \alpha \right) - 2\sqrt{\frac{2\varepsilon}{\lambda |\sin\theta| |\tilde{y}|^{\frac{1}{2}}}} - \frac{3\sqrt{2}\varepsilon}{\lambda |\sin\theta \tilde{y}|} \ln\left(\frac{\lambda |\sin\theta| |\tilde{y}|^{\frac{3}{2}}}{\sqrt{2}\varepsilon} \right).
\end{aligned}$$

Finally, using $\tilde{y} = \lambda^{-\frac{1}{2}} y$, we get

$$\begin{aligned}
\Delta &\geq (|\cos\theta| - |\sin\theta|) \left((2\lambda)^{-\frac{1}{4}} |y|^{\frac{1}{2}} - \alpha \right) - 2\sqrt{\frac{2\varepsilon}{\lambda^{\frac{3}{4}} |\sin\theta| |y|^{\frac{1}{2}}}} - \frac{3\sqrt{2}\varepsilon}{\lambda^{\frac{1}{2}} |\sin\theta y|} \ln\left(\frac{\lambda^{\frac{1}{4}} |\sin\theta| |y|^{\frac{3}{2}}}{\sqrt{2}\varepsilon} \right) \\
&= M_-.
\end{aligned}$$

Therefore, $\Delta \geq M_-$. □

Lemma B.4. *Let us consider $M_-(\varepsilon)$ as a function of ε with $M_-(0) := \lim_{\varepsilon \to 0^+} M_-(\varepsilon)$. We have $M_-(0) > 0$ and $M_-(\varepsilon)$ is strictly decreasing for $0 \leq \varepsilon \leq \frac{1}{9} \frac{\lambda |\sin\theta| |\tilde{y}|^{\frac{3}{2}}}{\sqrt{2}(|\cos\theta| + |\sin\theta|)}$.*

Proof. We write $M_-(\varepsilon) = N_0 + N_1(\varepsilon) + N_2(\varepsilon)$, where

$$\begin{aligned}
N_0 &= (|\cos\theta| - |\sin\theta|) \left(2^{-\frac{1}{4}} \sqrt{|\tilde{y}|} - \alpha \right), \\
N_1(\varepsilon) &= -2\sqrt{\frac{2\varepsilon}{\lambda |\sin\theta| |\tilde{y}|^{\frac{1}{2}}}}, \\
N_2(\varepsilon) &= -\frac{3\sqrt{2}\varepsilon}{\lambda |\sin\theta \tilde{y}|} \ln\left(|\tilde{y}|^{\frac{1}{2}} \frac{\lambda |\sin\theta \tilde{y}|}{\sqrt{2}\varepsilon} \right).
\end{aligned}$$

Notice that N_0 does not depend on ε, and $N_1(\varepsilon)$ is decreasing in ε for all $\varepsilon \geq 0$. If $\varepsilon' := \frac{\sqrt{2}\varepsilon}{\lambda |\sin\theta \tilde{y}|}$, then

$$\begin{aligned}
N_2(\varepsilon') &= -3\varepsilon' \ln\left(\frac{|\tilde{y}|^{\frac{1}{2}}}{\varepsilon'} \right), \\
\frac{dN_2(\varepsilon')}{d\varepsilon'} &= -3\left(\ln\left(\frac{|\tilde{y}|^{\frac{1}{2}}}{\varepsilon'} \right) - 1 \right).
\end{aligned}$$

Since $0 \leq \varepsilon \leq \frac{1}{9}\frac{\lambda|\sin\theta||\tilde{y}|^{\frac{3}{2}}}{\sqrt{2(|\cos\theta|+|\sin\theta|)}}$, we have $0 \leq \varepsilon' \leq \frac{1}{9}\frac{|\tilde{y}|^{\frac{1}{2}}}{\sqrt{|\cos\theta|+|\sin\theta|}}$. As a result,

$$\frac{|\tilde{y}|^{\frac{1}{2}}}{\varepsilon'} \geq 9\sqrt{|\cos\theta|+|\sin\theta|} \geq 9 > e.$$

