A. Preliminaries

In this section, we provide some definitions and basic lemmas which will be used in the proof. The section is organized as follows: in Subsection A.1, we show general properties which exponential loss and logistic loss share; in Subsection A.2, (approximate) KKT conditions is defined and sufficient conditions of being an approximate KKT point is given; in Subsection A.3, we show how conditioners of AdaGrad, RMSProp, and Adam in continuous flow is formulated; in Subsection A.4, we introduce o-Minimal structure, definable set and definable functions, and show two Kurdyka-Lojasiewicz inequalities; in Subsection A.6, we show some basic definitions from Measure Theory, including measurable set and Lebesgue Integrability.

A.1. Property of Exponential and Logistic Loss

In this subsection, we provide several properties which both exponential and logistic loss possess. The properties of exponential and logistic loss can be described as the following proposition:

Proposition 1. *For $\ell \in \{\ell_{exp}, \ell_{\log}\}$:*

- *There exists a C^1 function f, such that $\ell = e^{-f}$;*
- *For any $x \in \mathbb{R}$, $f' > 0$. Therefore, f is reversible, and $f^{-1} \in C^1$;*
- *$f'(x)x$ is non-decreasing for $x \in (0, \infty)$, and $\lim_{x\to\infty} f'(x)x = \infty$;*
- *There exists a large enough x_f and a constant $K \geq 1$,such that,*
  - *$\forall \theta \in [\frac{1}{2}, 1), \forall x \in (x_f, \infty)$, and $\forall y \in f^{-1}(x_f, \infty)$: $(f^{-1})'(x) \leq K(f^{-1})'(\theta x)$ and $f'(y) \leq Kf'(\theta y)$;*
  - *For all $y \in [x_f, \infty)$, $\frac{f(x)}{f'(x)} \in [\frac{1}{2K}x, 2Kx]$;*
  - *For all $x \in [f^{-1}(x_f), \infty)$, $\frac{f^{-1}(x)}{(f^{-1})'(x)} \in [\frac{1}{2K}x, 2Kx]$.*
  - *$f(x) = \Theta(x)$ as $x \to \infty$.*

All properties are easy to verify in Proposition 1 and we omit it here. For brevity, we will use $g(x) = f^{-1}(x)$ in the following proofs.

A.2. KKT Condition

Being a KKT point is a first order necessary condition for being an optimal point. We first give the definition of approximate KKT point for general optimization problem (Q).

Definition 2. *Consider the following optimization problem (Q) for $\boldsymbol{x} \in \mathbb{R}^d$:*

$$\min f(\boldsymbol{x})$$
$$\text{subject to } g_i(x) \leq 0, \forall i \in [N],$$

where f, g_i $(i = 1, \cdots, N)$: $\mathbb{R}^d \to \mathbb{R}$ are locally Lipschitz functions. We say that $x \in \mathbb{R}^d$ is a feasible point of (P) if x satisfies $g_i(x) \leq 0$ for all $i \in [N]$.

For any $\varepsilon, \delta > 0$, a feasible point of (Q) is an (ε, δ)-KKT point if there exists $\lambda_i \geq 0$, $\mathbf{k} \in \bar{\partial} f(\boldsymbol{x})$, and $\boldsymbol{h}_i \in \bar{\partial} g_i(\boldsymbol{x})$ for all $i \in [N]$ (we will slightly abuse $\bar{\partial} f(\boldsymbol{x})$ to respresent a element in $\bar{\partial} f(\boldsymbol{x})$) such that

1. $\left\| \boldsymbol{k} + \sum_{i\in[N]} \lambda_i \boldsymbol{h}_i(\boldsymbol{x}) \right\|_2 \leq \varepsilon$;

2. $\forall i \in [N] : \lambda_i g_i(\boldsymbol{x}) \geq -\delta$.

Specifically, when $\varepsilon = \delta = 0$, we call $\boldsymbol{x}$ a KKT point of (Q).

The following Mangasarian-Fromovitz constraint qualification (MFCQ) bridges (ε, δ) KKT points with KKT points.

Definition 3. *A feasible point $\boldsymbol{x}$ of (Q) is said to satisfy MFCQ if there exists $\boldsymbol{a} \in \mathbb{R}^d$ such that for every $i \in [N]$ with $g_i(\boldsymbol{x}) = 0$,*

$$\forall \boldsymbol{h} \in \bar{\partial} g_i(\boldsymbol{x}) : \langle \boldsymbol{h}, \boldsymbol{a} \rangle > 0.$$

MFCQ guarantees that the limit of approximate KKT point with convergent ε and δ is a KKT point.

Lemma 7. *Suppose for any $k \in \mathbb{N}$, $\boldsymbol{x}_k$ is a $(\varepsilon_k, \delta_k)$-KKT point of (Q) defined in Definition 2. If $\lim_{k\to\infty} \varepsilon_k = 0$, $\lim_{k\to\infty} \delta_k = 0$, and $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}$, where the limit point $\boldsymbol{x}$ satisfies MFCQ, then $\boldsymbol{x}$ is a KKT point of (Q).*

A.3. How is the Continuous Form of Conditioner Formulated?

In this subsection, we show how conditioners of the continuous case for AdaGrad, RMSProp, Adam (w/m) are derived. Both discrete updates of these optimizers can be written as

$$\boldsymbol{w}(t+1) - \boldsymbol{w}(t) = -\eta \frac{1}{\sqrt{\varepsilon + \psi(\boldsymbol{m}(t), t)}} \odot \bar{\partial}\mathcal{L}(\boldsymbol{w}(t)), \tag{6}$$

$$\boldsymbol{m}(t+1) - \boldsymbol{m}(t) = \phi(\boldsymbol{m}(t), \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))), \tag{7}$$

$$\boldsymbol{m}(0) = \boldsymbol{0}.$$

where for AdaGrad, $\phi(\boldsymbol{m}(t), \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))) = \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2$, $\psi(\boldsymbol{m}(t), t) = \boldsymbol{m}(t)$; for RMSProp, $\phi(\boldsymbol{m}(t), \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))) = (1-b)(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2 - \boldsymbol{m}(t))$, $\psi(\boldsymbol{m}(t), t) = \boldsymbol{m}(t)$; for Adam (w/m), $\phi(\boldsymbol{m}(t), \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))) = (1-b)(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2 - \boldsymbol{m}(t))$, $\psi(\boldsymbol{m}(t), t) = \frac{\boldsymbol{m}(t)}{1-b^t}$.

One can easily observe that eqs. (6) and (7) is a discretization of the following equations:

$$\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t} = -\frac{1}{\sqrt{\varepsilon + \psi(\boldsymbol{m}(t), t)}} \odot \bar{\partial}\mathcal{L}(\boldsymbol{w}(t)), \tag{8}$$

$$\frac{\mathrm{d}\boldsymbol{m}(t)}{\mathrm{d}t} = \phi(\boldsymbol{m}(t), \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))), \tag{9}$$

$$\boldsymbol{m}(0) = \boldsymbol{0}.$$

As for AdaGrad,

$$\frac{\mathrm{d}\boldsymbol{m}(t)}{\mathrm{d}t} = \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2,$$

which leads to

$$\boldsymbol{m}(t) = \int_0^t \bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2 \mathrm{d}\tau,$$

and

$$\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t} = -\frac{1}{\sqrt{\varepsilon + \int_0^t \bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2 \mathrm{d}\tau}} \odot \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))$$

As for RMSProp and Adam

$$\frac{\mathrm{d}\boldsymbol{m}(t)}{\mathrm{d}t} = (1-b)(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2 - \boldsymbol{m}(t)).$$

By solving the above differential equation, we have

$$\frac{\mathrm{d}e^{(1-b)t}\boldsymbol{m}(t)}{\mathrm{d}t} = e^{(1-b)t}(1-b)\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2,$$

which by integration implies

$$\boldsymbol{m}(t) = \int_0^t e^{-(1-b)(t-\tau)}(1-b)\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2 \mathrm{d}\tau.$$

Therefore, for RMSProp, the continuous flow is

$$\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t} = -\frac{1}{\sqrt{\varepsilon + \int_0^t e^{-(1-b)(t-\tau)}(1-b)\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2 \mathrm{d}\tau}} \odot \bar{\partial}\mathcal{L}(\boldsymbol{w}(t));$$

while for Adam (w/m), the continuous flow is

$$\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t} = -\frac{1}{\sqrt{\varepsilon + \frac{\int_0^t e^{-(1-b)(t-\tau)}(1-b)\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2 \mathrm{d}\tau}{1-b^t}}} \odot \bar{\partial}\mathcal{L}(\boldsymbol{w}(t)).$$

A.4. o-Minimal Structure and Definable functions

Here we define o-Minimal structure and definable functions which we omit in Theorem 3.

Definition 4 (Appendix B, Ji & Telgarsky (2020)). *An o-minimal structure is a collection $\mathcal{S} = \{\mathcal{S}_n\}_{n=1}^{\infty}$, where each $\mathcal{S}_n$ is a set of subsets of $\mathbb{R}_n$ satisfying the following conditions: 1. $\mathcal{S}_1$ is the collection of all finite unions of open intervals and points;*
2. $\mathcal{S}_n$ includes the zero sets of all polynomials on $\mathbb{R}_n$;
3. $\mathcal{S}_n$ is closed under finite union, finite intersection, and complement;
4. $\mathcal{S}$ is closed under Cartesian products: if $A \in \mathcal{S}_m$ and $B \in \mathcal{S}_n$, then $A \times B \in \mathcal{S}_{m+n}$;
5. $\mathcal{S}$ is closed under projection Π_n onto the first n coordinates: if $A \in \mathcal{S}_{n+1}$, then $\Pi_n(A) \in \mathcal{S}_n$.

A definable function on above o-Minimal Structure can be defined as follows:

Definition 5 (Appendix B, Ji & Telgarsky (2020)). *A function $f : D \to \mathbb{R}^m$ with $D \subset \mathbb{R}^n$ is definable if the graph of f is in $\mathcal{S}_{n+m}$.*

A natural question is: which function is definable? The next Lemma helps to solve this question.

Lemma 8 (Lemma B.2, Ji & Telgarsky (2020)).

- *All polynomials are definable, therefore, linear or other polynomial activation is definable;*
- *If both $f(x)$ and $g(x)$ are definable, $\min f(x), g(x)$ and $\max f(x), g(x)$ are definable, therefore, ReLU activation is definable;*
- *If $f_i(x) : D \to \mathbb{R}$ $(i = 1, 2, \cdots, n)$ is definable, then $\boldsymbol{f}(x) = (f_1(x), f_2(x), \cdots, f_n(x))$ is definable.*
- *Suppose there exists $k, d_0, d_1, \cdots, d_L > 0$, and L definable functions $(g_1, g_2, \cdots, g_L)$, where $g_j :\to \mathbb{R}^{d_0} \times \cdots \times \mathbb{R}^{d_{j-1}} \times \mathbb{R}^k \to \mathbb{R}^{d_j}$. Let $h_1(x, W) \triangleq g_1(x, W)$, and for $2 \le j \le L$,*

$$h_j(x, W) := g_j\left(x, h_1(x, W), \ldots, h_{j-1}(x, W), W\right),$$

then all h_j are definable. Therefore, neural networks with polynomial and ReLU activation, convolutional and max-pooling layers, and skip connections are definable.

An important property for definable function is Kurdyka-Lojasiewicz inequality, which can bound gradient of definable function in a small region. Here we present two Kurdyka-Lojasiewicz inequalities given by (Ji & Telgarsky, 2020):

Lemma 9 (Lemma 3.6, Ji & Telgarsky (2020)). *Given a locally Lipschitz definable function f with an open domain $D \in \{x | \|x\| > 1\}$, for any c, $\eta > 0$, there exists $a > 0$ and a definable desingularizing function Ψ on $[0, a)$ (that is, $\Psi(x) \in C^1((0, a)) \cap C^0([0, a))$ with $\Psi(0) = 0$), such that,*

$$\Psi'(f(x))\|x\|\|\bar{\partial} f(x)\| \ge 1, \text{ if } f(x) \in (0, a), \text{ and } \left\|\bar{\partial}_{\perp} f(x)\right\| \ge c\|x\|^{\eta} \left\|\bar{\partial}_{\setminus\setminus} f(x)\right\|.$$

Lemma 10 (Lemma 3.7, Ji & Telgarsky (2020)). *Given a locally Lipschitz definable function f with an open domain $D \subset \{x | \|x\| > 1\}$, for any $\lambda > 0$, there exists $a > 0$, and a definable desingularizing function Ψ on $[0, a)$ such that*

$$\max\left\{1, \frac{2}{\lambda}\right\} \Psi'(f(x))\|x\|^{1+\lambda}\|\bar{\partial} f(x)\| \ge 1, \text{ if } f(x) \in (0, a).$$

At the end of this subsection, we show that definability actually guarantees that Φ admits a chain rule, which is formally stated as following:

Lemma 11 (Lemma B.9, Ji & Telgarsky (2020)). *Given a locally Lipschitz definable $f : D \to \mathbb{R}$ with an open domain D, for any interval I and any arc $z : I \to D$, it holds for a.e. $t \in I$ that*

$$\frac{\mathrm{d} f(z_t)}{\mathrm{d} t} = \left\langle z_t^*, \frac{\mathrm{d} z_t}{\mathrm{d} t} \right\rangle, \quad \text{for all } z_t^* \in \partial f(z_t).$$

Therefore, if we are deal with definable neural networks Φ as in Theorem 3, we no longer need to assume Φ admits a chain rule which is already guaranteed by lemma 11.

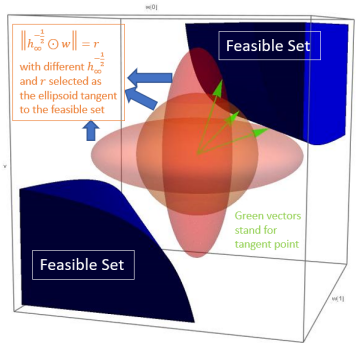


Figure 3. How $\boldsymbol{h}_\infty$ influence convergence direction of $\boldsymbol{w}$

A.5. Discussion of the influence of initialization on the solution of P^A

For AdaGrad, $\boldsymbol{h}_\infty^{-2}$ is defined as $\varepsilon \mathbf{1}_p + \sum_{t=0}^{\infty} \nabla \mathcal{L}(\boldsymbol{w}(t))^2$, which is the sum of squared gradients along the trajectory. Intuitively, as the initialization changes, the trajectory changes respectively, and so does the direction of $\boldsymbol{h}_\infty$. This intuition can be further verified by Experiment in Section 6.2, where we plot the direction of $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ as the initialization changes. Furthermore, how $\boldsymbol{h}_\infty$ influence the max-margin problem can be interpreted as follows: optimizing $\|\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \boldsymbol{w}\|^2$ with constraints is equivalent to find the radius r of ellipsoid $\|\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \boldsymbol{w}\|^2 = r^2$ when the ellipsoid is tangent to the feasible set. This intuition is visualized in Figure 3. One can easily observe that as the direction of $\boldsymbol{h}_\infty$ changes, the direction of the tangent point changes.

A.6. Basic knowledge from Measure Theory

In this section, we present basic definitions of measurable set, measurable functions and Lebesgue Integrability. These definition involves use of exterior measure and Borel set in Euclidean space, which we omit them here. Readers interested in measure theory can refer to (Stein & Shakarchi, 2009) for details.

Definition 6 (Stein & Shakarchi (2009), Chapter 1, page 16)**.** *A subset E of $\mathbb{R}^d$is Lesbesgue measurable, or simply measurable, if for any $\varepsilon > 0$, there exists an open set G, with $E \subset G$, and*

$$m_*(G/E) \leq \varepsilon,$$

where m_* is the exterior measure on $\mathbb{R}^d$.

Definition 7 (Stein & Shakarchi (2009), Chapter 1, page 28)**.** *A function f on a measurable subset E of $\mathbb{R}^d$ is measurable if for all $a \in \mathbb{R}$, the set*

$$f^{-1}([-\infty, a)) = \{x \in E : f(x) < a\}$$

is measurable.

Definition 8 (Stein & Shakarchi (2009), Chapter 1, page 64)**.** *A measurable function f defined on a measurable subset of $\mathbb{R}^d$ is Lesbesgue integrable if*

$$\int_E |f(x)| \mathrm{d}m(x) < \infty.$$

B. Proof of Results for Adaptive Algorithms in Continuous Case

This section collects proof of Theorem 2, Theorem 4, Theorem 5, and also contains proof of Theorem 6 and Theorem 7. Organization of this section is as follows: In Subsection B.1, we present proof of Theorems 4 and Theorem 5; in Subsection B.2, we present proof of Theorem 2 based on the proof skeleton in Section 5; in Subsection B.3, we prove Theorem 6 and Theorem 7 based on 2, Theorem 4, Theorem 5; finally, in Subsection B.4, we provide tight convergence rate of loss and parameter norm in adaptive gradient flows.

B.1. Proof of Theorem 4 and Theorem 5: Transition from Continuous Adaptive Algorithms to Adaptive Gradient Flow

B.1.1. Proof of Theorem 4

The proof of Theorem 4 is divided into two stages: we first prove convergence of $\boldsymbol{h}^A(t)$ and $\boldsymbol{\beta}^A(t)$; then we show $\frac{\mathrm{d}\boldsymbol{v}^A(t)}{\mathrm{d}t} = -\boldsymbol{\beta}^A(t) \odot \bar{\partial}\tilde{\mathcal{L}}^A(\boldsymbol{v}^A(t))$ satisfies adaptive gradient flow, and is equivalent to AdaGrad flow.

We first show $\int_0^\infty \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2 \mathrm{d}\tau$ is bounded.

Lemma 12. *For AdaGrad flow defined as eq. (4) with $\boldsymbol{h} = \boldsymbol{h}^A$,*

$$\int_0^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau < \infty, i = 1, \cdots, p.$$

Proof. We use reduction of absurdity. If there exists an i, such that, $\int_0^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau$ diverges, by equivalence of integral convergence, then

$$\int_0^\infty \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\varepsilon + \int_0^t (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau} \mathrm{d}t = \infty.$$

Since $\int_0^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau = \infty$, when t is large enough,

$$\varepsilon + \int_0^t (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau > \sqrt{\varepsilon + \int_0^t (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau}.$$

Therefore,

$$\int_0^\infty \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon + \int_0^t (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau}} \mathrm{d}t = \infty.$$

By integrating $\frac{\mathrm{d}\mathcal{L}(\boldsymbol{w}(t))}{\mathrm{d}t}$,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}(0)) - \mathcal{L}(\boldsymbol{w}(t)) &= -\int_0^t \frac{\mathrm{d}\mathcal{L}(\boldsymbol{w}(\tau))}{\mathrm{d}\tau} \mathrm{d}\tau \\
&= -\int_0^t \left\langle \bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)), \frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} \right\rangle \mathrm{d}\tau \\
&= \int_0^t \langle \bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)), \boldsymbol{h}^A(\tau) \odot \bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)) \rangle \mathrm{d}\tau \\
&= \sum_{i=1}^p \int_0^t \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(s)))_i^2}{\sqrt{\varepsilon + \int_0^t (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau}} \mathrm{d}s \\
&= \infty,
\end{aligned}$$

which leads to a contradictory, since $\mathcal{L}(\boldsymbol{\omega}(0)) - \mathcal{L}(\boldsymbol{\omega}(t))$ is upper bounded by $\mathcal{L}(\boldsymbol{\omega}(0))$.

The proof is completed. □

Now we are ready to prove Theorem 4.

Theorem 9 (Theorem 4, restated)**.** *Define $\boldsymbol{h}_\infty = \lim_{t\to\infty} \boldsymbol{h}^A(t)$. Then $\boldsymbol{h}_\infty$ has no zero elements. Let*

$$\begin{aligned}
\boldsymbol{v}^A(t) &= \boldsymbol{h}_\infty^{-1/2} \odot \boldsymbol{w}(t), \\
\boldsymbol{\beta}^A(t) &= \boldsymbol{h}_\infty^{-1} \odot \boldsymbol{h}^A(t), \\
\tilde{\mathcal{L}}^A(\boldsymbol{v}) &= \mathcal{L}(\boldsymbol{h}_\infty^{\frac{1}{2}} \odot \boldsymbol{v}).
\end{aligned}$$

We have

$$\frac{\mathrm{d}\boldsymbol{v}^A(t)}{\mathrm{d}t} = -\boldsymbol{\beta}^A(t) \odot \bar{\partial}\tilde{\mathcal{L}}^A(\boldsymbol{v}^A(t)), \tag{10}$$

while $\lim_{t\to\infty} \boldsymbol{\beta}^A(t) = 1$, *and* $\frac{\mathrm{d}\log \boldsymbol{\beta}^A(t)}{\mathrm{d}t}$ *is Lebesgue integrable.*

Proof. Since

$$\boldsymbol{h}_\infty = \frac{1}{\sqrt{\varepsilon \mathbf{1}_p + \int_0^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau}},$$

by Lemma 12, $\boldsymbol{h}_\infty$ has no zero elements.

We then prove eq. (10) by direct calculation.

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{v}^A(t)}{\mathrm{d}t} &= \boldsymbol{h}_\infty^{-1/2} \odot \frac{d\boldsymbol{w}(t)}{\mathrm{d}t} \\
&= -\boldsymbol{h}_\infty^{-1/2} \odot \boldsymbol{h}^A(t) \odot \bar{\partial}\mathcal{L}(\boldsymbol{w}(t)) \\
&= -\boldsymbol{h}_\infty^{-1/2} \odot \boldsymbol{h}^A(t) \odot \boldsymbol{h}_\infty^{-1/2} \odot \boldsymbol{h}_\infty^{1/2} \odot \bar{\partial}_{\boldsymbol{w}}\tilde{\mathcal{L}}^A(\boldsymbol{h}_\infty^{-1/2} \odot \boldsymbol{w}(t)) \\
&= -\boldsymbol{\beta}^A(t) \odot \bar{\partial}_{\boldsymbol{h}_\infty^{-1/2}\odot \boldsymbol{w}(t)}\tilde{\mathcal{L}}^A(\boldsymbol{h}_\infty^{-1/2} \odot \boldsymbol{w}(t)) \\
&= -\boldsymbol{\beta}^A(t) \odot \bar{\partial}_{\boldsymbol{v}(t)}\tilde{\mathcal{L}}^A(\boldsymbol{v}(t)),
\end{aligned}$$

which completes the proof of eq. (10).

$\lim_{t\to\infty} \boldsymbol{\beta}^A(t) = 1$ can be derived directly by convergence of $\boldsymbol{h}^A(t)$, while since

$$\frac{\mathrm{d}\log \boldsymbol{\beta}^A(t)}{\mathrm{d}t} = \frac{\mathrm{d}\log \boldsymbol{h}^A(t)}{\mathrm{d}t} \overset{(*)}{\leq} \mathbf{0},$$

where inequality $(*)$ is due to $\boldsymbol{h}^A$ is non-increasing.

Therefore,

$$\int_0^\infty \left|\frac{\mathrm{d}\log \boldsymbol{\beta}^A(t)}{\mathrm{d}t}\right| \mathrm{d}t = -\int_0^\infty \frac{\mathrm{d}\log \boldsymbol{\beta}^A(t)}{\mathrm{d}t} \mathrm{d}t = \log \boldsymbol{\beta}^A(0) < \infty.$$

The proof is completed. □

B.1.2. Proof of Theorem 5

We first prove Theorem 5 for RMSProp, and then extend the proof for Adam (w/m) and other Adam-like optimizers. The proof strategy is similar with AdaGrad: we first prove convergence of $\boldsymbol{h}^R(t)$ and integrability $\frac{\mathrm{d}\log \boldsymbol{\beta}^R(t)}{\mathrm{d}t}$; then we show $\frac{\mathrm{d}\boldsymbol{v}^R(t)}{\mathrm{d}t} = -\boldsymbol{\beta}^R(t) \odot \bar{\partial}\tilde{\mathcal{L}}^R(\boldsymbol{v}^R(t))$ is equivalent to RMSProp flow, and satisfies adaptive gradient flow.