Therefore, $\ln\left(\frac{|\tilde{y}|^{\frac{1}{2}}}{\varepsilon'}\right) > 1$ and $\frac{dN_2(\varepsilon')}{d\varepsilon'} < 0$. It follows that $N_2(\varepsilon)$ is decreasing in ε on the given interval. Combining N_0, N_1 and N_2, we conclude that $M_-(\varepsilon)$ is decreasing in ε on the given interval. We obtain $\lim_{\varepsilon'\to 0^+}\varepsilon'\ln(\frac{|\tilde{y}|^{\frac{1}{2}}}{\varepsilon'}) = 0$ using L'Hopital's rule, so $\lim_{\varepsilon\to 0^+} N_2(\varepsilon) = 0$. Applying Assumption 3.3 yields

$$M_-(0) = (|\cos\theta| - |\sin\theta|)\left(2^{-\frac{1}{4}}\sqrt{|\tilde{y}|} - \alpha\right) > 0.$$

□

B.1 PROOF OF THEOREM 3.9

Proof. By Lemma A.3, we know that the weights converge and the residual $r(t)$ converges to zero. It follows that $\boldsymbol{\beta}^\infty := \lim_{t\to\infty}\boldsymbol{\beta}(t)$ exists and is finite. Using (30), we get

$$0 = \lim_{t\to\infty} r(t) = y - \left(\sqrt{\lambda}\cos\theta\beta_1^\infty + \sqrt{\lambda}\sin\theta\beta_2^\infty\right) = \boldsymbol{y} - X\boldsymbol{\beta}^\infty.$$

Therefore, the convergent solution $\boldsymbol{\beta}^\infty$ is an interpolating solution.

Next, we derive the stationary condition for the optimization problem (23) as

$$\nabla_{\boldsymbol{\beta}}\left(X\boldsymbol{\beta}^\infty\right) = \begin{bmatrix}\sqrt{\lambda}\cos\theta\\ \sqrt{\lambda}\sin\theta\end{bmatrix}, \tag{78}$$

$$\nabla_{\boldsymbol{\beta}} E\left(\boldsymbol{\beta}^\infty\right) = \nabla\Phi_\infty\left(\boldsymbol{\beta}^\infty\right) - \nabla\Phi_0\left(\boldsymbol{\beta}(0)\right). \tag{79}$$

The gradient in the left-hand side of (79) is equal to the difference between the convergent point and the starting point of the dual variable. We use the result of the dual dynamics in Proposition 3.6 to calculate the gradient. Recall that the dual dynamics follow

$$\frac{d\nabla\Phi_t(\boldsymbol{\beta}(t))}{dt} = -\operatorname{sgn}(\boldsymbol{\beta}(t)) \odot \frac{\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))}{|\nabla_{\boldsymbol{u}} L(\boldsymbol{w}(t))| + \varepsilon}.$$

From (29) we note that $\operatorname{sgn}(\beta_i(t)) = \operatorname{sgn}(x_i y)$. Therefore, $\operatorname{sgn}(\boldsymbol{\beta}(t))$ remains the same for all t. Integrating both sides with respect to t from 0 to infinity, it follows that

$$\nabla\Phi_\infty(\boldsymbol{\beta}^\infty) - \nabla\Phi_0(\boldsymbol{\beta}(0)) = \begin{bmatrix}\operatorname{sgn}(\cos\theta\tilde{y})\\ \operatorname{sgn}(\sin\theta\tilde{y})\end{bmatrix} \odot \left(\boldsymbol{u}^\infty - \boldsymbol{u}(0)\right).$$