However, for RMSProp flow, the convergence of $\boldsymbol{h}^R$ requires more effort. We start from the following lemma, which bounds $F_i(t) \triangleq \int_0^t (1-b)b^{t-\tau}(\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau)))_i^2 \mathrm{d}\tau$ ($i \in [p]$).

Lemma 13. *For RMSProp flow defined as eq. (4) with* $\boldsymbol{h} = \boldsymbol{h}^R$, $F_i(t)$ *is bounded* ($i = 1, 2, \cdots, p$), *that is,*

$$\overline{\lim}_{t\to\infty} F_i(t) < \infty.$$

Proof. When $b = 1$, $F_i(t) = 0$ for all t and i, which trivially yields the claim. When $b \neq 1$, we use reduction of absurdity. If there exists an i, such that, $\overline{\lim}_{t\to\infty} F_i(t) = \infty$, then $t_k \triangleq \inf\{t : F_i(t) \geq k\} < \infty$ holds. Furthermore, since $\mathcal{L}$ is locally Lipschitz with respect to $\boldsymbol{w}$, $g_i(t)$ is locally bounded for any t, which leads to the absolute continuity of F_i. Therefore, since $F_i(0) = 0$, t_k monotonously increases.

Therefore, we have that

$$\begin{aligned}\int_{t_k}^{t_{k+1}} \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon+F_i(t)}}\mathrm{d}t &\geq \int_{t_k}^{t_{k+1}} \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon+k+1}}\mathrm{d}t \\ &= \frac{1}{\sqrt{\varepsilon+k+1}}\int_{t_k}^{t_{k+1}} (\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t \\ &\geq \frac{1}{\sqrt{\varepsilon+k+1}}\int_{t_k}^{t_{k+1}} (1-b)e^{-(1-b)(t_{k+1}-t)}(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t \\ &= \frac{1-b}{\sqrt{\varepsilon+k+1}}(F_i(t_{k+1}) - e^{-(1-b)(t_{k+1}-t_k)}F_i(t_k)) \\ &\geq \frac{1-b}{\sqrt{\varepsilon+k+1}}(k+1-e^{-(1-b)(t_{k+1}-t_k)}k) \\ &\geq \frac{1-b}{\sqrt{\varepsilon+k+1}}.\end{aligned}$$

Adding all $k \geq 1$, we then have

$$\mathcal{L}(t_1) \geq \int_{t_1}^{\infty} \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon+F_i(t)}}\mathrm{d}t = \infty,$$

which leads to a contradictory.

The proof is completed. □

The next lemma shows that $\int_0^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t$ converges, which indicates that $\lim_{t\to\infty} F_i(t) = 0$.

Lemma 14. *For RMSProp flow defined as eq. (4) with $\boldsymbol{h} = \boldsymbol{h}^R$, $\int_0^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t$ converges, which indicates that $\lim_{t\to\infty} F_i(t) = 0$ $(i = 1, 2, \cdots, p)$. Consequently, $\lim_{t\to\infty} F_i(t) = 0$ and $\lim_{t\to\infty} \boldsymbol{h}^R(t) = \frac{1}{\sqrt{\varepsilon}}\mathbf{1}_p$.*

Proof. Similar to Lemma 13, when $b = 1$, the claim trivially holds. When $b \neq 1$, by Lemma 13, there exist $M_i > 0$ $(i = 1, 2, \cdots, p)$, such that, $F_i(t) \leq M_i$ for any $t > 0$. Therefore,

$$\mathcal{L}(0) \geq \sum_{i=1}^{p}\int_0^\infty \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon+F_i(t)}}\mathrm{d}t \geq \sum_{i=1}^{p}\int_0^\infty \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{M_i+\varepsilon}}\mathrm{d}t,$$

which proves $\int_0^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t < \infty$.

Therefore, for any positive real $\varepsilon > 0$ and a fixed index $i \in [p]$, there exists a time T, such that,

$$\int_T^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t \leq \varepsilon.$$

Thus, for any $t \geq T$,

$$F_i(t) = e^{-(t-T)(1-b)}F_i(T) + \int_T^\infty (\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t \leq e^{-(t-T)(1-b)}F_i(T) + \varepsilon,$$

which leads to

$$\overline{\lim}_{t\to\infty} F_i(t) \leq \varepsilon.$$

Since ε and i can be picked arbitrarily, the proof is completed. □

Similar to the proof of Lemma 9, we can rewrite the RMSProp flow as

$$\frac{\mathrm{d}\boldsymbol{v}^R(t)}{\mathrm{d}t} = -\boldsymbol{\beta}^R(t) \odot \bar{\partial}\tilde{\mathcal{L}}^R(\boldsymbol{v}^R(t)),$$

where

$$\begin{aligned}\boldsymbol{v}^R(t) &= \sqrt[4]{\varepsilon}\boldsymbol{w}(t),\\ \boldsymbol{\beta}^R(t) &= \sqrt{\varepsilon}\boldsymbol{h}(t),\\ \tilde{\mathcal{L}}^R(\boldsymbol{v}) &= \mathcal{L}(\sqrt[4]{\varepsilon^{-1}}\boldsymbol{v}),\end{aligned}$$

and $\lim_{t\to\infty}\boldsymbol{\beta}^R(t) = \mathbf{1}_p$. We only need to prove $\frac{d\log\boldsymbol{\beta}^R(t)}{\mathrm{d}t}$ is Lebesgue integrable to complete the proof of Theorem 5.

Proof of Theorem 5 for RMSProp. For any fixed $i = 1, 2, \cdots, p$, by Lemma 14,

$$\int_0^\infty \frac{\mathrm{d}\log\boldsymbol{\beta}^R(t)}{\mathrm{d}t}\mathrm{d}t = -\int_0^\infty \frac{\mathrm{d}\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}\mathrm{d}t = 0.$$

Therefore,

$$0 = \int_0^\infty \frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}\mathrm{d}t = \int_0^\infty \left(\frac{\mathrm{d}\log\boldsymbol{\beta}^R(t)}{\mathrm{d}t}\right)_+\mathrm{d}t + \int_0^\infty \left(\frac{\mathrm{d}\log\boldsymbol{\beta}^R(t)}{\mathrm{d}t}\right)_-\mathrm{d}t,$$

we only need to prove the convergence of $\int_0^\infty \left(\frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}\right)_-\mathrm{d}t$, is equivalent to convergence of $\int_0^\infty (\frac{\mathrm{d}\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t})_+\mathrm{d}t$.

For any $k\in\mathcal{Z}^+$, denote $B_k = (k, k+1)\cap\{t : \frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t} > 0\}$. Since $\frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}$ is measurable, we have that B_k is a measurable set. Denote M_k as an upper bound for $\frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}$ on $[k, k+1)$ (which is guaranteed since $\mathcal{L}$ is locally Lipschitz). Since B_k is a measurable set, there exists an open set $\tilde{B}_k\subset(k, k+1)$, such that $m(\tilde{B}_k/B_k)\le\frac{1}{k^2M_k}$, where m is Lebesgue measure on $\mathbb{R}$.

Therefore, we have

$$\begin{aligned}\int_{t\in[k,k+1)}\left(\frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}\right)_+\mathrm{d}t &= \int_{t\in B_k}\left(\frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}\right)\mathrm{d}t\\ &\le \int_{t\in\tilde{B}_k}\left(\frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}\right)\mathrm{d}t + M_k\cdot\frac{1}{k^2M_k}\\ &= \int_{t\in\tilde{B}_k}\left(\frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}\right)\mathrm{d}t + \frac{1}{k^2}.\end{aligned}$$

Furthermore, let $\tilde{B}_k = \cup_{j=1}^{\infty}(t_{k,j}, t'_{k,j})$. Then we have

$$
\begin{aligned}
&\int_{t\in\tilde{B}_k}\left(\frac{d\log\sqrt{\frac{\varepsilon+F_i(t)}{\varepsilon}}}{\mathrm{d}t}\right)\mathrm{d}t\\
=&\sum_{j=1}^{\infty}\int_{t_{k,j}}^{t'_{k,j}}\frac{d\log\sqrt{\frac{\varepsilon+lF_i(t)}{\varepsilon}}}{\mathrm{d}t}\mathrm{d}t\\
=&\sum_{j=1}^{\infty}\left(\log\sqrt{\frac{\varepsilon+F_i(t'_{k,j})}{\varepsilon}}-\log\sqrt{\frac{\varepsilon+F_i(t_{k,j})}{\varepsilon}}\right)\\
=&\frac{1}{2}\sum_{j=1}^{\infty}\log\frac{\varepsilon+F_i(t'_{k,j})}{\varepsilon+F_i(t_{k,j})}\\
=&\frac{1}{2}\sum_{j=1}^{\infty}\log\frac{\varepsilon+e^{-(t'_{k,j}-t_{k,j})(1-b)}F_i(t_{k,j})+\int_{t_{k,j}}^{t'_{k,j}}(1-b)e^{-(t'_{k,j}-t)(1-b)}(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t}{\varepsilon+F_i(t_{k,j})}\\
\leq&\frac{1}{2}\sum_{j=1}^{\infty}\log\frac{\varepsilon+F_i(t_{k,j})+\int_{t_{k,j}}^{t'_{k,j}}(1-b)e^{-(t'_{k,j}-t)(1-b)}(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t}{\varepsilon+F_i(t_{k,j})}\\
\leq&\frac{1}{2}\sum_{j=1}^{\infty}\frac{\int_{t_{k,j}}^{t'_{k,j}}(1-b)e^{-(t'_{k,j}-t)(1-b)}(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t}{\varepsilon+F_i(t_{k,j})}\\
\leq&\frac{1}{2}\sum_{j=1}^{\infty}\frac{\int_{t_{k,j}}^{t'_{k,j}}(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t}{\varepsilon}\\
\leq&\frac{1}{2}\frac{\int_{t\in\tilde{B}_k}(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}t}{\varepsilon}<\infty.
\end{aligned}
$$

The proof is completed. □

In the rest of the section, we extend the proof of Theorem 5 from RMSProp to Adam.

Conditioner for Adam is the same as RMSProp, except that Adam will divide a bias-corrected term $1-b^t$ for conditioner each step, that is,

$$
\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t}=-\frac{\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))}{\sqrt{\varepsilon\mathbf{1}_d+\frac{\int_0^t e^{-(1-b)(t-\tau)}(1-b)\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2\mathrm{d}\tau}{(1-b^t)}}}.
$$

Generally, updates of RMSProp and Adam (w/m) can be both expressed as

$$
\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t}=-\frac{\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))}{\sqrt{\varepsilon\mathbf{1}_d+\frac{\int_0^t e^{-(1-b)(t-\tau)}(1-b)\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2\mathrm{d}\tau}{(1-a^t)}}},
$$

where for RMSProp $a=0$, and for Adam $a=b$. For any $0\leq a<1$, $0\leq b\leq 1$, define $F_i(t)\triangleq\int_0^t(1-b)b^{t-\tau}(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2\mathrm{d}\tau$ as in RMSProp case. We will show Lemmas 13 and 14.

Lemma 15. *For adaptive gradient flow defined as eq. (4) with* $\boldsymbol{h}^{-1}(t)=\sqrt{\varepsilon\mathbf{1}_d+\frac{\int_0^t e^{-(1-b)(t-\tau)}(1-b)\bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2\mathrm{d}\tau}{(1-a^t)}}$ *with* $0\leq a<1$ *and* $0\leq b\leq 1$, $F_i(t)$ *is bounded* ($i=1,2,\cdots,p$), *that is,*

$$
\overline{\lim}_{t\to\infty}F_i(t)<\infty.
$$

Proof. The proof follows the same routine as proof of Lemma 13, except in this case we have

$$\int_{t_k}^{t_{k+1}} \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon + F_i(t)/(1-a^\tau)}}\mathrm{d}t \geq \int_{t_k}^{t_{k+1}} \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon + (k+1)/(1-a^{t_k})}}\mathrm{d}t.$$

For $k \geq 1$, we further have

$$\int_{t_k}^{t_{k+1}} \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon + (k+1)/(1-a^{t_k})}}\mathrm{d}t \geq \int_{t_k}^{t_{k+1}} \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon + (k+1)/(1-a^{t_1})}}\mathrm{d}t \geq \frac{1-b}{\sqrt{\varepsilon + (k+1)/(1-a^{t_1})}},$$

sum of which diverges.

The proof is completed. □

Lemma 16. *For adaptive gradient flow, $\int_0^\infty g_i(t)^2\mathrm{d}t$ converges ($i = 1, 2, \cdots, p$). Consequently, $\lim_{t\to\infty} F_i(t) = 0$ and $\lim_{t\to\infty} \boldsymbol{h}(t) = \frac{1}{\sqrt{\varepsilon}}\mathbf{1}_p$.*

Proof. The proof is the same as proof of Lemma 14, except that

$$\mathcal{L}(0) \geq \sum_{i=1}^p \int_1^\infty \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{\varepsilon + F_i(t)/(1-a^t)}}\mathrm{d}t \geq \sum_{i=1}^p \int_1^\infty \frac{(\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)))_i^2}{\sqrt{M_i/(1-a^1) + \varepsilon}}\mathrm{d}t.$$

The proof is completed. □

Proof of Theorem 5 for Adam flow. We only need to prove Lebesgue integrability of $\frac{\mathrm{d}\log \boldsymbol{\beta}(t)}{\mathrm{d}t}$.

The proof is the same as proof of Theorem 5 for RMSProp flow, except that

$$\begin{aligned}
&F_i(t'_{k,j})/(1-a^{t'_{k,j}}) - F_i(t_{k,j})/(1-a^{t_{k,j}}) \\
\leq &F_i(t'_{k,j})/(1-a^{t'_{k,j}}) - F_i(t_{k,j})/(1-a^{t'_{k,j}}) \\
\leq &\frac{1}{1-a}(F_i(t'_{k,j}) - F_i(t_{k,j})).
\end{aligned}$$

The proof is completed. □

It is worth noting that the current framework of adaptive gradient flow can not cover Adam with a decaying ε or without ε. It will be interesting to see if the framework can be modified to analyze these optimizers, and we leave this as a future work.

B.2. Proof of Theorem 2

B.2.1. Proof of surrogate margin Lemmas: Lemma 1, Lemma 2, and Lemma 3

In the beginning, we first prove a basic lemma for normalized margin, i.e., the normalized margin γ and normalized gradients are upper bounded:

Lemma 17. *For any $\boldsymbol{v} \in \mathbb{R}^p/\{\mathbf{0}\}$, the normalized margin $\gamma = \frac{\tilde{q}_{\min}(\boldsymbol{v})}{\|\boldsymbol{v}\|^L}$ and normalized gradients $\|\frac{\bar{\partial}\tilde{q}_i(\boldsymbol{v})}{\|\boldsymbol{v}\|^{L-1}}\|$ ($i = 1, 2, \cdots, p$) are upper bounded universally.*

Proof. By homogeneity of $\tilde{q}_i$ ($i = 1, 2, \cdots, p$), only parameters with unit norm needed to be considered. That is,

$$\begin{aligned}
&\{\gamma(\boldsymbol{v}) : \boldsymbol{v} \in \mathbb{R}^p/\{\mathbf{0}\} \\
= &\{\gamma(\boldsymbol{v}) : \|\boldsymbol{v}\| = 1\}.
\end{aligned}$$

Since $\tilde{q}_i$ is continuous and $\{\boldsymbol{v} : \|\boldsymbol{v}\| = 1\}$ is a compact set, normalized margin is upper bounded.

Normalized gradients $\|\frac{\bar{\partial}\tilde{q}_i(\boldsymbol{v})}{\|\boldsymbol{v}\|^{L-1}}\|$ are also bounded following similar routine since $\tilde{q}_i$ is locally Lipschitz. □

We then formally define surrogate norm $\rho(t)$ and surrogate margin $\tilde{\gamma}$ as follows:

Definition 9 (Surrogate norm and surrogate margin)**.** *Let $\boldsymbol{v}(t)$ obey an adaptive gradient flow $\mathcal{F}$ which satisfies Assumption 1, with loss $\tilde{\mathcal{L}}$ and component learning rate $\boldsymbol{\beta}(t)$. The surrogate margin $\rho(t)$ along $\mathcal{F}$ is defined as*

$$\rho(t) = \|\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}(t)\|,$$

and surrogate margin $\tilde{\gamma}(t)$ is defined as

$$\tilde{\gamma}(t) = \frac{f^{-1}(\log \frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))})}{\rho(t)^L}.$$

We can now restate Lemma 1 as follows:

Lemma 18 (Lemma 1 restated)**.** *Let $\boldsymbol{v}(t)$ obey an adaptive gradient flow $\mathcal{F}$ which satisfies Assumption 1, with loss $\tilde{\mathcal{L}}$ and component learning rate $\boldsymbol{\beta}(t)$. Then we have $\lim_{t\to\infty} \frac{\rho(t)}{\|\boldsymbol{v}(t)\|} = 1$. Furthermore, if further $\lim_{t\to\infty} \tilde{\mathcal{L}}(t) = \infty$, we have $\lim_{t\to\infty} \frac{\gamma(t)}{\tilde{\gamma}(t)} = 1$*

Proof. By the definition of approximate norm $\rho(t) = \|\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}(t)\|$ and $\lim_{t\to\infty} \boldsymbol{\beta}(t) = \mathbf{1}_p$, we have that

$$1 = \underline{\lim}_{t\to\infty} \min_i \boldsymbol{\beta}_i^{-\frac{1}{2}}(t) \le \underline{\lim}_{t\to\infty} \frac{\rho(t)}{\|\boldsymbol{v}(t)\|} \le \overline{\lim}_{t\to\infty} \frac{\rho(t)}{\|\boldsymbol{v}(t)\|} \le \overline{\lim}_{t\to\infty} \max_i \boldsymbol{\beta}_i^{-\frac{1}{2}}(t) = 1,$$

which leads to $\lim_{t\to\infty} \frac{\rho(t)}{\|\boldsymbol{v}(t)\|} = 1$.

By the definition of $\tilde{\mathcal{L}}(\boldsymbol{v})$, we have

$$e^{-f(\tilde{q}_{\min}(\boldsymbol{v}))} \le \tilde{\mathcal{L}}(\boldsymbol{v}) = \sum_{i=1}^{N} e^{-f(\tilde{q}_i(\boldsymbol{v}))} \le N e^{-f(\tilde{q}_{\min}(\boldsymbol{v}))}.$$

Rearranging the above equation, we have

$$\begin{aligned}\tilde{q}_{\min}(\boldsymbol{v}) \ge& g\left(\log \frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})}\right) \ge g(f(\tilde{q}_{\min}(\boldsymbol{v})) - \log N)\\ =& g(f(\tilde{q}_{\min}(\boldsymbol{v}))) - \log N g'(\xi) = \tilde{q}_{\min}(\boldsymbol{v}) - \log N g'(\xi),\end{aligned}$$

where $\xi \in (f(\tilde{q}_{\min}(\boldsymbol{v})) - \log N, f(\tilde{q}_{\min}(\boldsymbol{v})))$.

Therefore, the surrogate margin can be bounded as

$$\begin{aligned}&\frac{\tilde{q}_{\min}(\boldsymbol{v}(t)) - \log N g'(\xi)}{\|\boldsymbol{v}(t)\|^L} \frac{\|\boldsymbol{v}(t)\|^L}{\rho(t)^L} = \frac{\tilde{q}_{\min}(\boldsymbol{v}(t)) - \log N g'(\xi)}{\rho(t)^L}\\ \le& \tilde{\gamma}(t) \le \frac{\tilde{q}_{\min}}{\rho(t)^L} = \gamma(t) \left(\frac{\|\boldsymbol{v}(t)\|}{\rho(t)}\right)^L.\end{aligned} \tag{11}$$

By the assumption that $\lim_{t\to\infty} \tilde{\mathcal{L}} = 0$, there exists a large enough time $\tilde{t}$, such that, any time $t \ge \tilde{t}$, $g(\xi) > 0$. The left side of the above equation 11 can be further rearranged as

$$\begin{aligned}\frac{\tilde{q}_{\min}(\boldsymbol{v}(t)) - \log N g'(\xi)}{\|\boldsymbol{v}(t)\|^L} \frac{\|\boldsymbol{v}(t)\|^L}{\rho(t)^L} &= \left(\gamma(t) - \frac{\log N g'(\xi)}{\|\boldsymbol{v}(t)\|^L}\right) \frac{\|\boldsymbol{v}(t)\|^L}{\rho(t)^L}\\ &= \left(\gamma(t) - \frac{\log N g'(\xi) g(\xi)}{\|\boldsymbol{v}(t)\|^L g(\xi)}\right) \frac{\|\boldsymbol{v}(t)\|^L}{\rho(t)^L}\\ &\ge \left(\gamma(t) - \frac{\log N g'(\xi)\gamma(t)}{g(\xi)}\right) \frac{\|\boldsymbol{v}(t)\|^L}{\rho(t)^L}\\ &= \left(\gamma(t) - \frac{\log N\gamma(t)}{g(\xi) f'(g(\xi))}\right) \frac{\|\boldsymbol{v}(t)\|^L}{\rho(t)^L}.\end{aligned}$$

By the assumption that $\lim_{t\to\infty}\tilde{\mathcal{L}}(t) = 0$, $\underline{\lim}_{t\to\infty}\xi \geq \underline{\lim}_{t\to\infty}\log\frac{1}{\tilde{\mathcal{L}}(t)} - \log N = \infty$, which further indicates $\lim_{t\to\infty} g(\xi)f'(g(\xi)) = \infty$ by the third item of Proposition 1. The proof is completed by taking t to infinity of eq. (11).

□

By Lemmas 17 and 18, we have that if $\lim_{t\to\infty}\tilde{\mathcal{L}}(t)\to 0$, $\tilde{\gamma}(t)$ is upper bounded. We then lower bound $\tilde{\gamma}(t)$ by proving Lemma 2. As a warm-up, we first calculate the derivative of ρ^2.

Lemma 19. *The derivative of ρ^2 is as follows:*

$$\frac{1}{2}\frac{\mathrm{d}\rho(t)^2}{\mathrm{d}t} = L\nu(t) + \left\langle \boldsymbol{v}(t), \boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\frac{\mathrm{d}\boldsymbol{\beta}^{-\frac{1}{2}}}{\mathrm{d}t}(t)\odot\boldsymbol{v}(t)\right\rangle,$$

where $\nu(t)$ is defined as $\sum_{i=1}^N e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\tilde{q}_i(\boldsymbol{v}(t))$. Furthermore, we have that $\nu(t) > \frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\tilde{\mathcal{L}}(\boldsymbol{v}(t))$.

Proof. By taking derivative directly, we have that

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}\rho(t)^2}{\mathrm{d}t} &= \left\langle \boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t), \frac{\mathrm{d}\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t)}{\mathrm{d}t}\right\rangle \\
&= \left\langle \boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t), \boldsymbol{v}(t)\odot\frac{\mathrm{d}\boldsymbol{\beta}(t)^{-\frac{1}{2}}}{\mathrm{d}t} + \boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\rangle \\
&= -\left\langle \boldsymbol{v}(t), \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle + \left\langle \boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t), \boldsymbol{v}(t)\odot\frac{\mathrm{d}\boldsymbol{\beta}(t)^{-\frac{1}{2}}}{\mathrm{d}t}\right\rangle \\
&\overset{(*)}{=} L\sum_{i=1}^N e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\tilde{q}_i(\boldsymbol{v}(t)) + \left\langle \boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t), \boldsymbol{v}(t)\odot\frac{\mathrm{d}\boldsymbol{\beta}(t)^{-\frac{1}{2}}}{\mathrm{d}t}\right\rangle,
\end{aligned}$$

where eq. $(*)$ comes from Homogeneity Assumption 1. I.

Furthermore, since $\tilde{q}_i(\boldsymbol{v}(t)) \geq \tilde{q}_{\min}(\boldsymbol{v}(t)) \geq g(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))})$ for all $i\in[N]$ and $f'(x)x$ keeps increasing on $(0,\infty)$ by Proposition 1,

$$f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\tilde{q}_i(\boldsymbol{v}(t)) \geq f'\left(g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right)\cdot g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right) = \frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}.$$

Therefore,

$$\begin{aligned}
&\sum_{i=1}^N e^{-f(\tilde{q}_i(\boldsymbol{v}(t))(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\tilde{q}_i(\boldsymbol{v}(t)) \\
\geq &\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\sum_{i=1}^N e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))} = \frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\tilde{\mathcal{L}}(\boldsymbol{v}(t)).
\end{aligned}$$

The proof is completed. □

With the estimation of $\frac{\mathrm{d}\rho^2}{\mathrm{d}t}$ above, we come to the proof of Lemma 2.

Proof of Lemma 2. We first construct time t_1 as follows: by properties of $\boldsymbol{\beta}(t)$ in Definition 1, there exists some large enough time $t_1 > t_0$, such that for any $t > t_1$,

$$\sum_{i=1}^{p}\int_{t_1}^{\infty}\left(\frac{\mathrm{d}\log\left(\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)\right)}{\mathrm{d}t}\right)_+ \leq \min\left\{\frac{1}{2L},\frac{1}{4}\right\},$$
$$\sum_{i=1}^{p}\int_{t_1}^{\infty}\left(\frac{\mathrm{d}\log\left(\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)\right)}{\mathrm{d}t}\right)_- \geq -\min\left\{\frac{1}{2L},\frac{1}{4}\right\},$$

and

$$\frac{1}{2} \leq \|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\|_\infty \leq \frac{3}{2}.$$

Taking logarithmic derivative to $\tilde{\gamma}(\boldsymbol{v}(t))$, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\log\tilde{\gamma}(t)\\
=&\frac{\mathrm{d}}{\mathrm{d}t}\left(\log\left(g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right)-L\log\rho(t)\right)\\
=&\frac{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\cdot\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\cdot\left(-\frac{\mathrm{d}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\mathrm{d}t}\right)-L^2\cdot\frac{\nu(t)}{\rho(t)^2}-\frac{L\left\langle\boldsymbol{v}(t),\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\frac{\mathrm{d}\boldsymbol{\beta}^{-\frac{1}{2}}}{\mathrm{d}t}(t)\odot\boldsymbol{v}(t)\right\rangle}{\rho(t)^2}.
\end{aligned}$$

Let $A=\frac{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\cdot\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\cdot\left(-\frac{\mathrm{d}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\mathrm{d}t}\right)-L^2\cdot\frac{\nu(t)}{\rho(t)^2}$ and $B=\frac{L\left\langle\boldsymbol{v}(t),\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\frac{\mathrm{d}\boldsymbol{\beta}^{-\frac{1}{2}}}{\mathrm{d}t}(t)\odot\boldsymbol{v}(t)\right\rangle}{\rho(t)^2}$. We then have

$$\begin{aligned}
A&\geq\frac{1}{\nu(t)}\cdot\left(-\frac{\mathrm{d}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\mathrm{d}t}\right)-L^2\cdot\frac{\nu(t)}{\rho(t)^2}\\
&\geq\frac{1}{\nu(t)}\cdot\left(-\frac{\mathrm{d}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\mathrm{d}t}-\frac{L^2\nu(t)^2}{\rho(t)^2}\right)\\
&=\frac{1}{\nu(t)}\left(\left\langle\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t},\boldsymbol{\beta}(t)^{-1}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\rangle-\left\langle\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t},\widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t)}\right\rangle^2\right)\\
&\overset{(*)}{\geq}0,
\end{aligned}$$

where inequality $(*)$ comes from Cauchy-Schwarz inequality.

As for B, we have that

$$\begin{aligned}
B&=-\frac{L\left\langle\boldsymbol{v}(t),\boldsymbol{\beta}^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{\beta}^{-\frac{1}{2}}(t)}{\mathrm{d}t}(t)\odot\boldsymbol{v}(t)\right\rangle}{\rho^2}\\
&=-L\frac{\sum_{i=1}^{p}v_i^2(t)\beta_i^{-\frac{1}{2}}(t)\frac{\mathrm{d}\beta_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}}{\sum_{i=1}^{p}v_i^2(t)\beta_i^{-1}(t)}\\
&\geq-L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\right)_+.
\end{aligned}$$

Combining the estimation of A and B, we then have

$$\frac{\mathrm{d}}{\mathrm{d}t}\log\tilde{\gamma}(t) \geq A + B \geq -L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log \boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\right)_+, \tag{12}$$

and integrating both sides leads to

$$\log\tilde{\gamma}(t) - \log\tilde{\gamma}(t_1) \geq -\frac{1}{2}.$$

The proof is completed.

□

By the proof of Lemma 2, we can then prove convergence of surrogate margin $\tilde{\gamma}$.

Proof of Lemma 3. By eq. (12), $\log\tilde{\gamma}(t) + \int_{t_1}^{t} L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log \boldsymbol{\beta}_i^{-\frac{1}{2}}(\tau)}{\mathrm{d}\tau}\right)_+ \mathrm{d}\tau$ is non-decreasing. Furthermore, since $g(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}) \leq \tilde{q}_{\min}(\boldsymbol{v}(t))$, we have that

$$\overline{\lim}_{t\to\infty}\tilde{\gamma}(t) \leq \overline{\lim}_{t\to\infty}\frac{\tilde{q}_{\min}(\boldsymbol{v}(t))}{\rho(t)^L} = \overline{\lim}_{t\to\infty}\frac{\tilde{q}_{\min}(\boldsymbol{v}(t))}{\|\boldsymbol{v}(t)\|^L},$$

which by Lemma 17, the last term is bounded.

Therefore, $\log\tilde{\gamma}(t) + \int_{t_1}^{t} L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log \boldsymbol{\beta}_i^{-\frac{1}{2}}(\tau)}{\mathrm{d}t}\right)_+ \mathrm{d}\tau$ is upper bounded, which further indicates that it converges due to its monotony. The proof is completed by the convergence of $\int_{t_1}^{t} L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log \boldsymbol{\beta}_i^{-\frac{1}{2}}(\tau)}{\mathrm{d}\tau}\right)_+ \mathrm{d}\tau$.

□

B.2.2. Convergence of $\tilde{\mathcal{L}}$ and ρ: Proof of Lemma 4

Here we restate the complete Lemma 4.

Lemma 20. *Let $\boldsymbol{v}$ obey an adaptive gradient flow which satisfies Assumption 1. Let t_1 be constructed as Lemma 2. Then, for any $t \geq t_1$, define*

$$G(t) = \int_{\frac{1}{\tilde{\mathcal{L}}(t_1)}}^{\frac{1}{\tilde{\mathcal{L}}(t)}} \frac{g'\left(\log x\right)^2}{g\left(\log x\right)^{2-2/L}}\cdot dx.$$

Then, for any $t \geq t_1$, the following inequality holds:

$$G(t) - G(t_1) \geq \int_{\frac{1}{\tilde{\mathcal{L}}(t_1)}}^{\frac{1}{\tilde{\mathcal{L}}(t)}} \frac{g'\left(\log x\right)^2}{g\left(\log x\right)^{2-2/L}}\cdot dx \geq (t-t_1)e^{-\frac{1}{L}}L^2\tilde{\gamma}\left(t_1\right)^{2/L}.$$

Consequently, $\lim_{t\to\infty}\tilde{\mathcal{L}} = 0$ *and* $\lim_{t\to\infty}\rho = \infty$.

Proof.

$$-\frac{\mathrm{d}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\mathrm{d}t} = \left\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\|_2^2 \geq \langle\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)}\rangle^2 = L^2\cdot\frac{\nu(t)^2}{\rho(t)^2}.$$

Furthermore, we have that

$$\begin{aligned}
L^2\frac{\nu(t)^2}{\rho(t)^2} &\geq L^2\cdot\left(\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)^2\cdot\left(\frac{\tilde{\gamma}(t)}{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\right)^{2/L} \\
&\geq L^2\cdot\left(\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)^2\cdot\left(\frac{e^{-1/2}\tilde{\gamma}(t_1)}{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\right)^{2/L}.
\end{aligned}$$

By simple calculation, we have that

$$\frac{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)^2}{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)^{2-2/L}}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\geq e^{-\frac{1}{L}}L^2\tilde{\gamma}\left(t_1\right)^{2/L}.$$

Taking integration to both sides, we have

$$\int_{\frac{1}{\tilde{\mathcal{L}}(t_1)}}^{\frac{1}{\tilde{\mathcal{L}}(t)}}\frac{g'\left(\log x\right)^2}{g\left(\log x\right)^{2-2/L}}\cdot dx\geq(t-t_1)e^{-\frac{1}{L}}L^2\tilde{\gamma}\left(t_1\right)^{2/L}.$$

Since $\lim_{t\to\infty}(t-t_1)e^{-\frac{1}{L}}L^2\tilde{\gamma}\left(t_1\right)^{2/L}=\infty$, we have that $\lim_{t\to\infty}\frac{1}{\tilde{\mathcal{L}}(t)}=\infty$.

The proof is completed. □

B.2.3. Verification of KKT Condition

In Lemma 5, we omit the construction of coefficients λ_i to highlight the key factors of coefficients $(\mathcal{O}(1-\langle\hat{v}(t),-\widehat{\bar{\partial}\tilde{\mathcal{L}}(t)}\rangle),\mathcal{O}\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right))$. We restate Lemma 5 and provide the detailed construction as follows:

Lemma 21 (Lemma 5 restated)**.** *Let $\boldsymbol{v}$ obey adaptive gradient flow $\mathcal{F}$ with empirical loss $\tilde{\mathcal{L}}$ satisfying Assumption 1. Let time t_1 be constructed as Lemma 2. Then, define coefficients in Definition 2 as $\lambda_i(t)=\tilde{q}_{\min}(\boldsymbol{v}(t))^{1-2/L}\|\boldsymbol{v}(t)\|\cdot e^{-f(\bar{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)/\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|_2$. Then, for any time $t\geq t_1$, $\tilde{\boldsymbol{v}}(t)=\tilde{q}_{\min}(\boldsymbol{v}(t))^{-\frac{1}{L}}\boldsymbol{v}(t)$ is an $(\varepsilon(t),\delta(t))$ KKT point of L^2 max-margin problem (P) defined in Theorem 2 , where $\varepsilon(t)$, $\delta(t)$ are defined as follows:*

$$\begin{aligned}\varepsilon(t)&=8e^{\frac{1}{L}}\frac{1}{\tilde{\gamma}(t_1)^{2/L}}(1-\cos(\boldsymbol{\theta}(t)))\\\delta(t)&=\frac{2e^{-1+\frac{1}{L}}KN}{L}K^{\log_2\left(2^{L+1}e^{\frac{1}{2}}B_1/\tilde{\gamma}(t_1)\right)}\tilde{\gamma}(t_1)^{-2/L}\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}},\end{aligned}$$

where $\cos(\boldsymbol{\theta}(t))$ is defined as the cosine of angle between $\boldsymbol{v}(t)$ and $-\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))$, i.e.,

$$\cos(\boldsymbol{\theta}(t))=\left\langle\hat{\boldsymbol{v}}(t),-\widehat{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right\rangle.$$

Proof. We verify the definition of approximate KKT point directly.

$$
\begin{aligned}
&\left\|\tilde{\boldsymbol{v}}(t)-\sum_{i=1}^{N}\lambda_i(t)\bar{\partial}\tilde{q}_i\left(\tilde{\boldsymbol{v}}(t)\right)\right\|_2^2\\
=&\left\|\tilde{q}_{\min}(\boldsymbol{v}(t))^{-1/L}\|\boldsymbol{v}(t)\|\hat{\boldsymbol{v}}(t)-\frac{\tilde{q}_{\min}(\boldsymbol{v}(t))^{-1/L}\|\boldsymbol{v}(t)\|e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\bar{\partial}\tilde{q}_i\left(\boldsymbol{v}(t)\right)}{\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|_2}\right\|_2^2\\
=&\,\tilde{q}_{\min}(\boldsymbol{v}(t))^{-2/L}\|\boldsymbol{v}(t)\|^2\left\|\hat{\boldsymbol{v}}(t)+\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|_2}\right\|_2^2\\
=&\,\tilde{q}_{\min}(\boldsymbol{v}(t))^{-2/L}\|\boldsymbol{v}(t)\|^2\left(2-2\left\langle\hat{\boldsymbol{v}}(t),-\widehat{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right\rangle\right)\\
\leq&\,2g\left(\log\left(\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right)^{-2/L}\|\boldsymbol{v}(t)\|^2(1-\cos(\boldsymbol{\theta}(t)))\\
\leq&\,8g\left(\log\left(\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right)^{-2/L}\rho(t)^2(1-\cos(\boldsymbol{\theta}(t)))\\
=&\,8\frac{1}{\tilde{\gamma}(t)^{2/L}}(1-\cos(\boldsymbol{\theta}(t)))\leq 8\frac{1}{\tilde{\gamma}(t)^{2/L}}(1-\cos(\boldsymbol{\theta}(t)))\\
\leq&\,8e^{\frac{1}{L}}\frac{1}{\tilde{\gamma}(t_1)^{2/L}}(1-\cos(\boldsymbol{\theta}(t))).
\end{aligned}
$$

As for δ, we have that

$$
\begin{aligned}
&\sum_{i=1}^{N}\lambda_i\left(\tilde{q}_i(\tilde{\boldsymbol{v}}(t))-1\right)\\
=&\frac{\sum_{i=1}^{N}\tilde{q}_{\min}^{1-2/L}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|\cdot e^{-f(\tilde{q}_i(\boldsymbol{v}(t))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\left(\frac{\tilde{q}_i(\boldsymbol{v}(t))}{\tilde{q}_{\min}(\boldsymbol{v}(t))}-1\right)\right.}{\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)\|_2}\\
=&\frac{\sum_{i=1}^{N}\tilde{q}_{\min}(\boldsymbol{v}(t))^{-2/L}\|\boldsymbol{v}(t)\|\cdot e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)}{\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|_2}\\
\overset{(*)}{\leq}&\sum_{i=1}^{N}\frac{2K}{L}\tilde{q}_{\min}(\boldsymbol{v}(t))^{-2/L}\|\boldsymbol{v}(t)\|^2e^{f(\tilde{q}_{\min}(\boldsymbol{v}(t)))-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\\
&\cdot\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}},
\end{aligned}
$$

where inequality $(*)$ is because

$$
\begin{aligned}
\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\geq&\left\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\right\rangle=\frac{L\nu(t)}{\|\boldsymbol{v}(t)\|}\\
\geq&\frac{L}{\|\boldsymbol{v}(t)\|}\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\geq\frac{L}{\|\boldsymbol{v}(t)\|}\frac{1}{2K}\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\cdot\tilde{\mathcal{L}}(\boldsymbol{v}(t))\\
\geq&\frac{L}{\|\boldsymbol{v}(t)\|}\frac{1}{2K}e^{-f(\tilde{q}_{\min}(\boldsymbol{v}(t)))}\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}.
\end{aligned}
$$

Bounding $\|\boldsymbol{v}\|$ using ρ and applying the definition of surrogate margin $\tilde{\gamma}$, we further have

$$
\begin{aligned}
&\sum_{i=1}^{N}\lambda_i\left(\tilde{q}_i(\tilde{\boldsymbol{v}}(t))-1\right)\\
\leq&\sum_{i=1}^{N}\frac{8K}{L}\tilde{q}_{\min}^{-2/L}(\boldsymbol{v}(t))\rho(t)^2 e^{f(\tilde{q}_{\min}(\boldsymbol{v}(t)))-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}\\
\leq&\sum_{i=1}^{N}\frac{8K}{L}\tilde{\gamma}(t)^{-2/L}\cdot e^{f(\tilde{q}_{\min}(\boldsymbol{v}(t)))-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}\\
\leq&\sum_{i=1}^{N}\frac{8e^{\frac{1}{L}}K}{L}\tilde{\gamma}(t_1)^{-2/L}\cdot e^{f(\tilde{q}_{\min}(\boldsymbol{v}(t)))-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}.
\end{aligned}
$$

Since $f(\tilde{q}_{\min}(\boldsymbol{v}(t))(\boldsymbol{v}(t)))-f(\tilde{q}_i(\boldsymbol{v}(t)))=(\tilde{q}_{\min}(\boldsymbol{v}(t))-\tilde{q}_i(\boldsymbol{v}(t)))f'(\xi_i(t))$ ($\xi_i(t)\in[\tilde{q}_{\min}(\boldsymbol{v}(t)),\tilde{q}_i(\boldsymbol{v}(t))]$ is guaranteed by Mean Value Theorem), we then have

$$
\begin{aligned}
&\sum_{i=1}^{N}\lambda_i\left(\tilde{q}_i(\tilde{\boldsymbol{v}}(t))-1\right)\\
\leq&\sum_{i=1}^{N}\frac{2e^{\frac{1}{L}}K}{L}\tilde{\gamma}(t_1)^{-2/L}\cdot e^{f(\tilde{q}_{\min}(\boldsymbol{v}(t)))-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\\
&\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}\\
=&\sum_{i=1}^{N}\frac{2e^{\frac{1}{L}}K}{L}\tilde{\gamma}(t_1)^{-2/L}\cdot e^{(\tilde{q}_{\min}(\boldsymbol{v}(t))-\tilde{q}_i(\boldsymbol{v}(t)))f'(\xi_i(t))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\\
&\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}\\
\leq&\sum_{i=1}^{N}\frac{2e^{\frac{1}{L}}K}{L}\tilde{\gamma}(t_1)^{-2/L}\cdot e^{(\tilde{q}_{\min}(\boldsymbol{v}(t))-\tilde{q}_i(\boldsymbol{v}(t)))f'(\xi_i(t))}K^{\lceil\log_2(\tilde{q}_i(\boldsymbol{v}(t))/\xi_i(t))\rceil}f'\left(\xi_i(t)\right)\\
&\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}\\
\overset{(*)}{\leq}&\sum_{i=1}^{N}\frac{2e^{\frac{1}{L}}K}{L}\tilde{\gamma}(t_1)^{-2/L}\cdot e^{(\tilde{q}_{\min}(\boldsymbol{v}(t))-\tilde{q}_i(\boldsymbol{v}(t)))f'(\xi_i(t))}K^{\log_2\left(2e^{\frac{1}{2}}B_1/\tilde{\gamma}(t_1)\right)}f'\left(\xi_i(t)\right)\\
&\left(\tilde{q}_i(\boldsymbol{v}(t))-\tilde{q}_{\min}(\boldsymbol{v}(t))\right)\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}\\
\overset{(**)}{\leq}&\frac{2e^{-1+\frac{1}{L}}KN}{L}K^{\log_2\left(2^{L+1}e^{\frac{1}{2}}B_1/\tilde{\gamma}(t_1)\right)}\tilde{\gamma}(t_1)^{-2/L}\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}},
\end{aligned}
$$

where eq. $(*)$ is because

$$
\begin{aligned}
\lceil\log_2\left(\tilde{q}_i(\boldsymbol{v}(t))/\xi_i(t)\right)\rceil&\leq\log_2\left(\tilde{q}_i(\boldsymbol{v}(t))/\tilde{q}_{\min}(\boldsymbol{v}(t))\right)+1\\
&\leq\log_2\left(\tilde{q}_i(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L/\|\boldsymbol{v}(t)\|^L\tilde{q}_{\min}(\boldsymbol{v}(t))\right)+1\\
&\leq\log_2\left(2^{L+1}B_1\rho(t)^L/\tilde{q}_{\min}\right)\leq\log_2\left(2^{L+1}B_1/\tilde{\gamma}(t)\right)\\
&\leq\log_2\left(2^{L+1}e^{\frac{1}{2}}B_1/\tilde{\gamma}(t_1)\right),
\end{aligned}
$$

which further leads to

$$
f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\leq K^{\log_2\left(2e^{\frac{1}{2}}B_1/\tilde{\gamma}(t_1)\right)}f'\left(\xi_i(t)\right),
$$

and eq. $(**)$ is because $e^{-x}x \leq e^{-1}$.

The proof is completed. □

By Lemma 4, we have proved that $\lim_{t\to\infty} \tilde{\mathcal{L}}(t) = 0$, which leads to $\lim_{t\to\infty} \delta(t) = 0$. As stated in the main text, we only need to bound $\varepsilon(t)$, or equivalently $(1 - \cos(\boldsymbol{\theta}(t)))$.

Before moving forward, we introduce an equivalent proposition of that $(1 - \cos(\boldsymbol{\theta}))$ goes to zero.

Lemma 22. *If there exists a time sequence $\{t_i\}_{i=1}^{\infty}$, such that, $\lim_{i\to\infty} t_i = \infty$ and $\lim_{i\to\infty} \left\langle \widehat{\boldsymbol{\beta}^{-\frac{1}{2}} \odot \boldsymbol{v}}, \widehat{\boldsymbol{\beta}^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}} \right\rangle = -1$, then*

$$\lim_{i\to\infty} \cos(\boldsymbol{\theta}(t_i)) = 1.$$

Proof.

$$\begin{aligned}
&- \cos(\boldsymbol{\theta}(t)) \\
=&\left\langle \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t) + \left(\hat{\boldsymbol{v}}(t) - \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t)\right), \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t) + \left(\widehat{\bar{\partial}\tilde{\mathcal{L}}(t)} - \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t)\right) \right\rangle \\
=&\left\langle \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t), \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t) \right\rangle + \left\langle \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t), \left(\widehat{\bar{\partial}\tilde{\mathcal{L}}(t)} - \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t)\right) \right\rangle \\
+&\left\langle \left(\hat{\boldsymbol{v}}(t) - \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t)\right), \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t) \right\rangle \\
+&\left\langle \left(\hat{\boldsymbol{v}}(t) - \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t)\right), \left(\widehat{\bar{\partial}\tilde{\mathcal{L}}(t)} - \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t)\right) \right\rangle.
\end{aligned}$$

Furthermore,

$$\lim_{i\to\infty} \left\| \widehat{\bar{\partial}\tilde{\mathcal{L}}(t)} - \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t) \right\| = 2 - 2 \lim_{i\to\infty} \left\langle \widehat{\bar{\partial}\tilde{\mathcal{L}}(t)}, \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t) \right\rangle = 0.$$

Following the same routine, we have $\lim_{i\to\infty} \left\| \hat{\boldsymbol{v}}(t) - \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t) \right\| = 0$.