Next, we want to compute the extent of the deviation from the exact KKT point. To this end, we have

$$\delta := \min_{\nu\in\mathbb{R}} \left\|\nabla_{\boldsymbol{\beta}} E\left(\boldsymbol{\beta}\right) - \nu\cdot\nabla_{\boldsymbol{\beta}}(X\boldsymbol{\beta})\right\| = \min_{\nu\in\mathbb{R}} \left\|\begin{bmatrix}\operatorname{sgn}(\cos\theta\tilde{y})\\ \operatorname{sgn}(\sin\theta\tilde{y})\end{bmatrix} \odot \left(\boldsymbol{u}^\infty - \boldsymbol{u}(0)\right) - \nu\cdot\begin{bmatrix}\sqrt{\lambda}\cos\theta\\ \sqrt{\lambda}\sin\theta\end{bmatrix}\right\|.$$

Let $V := \begin{bmatrix}\operatorname{sgn}(\cos\theta\tilde{y})\left(u_1^\infty - u_1(0)\right)\\ \operatorname{sgn}(\sin\theta\tilde{y})\left(u_2^\infty - u_2(0)\right)\end{bmatrix}$. Using orthogonal projection, we derive that

$$\begin{aligned}
\min_{\nu\in\mathbb{R}} \left\|V - \nu\cdot\begin{bmatrix}\sqrt{\lambda}\cos\theta\\ \sqrt{\lambda}\sin\theta\end{bmatrix}\right\| &= \left|\left\langle V,\ \begin{bmatrix}-\sin\theta\\ \cos\theta\end{bmatrix}\right\rangle\right| \\
&= \left|-\operatorname{sgn}(\cos\theta\tilde{y})\sin\theta\left(u_1^\infty - u_1(0)\right) + \operatorname{sgn}(\sin\theta\tilde{y})\cos\theta\left(u_2^\infty - u_2(0)\right)\right| \\
&= \left|\operatorname{sgn}(\sin\theta\cos\theta\tilde{y})\cdot\left(-|\sin\theta|\left(u_1^\infty - u_1(0)\right) + |\cos\theta|\left(u_2^\infty - u_2(0)\right)\right)\right| \\
&= \left||\cos\theta|\left(u_2^\infty - u_2(0)\right) - |\sin\theta|\left(u_1^\infty - u_1(0)\right)\right|.
\end{aligned}$$

Therefore, $\delta = |\Delta|$, where $\Delta := |\cos\theta|\,(u_2^\infty - u_2(0)) - |\sin\theta|\,(u_1^\infty - u_1(0))$. Using Lemma B.2 and Lemma B.3, we get

$$M_- \le \Delta \le M_+,$$

where

$$M_- := (|\cos\theta| - |\sin\theta|)\left((2\lambda)^{-\frac{1}{4}}\,|y|^{\frac{1}{2}} - \alpha\right) - 2\sqrt{\frac{2\varepsilon}{\lambda^{\frac{3}{4}}|\sin\theta||y|^{\frac{1}{2}}}} - \frac{3\sqrt{2}\varepsilon}{\lambda^{\frac{1}{2}}|\sin\theta y|}\ln\left(\frac{\lambda^{\frac{1}{4}}|\sin\theta||y|^{\frac{3}{2}}}{\sqrt{2}\varepsilon}\right),$$

$$M_+ := (|\cos\theta| - |\sin\theta|)\left(\lambda^{-\frac{1}{4}}|y|^{\frac{1}{2}} - \frac{\sqrt{2}\varepsilon}{4\lambda^{\frac{1}{2}}|y|}\right).$$

We can further simplify the expressions to

$$M_- = (|\cos\theta_1| - |\sin\theta_1|)\left((2\lambda_1)^{-\frac{1}{4}}|y^{(1)}|^{\frac{1}{2}} - \alpha\right) + \mathcal{O}(\sqrt{\varepsilon}),$$

$$M_+ = (|\cos\theta_1| - |\sin\theta_1|)\,\lambda_1^{-\frac{1}{4}}|y^{(1)}|^{\frac{1}{2}} + \mathcal{O}(\varepsilon).$$