The proof is completed. □

Let $\cos(\tilde{\boldsymbol{\theta}}(t)) = -\left\langle \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t), \widehat{\boldsymbol{\beta}(t)^{\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}}(t) \right\rangle = \left\langle \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}}(t), \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}} \right\rangle$. Then we only need to bound $1 - \cos(\tilde{\boldsymbol{\theta}}(t))$.

In Section 5.3, we briefly state the methodology of proving the convergence of $\varepsilon(t)$. We will make it more clear here: In the

proof Lemma 2, we show that sum of the derivative of $g(\log \frac{1}{\tilde{\mathcal{L}}(t)})$ and $\frac{1}{\rho(t)^L}$ with $\boldsymbol{\beta}(t)$ fixed can be bounded as

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left(\log\left(g\left(\log\frac{1}{\tilde{\mathcal{L}}(t)}\right)\right)\right) - L\left\langle \bar{\partial}_{\boldsymbol{v}}\log\rho(t), \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\rangle \\
=&\frac{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(t)}\right)}{g\left(\log\frac{1}{\tilde{\mathcal{L}}(t)}\right)}\cdot\frac{1}{\tilde{\mathcal{L}}(t)}\cdot\left(-\frac{\mathrm{d}\tilde{\mathcal{L}}(t)}{\mathrm{d}t}\right) - L^2\cdot\frac{\nu(t)}{\rho(t)^2} \\
\geq&\frac{1}{\nu(t)}\left(\left\langle\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \boldsymbol{\beta}(t)^{-1}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\rangle - \left\langle\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t)}\right\rangle^2\right) \\
=&\frac{\left\langle\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t)}\right\rangle^2}{\nu(t)}(\cos(\tilde{\boldsymbol{\theta}}(t))^{-2}-1) \\
=&\frac{L^2\nu(t)}{\rho(t)^2}(\cos(\tilde{\boldsymbol{\theta}}(t))^{-2}-1) \\
=&L\left\langle\bar{\partial}_{\boldsymbol{v}}\log\rho(t), \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}\right\rangle(\cos(\tilde{\boldsymbol{\theta}}(t))^{-2}-1).
\end{aligned}
\tag{13}
$$

Eq. (13) indicates that, intuitively, $(\cos(\tilde{\boldsymbol{\theta}}(t))^{-2}-1)$ can be bounded by the division of change of $\tilde{\gamma}$ to change of parameter part $\boldsymbol{v}$ in ρ. For this purpose, we define $\tilde{\rho}$ in Section 5.3 to describe the accumulated change of $\boldsymbol{v}$ in ρ. The following lemma describe the basic property of $\tilde{\rho}$ and its relationship with ρ.

Lemma 23. *(1). The derivative of $\tilde{\rho}^2$ is as follows:*

$$\frac{1}{2}\frac{d\tilde{\rho}(t)^2}{\mathrm{d}t} = \langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \boldsymbol{v}(t)\rangle = L\nu(t);$$

(2). $\rho(t)$ satisfies that, for any $t^2 > t^1 > t_1$,

$$\rho(t_2) \geq e^{-\frac{1}{2}}\rho(t_1);$$

(3). For any $t > t_1$, $\sqrt{1-\frac{e^{\frac{1}{2}}}{2}} \leq \frac{\rho(t)}{\tilde{\rho}(t)} \leq \sqrt{1+\frac{e^{\frac{1}{2}}}{2}}$.

Proof. (1). can be directly verified similar to Lemma 19. As for (2)., since $\frac{d\log\rho(t)}{\mathrm{d}t} = \frac{1}{2}\frac{\frac{d\rho(t)^2}{\mathrm{d}t}}{\rho(t)^2}$, we have that

$$
\begin{aligned}
\frac{d\log\rho(t)}{\mathrm{d}t} &= \frac{L\sum_{i=1}^N e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))}f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\tilde{q}_i(\boldsymbol{v}(t)) + \left\langle\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t), \boldsymbol{v}(t)\odot\frac{\mathrm{d}\boldsymbol{\beta}(t)^{-\frac{1}{2}}}{\mathrm{d}t}\right\rangle}{\rho(t)^2} \\
&\geq \frac{\left\langle\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t), \boldsymbol{v}(t)\odot\frac{\mathrm{d}\boldsymbol{\beta}(t)^{-\frac{1}{2}}}{\mathrm{d}t}\right\rangle}{\rho(t)^2} \\
&\geq \frac{\sum_{i=1}^p\boldsymbol{\beta}_i(t)^{-\frac{1}{2}}(t)\frac{\mathrm{d}\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\boldsymbol{v}_i(t)^2}{\sum_{i=1}^p\boldsymbol{\beta}_i^{-1}(t)\boldsymbol{v}_i(t)^2} \\
&= \frac{\sum_{i=1}^p\boldsymbol{\beta}_i^{\frac{1}{2}}(t)\frac{\mathrm{d}\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\boldsymbol{\beta}_i^{-1}(t)\boldsymbol{v}_i(t)^2}{\sum_{i=1}^p\boldsymbol{\beta}_i^{-1}(t)\boldsymbol{v}_i(t)^2} \\
&= \frac{\sum_{i=1}^p\frac{d\log\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\boldsymbol{\beta}_i^{-1}(t)\boldsymbol{v}_i(t)^2}{\sum_{i=1}^p\boldsymbol{\beta}_i^{-1}(t)\boldsymbol{v}_i(t)^2} \\
&\geq \sum_{i=1}^p\left(\frac{d\log\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\right)_-.
\end{aligned}
$$

The proof for (2). is completed by integration.

As for (3)., by expanding $\langle \boldsymbol{v}(t), \boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \frac{\mathrm{d}\boldsymbol{\beta}^{-\frac{1}{2}}(t)}{\mathrm{d}t} \odot \boldsymbol{v}(t)\rangle$, we have that

$$
\begin{aligned}
\tilde{\rho}(t)^2 =& \int_{t_1}^{t} \langle \boldsymbol{v}(\tau), \boldsymbol{\beta}^{-\frac{1}{2}}(\tau) \odot \frac{\mathrm{d}\boldsymbol{\beta}^{-\frac{1}{2}}(\tau)}{\mathrm{d}\tau} \odot \boldsymbol{v}(\tau)\rangle \mathrm{d}\tau \\
=& \sum_{i=1}^{p} \int_{t_1}^{t} \boldsymbol{v}_i^2(\tau) \boldsymbol{\beta}_i^{-1}(\tau) \boldsymbol{\beta}_i^{\frac{1}{2}}(\tau) \frac{\mathrm{d}\boldsymbol{\beta}_i^{-\frac{1}{2}}(\tau)}{\mathrm{d}t} \mathrm{d}\tau \\
\leq& \sum_{i=1}^{p} \int_{t_1}^{t} \boldsymbol{\beta}_i^{-1}(\tau) \boldsymbol{v}_i^2(\tau) \left(\boldsymbol{\beta}_i^{\frac{1}{2}}(\tau) \frac{\mathrm{d}\boldsymbol{\beta}_i^{-\frac{1}{2}}(\tau)}{\mathrm{d}t} \right)_+ \mathrm{d}\tau \\
\leq& \int_{t_1}^{t} \left(\sum_{i=1}^{p} \boldsymbol{\beta}_i^{-1} \boldsymbol{v}_i^2(\tau) \right) \left(\sum_{i=1}^{p} \left(\boldsymbol{\beta}_i^{\frac{1}{2}}(\tau) \frac{\mathrm{d}\boldsymbol{\beta}_i^{-\frac{1}{2}}(\tau)}{\mathrm{d}t} \right)_+ \right) \mathrm{d}\tau \\
\overset{(*)}{=}& \|\boldsymbol{\beta}(\tau_0)^{-\frac{1}{2}} \odot \boldsymbol{v}(\tau_0)\|^2 \int_{t_1}^{t} \left(\sum_{i=1}^{p} \left(\boldsymbol{\beta}_i^{\frac{1}{2}}(\tau) \frac{\mathrm{d}\boldsymbol{\beta}_i^{-\frac{1}{2}}(\tau)}{\mathrm{d}t} \right)_+ \right) \mathrm{d}\tau \\
\leq& \frac{1}{4} \|\boldsymbol{\beta}(\tau_0)^{-\frac{1}{2}} \odot \boldsymbol{v}(\tau_0)\|^2 \\
\leq& \frac{e^{\frac{1}{2}}}{4} \|\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}(t)\|^2,
\end{aligned}
$$

where τ_0 in eq. $(*)$ is in $[t_1, t]$ and First Mean Value Theorem guarantees its existence.

The second inequality follows by lower bound $\int_{t_1}^{t} \langle \boldsymbol{v}(\tau), \boldsymbol{\beta}^{-\frac{1}{2}}(\tau) \odot \frac{\mathrm{d}\boldsymbol{\beta}^{-\frac{1}{2}}(\tau)}{\mathrm{d}\tau} \odot \boldsymbol{v}(\tau)\rangle \mathrm{d}\tau$ similarly.

The proof is completed. □

With Lemma 23, we integrate the analysis of the change of $\tilde{\gamma}$ into the following Lemma.

Lemma 24. *For any time $t_3 > t_2 \geq t_1$,*

$$
\begin{aligned}
\int_{t_2}^{t_3} \left(\cos(\tilde{\boldsymbol{\theta}}(\tau))^{-2} - 1 \right) \cdot \frac{\mathrm{d}}{d\tau} \log \tilde{\rho}(\tau) d\tau \leq & \frac{1}{L\left(1 - \frac{e^{\frac{1}{2}}}{2}\right)} \log \frac{\tilde{\gamma}(t_3)}{\tilde{\gamma}(t_2)} \\
& + \frac{L}{\left(1 - \frac{e^{\frac{1}{2}}}{2}\right)} \sum_{i=1}^{p} \int_{t_2}^{t_3} \sum_{i=1}^{p} \left(\frac{\mathrm{d} \log \boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t} \right)_+ \mathrm{d}t.
\end{aligned}
$$

Proof. Recall that

$$
\frac{d \log \tilde{\rho}(t)}{\mathrm{d}t} = \frac{L\nu(t)}{\tilde{\rho}(t)^2} = \frac{\left\langle \boldsymbol{v}(t), \boldsymbol{\beta}(t)^{-1} \odot \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t} \right\rangle}{\tilde{\rho}(t)^2},
$$

and

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \log \tilde{\gamma}(t) \geq & \frac{1}{\nu(t)} \left(\left\langle \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \boldsymbol{\beta}(t)^{-1} \odot \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t} \right\rangle - \left\langle \boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \boldsymbol{v}(t)} \right\rangle \right) \\
& - L \sum_{i=1}^{p} \left(\frac{\mathrm{d} \log \boldsymbol{\beta}_i(t)^{-\frac{1}{2}}(t)}{\mathrm{d}t} \right)_+.
\end{aligned}
$$

We then have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\log\tilde{\gamma}(t) \geq & \frac{1}{\nu(t)}\left(\left\langle \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \boldsymbol{\beta}(t)^{-1}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\rangle - \left\langle \boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t)}\right\rangle^2\right) \\
& -L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log\boldsymbol{\beta}(t)_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\right)_+ \\
= & L\frac{\tilde{\rho}(t)^2}{L^2\nu(t)^2}\frac{d\log\tilde{\rho}(t)}{\mathrm{d}t}\left(\left\langle \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \boldsymbol{\beta}(t)^{-1}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\rangle - \left\langle \boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \widehat{\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t)}\right\rangle^2\right) \\
& -L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log\boldsymbol{\beta}(t)_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\right)_+ \\
\geq & L\left(1-\frac{e^{\frac{1}{2}}}{2}\right)\frac{\rho(t)^2}{L^2\nu(t)^2}\frac{d\log\tilde{\rho}(t)}{\mathrm{d}t}\left(\left\langle \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \boldsymbol{\beta}(t)^{-1}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\rangle\right. \\
& \left. - \left\langle \boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}, \widehat{\boldsymbol{\beta}^{-\frac{1}{2}}\odot\boldsymbol{v}(t)}\right\rangle^2\right) \\
& -L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\right)_+ \\
= & L\left(1-\frac{e^{\frac{1}{2}}}{2}\right)\frac{d\log\tilde{\rho}}{\mathrm{d}t}(\cos(\tilde{\boldsymbol{\theta}}(t))^{-2}-1) - L\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\right)_+.
\end{aligned}$$

The proof is completed. □

Applying the First Mean Value Theorem together with Lemmas 22 and 24, one can easily obtain the first inequality in Lemma 6. To make the following proof simpler, we restate Lemma 6 as the following corollary, while using $\cos(\tilde{\boldsymbol{\theta}}(t))$ instead of $\cos(\boldsymbol{\theta}(t))$.

Corollary 1 (First inequality in Lemma 6, restated)**.** *For any time $t_3 > t_2 \geq t_1$, there exists a time $\xi \in (t_2, t_3)$, such that,*

$$\begin{aligned}
\left(\cos(\tilde{\boldsymbol{\theta}}(\xi))^{-2}-1\right)\left(\log\tilde{\rho}(t_2)-\log\tilde{\rho}(t_1)\right) \leq & \frac{1}{L\left(1-\frac{e^{\frac{1}{2}}}{2}\right)}\log\frac{\tilde{\gamma}(t_3)}{\tilde{\gamma}(t_2)} \\
& + \frac{L}{\left(1-\frac{e^{\frac{1}{2}}}{2}\right)}\sum_{i=1}^{p}\int_{t_2}^{t_3}\sum_{i=1}^{p}\left(\frac{\mathrm{d}\log\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\mathrm{d}t}\right)_+\mathrm{d}t.
\end{aligned}$$

We then prove the second inequality in Lemma 6 to complete the proof of Lemma 6.

Proof of Lemma 6. By direct calculation, we have that

$$\left\|\frac{d\hat{\boldsymbol{v}}(t)}{\mathrm{d}t}\right\| = \frac{1}{\|\boldsymbol{v}(t)\|}\left\|\left(\boldsymbol{I}-\hat{\boldsymbol{v}}(t)\hat{\boldsymbol{v}}(t)^\top\right)\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\| \leq \frac{1}{\|\boldsymbol{v}(t)\|}\left\|\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\| \leq \frac{2}{\|\boldsymbol{v}(t)\|}\left\|\boldsymbol{\beta}(t)^{-1}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\|.$$

Furthermore,

$$\begin{aligned}
&\left\|\boldsymbol{\beta}(t)^{-1}\odot\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t}\right\| = \left\|\bar{\partial}\tilde{\mathcal{L}}(t)\right\| \\
\leq& \sum_{i\in[N]} e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))} f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\left\|\bar{\partial}\tilde{q}_i(\boldsymbol{v}(t))\right\| \\
\leq& \sum_{i\in[N]} e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))} f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\tilde{q}_i(\boldsymbol{v}(t))\frac{1}{\tilde{q}_i(\boldsymbol{v}(t))}B_1\|\boldsymbol{v}(t)\|^{L-1} \\
\leq& \sum_{i\in[N]} 2^{L-1}e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))} f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\tilde{q}_i(\boldsymbol{v}(t))\frac{1}{\tilde{q}_i(\boldsymbol{v}(t))}B_1\rho(t)^{L-1} \\
\leq& \sum_{i\in[N]} 2^{L-1}e^{-f(\tilde{q}_i(\boldsymbol{v}(t)))} f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\tilde{q}_i(\boldsymbol{v}(t))\frac{1}{g\left(\log\frac{1}{\tilde{\mathcal{L}}(t)}\right)}B_1\rho(t)^{L-1} \\
=& 2^{L-1}\frac{\nu(t)}{\rho(t)}\frac{\rho(t)^L}{g\left(\log\frac{1}{\tilde{\mathcal{L}}(t)}\right)}B_1 = 2^{L-1}\frac{\nu(t)}{\rho(t)}\frac{1}{\tilde{\gamma}(t)}B_1
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left\|\frac{d\hat{\boldsymbol{v}}(t)}{\mathrm{d}t}\right\| &\leq 2^L\frac{\nu(t)}{\rho(t)\|\boldsymbol{v}(t)\|}\frac{1}{\tilde{\gamma}(t)}B_1 \leq 3\cdot 2^{L-1}\frac{\nu(t)}{\rho(t)^2}\frac{1}{\tilde{\gamma}(t)}B_1 \leq \frac{3\cdot 2^{L-1}}{(1-\frac{e^{\frac{1}{2}}}{2})}\frac{\nu(t)}{\tilde{\rho}(t)^2}\frac{1}{\tilde{\gamma}(t)}B_1 \\
&= \frac{3\cdot 2^{L-1}}{(1-\frac{e^{\frac{1}{2}}}{2})}\frac{B_1}{L}\frac{\mathrm{d}\log\tilde{\rho}(t)}{\mathrm{d}t}\frac{1}{\tilde{\gamma}(t)} \leq \frac{3\cdot 2^{L-1}}{(1-\frac{e^{\frac{1}{2}}}{2})}\frac{e^{\frac{1}{2}}B_1}{L}\frac{\mathrm{d}\log\tilde{\rho}(t)}{\mathrm{d}t}\frac{1}{\tilde{\gamma}(t_1)}.
\end{aligned}$$

The proof is completed. □

Applying Lemma 5 and Lemma 6, we can then prove Theorem 2.

Proof of Theorem 2. Let $\bar{\boldsymbol{v}}$ be any limit point of series $\{\boldsymbol{v}(t)\}_{t=0}^{\infty}$. We construct a series of approximate KKT point which converges to $\bar{\boldsymbol{v}}$ by induction.

Let $t^1 = t_1$. Now suppose t^{k-1} has been constructed. By Lemma 3 and that $\bar{\boldsymbol{v}}$ is a limit point, there exists $s_k > t^{k-1}$ such that, for any $t > s_k$

$$\left\|\hat{\boldsymbol{v}}\left(s_k\right)-\bar{\boldsymbol{v}}\right\|_2 \leq \frac{1}{k},\ \frac{2}{L}\log\frac{\tilde{\gamma}(t)}{\tilde{\gamma}(s_k)} \leq \frac{1}{2k^3},\ \text{and},\ L\sum_{i=1}^{p}\int_{s_m}^{\infty}\left(\frac{\mathrm{d}\log\boldsymbol{\beta}_i(t)^{-\frac{1}{2}}}{\mathrm{d}t}\right)_+\mathrm{d}t \leq \frac{1}{2k^3}.$$

Let $s_k' > s_k$ satisfy $\log\tilde{\rho}\left(s_k'\right) = \log\tilde{\rho}\left(s_k\right)+\frac{1}{k}$ (which is guaranteed as $\lim_{t\to\infty}\rho = \infty$ and Lemma 23). Therefore, by Corollary 1, there exists $t^k \in (s_k, s_k')$, such that

$$\cos(\tilde{\boldsymbol{\theta}}\left(t^k\right))^{-2}-1 \leq \frac{1}{k^2(1-\frac{e^{\frac{1}{2}}}{2})}. \tag{14}$$

Furthermore,

$$\left\|\hat{\boldsymbol{v}}\left(t^k\right)-\bar{\boldsymbol{v}}\right\|_2 \leq \left\|\hat{\boldsymbol{v}}\left(t^k\right)-\hat{\boldsymbol{v}}\left(s_k\right)\right\|_2+\left\|\hat{\boldsymbol{v}}\left(s_k\right)-\overline{\boldsymbol{v}}\right\|_2 \leq \frac{3\cdot 2^{L-1}}{(1-\frac{e^{\frac{1}{2}}}{2})}\frac{e^{\frac{1}{2}}}{L}\frac{1}{\tilde{\gamma}(t_1)}\frac{1}{k}+\frac{1}{k} \to 0.$$

Therefore, $\lim_{t\to\infty}\hat{\boldsymbol{v}}(t^k) = \bar{\boldsymbol{v}}$. Furthermore, by eq. (14) and Lemma 21, we have that $\boldsymbol{v}(t^k)/\tilde{q}_{\min}^{\frac{1}{L}}(\boldsymbol{v}(t^k))$ is an $(\varepsilon_i, \delta_i)$ KKT point with $\lim_{i\to\infty}\varepsilon_i = \lim_{i\to\infty}\delta_i = 0$. Since

$$\boldsymbol{v}(t^k)/\tilde{q}_{\min}^{\frac{1}{L}}(\boldsymbol{v}(t^k)) = \hat{\boldsymbol{v}}(t^k)/\gamma(t^k),$$

and $\gamma(t^k)$ converges to a positive number, we further have

$$\lim_{k\to\infty} \boldsymbol{v}(t^k)/\tilde{q}_{\min}^{\frac{1}{L}}(\boldsymbol{v}(t^k)) = \bar{\boldsymbol{v}}/\lim_{k\to\infty}\gamma(t^k)^{\frac{1}{L}},$$

is a KKT point of (P), and along the same direction of $\bar{\boldsymbol{v}}$.

The proof is completed. □

B.3. Convergent Direction of AdaGrad, RMSProp and Adam (w/m): proof of Theorems 6 and 7

First of all, we prove Theorem 6 by substitute $\boldsymbol{v}$ in Theorem 2 with $\boldsymbol{h}_\infty \odot \boldsymbol{w}$.

Proof of Theorem 6. Let $\bar{\boldsymbol{w}}$ be any limit point of series $\{\hat{\boldsymbol{w}}(t)\}_{t=0}^{\infty}$. Since $\boldsymbol{v}(t) = \boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \boldsymbol{w}(t)$, $\widehat{\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \bar{\boldsymbol{w}}}$ is a limit point of $\{\hat{\boldsymbol{v}}(t)\}$. By Theorem 2, $\widehat{\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \bar{\boldsymbol{w}}}/\tilde{q}_{\min}\left(\widehat{\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \bar{\boldsymbol{w}}}\right)^{1/L} = \boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \bar{\boldsymbol{w}}/q_{\min}(\bar{\boldsymbol{w}})^{1/L}$ is a KKT point of $(\tilde{P})$. Therefore, there exist non-negative reals $\{\lambda_i\}_{i=1}^N$, such that

$$\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \bar{\boldsymbol{w}}/q_{\min}(\bar{\boldsymbol{w}})^{1/L} = \sum_{i=1}^N \lambda_i \bar{\partial}\tilde{q}_i(\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \bar{\boldsymbol{w}}/q_{\min}(\bar{\boldsymbol{w}})^{1/L}),$$

$$\sum_{i=1}^N \lambda_i(\tilde{q}_i(\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \bar{\boldsymbol{w}}/q_{\min}(\bar{\boldsymbol{w}})^{1/L}) - 1) = 0.$$

Applying the relationship between $\tilde{q}_i$ and q_i, we then have

$$\boldsymbol{h}_\infty^{-1} \odot \bar{\boldsymbol{w}}/q_{\min}(\bar{\boldsymbol{w}})^{1/L} = \sum_{i=1}^N \lambda_i \bar{\partial}q_i(\bar{\boldsymbol{w}}/q_{\min}(\bar{\boldsymbol{w}})^{1/L}),$$

$$\sum_{i=1}^N \lambda_i(q_i(\bar{\boldsymbol{w}}/q_{\min}(\bar{\boldsymbol{w}})^{1/L}) - 1) = 0.$$

The proof is completed. □

Theorem 7 can be obtained in the same way.

Proof of Theorem 7. The claim holds since $\boldsymbol{v}^R$ is just $\boldsymbol{w}$ with a positive scaling factor, $\boldsymbol{v}^R$ and $\boldsymbol{w}$ share the same direction.

The proof is completed.

□

B.4. Convergence Rate of Empirical Loss and Parameter Norm

In the end of this section, we will give a tight bound for the convergence rate of empirical loss and parameter norm, which is derived by estimating $G(x)$ in Lemma 20. These results will further be used in Appendix C.

Theorem 10. *Let $\boldsymbol{v}$ obey an adaptive gradient flow $\mathcal{F}$ with empirical loss $\tilde{L}$ satisfying Assumption 1. Let $G(x)$ be defined as Lemma 20, i.e.,*

$$G(x) = \int_{\frac{1}{\tilde{\mathcal{L}}(t_1)}}^{x} \frac{g'(\log z)^2}{g(\log z)^{2-2/L}} \cdot dz.$$

Then $G(x) = \Theta(x(\log x)^{\frac{2}{L}-2})$, and $G^{-1}(x) = \Theta(x(\log x)^{2-\frac{2}{L}})$. Consequently,

$$\tilde{\mathcal{L}}(t) = \Theta\left(\frac{1}{t\log t^{2-\frac{2}{L}}}\right), \text{ and } \|\boldsymbol{v}(t)\| = \Theta\left(\frac{1}{\log t^{\frac{1}{L}}}\right).$$

Proof. For any large enough x,

$$\begin{aligned} G(x) &= \int_{\frac{1}{\tilde{\mathcal{L}}(t_1)}}^{x} \frac{g'\left(\log z\right)^2}{g\left(\log z\right)^{2-2/L}} \cdot dz \\ &\overset{(*)}{=} \Theta\left(\int_{\frac{1}{\tilde{\mathcal{L}}(t_1)}}^{x} \frac{g\left(\log z\right)^{2/L}}{\left(\log z\right)^2} \cdot dz\right) \\ &= \Theta\left(\int_{\frac{1}{\tilde{\mathcal{L}}(t_1)}}^{x} \frac{1}{\left(\log z\right)^{2-\frac{2}{L}}} \cdot dz\right) \\ &= \Theta\left(x(\log x)^{\frac{2}{L}-2}\right), \end{aligned}$$

where eq. $(*)$ is due to Proposition 1.

Since $G(x)$ is monotonously increasing, and $\lim_{x\to\infty} G(x) = \infty$, we have $x = \Theta\left(G^{-1}(x)(\log G^{-1}(x))^{\frac{2}{L}-2}\right)$, which further leads to $G^{-1}(x) = \Theta(x(\log x)^{2-\frac{2}{L}})$.

Since $\frac{1}{\tilde{\mathcal{L}}(t)} = G^{-1}(\Omega(t))$, we have $\frac{1}{\tilde{\mathcal{L}}(t)} = \Omega(t(\log t)^{2-\frac{2}{L}})$, which further leads to $\tilde{\mathcal{L}}(t) = \mathcal{O}\left(\frac{1}{t(\log t)^{2-\frac{2}{L}}}\right)$.

On the other hand,

$$-\frac{\mathrm{d}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\mathrm{d}t} = \left\|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(t)\right\|_2^2 \le 2\left\|\bar{\partial}\tilde{\mathcal{L}}(t)\right\|_2^2 = \mathcal{O}\left(\tilde{\mathcal{L}}(\boldsymbol{v}(t))g\left(\log\frac{1}{\tilde{\mathcal{L}}(t)}\right)^{1-\frac{1}{L}}\right),$$

which leads to $\frac{1}{\tilde{\mathcal{L}}(t)} = G^{-1}(\mathcal{O}(t))$, and $\tilde{\mathcal{L}}(t) = \Omega\left(\frac{1}{t(\log t)^{2-\frac{2}{L}}}\right)$. Therefore, $\tilde{\mathcal{L}}(t) = \Theta\left(\frac{1}{t(\log t)^{2-\frac{2}{L}}}\right)$

By Lemma 17, $\|\boldsymbol{v}(t)\|^L = \Theta\left(\log\frac{1}{\tilde{\mathcal{L}}(t)}\right)$, which leads to $\|\boldsymbol{v}(t)\| = \Theta\left(\frac{1}{\log t^{\frac{1}{L}}}\right)$.

The proof is completed. □

The convergent behavior of $\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))$ can be derived immediately by the above Theorem.

Corollary 2. *Let $\boldsymbol{v}$ obey an adaptive gradient flow $\mathcal{F}$ with empirical loss $\tilde{L}$ satisfying Assumption 1. Then, $\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\| = \Theta\left(\frac{1}{t(\log t)^{1-\frac{1}{L}}}\right)$.*

Proof. Since

$$\Omega\left(\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^{L-1}\right) = \langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \hat{\boldsymbol{v}}(t)\rangle \le \|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\| = \mathcal{O}\left(\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^{L-1}\right),$$

the proof is completed.

□

C. Proof of Theorem 3

In this section, we will prove that direction of parameters converges, that is, $\lim_{t\to\infty}\frac{\boldsymbol{v}(t)}{\|\boldsymbol{v}(t)\|}$ exists. Concretely, define the length swept by $\frac{\boldsymbol{v}(t)}{\|\boldsymbol{v}(t)\|}$ as $\zeta(t)$, i.e.,

$$\zeta(t) = \int_0^t \left\|\frac{\mathrm{d}\hat{\boldsymbol{v}}(\tau)}{\mathrm{d}\tau}\right\| \mathrm{d}\tau.$$

We will upper bound $\zeta(t)$ in the rest of this section.

To begin with, we define another surrogate margin $\bar{\gamma}$ as

$$\bar{\gamma}(\boldsymbol{v}) = \frac{g(\log \frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})})}{\|\boldsymbol{v}\|^L} + e^{-\tilde{\mathcal{L}}(\boldsymbol{v})}.$$

The following Lemma then lower bound the derivative of $\bar{\gamma}$.

Lemma 25. *For large enough t, we have*

$$\frac{\mathrm{d}\bar{\gamma}(\boldsymbol{v}(t))}{\mathrm{d}t} \geq \frac{1}{2}\left(\left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}\left(\boldsymbol{v}(t)\right)\right\|\left\|\bar{\partial}_{\backslash\backslash}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\| + \left\|\bar{\partial}_{\perp}\bar{\gamma}\left(\boldsymbol{v}(t)\right)\right\|\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\|\right),$$

and

$$\frac{\mathrm{d}\zeta(t)}{\mathrm{d}t} = \frac{\left\|(\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right))_{\perp}\right\|}{\|\boldsymbol{v}(t)\|}.$$

Proof. To begin with, we calculate the rate of $\boldsymbol{\beta}$ converging to $\mathbf{1}_p$. For AdaGrad, given a fixed index $i\in[N]$, we have that,

$$\begin{aligned}
0\leq \boldsymbol{\beta}_i(t)-1 =&(\boldsymbol{h}_\infty)_i^{-1}(\boldsymbol{h}_i(t)-(\boldsymbol{h}_\infty)_i) = \Theta(\boldsymbol{h}_i(t)-(\boldsymbol{h}_\infty)_i)\\
=&\Theta\left(\int_t^\infty \bar{\partial}\mathcal{L}(\boldsymbol{w}(\tau))^2 d\tau\right) = \Theta\left(\int_t^\infty \frac{1}{\tau^2(\log\tau)^{2-\frac{2}{L}}}d\tau\right)\\
=&\mathcal{O}\left(\frac{1}{t(\log t)^{2-\frac{2}{L}}}\right).
\end{aligned} \tag{15}$$

Similarly, for RMSProp, given a fixed index $i\in[N]$,

$$\begin{aligned}
&0\leq 1-\frac{\sqrt{\varepsilon}}{\sqrt{\varepsilon+(1-b)\int_0^t e^{-(1-b)(t-\tau)}(\bar{\partial}\mathcal{L}(\tau))_i^2\mathrm{d}\tau}} = \Theta\left((1-b)\int_0^t e^{-(1-b)(t-\tau)}(\bar{\partial}\mathcal{L}(\tau))_i^2\mathrm{d}\tau\right)\\
=&\Theta\left((1-b)\int_0^{t-\sqrt{t}} e^{-(1-b)(t-\tau)}(\bar{\partial}\mathcal{L}(\tau))_i^2\mathrm{d}\tau\right) + \Theta\left((1-b)\int_{t-\sqrt{t}}^{t} e^{-(1-b)(t-\tau)}(\bar{\partial}\mathcal{L}(\tau))_i^2\mathrm{d}\tau\right)\\
=&\mathcal{O}\left((1-b)e^{-(1-b)\sqrt{t}}\int_0^\infty(\bar{\partial}\tilde{\mathcal{L}}(\tau))_i^2\mathrm{d}\tau\right) + \mathcal{O}\left((1-b)\frac{\sqrt{t}}{(t-\sqrt{t})^2}\right)\\
=&\mathcal{O}\left(\frac{1}{t^{\frac{3}{2}}}\right) = \mathcal{O}\left(\frac{1}{t(\log t)^{2-\frac{2}{L}}}\right).
\end{aligned}$$

We then directly calculate the derivative of $\bar{\gamma}$:

$$\begin{aligned}
\frac{\mathrm{d}\bar{\gamma}(\boldsymbol{v}(t))}{\mathrm{d}t} =&\langle\bar{\partial}\bar{\gamma}(\boldsymbol{v}(t)), \frac{d\boldsymbol{v}(t)}{dt}\rangle\\
=&\left\langle -g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}\|\boldsymbol{v}(t)\|^L} - L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t) - e^{-\tilde{\mathcal{L}}}\bar{\partial}\tilde{\mathcal{L}}, \frac{d\boldsymbol{v}(t)}{dt}\right\rangle\\
=&\left\langle g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}\|\boldsymbol{v}(t)\|^L} + L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t) + e^{-\tilde{\mathcal{L}}}\bar{\partial}\tilde{\mathcal{L}}, \boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}\right\rangle\\
=&\left\langle g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}\|\boldsymbol{v}(t)\|^L} + L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t), \bar{\partial}\tilde{\mathcal{L}}\right\rangle + \left\langle e^{-\tilde{\mathcal{L}}}\bar{\partial}\tilde{\mathcal{L}}, \boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}\right\rangle\\
&+\left\langle g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}\|\boldsymbol{v}(t)\|^L} + L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t), (\mathbf{1}-\boldsymbol{\beta}(t))\odot\bar{\partial}\tilde{\mathcal{L}}\right\rangle.
\end{aligned}$$

The final term can be further bounded as follows:

$$\begin{aligned}
&\left|\left\langle g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}},(\mathbf{1}-\boldsymbol{\beta}(t))\odot\bar{\partial}\tilde{\mathcal{L}}\right\rangle\right| \\
\leq&\|\mathbf{1}-\boldsymbol{\beta}(t)\|_\infty\left\|g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}\right\|\|\bar{\partial}\tilde{\mathcal{L}}\| \\
\leq&\mathcal{O}\left(\frac{1}{t(\log t)^{2-\frac{2}{L}}}\right)\mathcal{O}\left(\frac{1}{\|\boldsymbol{v}(t)\|}\right)\|\bar{\partial}\tilde{\mathcal{L}}\| \\
=&\mathcal{O}\left(\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)\mathcal{O}\left(\frac{1}{\|\boldsymbol{v}(t)\|}\right)\|\bar{\partial}\tilde{\mathcal{L}}\| \\
=&\mathcal{O}\left(\frac{1}{\|\boldsymbol{v}(t)\|^L}\right)\|\bar{\partial}\tilde{\mathcal{L}}\|^2.
\end{aligned}$$

On the other hand,

$$\left\langle e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle=\Theta(\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2).$$

Therefore, there exists a large enough T, such that, any $t\geq T$,

$$\begin{aligned}
&\left\langle e^{-\tilde{\mathcal{L}}}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle+\left\langle g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}},(\mathbf{1}-\boldsymbol{\beta}(t))\odot\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle \\
\geq&\frac{1}{2}\left\langle e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle,
\end{aligned}$$

which further leads to

$$\frac{\mathrm{d}\bar{\gamma}(\boldsymbol{v}(t))}{\mathrm{d}t}\geq\frac{1}{2}\left\langle e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle+\left\langle\frac{g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}},\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle.$$

We then calculate axial component and radial component of $e^{-\tilde{\mathcal{L}}(\boldsymbol{v})}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})$, $g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})}{\tilde{\mathcal{L}}(\boldsymbol{v})\|\boldsymbol{v}\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})}\right)}{\|\boldsymbol{v}\|^{L+1}}\hat{\boldsymbol{v}}$, and $\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})$ respectively as follows:

$$\begin{aligned}
\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v})}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})\right)_{\backslash\backslash}&=\langle e^{-\tilde{\mathcal{L}}(\boldsymbol{v})}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}),\hat{\boldsymbol{v}}\rangle\hat{\boldsymbol{v}}, \\
\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v})}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})\right)_{\perp}&=e^{-\tilde{\mathcal{L}}(\boldsymbol{v})}(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})-\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}),\hat{\boldsymbol{v}}\rangle\hat{\boldsymbol{v}}), \\
\left(g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})}{\tilde{\mathcal{L}}(\boldsymbol{v})\|\boldsymbol{v}\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})}\right)}{\|\boldsymbol{v}\|^{L+1}}\hat{\boldsymbol{v}}\right)_{\backslash\backslash}&=L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})}\right)}{\|\boldsymbol{v}\|^{L+1}}\hat{\boldsymbol{v}}+g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}),\hat{\boldsymbol{v}}\rangle}{\tilde{\mathcal{L}}(\boldsymbol{v})\|\boldsymbol{v}\|^L}\hat{\boldsymbol{v}}, \\
\left(g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})}{\tilde{\mathcal{L}}(\boldsymbol{v})\|\boldsymbol{v}\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})}\right)}{\|\boldsymbol{v}\|^{L+1}}\hat{\boldsymbol{v}}\right)_{\perp}&=g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})-\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}),\hat{\boldsymbol{v}}\rangle\hat{\boldsymbol{v}}}{\tilde{\mathcal{L}}\|\boldsymbol{v}\|^L}, \\
\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})\right)_{\backslash\backslash}&=\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}),\hat{\boldsymbol{v}}\rangle\hat{\boldsymbol{v}} \\
\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})\right)_{\perp}&=\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})-\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}),\hat{\boldsymbol{v}}\rangle\hat{\boldsymbol{v}}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\frac{1}{2}\left\langle e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle+\left\langle g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}+L\frac{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}{\|\boldsymbol{v}\|^{L+1}}\hat{\boldsymbol{v}}(t),\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle\\
=&\frac{1}{2}\left\langle\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\backslash\backslash}+\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\perp},\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\backslash\backslash}+\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\perp}\right\rangle\\
&+\left\langle\left(g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t)\right)_{\backslash\backslash}+\left(g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t)\right)_{\perp},\right.\\
&\left.\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\backslash\backslash}+\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\perp}\right\rangle\\
\overset{(*)}{=}&\frac{1}{2}\left(\left\|\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\backslash\backslash}\right\|\left\|\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\backslash\backslash}\right\|+\left\|\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\perp}\right\|\left\|\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\perp}\right\|\right)\\
&+\left\|\left(g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t)\right)_{\backslash\backslash}\right\|\left\|\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\backslash\backslash}\right\|\\
&+\left\|\left(g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t)\right)_{\perp}\right\|\left\|\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\perp}\right\|\\
\geq&\frac{1}{2}\left\|\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))+g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t)\right)_{\backslash\backslash}\right\|\left\|\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\backslash\backslash}\right\|\\
&+\frac{1}{2}\left\|\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))+g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}+L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t)\right)_{\perp}\right\|\left\|\left(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right)_{\perp}\right\|,
\end{aligned}
$$

where eq. $(*)$ is because

$$
\begin{aligned}
\langle e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle&=-e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\langle-\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle<0,\\
e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}&>0,\\
L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}+g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\frac{\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}&\\
=-\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^{L+1}}\left(L\nu(t)g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)-\tilde{\mathcal{L}}(\boldsymbol{v}(t))g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right)&<0,\\
\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}&>0,\\
\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle=-\langle-\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle&<0.
\end{aligned}
$$

The proof is completed since $\bar{\partial}\bar{\gamma}(\boldsymbol{v})=-e^{-\tilde{\mathcal{L}}(\boldsymbol{v})}\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})-g'\left(\log\frac{1}{\tilde{\mathcal{L}}}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})}{\tilde{\mathcal{L}}(\boldsymbol{v})\|\boldsymbol{v}\|^L}-L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})}\right)}{\|\boldsymbol{v}\|^{L+1}}\hat{\boldsymbol{v}}$.

□

The following lemma gives an equivalent proposition of that the curve length ζ is finite.

Lemma 26. *ζ is finite if and only if*

$$
\int_0^\infty\frac{\left\|\bar{\partial}_\perp\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\|}{\|\boldsymbol{v}(t)\|}dt<\infty.
$$

Proof. By definition,

$$\begin{aligned}&(\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right))_{\perp}\\=&\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)-\langle\hat{\boldsymbol{v}}(t),\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\rangle\hat{\boldsymbol{v}}(t)\\=&\bar{\partial}_{\perp}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)+(1-\boldsymbol{\beta}(t))\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)-\langle\hat{\boldsymbol{v}}(t),(1-\boldsymbol{\beta}(t))\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\rangle\hat{\boldsymbol{v}}(t).\end{aligned}$$

Furthermore, by eq. (15), $\|1-\boldsymbol{\beta}(t)\|_{\infty}=\mathcal{O}(\frac{1}{t(\log t)^{2-\frac{2}{L}}})$,

$$\begin{aligned}&\|(1-\boldsymbol{\beta}(t))\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)-\langle\hat{\boldsymbol{v}}(t),(1-\boldsymbol{\beta}(t))\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\rangle\hat{\boldsymbol{v}}(t)\|\\\leq&2\|(1-\boldsymbol{\beta}(t))\odot\bar{\partial}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\|\\=&\mathcal{O}\left(\frac{1}{t^2}\right).\end{aligned}$$

Therefore,

$$\int_0^{\infty}\frac{\mathrm{d}\zeta(t)}{\mathrm{d}t}\leq\int_0^{\infty}\frac{\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\|}{\|\boldsymbol{v}(t)\|}dt+\int_0^{\infty}\mathcal{O}\left(\frac{1}{t^2}\right)dt.$$

The proof is completed. □

We then prove

$$\int_0^{\infty}\frac{\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\|}{\|\boldsymbol{v}(t)\|}dt<\infty.$$

Theorem 11. *There exists $a,\gamma_0>0$ and a definable desingularizing function Φ on $[0,a)$, such that, for large enough t,*

$$\frac{\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\|}{\|\boldsymbol{v}(t)\|}\leq-\frac{\mathrm{d}\Psi(\gamma_0-\bar{\gamma}(t))}{\mathrm{d}t}.$$

Proof. Since $\frac{d\bar{\gamma}(t)}{dt}\geq 0$, and both $\frac{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}{\|\boldsymbol{v}(t)\|^L}$ and $e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}$ are upper bounded, $\bar{\gamma}(t)$ converges to a limit non-decreasingly. Define $\gamma_0=\lim_{t\to\infty}\bar{\gamma}(t)$. If $\bar{\gamma}(t)=\gamma_0$ for a finite time t^0, then $\frac{\mathrm{d}\bar{\gamma}(t)}{dt}=0$ for any $t\geq t^0$, which further leads to $\|(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})(t))_{\perp}\|=0$, and the proof is then completed by letting $\Psi(x)=x$. Therefore, we only consider the case where $\bar{\gamma}(t)<\gamma_0$ for any finite time t. For any large enough t, we further divide the proof into two cases.

Case I. $\left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\right\|\geq\|\boldsymbol{v}(t)\|^{\frac{L}{4}}\left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\right\|$**.**

Applying Lemma 9 to $\gamma_0-\bar{\gamma}(\boldsymbol{v})|_{\|\boldsymbol{v}\|>1}$, there exists an $a_1>0$ and a definable desingularizing function Ψ_1, such that if $\|\boldsymbol{v}\|>1$, $\bar{\gamma}(\boldsymbol{v})>\gamma_0-a_1$, and

$$\left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v})\right\|\geq\|\boldsymbol{v}\|^{\frac{L}{4}}\left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v})\right\|,$$

then

$$\Psi_1'(\gamma_0-\bar{\gamma}(\boldsymbol{v}))\|\boldsymbol{v}\|\|\bar{\partial}\bar{\gamma}(\boldsymbol{v}(t))\|\geq 1.$$

Since $\lim_{t\to\infty}\bar{\gamma}(t)=\gamma_0$, and $\lim_{t\to\infty}\|\boldsymbol{v}(t)\|=\infty$, there exists a large enough time T_1, such that, for every $t\geq T_1$, $\|\boldsymbol{v}(t)\|>1$, and $\bar{\gamma}(t)>\gamma_0-a_1$.

Therefore, for any $t\geq T_1$ which satisfies $\left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\right\|\geq\|\boldsymbol{v}(t)\|^{\frac{L}{4}}\left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\right\|$, we have

$$\left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\right\|\geq\left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\right\|,$$

which further indicates

$$\left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\right\|\geq\frac{1}{2}\left\|\bar{\partial}\bar{\gamma}(\boldsymbol{v}(t))\right\|.$$

Furthermore, by Lemma 25,

$$\begin{aligned}\frac{\mathrm{d}\bar{\gamma}(t)}{dt} \geq & \frac{1}{2}\left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\right\|\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\| \\ \geq & \frac{1}{4}\|\boldsymbol{v}(t)\|\left\|\bar{\partial}\bar{\gamma}(\boldsymbol{v}(t))\right\|\frac{\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\|}{\|\boldsymbol{v}(t)\|} \\ \geq & \frac{1}{4\Psi_1'(\gamma_0-\bar{\gamma}(t))}\frac{\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\|}{\|\boldsymbol{v}(t)\|}.\end{aligned}$$

Case II. $\left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\right\| < \|\boldsymbol{v}(t)\|^{\frac{L}{4}}\left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\right\|$**.**

Applying Lemma 10 to $\gamma_0-\bar{\gamma}(\boldsymbol{v})|_{\|\boldsymbol{v}\|>1}$, we have that there exists an $a_2>0$ and a desingularizing function on $[0,a_2)$, such that if $\boldsymbol{v}>1$, and $\bar{\gamma}(\boldsymbol{v})>\gamma_0-a_2$, then

$$\max\left\{1,\frac{4}{L}\right\}\Psi_2'(\gamma_0-\bar{\gamma}(\boldsymbol{v}))\|\boldsymbol{v}\|^{\frac{L}{2}+1}\|\bar{\partial}\bar{\gamma}(\boldsymbol{v})\|\geq 1.$$

Similar to Case I., there exists a large enough time T_2 and constants C_1, such that, for every $t\geq T_1$, $\|\boldsymbol{v}(t)\|>1$, $\bar{\gamma}(t)>\gamma_0-a_2$, $C_1\|\boldsymbol{v}(t)\|^L\leq g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)$, and $\tilde{\mathcal{L}}(\boldsymbol{v}(t))\leq\|\boldsymbol{v}(t)\|^{-2L}$.