We conclude that $\delta = |\Delta| \le \max\{|M_-|, |M_+|\}$. □

B.2 PROOF OF COROLLARY 3.10

Proof. Let us consider $\delta(\varepsilon), \Delta(\varepsilon), M_-(\varepsilon), M_+(\varepsilon)$ as functions of ε on the domain $\mathcal{I}_\varepsilon = [0,\ \bar{\varepsilon}]$ implied by Assumption 3.3. We define $M_-(0) := \lim_{\varepsilon\to 0^+} M_-(\varepsilon)$ so that $M_-(\varepsilon)$ is continuous on the domain. Theorem 3.9 shows that $M_+(\varepsilon)$ is linearly decreasing in ε. By Lemma B.4, $M_-(\varepsilon)$ is strictly decreasing in ε on the domain $\mathcal{I}_\varepsilon$. Lemma B.4 also shows that $M_-(0) > 0$. If $M_-(\bar{\varepsilon}) \ge 0$, then $\Delta(\varepsilon) \ge M_-(\varepsilon) \ge M_-(\bar{\varepsilon}) \ge 0$ for all $\varepsilon \in \mathcal{I}_\varepsilon$. It implies that only $M_+(\varepsilon)$ applies to the bound, i.e., $\delta(\varepsilon) \le M_+(\varepsilon)$. Let $\varepsilon^* = \bar{\varepsilon}$. It follows that for $\varepsilon \in [0,\ \varepsilon^*]$, we have

$$\delta(\varepsilon) \le \bar{M} - (|\cos\theta| - |\sin\theta|)\,\frac{\sqrt{2}\varepsilon}{4\lambda^{\frac{1}{2}}|y|}. \tag{80}$$

If $M_-(\bar{\varepsilon}) < 0$, since $M_-(0) > 0$, the monotonicity of $M_-(\varepsilon)$ ensures a unique $\hat{\varepsilon} \in (0, \bar{\varepsilon})$ such that $M_-(\hat{\varepsilon}) = 0$ and $\Delta(\varepsilon) \ge M_-(\varepsilon) \ge 0$ for $\varepsilon \in [0,\ \hat{\varepsilon}]$. Let $\varepsilon^* = \hat{\varepsilon}$. Notice that $\hat{\varepsilon}$ is positive, so $[0,\ \varepsilon^*]$ is non-degenerate. By a similar argument, we establish (80). We complete the proof by setting $\mathcal{I}' := [0,\ \varepsilon^*] \subseteq \mathcal{I}_\varepsilon$. □

B.3 PROOF OF COROLLARY 3.11

Proof. By Lemma A.3, $\lim_{t\to\infty} r^{(n)}(t) = 0$ for all $n \in \{1, \dots, N\}$ and the weights converge. We let $\bar{\boldsymbol{\beta}}^{(n)} := \lim_{t\to\infty} \boldsymbol{\beta}^{(n)}(t)$, $\bar{\boldsymbol{u}}^{(n)} := \lim_{t\to\infty} \boldsymbol{u}^{(n)}(t)$ and $\bar{\boldsymbol{v}}^{(n)} := \lim_{t\to\infty} \boldsymbol{v}^{(n)}(t)$ for each n. We also let $\boldsymbol{\beta}^\infty := \lim_{t\to\infty} \boldsymbol{\beta}(t) = \begin{bmatrix}\bar{\boldsymbol{\beta}}^{(1)} & \dots & \bar{\boldsymbol{\beta}}^{(n)}\end{bmatrix}^\top$. Using (30), we derive that for all n

$$0 = \lim_{t\to\infty} r^{(n)}(t) = y^{(n)} - x_1^{(n)}\bar{\beta}_1^{(n)} - x_2^{(n)}\bar{\beta}_2^{(n)}.$$

Therefore, $X\boldsymbol{\beta}^\infty = \boldsymbol{y}$, i.e., $\boldsymbol{\beta}^\infty$ is an interpolating solution.