Therefore,

$$\left\|\bar{\partial}_{\backslash\backslash}g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right\| = g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\frac{L\nu(t)}{\|\boldsymbol{v}(t)\|}\geq\frac{Lg\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|}\geq LC_1\|\boldsymbol{v}(t)\|^{L-1}, \tag{16}$$

and

$$\begin{aligned}&\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\| \\ =&\left\|L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}\hat{\boldsymbol{v}}(t)+g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\frac{\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle}{\tilde{\mathcal{L}}\|\boldsymbol{v}(t)\|^L}\hat{\boldsymbol{v}}(t)+e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle\hat{\boldsymbol{v}}(t)\right\| \\ =&-L\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}-g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\frac{\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}-e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle \\ \overset{(*)}{\leq}&\frac{Lg\left(\log\frac{1}{Ne^{-f(0)}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}+\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\| \\ \leq&\frac{Lg\left(\log\frac{1}{Ne^{-f(0)}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}}+B_1\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^{L-1} \\ \leq&\frac{B_1+Lg\left(\log\frac{1}{Ne^{-f(0)}}\right)}{\|\boldsymbol{v}(t)\|^{L+1}},\end{aligned} \tag{17}$$

where inequality $(*)$ is due to

$$\begin{aligned}
&- Lg\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right) - g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\frac{\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\boldsymbol{v}(t)\rangle}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\\
=&- Lg\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right) + g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\frac{L\nu(t)}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\\
=&- Lg\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right) + g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\frac{L}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\left(\sum_{i=1}^{N}e^{-\tilde{q}_i(\boldsymbol{v}(t))}\tilde{q}_i(\boldsymbol{v}(t))f'\left(\tilde{q}_i(\boldsymbol{v}(t))\right)\right)\\
=&- Lg\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right) + L\left\langle\nabla_{\tilde{q}}g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right),\tilde{q}\right\rangle\\
\overset{(*)}{\leq}&- Lg\left(\log\frac{1}{Ne^{-f(0)}}\right),
\end{aligned}$$

where $\tilde{q}=(\tilde{q}_i)_{i=1}^N$ inequality $(*)$ comes from Jensen Inequality and $g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)$ is concave with respect to $\tilde{q}$.

Combining eqs. (16) and (17), we have

$$\left\|\bar{\partial}_{\backslash\backslash}g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right\| \geq \frac{LC_1}{B_1+Lg\left(\log\frac{1}{Ne^{-f(0)}}\right)}\|\boldsymbol{v}(t)\|^{2L}\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\|. \tag{18}$$

On the other hand,

$$\left\|\bar{\partial}_{\perp}g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right\| = \frac{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)) - \langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle\hat{\boldsymbol{v}}(t)\|,$$

while

$$\begin{aligned}
&\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\|\\
=&\left\|e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))}(\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)) - \langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle\hat{\boldsymbol{v}}(t)) + g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\frac{\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)) - \langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle\hat{\boldsymbol{v}}(t)}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}\right\|\\
=&\left(e^{-\tilde{\mathcal{L}}(\boldsymbol{v}(t))} + \frac{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\boldsymbol{v}(t)\|^L}\right)\left\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)) - \langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\hat{\boldsymbol{v}}(t)\rangle\hat{\boldsymbol{v}}(t)\right\|\\
\geq&\frac{1}{\|\boldsymbol{v}(t)\|^L}\left\|\bar{\partial}_{\perp}g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right\|.
\end{aligned} \tag{19}$$

Since

$$\left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\right\| < \|\boldsymbol{v}\|^{\frac{L}{4}}\left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\right\|,$$

combining eqs. (18) and (19), we have that

$$\begin{aligned}
\left\|\bar{\partial}_{\backslash\backslash}g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right\| \geq&\frac{LC_1}{B_1+Lg\left(\log\frac{1}{Ne^{-f(0)}}\right)}\|\boldsymbol{v}(t)\|^{2L}\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\|\\
\geq&\frac{LC_1}{B_1+Lg\left(\log\frac{1}{Ne^{-f(0)}}\right)}\|\boldsymbol{v}(t)\|^{\frac{7}{4}L}\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\|\\
\geq&\frac{LC_1}{B_1+Lg\left(\log\frac{1}{Ne^{-f(0)}}\right)}\|\boldsymbol{v}(t)\|^{\frac{3}{4}L}\left\|\bar{\partial}_{\perp}g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)\right\|.
\end{aligned}$$

Since $\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))$ is parallel to $\bar{\partial}g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)$, we also have

$$\left\|\bar{\partial}_{\backslash\backslash}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\| \geq \frac{LC_1}{B_1+Lg\left(\log\frac{1}{Ne^{-f(0)}}\right)}\|\boldsymbol{v}(t)\|^{\frac{3}{4}L}\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\|.$$

Furthermore,

$$\left\|\bar{\partial}\bar{\gamma}(\boldsymbol{v}(t))\right\| = \left\|\bar{\partial}_{\perp}\bar{\gamma}(\boldsymbol{v}(t))\right\| + \left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\right\| \leq \|\boldsymbol{v}\|^{\frac{L}{4}} \left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}(\boldsymbol{v}(t))\right\|.$$

Thus, by Lemma 25,

$$\begin{aligned}
\frac{\mathrm{d}\bar{\gamma}(t)}{\mathrm{d}t} \geq & \frac{1}{2}\left\|\bar{\partial}_{\backslash\backslash}\bar{\gamma}\left(\boldsymbol{v}(t)\right)\right\| \left\|\bar{\partial}_{\backslash\backslash}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\| \\
\geq & \frac{LC_1}{2(B_1 + Lg\left(\log\frac{1}{Ne^{-f(0)}}\right))}\|\boldsymbol{v}(t)\|^{\frac{1}{2}L}\left\|\bar{\partial}\bar{\gamma}\left(\boldsymbol{v}(t)\right)\right\| \left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\| \\
= & \frac{LC_1}{2(B_1 + Lg\left(\log\frac{1}{Ne^{-f(0)}}\right))}\|\boldsymbol{v}(t)\|^{\frac{1}{2}L+1}\left\|\bar{\partial}\bar{\gamma}\left(\boldsymbol{v}(t)\right)\right\| \frac{\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\|}{\|\boldsymbol{v}(t)\|} \\
\geq & \frac{LC_1}{2(B_1 + Lg\left(\log\frac{1}{Ne^{-f(0)}}\right))\max\left\{1,\frac{4}{L}\right\}\Psi_2'(\gamma_0 - \bar{\gamma}(\boldsymbol{v}(t)))} \frac{\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\|}{\|\boldsymbol{v}(t)\|}.
\end{aligned}$$

Concluding **Case I.** and **Case II.**, for any $t \geq \max\{T_1, T_2\}$, and $\Psi(x) = \max\{4\Psi_1(x), \frac{2(B_1+Lg\left(\log\frac{1}{Ne^{-f(0)}}\right))\max\left\{1,\frac{4}{L}\right\}}{LC_1}\Psi_2(x)\}$, we have that

$$\Psi'(\gamma_0 - \bar{\gamma}(\boldsymbol{v}(t)))\frac{\mathrm{d}\bar{\gamma}(t)}{\mathrm{d}t} \geq \frac{\left\|\bar{\partial}_{\perp}\tilde{\mathcal{L}}\left(\boldsymbol{v}(t)\right)\right\|}{\|\boldsymbol{v}(t)\|}.$$

The proof is completed. □

D. Proof for the Discrete Case

We prove the result for AdaGrad and experiential loss, with the result for RMSProp and logistic loss follows exactly as the continuous case. We slightly change the order of four stages in the flow: First, in Section D.1, we prove that the conditioner has a limit with no zero entry; secondly, in Section D.2, we prove that the empirical loss converges to zero; then, in Section D.3, we construct a further smoothed approximate margin, and prove it has a lower bound; finally, in Section D.4, we prove that every limit point of AdaGrad is along some KKT point of optimization problem (P^A) defined in Theorem 6.

D.1. Convergence of conditioners

Before the proof, we give a formal definition of the learning rate bound $C(t)$ in Assumption 2: let M be the smooth constant in Assumption 2. I. Then, $C(t) = \max\{\min_i\{\boldsymbol{h}_i^A(t)^{-1}\}/M, 1, \frac{C_1^2}{2LNe^{-1}}\}$, where C_0 will be clear below. By the monotony of $\boldsymbol{h}_i^A(t)$, apparently $C(t)$ is non-decreasing. Now we can prove $\sum_{t=1}^{\infty}\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2 < \infty$.

Lemma 27. *Suppose $\mathcal{L}$ is M smooth with respect to $\boldsymbol{w}$. Then, for $\{\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))\}_{t=1}^{\infty}$ updated by AdaGrad (eq. (3)), $\sum_{t=1}^{\infty}\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))^2 < \infty$.*

Proof. For any $t > t_0$,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}(t)) - \mathcal{L}(\boldsymbol{w}(t+1)) &= \eta_t\langle\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)), \boldsymbol{h}^A(t)\odot\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))\rangle - \frac{M}{2}\eta_t^2\|\boldsymbol{h}^A(t)\odot\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))\|^2 \\
&\geq \eta_t\frac{1}{2}\langle\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)), \boldsymbol{h}^A(t)\odot\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))\rangle \\
&\geq \tilde{\eta}\frac{1}{2}\langle\bar{\partial}\mathcal{L}(\boldsymbol{w}(t)), \boldsymbol{h}^A(t)\odot\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))\rangle.
\end{aligned}$$

Thus, since $\sum_{t=1}^{\infty}a_t$ share the same convergent behavior with $\sum_{t=1}^{\infty}\frac{a_t}{\sum_{\tau=1}^{t}a_\tau}$ ($a_t \geq 0$), by similar routine of Lemma 12, the proof is completed. □

Therefore, $\boldsymbol{h}_\infty \overset{\triangle}{=} \lim_{t\to\infty} \boldsymbol{h}^A(t)$ has no zero entry. We can then define a discrete version of adaptive gradient flow as

$$\begin{aligned}\boldsymbol{v}(t) &= \boldsymbol{h}_\infty^{-1/2} \odot \boldsymbol{w}(t),\\ \boldsymbol{\beta}(t) &= \boldsymbol{h}_\infty^{-1} \odot \boldsymbol{h}(t),\\ \tilde{\mathcal{L}}(\boldsymbol{v}) &= \mathcal{L}(\boldsymbol{h}_\infty^{\frac{1}{2}} \odot \boldsymbol{v}),\end{aligned}$$

and

$$\tilde{q}_i(\boldsymbol{v}) = q_i(\boldsymbol{h}_\infty^{\frac{1}{2}} \odot \boldsymbol{v}),$$

which further leads to

$$\boldsymbol{v}(t+1) - \boldsymbol{v}(t) = -\boldsymbol{\beta}(t) \odot \bar{\partial}\mathcal{L}_{ind}(\boldsymbol{v}(t)), \tag{20}$$

and $\boldsymbol{\beta}(t)$ decreases component-wisely to $\mathbf{1}_p$.

By Lemma 27, for any $t \geq t_0$, $\mathcal{L}(\boldsymbol{w}(t)) \leq \mathcal{L}(\boldsymbol{w}(t_0)) < N$. Therefore, there exists a positive real constant C_0 only depending on $\mathcal{L}(\boldsymbol{w}(t_0))$, such that, $\|\boldsymbol{w}(t)\| \geq C_0$. Furthermore, since $\boldsymbol{h}_\infty^{-\frac{1}{2}} \geq \boldsymbol{h}(t_0)^{-\frac{1}{2}}$, $\|\boldsymbol{v}(t)\| \geq C_0 \max_i\{\boldsymbol{h}_i(t_0)^{-\frac{1}{2}}\}$. Define $C_1 = C_0 \max_i\{\boldsymbol{h}_i(t_0)^{-\frac{1}{2}}\}$, which only depends on $\mathcal{L}(\boldsymbol{w}(t_0))$ and $\bar{\partial}\mathcal{L}(\boldsymbol{w}(t))$ $(t \leq t_0)$.

Moreover, similar to approximate flow, there exists a time t_1, such that, for any time $t \geq t_1$,

$$\begin{aligned}\sum_{i=1}^{p} \log \frac{1}{\boldsymbol{\beta}_i(t)^{-\frac{1}{2}}} &\leq \frac{1}{2},\\ \|\boldsymbol{\beta}^{-1}(t)\| &\geq \sqrt{\frac{1}{2}}.\end{aligned}$$

D.2. Convergence of Empirical Loss

Define function $\tilde{\gamma}'$ as the rate of $\log \frac{1}{\tilde{\mathcal{L}}}$ to $\|\boldsymbol{v}\|^{\frac{L}{4}}$:

$$\tilde{\gamma}'(t) = \frac{\log \frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}{\|\boldsymbol{v}(t)\|^{\frac{L}{4}}}.$$

Then, we have the following lemma.

Lemma 28. *For any $t \geq t_1$, $\tilde{\gamma}'(t)$ is non-decreasing.*

Proof. Since by Lemma 27, $\tilde{\mathcal{L}}$ is non-increasing, if $\|\boldsymbol{v}(t+1)\| \leq \|\boldsymbol{v}(t)\|$, the proposition trivially holds. Therefore, we consider the case that $\|\boldsymbol{v}(t+1)\| > \|\boldsymbol{v}(t)\|$ in the following proof.

The change of $\|\boldsymbol{v}(t)\|$ can be calculated as

$$\begin{aligned}&\|\boldsymbol{v}(t+1)\|^2 - \|\boldsymbol{v}(t)\|^2\\ =& -\eta_t\langle\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \boldsymbol{v}(t)\rangle + \eta_t^2\|\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2.\end{aligned}$$

Let $A = -\eta_t\langle\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \boldsymbol{v}(t)\rangle$, and $B = \eta_t^2\|\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2$. We estimate them separately.

As for A:

$$\begin{aligned}A &= -\eta_t\langle\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \boldsymbol{v}(t)\rangle\\ &\leq \eta_t\|\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\|\boldsymbol{v}(t)\|\\ &\leq 2\eta_t\|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\|\boldsymbol{v}(t)\|\\ &\leq 2\eta_t\|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\|\boldsymbol{v}(t)\|\frac{\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|\|\boldsymbol{v}(t)\|}{\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \boldsymbol{v}(t)\rangle}\\ &\leq 2\eta_t\frac{\|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2\|\boldsymbol{v}(t)\|^2}{\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \boldsymbol{v}(t)\rangle}.\end{aligned}$$

As for B:

$$\begin{aligned}
B &= \eta_t^2 \|\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2 \\
&\leq 2\eta_t^2 \|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2 \\
&\leq \frac{LNe^{-1}}{L\nu(t)} 2\eta_t^2 \|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2 \\
&\leq \frac{LNe^{-1}}{L\nu(t)C_1^2} 2\eta_t^2 \|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2 \|\boldsymbol{v}(t)\|^2 \\
&\leq 2\eta_t \frac{\|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2 \|\boldsymbol{v}(t)\|^2}{\langle \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \boldsymbol{v}(t)\rangle}.
\end{aligned}$$

Therefore,

$$\|\boldsymbol{v}(t+1)\|^2 - \|\boldsymbol{v}(t)\|^2 \leq 4\eta_t \frac{\|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2 \|\boldsymbol{v}(t)\|^2}{\langle \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \boldsymbol{v}(t)\rangle}. \tag{21}$$

On the other hand, by Lemma 27,

$$\begin{aligned}
\tilde{\mathcal{L}}(\boldsymbol{v}(t)) - \tilde{\mathcal{L}}(\boldsymbol{v}(t+1)) &\geq \frac{1}{2}\langle \bar{\partial}\mathcal{L}(\boldsymbol{w}(t)), \eta_t \boldsymbol{h}^A(t) \odot \bar{\partial}\mathcal{L}(\boldsymbol{w}(t))\rangle \\
&= \eta_t \frac{1}{2} \|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2.
\end{aligned} \tag{22}$$

Combining eqs. (21) and (22), we have

$$4\frac{1}{\nu(t)}(\tilde{\mathcal{L}}(\boldsymbol{v}(t)) - \tilde{\mathcal{L}}(\boldsymbol{v}(t+1))) \geq \frac{L}{2} \frac{\|\boldsymbol{v}(t+1)\|^2 - \|\boldsymbol{v}(t)\|^2}{\|\boldsymbol{v}(t)\|^2}.$$

Since $\nu(t) \geq \tilde{\mathcal{L}} \log \frac{1}{\tilde{\mathcal{L}}(t)}$, we further have

$$\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t)) \log \frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}} (\tilde{\mathcal{L}}(\boldsymbol{v}(t)) - \tilde{\mathcal{L}}(\boldsymbol{v}(t+1))) \geq \frac{L}{8} \frac{\|\boldsymbol{v}(t+1)\|^2 - \|\boldsymbol{v}(t)\|^2}{\|\boldsymbol{v}(t)\|^2}.$$

By the convexity of $\log\log\frac{1}{x}$ (when x is small) and $-\log x$,

$$\log\log\frac{1}{\mathcal{L}(\boldsymbol{v}(t+1))} - \log\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))} \geq \frac{L}{4}(\log\|\boldsymbol{v}(t+1)\| - \log\|\boldsymbol{v}(t)\|).$$

The proof is completed. □

Remark 3. *Actually, the convexity does not hold for $x \in [e^{-1}, \tilde{\mathcal{L}}(\boldsymbol{v}(t_1))]$ (if $\tilde{\mathcal{L}}(\boldsymbol{v}(t_1)) > e^{-1}$). However, we can instead define*

$$\Phi_0(x) = \log\log\frac{1}{x} + \int_0^x \inf\{-\frac{1}{w\log\frac{1}{w}} : w \in [x, \log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t_1))}]\} + \frac{1}{x\log\frac{1}{x}} dw.$$

Which satisfies for $x \in [0, \tilde{\mathcal{L}}(\boldsymbol{v}(t_1))]$, $\frac{1}{x\log\frac{1}{x}} \leq -\Phi_0'(x)$, $\Phi_0(x) \leq \log\log x$, and $\Phi_0(x)$ is convex. We can then replace $\log x$ by $e^{-\Phi_0(x)}$ and prove the above theorem in exactly the same way.0

With the relationship between $\|\boldsymbol{v}\|$ and $\tilde{\mathcal{L}}$, we can now prove that the empirical loss goes to zero.

Theorem 12. $\lim_{t\to\infty} \tilde{\mathcal{L}}(t) = 0$. *Furthermore,* $\lim_{t\to\infty} \rho(t) = \infty$.

Proof. By Lemma 31, for any integer time $t \geq t_0$

$$\tilde{\mathcal{L}}(t+1) - \tilde{\mathcal{L}}(t) \leq -\frac{1}{2}\eta_t \frac{L^2 \nu(t)^2}{\|\boldsymbol{v}(t)\|^2} \leq -\frac{1}{2}\eta_t \frac{L^2 \tilde{\mathcal{L}}(t)^2 \log \frac{1}{\tilde{\mathcal{L}}(t)^2}}{\|\boldsymbol{v}(t)\|^2}.$$

Furthermore, since $\tilde{\gamma}'(t) \geq \tilde{\gamma}'(t_1)$, we have that

$$\left(\frac{\log(\frac{1}{\tilde{\mathcal{L}}(t)})}{M_1^L \hat{\gamma}(t_0)}\right)^{\frac{1}{L}} \geq \rho.$$

Therefore,

$$\begin{aligned}
\tilde{\mathcal{L}}(t+1) - \tilde{\mathcal{L}}(t) &\leq -\frac{1}{2}\eta_t \frac{L^2 \tilde{\mathcal{L}}(t)^2 (\log \frac{1}{\tilde{\mathcal{L}}(t)})^2}{\|\boldsymbol{v}(t)\|^2} \\
&\leq -\frac{1}{2}\eta_t L^2 \tilde{\mathcal{L}}(t)^2 \left(\log \frac{1}{\tilde{\mathcal{L}}(t)}\right)^2 \left(\frac{\tilde{\gamma}'(t_1)}{\log(\frac{1}{\tilde{\mathcal{L}}(t)})}\right)^{\frac{8}{L}} \\
&= -\frac{1}{2}\eta_t L^2 \tilde{\mathcal{L}}(t)^2 \left(\log \frac{1}{\tilde{\mathcal{L}}(t)}\right)^{2-\frac{8}{L}} (\tilde{\gamma}'(t_1))^{\frac{8}{L}}.
\end{aligned}$$

Let $E_0 = \tilde{\mathcal{L}}(t_1)^2 \left(\log \frac{1}{\tilde{\mathcal{L}}(t_1)}\right)^{2-8/L}$, then $\psi(x) = \min\{x^2 (\log \frac{1}{x})^{2-\frac{8}{L}}, E_0\}$. Apparently, $\psi(x)$ is non-decreasing in $(0, \tilde{\mathcal{L}}(t_1)]$. Therefore, $E(x) = \int_x^{\tilde{\mathcal{L}}(t_1)} \psi(s) ds$ is convex with respect to x, and

$$\begin{aligned}
E(\tilde{\mathcal{L}}(t+1)) - E(\tilde{\mathcal{L}}(t)) &\geq E'(\tilde{\mathcal{L}}(t))(\tilde{\mathcal{L}}(t+1) - \tilde{\mathcal{L}}(t)) \\
&\geq \frac{1}{2}\eta_t L^2 (\tilde{\gamma}'(t_1))^{\frac{8}{L}},
\end{aligned}$$

which further implies

$$E(\tilde{\mathcal{L}}(t)) - E(\tilde{\mathcal{L}}(t_1)) \geq \sum_{\tau=t_0}^{t-1} \frac{1}{2}\eta_\tau L^2 (\tilde{\gamma}'(t_1))^{\frac{2}{L}}.$$

Since $\lim_{t\to\infty} \sum_{\tau=t_0}^{t-1} \frac{1}{2}\eta_\tau L^2 (\tilde{\gamma}'(t_1))^{\frac{8}{L}} = \infty$, we then have

$$\lim_{t\to\infty} E(\tilde{\mathcal{L}}(t)) = \infty,$$

and as a result,

$$\lim_{t\to\infty} \tilde{\mathcal{L}}(t) = 0.$$

□

D.3. Convergence of surrogate margin

As a preparation, define $B_1 = \max\{\|\bar{\partial}\tilde{q}_i(\boldsymbol{v})\| : i \in [N], \boldsymbol{v} \in \mathcal{B}(1)\}$, and $B_2 = \max_i \{\|\mathcal{H}\tilde{q}_i(\boldsymbol{v})\| : i \in [N], \boldsymbol{v} \in \mathcal{B}(1)\}$.

Furthermore, we define $\lambda(x) = (\log \frac{1}{x})^{-1}$, and $\mu(x) = \frac{\log \frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t_1))}}{\log \frac{1}{x}}$. Since $\lim_{t\to\infty} \tilde{\mathcal{L}}(t) = 0$, we have the following lemma.

Lemma 29. *There exists a large enough time $t_2 \geq t_1$, such that, for any $t \geq t_2$,*

$$B_1^2 \tilde{\mathcal{L}}(\boldsymbol{v}) \log^{7-\frac{8}{L}} \frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})} / \tilde{\gamma}'(t_1)^{\frac{8}{L}} \leq \lambda(\tilde{\mathcal{L}}(\boldsymbol{v}(t)))\mu(\tilde{\mathcal{L}}(\boldsymbol{v}(t))),$$

and,

$$\frac{1}{\tilde{\gamma}'(t_1)^{8-8/L}} \tilde{\mathcal{L}}(\boldsymbol{v}(t)) \log^{8-8/L} \frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))} \left(B_1^2 + C_1^{-L} B_2\right) \leq \mu(t).$$

Proof. The proposition is obvious since $\log^i(\frac{1}{x}) = \mathbf{o}(x), \forall i$ as $x \to 0$. □

Then, we define a further surrogate margin $\hat{\gamma}$ of the discrete case as following:

$$\hat{\gamma}(t) := \frac{e^{\Phi(\tilde{\mathcal{L}})}}{\rho^L},$$

where $\rho(t)$ is defined as $\|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t)\|$, and $\Phi(x)$ is defined as

$$\Phi(x) = \log\log\frac{1}{x} + \int_0^x \left(-\sup\left\{\frac{1+2(1+\lambda(\tilde{w})/L)\mu(\tilde{w})}{\tilde{w}\log\frac{1}{\tilde{w}}} : \tilde{w} \in \left[w, \tilde{\mathcal{L}}\left(t_2\right)\right]\right\} + \frac{1}{w\log\frac{1}{w}}\right) dw.$$

The following properties hold for $\hat{\gamma}$.

Lemma 30. • *Let a series of $\boldsymbol{v}_i$ satisfies $\lim_{i\to\infty}\|\boldsymbol{v}_i\| = \infty$. Then, $\lim_{i\to\infty}\frac{\hat{\gamma}(\boldsymbol{v}_i)}{\tilde{\gamma}(\boldsymbol{v}_i)} = 1$;*

• *If $\tilde{\mathcal{L}}(\boldsymbol{v}(t)) \le \tilde{\mathcal{L}}(\boldsymbol{v}(t_2))$, then $\hat{\gamma}(t) < \tilde{\gamma}(t) \le \bar{\gamma}(t)$.*

Proof. As beginning, we verify the existence of Φ. Actually, when x is small enough, $\frac{1+2(1+\lambda(x)/L)\mu(x)}{x\log\frac{1}{x}}$ decreases, and $\lim_{x\to 0}\frac{1+2(1+\lambda(x)/L)\mu(x)}{x\log\frac{1}{x}} = \infty$. Therefore, there exists a small enough ε, such that, for any $w < \varepsilon$,

$$\sup\left\{\frac{1+2(1+\lambda(\tilde{w})/L)\mu(\tilde{w})}{\tilde{w}\log\frac{1}{\tilde{w}}} : \tilde{w} \in \left[w, \tilde{\mathcal{L}}\left(t_2\right)\right]\right\} = \frac{1+2(1+\lambda(w)/L)\mu(w)}{w\log\frac{1}{w}},$$

which further leads to

$$\begin{aligned}
&-\sup\left\{\frac{1+2(1+\lambda(\tilde{w})/L)\mu(\tilde{w})}{\tilde{w}\log\frac{1}{\tilde{w}}} : \tilde{w} \in \left[w, \tilde{\mathcal{L}}\left(t_2\right)\right]\right\} + \frac{1}{w\log\frac{1}{w}} \\
&= -\frac{2(1+\lambda(w)/L)\mu(w)}{w\log\frac{1}{w}},
\end{aligned}$$

which is integrable as $w \to 0$. Concretely, for any $x < \varepsilon$,

$$\begin{aligned}
&\int_0^x \left(-\sup\left\{\frac{1+2(1+\lambda(\tilde{w})/L)\mu(\tilde{w})}{\tilde{w}\log\frac{1}{\tilde{w}}} : \tilde{w} \in \left[w, \tilde{\mathcal{L}}\left(t_2\right)\right]\right\} + \frac{1}{w\log\frac{1}{w}}\right) dw \\
&= \int_0^x -\frac{2(1+\lambda(w)/L)\mu(w)}{w\log\frac{1}{w}} dw \\
&= -\log\frac{1}{\mathcal{L}(t_1)}\left(\frac{1}{\log\frac{1}{x}} + \frac{1}{2L\log^2\frac{1}{x}}\right).
\end{aligned}$$

Therefore, for a series $\{\boldsymbol{v}_i\}_{i=1}^{\infty}$ satisfying $\lim_{i\to\infty}\|\boldsymbol{v}_i\| = \infty$,

$$\lim_{i\to\infty}\frac{\tilde{\gamma}(\boldsymbol{v}_i)}{\hat{\gamma}(\boldsymbol{v}_i)} = \lim_{i\to\infty} e^{-\mathcal{O}\left(\frac{1}{\log^2\frac{1}{\mathcal{L}(\boldsymbol{v}_i)}}\right)} = 1.$$

Furthermore, if $x \le \tilde{\mathcal{L}}(\boldsymbol{v}(t_2))$,

$$\Phi'(x) \le \frac{1}{w\log\frac{1}{w}} - \frac{1+2(1+\lambda(\tilde{w})/L)\mu(\tilde{w})}{\tilde{w}\log\frac{1}{\tilde{w}}} < 0,$$

which proves that $\hat{\gamma}(t) < \tilde{\gamma}(t)$.

The proof is completed. □

To bound the norm of first and second derivatives of $\tilde{\mathcal{L}}$, we further need the following lemma.

The next lemma characterizes the behavior of surrogate margin $\hat{\gamma}(t)$.

Lemma 31. *For positive integer time $t \geq t_2$, $\hat{\gamma}(t) \geq e^{-\frac{1}{2}}\hat{\gamma}(t_2)$.*

Proof. For any time $t \geq t_2$,

$$\rho(t+1)^2 - \rho(t)^2 = (\|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t+1)\|^2 - \rho(t)^2) + (\rho(t+1)^2 - \|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t+1)\|^2).$$

We calculate two parts separately

$$\begin{aligned}
&\|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t+1)\|^2 - \|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t)\|^2 \\
=&\eta_t^2 \|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2 + 2L\eta_t\nu(t) \\
\geq& 0.
\end{aligned}$$

On the other hand, since $\boldsymbol{\beta}^{-\frac{1}{2}}$ is non-decreasing,

$$\rho(t+1) = \|\boldsymbol{\beta}^{-\frac{1}{2}}(t+1) \odot \boldsymbol{v}(t+1)\| \geq \|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t+1)\|.$$

$\|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t+1)\|^2 - \|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t)\|^2$ can also be upper bounded as follows:

$$\begin{aligned}
&\|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t+1)\|^2 - \|\boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \boldsymbol{v}(t)\|^2 \\
=&\eta_t^2 \|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2 + 2L\eta_t\nu(t) \\
=&2L\eta_t\nu(t)\left(\frac{\eta_t \|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2}{2L\nu(t)} + 1\right) \\
\overset{(*)}{\leq}&2L\eta_t\nu(t)\left(\frac{\lambda(\tilde{\mathcal{L}}(\boldsymbol{v}(t)))\mu(\tilde{\mathcal{L}}(\boldsymbol{v}(t)))}{L} + 1\right),
\end{aligned}$$

where inequality (*) comes from the estimation of $\|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2$ as follows: by the homogeneity of $\tilde{q}_i$, we have

$$\begin{aligned}
\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})\| = \left\|-\sum_{i=1}^{N} e^{-\tilde{q}_i(\boldsymbol{v})}\bar{\partial}\tilde{q}_i(\boldsymbol{v})\right\| &\overset{(**)}{\leq} B_1\tilde{\mathcal{L}}(\boldsymbol{v})\|\boldsymbol{v}\|^{L-1} \\
&\overset{(***)}{\leq} B_1 \frac{1}{\tilde{\gamma}'(t_1)^{\frac{4}{L}}}\tilde{\mathcal{L}}(\boldsymbol{v})\log^{4-\frac{4}{L}}\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})},
\end{aligned}$$

where inequality $(**)$ is due to $\bar{\partial}\tilde{q}_i$ is $(L-1)$ homogeneous, and inequality $(***)$ holds by Lemma 28. On the other hand, $\nu(t) \geq \tilde{\mathcal{L}}(\boldsymbol{v}(t))\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}$. Combining the estimation of ν and $\|\boldsymbol{\beta}(t)^{-\frac{1}{2}} \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{\beta}(t))\|$, we have

$$\begin{aligned}
\frac{\eta_t \|\boldsymbol{\beta}^{\frac{1}{2}}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|^2}{2L\nu(t)} &\leq \frac{B_1^2\tilde{\mathcal{L}}(\boldsymbol{v})\log^{7-\frac{8}{L}}\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v})}}{L\tilde{\gamma}'(t_1)^{\frac{8}{L}}} \\
&\leq \frac{\lambda(\tilde{\mathcal{L}}(\boldsymbol{v}(t)))\mu(\tilde{\mathcal{L}}(\boldsymbol{v}(t)))}{L}.
\end{aligned}$$

Similar to Lemma 27, the decrease of $\tilde{\mathcal{L}}(\boldsymbol{v})$ can be calculated by second order Taylor Expansion:

$$\begin{aligned}
\tilde{\mathcal{L}}(t+1) - \tilde{\mathcal{L}}(t) =& -\eta_t\left\langle \boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\right\rangle \\
&+\frac{1}{2}\eta_t^2(\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}})^T\mathcal{H}(\tilde{\mathcal{L}}(\boldsymbol{v}(\xi)))(\boldsymbol{\beta}(t) \odot \bar{\partial}\tilde{\mathcal{L}}),
\end{aligned} \tag{23}$$

where $\xi \in (0,1)$.

By homogeneity of $\tilde{q}_i$, the norm of Hessian matrix $\|\mathcal{H}(\tilde{\mathcal{L}}(\boldsymbol{v}(\xi)))\|$ can be bounded as

$$\begin{aligned}\|\mathcal{H}(\tilde{\mathcal{L}}(\boldsymbol{v}(\xi)))\| &= \left\|\sum_{i=1}^{N} e^{-\tilde{q}_i}\left(\bar{\partial}\tilde{q}_i\bar{\partial}\tilde{q}_i^\top - \mathcal{H}\tilde{q}_i\right)\right\|_2 \\ &\overset{(*)}{\leq} \sum_{i=1}^{N} e^{-\tilde{q}_i}\left(B_1^2\|\boldsymbol{v}(\xi)\|^{2L-2} + B_2\|\boldsymbol{v}(\xi)\|^{L-2}\right) \\ &\leq \tilde{\mathcal{L}}(\boldsymbol{v}(\xi))\|\boldsymbol{v}(\xi)\|^{2L-2}\left(B_1^2 + C_1^{-L}B_2\right) \\ &\leq \frac{1}{\tilde{\gamma}'(t_1)^{8-8/L}}\tilde{\mathcal{L}}(\boldsymbol{v}(\xi))\log^{8-8/L}\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(\xi))}\left(B_1^2 + C_1^{-L}B_2\right).\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{1}{2}\eta_t^2(\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}})^T\mathcal{H}(\tilde{\mathcal{L}}(\boldsymbol{v}(\xi)))(\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}) \\ \leq&\eta_t\|\boldsymbol{\beta}(t)^{\frac{1}{2}}\odot\bar{\partial}\tilde{\mathcal{L}})\|^2\frac{1}{\tilde{\gamma}'(t_1)^{8-8/L}}\tilde{\mathcal{L}}(\boldsymbol{v}(\xi))\log^{8-8/L}\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(\xi))}\left(B_1^2 + C_1^{-L}B_2\right) \\ \leq&\eta_t\|\boldsymbol{\beta}(t)^{\frac{1}{2}}\odot\bar{\partial}\tilde{\mathcal{L}})\|^2\mu(t).\end{aligned} \tag{24}$$

Taking the estimation eq. (24) back to eq. (23), we have

$$\tilde{\mathcal{L}}(t) - \tilde{\mathcal{L}}(t+1) \geq (1-\mu(t))\eta_t\|\boldsymbol{\beta}(t)^{\frac{1}{2}}\odot\bar{\partial}\tilde{\mathcal{L}})\|^2 \tag{25}$$

By multiplying $\frac{1+\lambda(\tilde{\mathcal{L}}(t))\mu(\tilde{\mathcal{L}}(t))/L}{(1-\mu(\tilde{\mathcal{L}}(t)))\nu(t)}$ to both sides of eq. (25), we then have

$$\begin{aligned}&\frac{1+\lambda(\tilde{\mathcal{L}}(t))\mu(\tilde{\mathcal{L}}(t))/L}{(1-\mu(\tilde{\mathcal{L}}(t)))\nu(t)}(\tilde{\mathcal{L}}(t+1) - \tilde{\mathcal{L}}(t)) \\ \leq& -\eta_t\left(1 + \frac{\lambda(\tilde{\mathcal{L}}(t))\mu(\tilde{\mathcal{L}}(t))}{L}\right)\frac{L^2\nu(t)}{\rho(t)^2} \\ \leq& -\frac{L}{2}\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|^2 - \|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)\|^2}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)\|^2}.\end{aligned}$$

Furthermore, since $-\Phi'(\tilde{\mathcal{L}}(t)) \geq \frac{1+\lambda(\tilde{\mathcal{L}}(t))\mu(\tilde{\mathcal{L}}(t))/L}{(1-\mu(\tilde{\mathcal{L}}(t)))\tilde{\mathcal{L}}(t)/\lambda(\tilde{\mathcal{L}}(t))}$, by the convexity of Φ and $-\log x$, we have that

$$\begin{aligned}&L\left(\log\frac{1}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|} - \log\frac{1}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)\|}\right) \\ &+(\Phi(\tilde{\mathcal{L}}(t+1)) - \Phi(\tilde{\mathcal{L}}(t))) \geq 0.\end{aligned}$$

Therefore,

$$\log\frac{\Phi(\tilde{\mathcal{L}}(t+1))}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|^L} - \log\frac{\Phi(\tilde{\mathcal{L}}(t)}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)\|^L} \geq 0.$$

Furthermore, since

$$\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot\boldsymbol{v}(t+1)\|^L}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot\boldsymbol{v}(t)\|^L} \geq \Pi_{i=1}^{p}\frac{\beta_i^{-L/2}(t)}{\beta_i^{-L/2}(t+1)},$$

we have

$$\hat{\gamma}(t+1) = \frac{e^{-\Phi(t+1)}}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot \boldsymbol{v}(t+1)\|^L} \geq \frac{e^{-\Phi(t+1)}}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot \boldsymbol{v}(t+1)\|^L}\Pi_{i=1}^p \frac{\beta_i^{-L/2}(t)}{\beta_i^{-L/2}(t+1)}$$
$$\geq \hat{\gamma}(t)\Pi_{i=1}^p \frac{\beta_i^{-L/2}(t)}{\beta_i^{-L/2}(t+1)}.$$

Thus, by induction,

$$\hat{\gamma}(t+1) \geq \hat{\gamma}(t)\Pi_{i=1}^p \frac{\beta_i^{-L/2}(t_0)}{\beta_i^{-L/2}(t+1)} \geq e^{-\frac{1}{2}}\hat{\gamma}(t_1).$$

The proof is completed. □

Similar to the flow case, we can then prove the convergence of $\hat{\gamma}$.

Lemma 32. *There exists a positive real $\hat{\gamma}_\infty$, such that*

$$\lim_{t\to\infty} \hat{\gamma}(t) = \hat{\gamma}_\infty.$$

Proof. Since for any $t \geq t_2$

$$\log \frac{\hat{\gamma}(t+1)}{\hat{\gamma}(t)} \geq \log \Pi_{i=1}^p \frac{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t+1)},$$

we have that $\hat{\gamma}(t)\Pi_{i=1}^p \frac{1}{\boldsymbol{\beta}_i^{\frac{1}{2}}(t)}$ monotonously increases. Furthermore, since $\hat{\gamma}(t) < \gamma(t)$ is bounded, so does $\hat{\gamma}(t)\Pi_{i=1}^p \frac{1}{\boldsymbol{\beta}_i^{\frac{1}{2}}(t)}$. Therefore, $\hat{\gamma}(t)\Pi_{i=1}^p \frac{1}{\boldsymbol{\beta}_i^{\frac{1}{2}}(t)}$ converges to a positive real. Since $\lim_{t\to\infty} \Pi_{i=1}^p \frac{1}{\boldsymbol{\beta}_i^{\frac{1}{2}}(t)} = 1$, the proof is completed. □

D.4. Verification of KKT point

Similar to the flow case, we have the following construction of (ε, δ) KKT point. The proof is exactly the same as Lemma 5, and we omit it here.

Lemma 33. *Let $\lambda_i = q_{\min}^{1-2/L}\|\boldsymbol{v}\| \cdot e^{-f(q_i)} f'(q_i) / \|\bar{\partial}\tilde{\mathcal{L}}\|_2$. Then $\tilde{\boldsymbol{v}}(t)$ is a (ε, δ) KKT point of $(\tilde{P})$, where ε, δ are defined as follows:*

$$\varepsilon = C_1(1+\cos(\boldsymbol{\theta}))$$
$$\delta = C_2 \frac{1}{\log \frac{1}{\mathcal{L}}},$$

where $\cos(\boldsymbol{\theta})$ is defined as $\langle \hat{v}(t), \widehat{\bar{\partial}\tilde{\mathcal{L}}}(t)\rangle$, and C_1, C_2 are positive real constants.

By Lemma 22, we only need to prove that $\lim_{t\to\infty} \tilde{\cos}(\boldsymbol{\theta}) = 1$. Furthermore, the estimation of $\cos(\tilde{\boldsymbol{\theta}})$ can be given by the following lemma.

Lemma 34. *For any $t_3 > t_4 \geq t_2$,*

$$\sum_{\tau=t_3}^{t_4-1} \left(\tilde{\cos}(\boldsymbol{\theta})(\tau)^{-2} - 1\right)\left(\log \frac{1}{\rho(t)} - \log \frac{1}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(\tau)\odot \boldsymbol{v}(\tau+1)\|}\right) \leq \frac{1}{L}\log \frac{\hat{\gamma}(t_4)}{\hat{\gamma}(t_3)} + \log\left(\Pi_{i=1}^p \frac{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t_3)}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t_4)}\right).$$

Proof. By Lemma 31, for any $t \geq t_0$

$$\begin{aligned}\frac{1}{L}\log\frac{\hat{\gamma}(t+1)}{\hat{\gamma}(t)} &= \frac{1}{L}\left(\Phi(t+1)-\Phi(t)\right)+\left(\log\frac{1}{\rho(t+1)}-\log\frac{1}{\rho(t)}\right)\\ &\geq \left(\log\frac{1}{\rho(t)}-\log\frac{1}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|}\right)\frac{\rho(t)^2}{L^2\nu(t)^2}\|\boldsymbol{\beta}^{\frac{1}{2}}(t)\odot\bar{\partial}\tilde{\mathcal{L}}(t)\|^2\\ &\quad+\left(\log\frac{1}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|}-\log\frac{1}{\rho(t)}\right)-\log\left(\Pi_{i=1}^{p}\frac{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t+1)}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}\right).\end{aligned}$$

Furthermore, since $L\nu(t)=\left\langle\widehat{\boldsymbol{\beta}^{\frac{1}{2}}(t)\odot\bar{\partial}\tilde{\mathcal{L}}(t)},\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}}\right\rangle$, we have

$$\begin{aligned}\frac{1}{L}\log\frac{\hat{\gamma}(t+1)}{\hat{\gamma}(t)} &\geq \left(\log\frac{1}{\rho(t)}-\log\frac{1}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|}\right)\cos(\boldsymbol{\theta})^{-2}\\ &\quad+\left(\log\frac{1}{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|}-\log\frac{1}{\rho(t)}\right)-\log\left(\Pi_{i=1}^{p}\frac{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t+1)}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}\right).\end{aligned}$$

The proof is completed. □

We still need a lemma to bound the change of the direction of $\boldsymbol{\beta}^{-\frac{1}{2}}\odot\boldsymbol{v}$.

Lemma 35. *For any $t \geq t_2$,*

$$\begin{aligned}&\|\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)}\odot\boldsymbol{v}(t+1)-\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}}(t)\|\\ \leq&\mathcal{O}(1)\sum_{i=1}^{p}\left(\boldsymbol{\beta}_i^{-1}(t+1)-\boldsymbol{\beta}_i^{-1}(t)\right)+\left(\mathcal{O}(1)\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}\odot\boldsymbol{v}(t+1)\|}{\rho(t)}+1\right)\log\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|}{\rho(t)}\\ &+\left(\mathcal{O}(1)\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}\odot\boldsymbol{v}(t+1)\|}{\rho(t)}+1\right)\log\Pi_{i=1}^{p}\frac{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t+1)}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}.\end{aligned}$$

Proof. Since triangular inequality,

$$\begin{aligned}&\|\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)}\odot\boldsymbol{v}(t+1)-\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}}(t)\|\\ \leq&\left\|\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot\boldsymbol{v}(t+1)-\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot\boldsymbol{v}(t)\right\|\\ &+\left\|\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot\boldsymbol{v}(t)-\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)\right\|\\ &+\left\|\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)-\frac{1}{\rho(t)}\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)\right\|.\end{aligned}$$

Let

$$\begin{aligned}A&=\left\|\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot\boldsymbol{v}(t+1)-\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot\boldsymbol{v}(t)\right\|;\\ B&=\left\|\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)\odot\boldsymbol{v}(t)-\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)\right\|;\\ C&=\left\|\frac{1}{\rho(t+1)}\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)-\frac{1}{\rho(t)}\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)\right\|.\end{aligned}$$

Then,

$$\begin{aligned}A =&\mathcal{O}(1)\frac{1}{\rho(t+1)}\|\boldsymbol{v}(t+1)-\boldsymbol{v}(t)\|\\ =&\mathcal{O}(1)\frac{1}{\rho(t+1)}\|\eta_t\boldsymbol{\beta}(t)\odot\bar{\partial}\tilde{\mathcal{L}}\|=\mathcal{O}(1)\frac{\eta_t}{\rho(t+1)}\|\bar{\partial}\tilde{\mathcal{L}}\|\\ \overset{(*)}{\leq}&\mathcal{O}(1)\frac{\eta_t\nu(t)}{\hat{\gamma}(t_2)\rho(t+1)\rho(t)}\leq\mathcal{O}(1)\frac{\rho(t+1)^2-\rho(t)^2}{\rho(t+1)\rho(t)}\\ \leq&\mathcal{O}(1)\frac{\rho(t+1)^2-\rho(t)^2}{\rho(t+1)^2}\frac{\rho(t+1)}{\rho(t)}\leq\mathcal{O}(1)\frac{\rho(t+1)}{\rho(t)}\log\frac{\rho(t+1)}{\rho(t)},\end{aligned}$$

where eq. $(*)$ can be derived in the same way as Lemma 6;

$$\begin{aligned}B&\leq\sum_{i=1}^{p}(\boldsymbol{\beta}_i^{-\frac{1}{2}}(t+1)-\boldsymbol{\beta}_i^{-\frac{1}{2}}(t))^2\frac{\|\boldsymbol{v}(t)\|}{\rho(t+1)}\\ &\leq\mathcal{O}(1)\sum_{i=1}^{p}\boldsymbol{\beta}_i^{-1}(t+1)-\boldsymbol{\beta}_i^{-1}(t);\end{aligned}$$

and

$$C=1-\frac{\rho(t)}{\rho(t+1)}.$$

Therefore,

$$\begin{aligned}&\|\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t+1)}\odot\boldsymbol{v}(t+1)-\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t)}\|\\ \leq&\mathcal{O}(1)\sum_{i=1}^{p}\left(\boldsymbol{\beta}_i^{-1}(t+1)-\boldsymbol{\beta}_i^{-1}(t)\right)+\left(\mathcal{O}(1)\frac{\rho(t+1)}{\rho(t)}+1\right)\log\frac{\rho(t+1)}{\rho(t)}\\ \leq&\mathcal{O}(1)\sum_{i=1}^{p}\left(\boldsymbol{\beta}_i^{-1}(t+1)-\boldsymbol{\beta}_i^{-1}(t)\right)+\left(\mathcal{O}(1)\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}\odot\boldsymbol{v}(t+1)\|}{\rho(t)}+1\right)\log\frac{\|\boldsymbol{\beta}(t)^{-\frac{1}{2}}(t)\odot\boldsymbol{v}(t+1)\|}{\rho(t)}\\ &+\left(\mathcal{O}(1)\frac{\|\boldsymbol{\beta}(t)^{-\frac{1}{2}}\odot\boldsymbol{v}(t+1)\|}{\rho(t)}+1\right)\log\Pi_{i=1}^{p}\frac{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t+1)}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}\end{aligned}$$

□

We then prove that $\sum_{\tau=t_2}^{\infty}\log\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}(\tau)\odot\boldsymbol{v}(\tau+1)\|}{\rho(\tau)}=\infty$.