Each block of $X^\top X$ is parameterized by θ_n and λ_n as

$$B^{(n)} = \begin{bmatrix}\cos\theta_n & -\sin\theta_n \\ \sin\theta_n & \cos\theta_n\end{bmatrix}\begin{bmatrix}\lambda_n & 0 \\ 0 & 0\end{bmatrix}\begin{bmatrix}\cos\theta_n & \sin\theta_n \\ -\sin\theta_n & \cos\theta_n\end{bmatrix},$$

where $|\cos\theta_n| \ge |\sin\theta_n| > 0$. Matrix $B^{(n)}$ is positive semi-definite and has rank 1, so $\lambda_n > 0$. We let $\tilde{y}^{(n)} := \frac{y^{(n)}}{\sqrt{\lambda_n}}$. The constraint $X\boldsymbol{\beta}^\infty = \boldsymbol{y}$ consists of N equality conditions

$$\left\langle \boldsymbol{x}^{(1)},\, \boldsymbol{\beta}^\infty \right\rangle = y^{(1)},$$

$$\dots$$

$$\left\langle \boldsymbol{x}^{(N)},\, \boldsymbol{\beta}^\infty \right\rangle = y^{(N)}.$$

By integrating both sides of (20), we get

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} E\left(\boldsymbol{\beta}^{\infty}\right) &= \nabla\Phi_{\infty}\left(\boldsymbol{\beta}^{\infty}\right) - \nabla\Phi_0\left(\boldsymbol{\beta}(0)\right) \\ &= \begin{bmatrix} \operatorname{sgn}(\cos\theta_1\tilde{y}^{(1)})\left(\bar{u}_1^{(1)} - u_1^{(1)}(0)\right) \\ \operatorname{sgn}(\sin\theta_1\tilde{y}^{(1)})\left(\bar{u}_2^{(1)} - u_2^{(1)}(0)\right) \\ \cdots \\ \operatorname{sgn}(\cos\theta_N\tilde{y}^{(N)})\left(\bar{u}_1^{(N)} - u_1^{(N)}(0)\right) \\ \operatorname{sgn}(\sin\theta_N\tilde{y}^{(N)})\left(\bar{u}_2^{(N)} - u_2^{(N)}(0)\right) \end{bmatrix}.\end{aligned}$$

We let $\boldsymbol{\mu} := \begin{bmatrix}\mu_1 & \dots & \mu_N\end{bmatrix}$, and then we have

$$\begin{aligned}\bar{\delta} &:= \min_{\boldsymbol{\mu}\in\mathbb{R}^N} \left\| \nabla_{\boldsymbol{\beta}} E\left(\boldsymbol{\beta}^{\infty}\right) - \sum_{n=1}^{N} \mu_n \boldsymbol{x}^{(n)} \right\| \\ &= \min_{\boldsymbol{\mu}\in\mathbb{R}^N} \left\| \nabla_{\boldsymbol{\beta}} E\left(\boldsymbol{\beta}^{\infty}\right) - \begin{bmatrix} \mu_1\sqrt{\lambda_1}\cos\theta_1 \\ \mu_1\sqrt{\lambda_1}\sin\theta_1 \\ \cdots \\ \mu_N\sqrt{\lambda_N}\cos\theta_N \\ \mu_N\sqrt{\lambda_N}\sin\theta_N \end{bmatrix} \right\| \\ &= \min_{\boldsymbol{\mu}\in\mathbb{R}^N} \left(\sum_{n=1}^{N} \left\| \begin{bmatrix} \operatorname{sgn}(\cos\theta_n\tilde{y}^{(n)})\left(\bar{u}_1^{(n)} - u_1^{(n)}(0)\right) \\ \operatorname{sgn}(\sin\theta_n\tilde{y}^{(n)})\left(\bar{u}_2^{(n)} - u_2^{(n)}(0)\right) \end{bmatrix} - \mu_n \cdot \begin{bmatrix} \sqrt{\lambda_n}\cos\theta_n \\ \sqrt{\lambda_n}\sin\theta_n \end{bmatrix} \right\|^2 \right)^{\frac{1}{2}} \\ &\le \sum_{n=1}^{N} \min_{\mu_n\in\mathbb{R}} \left\| \begin{bmatrix} \operatorname{sgn}(\cos\theta_n\tilde{y}^{(n)})\left(\bar{u}_1^{(n)} - u_1^{(n)}(0)\right) \\ \operatorname{sgn}(\sin\theta_n\tilde{y}^{(n)})\left(\bar{u}_2^{(n)} - u_2^{(n)}(0)\right) \end{bmatrix} - \mu_n \cdot \begin{bmatrix} \sqrt{\lambda_n}\cos\theta_n \\ \sqrt{\lambda_n}\sin\theta_n \end{bmatrix} \right\|.\end{aligned}$$