Lemma 36. *The sum of* $\log\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}(\tau)\odot\boldsymbol{v}(\tau+1)\|}{\rho(\tau)}$ *diverges, that is,* $\sum_{\tau=t_2}^{\infty}\log\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}(\tau)\odot\boldsymbol{v}(\tau+1)\|}{\rho(\tau)}=\infty$.

Proof.

$$\sum_{\tau=t}^{\infty}\log\frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}(\tau)\odot\boldsymbol{v}(\tau+1)\|}{\rho(\tau)}\geq\sum_{\tau=t}^{\infty}\log\frac{\rho(\tau+1)}{\rho(\tau)}-\log\Pi_{i=1}^{p}\frac{1}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}.$$

The proof is completed since $\lim_{t\to\infty}\rho(t)=\infty$ and $\log\Pi_{i=1}^{p}\frac{1}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(t)}$ is bounded. □

Now we can prove the following lemma.

Lemma 37. *Let* $\bar{\boldsymbol{v}}$ *be any limit point of* $\{\boldsymbol{v}(t)\}_{t=1}^{\infty}$. *Then* $\bar{\boldsymbol{v}}$ *is a KKT point of optimization problem* (P).

Proof. Let t^1 be any integer time larger than t_2. We construct a sequence $\{t^i\}_{i=1}^{\infty}$ by iteration. Suppose $t^1, \cdots, t^{k-1}$ have been constructed. Let $s^k > t^{k-1}$ be a large enough time which satisfies

$$\log \Pi_{i=1}^{p} \frac{1}{\boldsymbol{\beta}_i^{-\frac{1}{2}}(s^k)} \leq \frac{1}{k^3},$$
$$\|\hat{\boldsymbol{v}}(s^k) - \bar{\boldsymbol{v}}\| \leq \frac{1}{k},$$
$$\log\left(\frac{\hat{\gamma}_\infty}{\hat{\gamma}(s^k)}\right) \leq \frac{1}{k^3}$$
$$\|\boldsymbol{\beta}^{-1}(s^k)\| \leq 1 + \frac{1}{k-1}.$$

Then let $(s^k)'$ (guaranteed by Lemma 36) be the first time greater than s^k that $\sum_{\tau=s^k}^{(s^k)'-1} \log \frac{\|\boldsymbol{\beta}^{-\frac{1}{2}}(\tau)\odot \boldsymbol{v}(\tau+1)\|}{\rho(\tau)} \geq \frac{1}{k}$. By Lemma 34, there exists a time $t^k \in [s^k, (s^k)'-1]$, such that $\tilde{\cos}(\boldsymbol{\theta})^{-2} - 1 \leq \frac{1}{k^2}$.

Moreover,

$$\begin{aligned}
\|\hat{\boldsymbol{v}}(t^k) - \bar{\boldsymbol{v}}\|^2 &\leq 2\left(\|\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t^k)} \odot \boldsymbol{v}(t^k) - \bar{\boldsymbol{v}}\|^2 + \|\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t^k)} \odot \boldsymbol{v}(t^k) - \hat{\boldsymbol{v}}(t^k)\|^2\right) \\
&= 2\|\widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t^k)} \odot \boldsymbol{v}(t^k) - \bar{\boldsymbol{v}}\|^2 + 2\left(2 - 2\left\langle \widehat{\boldsymbol{\beta}^{-\frac{1}{2}}(t^k)} \odot \boldsymbol{v}(t^k), \hat{\boldsymbol{v}}(t^k)\right\rangle\right) \\
&\leq 2\|\widehat{\boldsymbol{\beta}(t^k) \odot \boldsymbol{v}}(t^k) - \bar{\boldsymbol{v}}\|^2 + O\left(\frac{1}{k}\right) \\
&\leq 2\|\widehat{\boldsymbol{\beta}(t^k) \odot \boldsymbol{v}}(t^k) - \widehat{\boldsymbol{\beta}(s^k) \odot \boldsymbol{v}}(s^k)\|^2 + O\left(\frac{1}{k}\right) \\
&\leq \mathcal{O}(1)\frac{1}{k} + O(e^{\frac{1}{k}})\frac{1}{k} \to 0.
\end{aligned}$$

The proof is completed.

□

Therefore, similar to the gradient flow case, we then have the following theorem.

Theorem 13. *Let $\bar{\boldsymbol{w}}$ be any limit point of $\{\hat{\boldsymbol{w}}(t)\}$. Then $\bar{\boldsymbol{w}}$ is along the direction of a KKT point of the following optimization problem.*

$$\begin{aligned}
&\textit{Minimize } \frac{1}{2}\|\boldsymbol{h}_\infty^{-\frac{1}{2}} \odot \boldsymbol{w}\| \\
&\textit{Subject to: } q_i(w) \geq 1.
\end{aligned}$$

E. Proof of Multi-class Classification with Logistic Loss

In this section, we prove the result for multi-class classification with logistic loss mentioned in Remark 1. Concretely, the dataset for this case can be represented as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $y_i \in [C]$ represents the class $\boldsymbol{x}_i$ belongs to. Unlike the binary classification case, neural network $\boldsymbol{\Phi}$ outputs a C-dimension vector as scores for C classes, and we use $\boldsymbol{\Phi}(\boldsymbol{w}, \boldsymbol{x}_i)_j$ as the j-th component of $\boldsymbol{\Phi}(\boldsymbol{w}, \boldsymbol{x}_i)$. The empirical loss can then be represented as

$$\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{N} -\log \frac{e^{\boldsymbol{\Phi}(\boldsymbol{w}, \boldsymbol{x}_i)_{y_i}}}{\sum_{j=1}^{C} e^{\boldsymbol{\Phi}(\boldsymbol{w}, \boldsymbol{x}_i)_j}}. \tag{26}$$

For AdaGrad, RMSProp, and Adam (w/m), limit $\boldsymbol{h}_\infty^A$, $\boldsymbol{h}_\infty^R$, $\boldsymbol{h}_\infty^M$ remains non-zero, and we can then define $\boldsymbol{v}$, $\tilde{\mathcal{L}}$, and $\boldsymbol{\beta}$ the same as Theorems 6 and 7 (we use $\boldsymbol{v}$ and $\tilde{\mathcal{L}}$ to represent all cases), and $\tilde{\boldsymbol{\Phi}}(\boldsymbol{v}, \boldsymbol{x}_i) = \boldsymbol{\Phi}(\boldsymbol{h}_\infty^{\frac{1}{2}} \odot \boldsymbol{v}, \boldsymbol{x}_i)$.

We can then define margins in the multi-class classification similarly as the binary case: surrogate norm and margin are defined exactly the same as the binary case; define $\tilde{q}_i(\boldsymbol{v}) = \tilde{\boldsymbol{\Phi}}(\boldsymbol{v}, \boldsymbol{x}_i)_{y_i} - \max_{j \neq y_i} \tilde{\boldsymbol{\Phi}}(\boldsymbol{v}, \boldsymbol{x}_i)_j$ and normalized margin can be still defined as $\frac{\tilde{q}_{\min}(\boldsymbol{v})}{\|\boldsymbol{v}\|^L}$. The corresponding convergent direction for adaptive gradient flow under multi-class setting can then be characterized by the following theorem:

Theorem 14. *Let $\boldsymbol{v}$ satisfy an adaptive gradient flow $\mathcal{F}$ which satisfies Assumption 1. Let $\bar{\boldsymbol{v}}$ be any limit point of $\{\hat{\boldsymbol{v}}(t)\}_{t=0}^{\infty}$ (where $\hat{\boldsymbol{v}}(t) = \frac{\boldsymbol{v}(t)}{\|\boldsymbol{v}(t)\|}$ is normalized parameter). Then $\bar{\boldsymbol{v}}$ is along the direction of a KKT point of the following L^2 max-margin problem (P):*

$$\min \frac{1}{2}\|\boldsymbol{v}\|^2$$
$$\textit{subject to } \tilde{q}_i(\boldsymbol{v}) \geq 1, \forall i \in [N].$$

Proof of Theorem 14 differs from that of Theorem 2 only by Lemma 18, Lemma 19, and the construction of λ_i in Lemma 21. We show modifications respectively.

First of all, we show normalized margin γ and surrogate margin $\tilde{\gamma}$ converge to the same limit:

Lemma 38. *Let a function $\boldsymbol{v}(t)$ obey an adaptive gradient flow $\mathcal{F}$ which satisfies Assumption 1, with loss $\tilde{\mathcal{L}}$ and component learning rate $\boldsymbol{\beta}(t)$. Then we have $\lim_{t\to\infty} \frac{\rho(t)}{\|\boldsymbol{v}(t)\|} = 1$. Furthermore, if further $\lim_{t\to\infty} \tilde{\mathcal{L}}(t) = \infty$, we have $\lim_{t\to\infty} \frac{\gamma(t)}{\tilde{\gamma}(t)} = 1$.*

Proof. By definition of empirical loss (eq. (26)),

$$\begin{aligned}\tilde{\mathcal{L}}(\boldsymbol{v}) =& \sum_{i=1}^{N} -\log \frac{e^{\tilde{\boldsymbol{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}}{\sum_{j=1}^{C} e^{\tilde{\boldsymbol{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j}} \\ =& \sum_{i=1}^{N} -\log \frac{1}{1+\sum_{j\neq y_i} e^{\tilde{\boldsymbol{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j - \tilde{\boldsymbol{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}}.\end{aligned}$$

Let $\tilde{q}_i'(\boldsymbol{v}) = \log \sum_{j\neq y_i} e^{\tilde{\boldsymbol{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j - \tilde{\boldsymbol{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}$. By exact the same routine of Lemma 18, we have

$$\lim_{t\to\infty} \frac{\gamma(t)}{\frac{\tilde{q}'_{\min}(\boldsymbol{v})}{\rho^L}} = 1. \tag{27}$$

On the other hand,

$$-\tilde{q}_{\min}(\boldsymbol{v}) \leq \log \sum_{j\neq y_i} e^{\tilde{\boldsymbol{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j - \tilde{\boldsymbol{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}} \leq -\tilde{q}_{\min}(\boldsymbol{v}) + \log N,$$

which leads to

$$\lim_{t\to\infty} \frac{\tilde{\gamma}(t)}{\frac{\tilde{q}'_{\min}(\boldsymbol{v})}{\rho^L}} = 1. \tag{28}$$

Combining eqs. (27) and (28), the proof is completed.

□

Secondly, we calculate derivative of surrogate norm ρ under multi-class classification setting.

Lemma 39. *The derivative of ρ^2 is as follows:*

$$\frac{1}{2}\frac{\mathrm{d}\rho(t)^2}{\mathrm{d}t} = L\nu(t) + \left\langle \boldsymbol{v}(t), \boldsymbol{\beta}^{-\frac{1}{2}}(t) \odot \frac{\mathrm{d}\boldsymbol{\beta}^{-\frac{1}{2}}}{\mathrm{d}t}(t) \odot \boldsymbol{v}(t) \right\rangle,$$

where $\nu(t)$ is defined as

$$\nu(t)=\sum_{i=1}^{N}\frac{\sum_{j\neq y_n}e^{\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}(\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j)}{1+\sum_{j\neq y_n}e^{\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}}.$$

Furthermore, we have that $\nu(t)>\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\tilde{\mathcal{L}}(\boldsymbol{v}(t))$.

Proof. We only need to show

$$-\langle\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t)),\boldsymbol{v}(t)\rangle=L\nu(t), \tag{29}$$

and

$$\nu(t)>\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\tilde{\mathcal{L}}(\boldsymbol{v}(t)), \tag{30}$$

while other parts of the proof follows exact the same as Lemma 19.

By chain rule,

$$\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v})=\sum_{i=1}^{N}\frac{\sum_{j\neq y_i}e^{\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}\bar{\partial}(\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i})}{1+\sum_{j\neq y_i}e^{\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}},$$

while by homogeneity of $\tilde{\mathbf{\Phi}}$,

$$\langle\bar{\partial}(\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}),\boldsymbol{v}\rangle=L(\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}),$$

which completes the proof of eq. (29).

As for eq. (30),

$$\begin{aligned}\nu(t)=&\sum_{i=1}^{N}e^{-f(\tilde{q}_i'(\boldsymbol{v}(t)))}f'(\tilde{q}_i'(\boldsymbol{v}(t)))\frac{\sum_{j\neq y_i}e^{\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}(\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j)}{\sum_{j\neq y_i}e^{\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}}\\ \geq&\sum_{i=1}^{N}e^{-f(\tilde{q}_i'(\boldsymbol{v}(t)))}f'(\tilde{q}_i'(\boldsymbol{v}(t)))\tilde{q}_i'(\boldsymbol{v}(t))\\ \geq&\frac{g\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}{g'\left(\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}\right)}\tilde{\mathcal{L}}(\boldsymbol{v}(t)).\end{aligned}$$

The proof is completed.

□

Finally, we provide construction of λ_i similar to Lemma 21. The proof follows the same routine as Lemma 21 and we omit it here.

Lemma 40. *Let $\boldsymbol{v}$ obey adaptive gradient flow $\mathcal{F}$ with empirical loss $\tilde{\mathcal{L}}$ satisfying Assumption 1. Let time t_1 be constructed as Lemma 2. Then, define coefficients in Definition 2 as*

$$\lambda_{i,j}(t)=\tilde{q}_{\min}(\boldsymbol{v}(t))^{1-2/L}\|\boldsymbol{v}(t)\|\cdot\frac{1}{\|\bar{\partial}\tilde{\mathcal{L}}(\boldsymbol{v}(t))\|_2}\frac{1+e^{\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}}{\sum_{j\neq y_i}e^{\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_j-\tilde{\mathbf{\Phi}}(\boldsymbol{v},\boldsymbol{x}_i)_{y_i}}},$$

(where $i\in[N]$, $j\in[C]/\{i\}$). Then, for any time $t\geq t_1$, $\tilde{\boldsymbol{v}}(t)=\tilde{q}_{\min}(\boldsymbol{v}(t))^{-\frac{1}{L}}\boldsymbol{v}(t)$ is an $(\varepsilon(t),\delta(t))$ KKT point of $(\tilde{P})$, where $\varepsilon(t)$, $\delta(t)$ are defined as follows:

$$\begin{aligned}\varepsilon(t)&=\mathcal{O}(1-\cos(\boldsymbol{\theta}(t)))\\ \delta(t)&=\mathcal{O}\left(\frac{1}{\log\frac{1}{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}}\right),\end{aligned}$$

where $\cos(\boldsymbol{\theta}(t))$ *is defined as inner product of* $\hat{\boldsymbol{v}}(t)$ *and* $-\bar{\partial}\widehat{\tilde{\mathcal{L}}(\boldsymbol{v}(t))}$.

F. Experiment Details

In this section, we provide detailed explanation of experiments showed in Section 6 [4]. This section is divided into two parts according to Section 6: in Section F.1, we provide details of structure of neural network we use and hyper-parameters. We also further plot two additional experiments of Adam and SGD to show the influence of momentum; in Section F.3, we show construction of dataset in Section 6.2 and choose of hyper-parameters. We also show how direction of $\boldsymbol{h}_{\infty}^{-\frac{1}{2}}$ influence convergent direction of parameters.

F.1. Experiment on MNIST

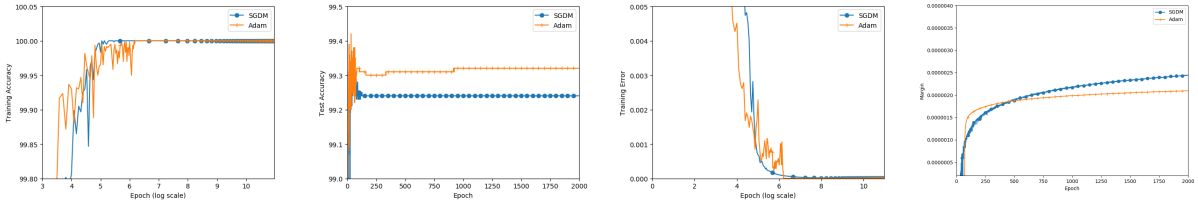
F.1.1. Construction of Neural Network and Choice of Hyper-parameters

We use the 4-layer convolutional neural network adopted by (Madry et al., 2018) as our model to conduct multi-class classification on MNIST (LeCun, 1998). Concretely, this convolutional neural network can be expressed in order as convolutional layer with 32 channel and filter size 5×5, max-pool layer with kernel size 2 and stride 2, convolutional layer with 64 channel and filter size 3×3, max-pool with kernel size 2, fully connected layer with width 1024, and fully connected layer with width 10. In order to guarantee this neural network is homogeneous, we further set bias in all layers to be zero. We use default method in Pytorch to initialize the neural network.

As for hyper-parameters, we set learning rate of AdaGrad to be the default value in Pytorch; while for RMSProp, we set learning rate and decay parameter b as 0.001 and 0.9, which is suggested by (Hinton et al., 2012) and used as a default value in Tensorflow; for Adam, we set the learning rate to 0.0001 as default value in Pytorch, and b to be the same as RMSProp.

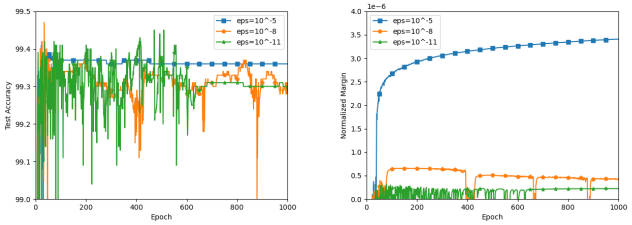
F.1.2. Influence of Momentum

We plot convergent behaviors for SGDm and Adam in this section. Figure 4 shows that adding momentum term will NOT keep normalized margin from lower bounded, which indicates our theory might be extended to gradient based optimization methods with momentum. Specifically, for SGD, we use learning rate 0.1 and momentum parameter 0.9; for Adam, we use the same setting as Adam (w/m) with momentum parameter 0.9.



(a) Training Accuracy (b) Test Accuracy (c) Training Loss (d) Normalized Margin

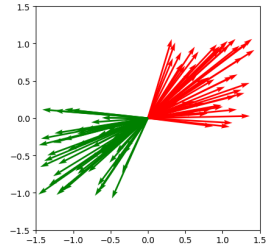
Figure 4. Observation of convergent behavior after adding momentum for SGD and Adam. One can observe that loss still converge to zero, and margin will keep lower bounded.



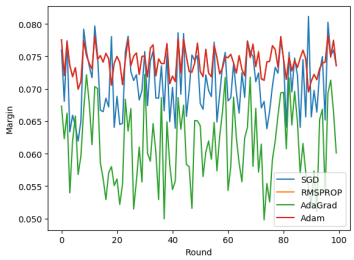
(a) Test Accuracy (b) Normalized margin

Figure 5. Observation of generalization behavior of RMSProp with different ε. One can observe that larger ε leads to larger margin and smaller generalization error.

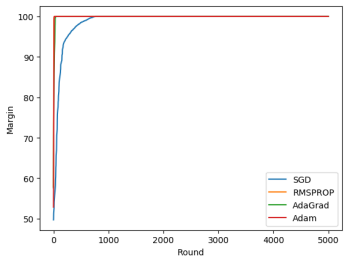
[4] https://github.com/bhwangfy/ICML-2021-Adaptive-Bias



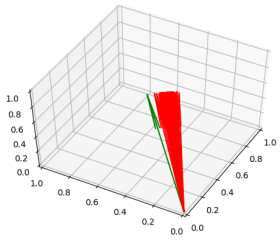
(a) Visualizing Constructed Dataset



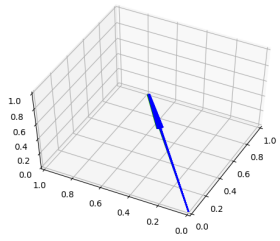
(b) Margin varies across runs



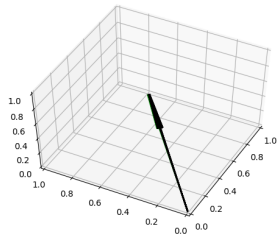
(c) Training accuracy



(d) Direction of $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ in AdaGrad



(e) Direction of $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ in RMSProp



(f) Direction of $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ in Adam

Figure 6. Experiment on two layer MLP. In (a), we visualize constructed dataset as vectors, with green vectors standing for data with label -1, and red stands for data with label 1. In (b), we plot convergent margin of different optimizers across runs. In (c), averaged training accuracy across runs is shown, and one can observe all training accuracy achieves 100%. (d)-(e) respectively picture $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ in AdaGrad (red vectors), RMSProp (blue vectors) and Adam (black vectors) across runs. While direction of $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ in AdaGrad varies across runs, RMSProp and Adam stay the same and coincide with isotropic direction (green vector) $(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$

F.2. Influence of ε

We compare the generalization behaviors of RMSProp with different ε selected in Figure 5. It is observed that as ε decreases, normalized margin gets smaller and the generalization error gets larger, which indicates the importance of ε on the generalization behavior. When ε is completed removed (i.e., is set to 0), the training does not converge. Therefore, we do not include the results for $\varepsilon = 0$ here.

F.3. Experiment on Two Layer MLP

F.3.1. Dataset Construction and Choice of Hyper-parameters

As mentioned in Section 6.2, we use a two layer MLP Φ with leaky ReLU activation σ defined as $\Phi(\boldsymbol{x}, \boldsymbol{w}, v) = v\sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$, where $x \in \mathbb{R}^2, \boldsymbol{w} \in \mathbb{R}^2$ and $v \in \mathbb{R}$ and $\sigma(t)$ is the Leaky ReLU activation function, i.e., $\sigma(t) = t$ for $t \geq 0$ and $\sigma(t) = \frac{t}{2}$ for $t < 0$. We construct binary classification dataset S as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{100}$ as follows:

$$(\boldsymbol{x}_i, y_i) = ((\cos(0.5), \sin 0.5) + \boldsymbol{\varepsilon}_i, 1),\ i \in \{1, 2, \cdots, 50\};$$
$$(\boldsymbol{x}_i, y_i) = ((-\cos(0.5), -\sin 0.5) + \boldsymbol{\varepsilon}_i, -1),\ i \in \{51, 52, \cdots, 100\},$$

where $\boldsymbol{\varepsilon}_i$ $(i = 1, 2, 3, \cdots, 100)$ are random variables sampled uniformly and i.i.d. from $[-0.6, 0.6] \times [-0.6, 0.6]$. We visualize the dataset in Figure 6a.

We then run SGD, AdaGrad, RMSProp (and Adam (w/m)) respectively with learning rates $\eta = 0.1$, while Weight-decay hyper-parameter b is set to be 0.9. For each round, we train the model for 5000 epochs to ensure that training accuracy achieves 100% (see Figure 6c for details); while for each optimizer, we conduct 100 rounds of experiments with random initialization, Convergent directions of square root of inverse conditioners $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ are plotted in Figures 6d, 6e, and 6f. Since $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ occurs in optimization target in (P^A), different direction of $\boldsymbol{h}_\infty^{-\frac{1}{2}}$ may lead to different convergent direction of parameters, which further indicates convergent direction of parameters in AdaGrad can be vulnerable to random initialization.