By Theorem 3.9, it follows that for each n,

$$\begin{aligned}\delta_n &:= \min_{\mu_n\in\mathbb{R}} \left\| \begin{bmatrix} \operatorname{sgn}(\cos\theta_n\tilde{y}^{(n)})\left(\bar{u}_1^{(n)} - u_1^{(n)}(0)\right) \\ \operatorname{sgn}(\sin\theta_n\tilde{y}^{(n)})\left(\bar{u}_2^{(n)} - u_2^{(n)}(0)\right) \end{bmatrix} - \mu_n \cdot \begin{bmatrix} \sqrt{\lambda_n}\cos\theta_n \\ \sqrt{\lambda_n}\sin\theta_n \end{bmatrix} \right\| \\ &\le \max\left\{ \left| M_+^{(n)} \right|, \left| M_-^{(n)} \right| \right\}.\end{aligned}$$

Therefore, $\bar{\delta} \le \sum_{n=1}^{N} \delta_n \le \sum_{n=1}^{N} \max\left\{ \left| M_+^{(n)} \right|, \left| M_-^{(n)} \right| \right\}$. □

B.4 PROOF OF COROLLARY 3.12

Proof. For each $n \in \{1, \dots, N\}$, we apply Corollary 3.10 and show that there exists a non-degenerate interval $\mathcal{I}_n' = [0,\ \varepsilon_n^*]$ such that for all $\varepsilon \in \mathcal{I}_n'$, we have

$$\delta_n(\varepsilon) \le \left(|\cos\theta_n| - |\sin\theta_n|\right)\left(\lambda_n^{-\frac{1}{4}} |y^{(n)}|^{\frac{1}{2}}\right) - \left(|\cos\theta_n| - |\sin\theta_n|\right)\frac{\sqrt{2}\varepsilon}{4\lambda_n^{\frac{1}{2}}|y^{(n)}|}. \tag{81}$$

We let $\tilde{\varepsilon} := \min_n\{\varepsilon_n^*\}$ and let $\mathcal{J} := \bigcap_{n=1}^{N} \mathcal{I}_n' = [0,\ \tilde{\varepsilon}]$. Since each $\mathcal{I}_n'$ is non-degenerate, we have $\varepsilon_n^* > 0$ for all n and $\tilde{\varepsilon} > 0$. Therefore, the interval $\mathcal{J}$ is non-degenerate. In turn, for all $\varepsilon \in \mathcal{J}$, the relation (81) holds. By Corollary 3.11, we have

$$\bar{\delta}(\varepsilon) \le \sum_{n=1}^{N} \delta_n(\varepsilon) \le \sum_{n=1}^{N} \left(|\cos\theta_n| - |\sin\theta_n|\right)\left(\lambda_n^{-\frac{1}{4}} |y^{(n)}|^{\frac{1}{2}}\right) - \left(\sum_{n=1}^{N} \left(|\cos\theta_n| - |\sin\theta_n|\right)\frac{\sqrt{2}}{4\lambda_n^{\frac{1}{2}}|y^{(n)}|} \right)\varepsilon.$$

□

C DERIVATION OF DUAL DYNAMICS FOR GRADIENT DESCENT

When applying GD to minimize loss (7) with respect to weights, in the continuous-time limit we have

$$\frac{d\boldsymbol{w}^+(t)}{dt} = -\boldsymbol{w}^+(t) \odot X^\top (X\boldsymbol{\beta}(t) - \boldsymbol{y}) \tag{82}$$

$$\frac{d\boldsymbol{w}^-(t)}{dt} = \boldsymbol{w}^-(t) \odot X^\top (X\boldsymbol{\beta}(t) - \boldsymbol{y}). \tag{83}$$

With initialization $\boldsymbol{w}^+(0) = \boldsymbol{w}^-(0) = \alpha\mathbf{1}$, we write implicit solutions to (82) and (83) as

$$\boldsymbol{w}^+(t) = \alpha \exp\left(-\int_0^t X^\top (X\boldsymbol{\beta}(s) - \boldsymbol{y})\, ds\right)$$

$$\boldsymbol{w}^-(t) = \alpha \exp\left(\int_0^t X^\top (X\boldsymbol{\beta}(s) - \boldsymbol{y})\, ds\right).$$

Therefore, we have

$$\begin{aligned}
\boldsymbol{\beta}(t) &= \boldsymbol{w}^+(t) \odot \boldsymbol{w}^+(t) - \boldsymbol{w}^-(t) \odot \boldsymbol{w}^-(t) \\
&= \alpha^2 \left[\exp\left(-2\int_0^t X^\top (X\boldsymbol{\beta}(s) - \boldsymbol{y})\, ds\right) - \exp\left(2\int_0^t X^\top (X\boldsymbol{\beta}(s) - \boldsymbol{y})\, ds\right)\right] \\
&= 2\alpha^2 \sinh\left(-2\int_0^t X^\top (X\boldsymbol{\beta}(s) - \boldsymbol{y})\, ds\right).
\end{aligned}$$

It follows that

$$\frac{1}{2\alpha^2}\boldsymbol{\beta}(t) = \sinh\left(-2\int_0^t X^\top (X\boldsymbol{\beta}(s) - \boldsymbol{y})\, ds\right)$$

$$\operatorname{arcsinh}\left(\frac{\boldsymbol{\beta}(t)}{2\alpha^2}\right) = -\int_0^t 2X^\top (X\boldsymbol{\beta}(s) - \boldsymbol{y})\, ds$$

$$\frac{d \operatorname{arcsinh}\left(\frac{\boldsymbol{\beta}(t)}{2\alpha^2}\right)}{dt} = -2X^\top (X\boldsymbol{\beta}(t) - \boldsymbol{y}).$$

We note that $\nabla_{\boldsymbol{\beta}} L(\boldsymbol{\beta}(t)) = \frac{1}{2} X^\top (X\boldsymbol{\beta}(t) - \boldsymbol{y})$. In turn, we have

$$\frac{d \operatorname{arcsinh}\left(\frac{\boldsymbol{\beta}(t)}{2\alpha^2}\right)}{dt} = -2X^\top (X\boldsymbol{\beta}(t) - \boldsymbol{y}) = -4\nabla_{\boldsymbol{\beta}} L(\boldsymbol{\beta}(t)). \tag{84}$$

Given the potential function $\Psi_\alpha(\boldsymbol{\beta}(t)) = \frac{1}{4}\left(\sum_{i=1}^D \beta_i \operatorname{arcsinh}(\frac{\beta_i}{2\alpha^2}) + \sqrt{\beta_i^2 + 4\alpha^4}\right)$, we have the mirror map

$$\nabla \Psi_\alpha(\boldsymbol{\beta}(t)) = \frac{1}{4} \operatorname{arcsinh}\left(\frac{\boldsymbol{\beta}(t)}{2\alpha^2}\right). \tag{85}$$

Combining (84) and (85), we get the dual dynamics for GD

$$\frac{d\nabla \Psi_\alpha(\boldsymbol{\beta}(t))}{dt} = -\nabla_{\boldsymbol{\beta}} L(\boldsymbol{\beta}(t)).$$