Characterizing Implicit Bias in Terms of Optimization Geometry

Suriya Gunasekar *TTI Chicago, USA* SURIYA@TTIC.EDU

Jason Lee *USC Los Angeles, USA* JASONLEE@MARSHALL.USC.EDU

Daniel Soudry *Technion, Israel* DANIEL.SOUDRY@GMAIL.COM

Nathan Srebro *TTI Chicago, USA* NATI@TTIC.EDU

Abstract

We study the implicit bias of generic optimization methods, such as mirror descent, natural gradient descent, and steepest descent with respect to different potentials and norms, when optimizing underdetermined linear regression or separable linear classification problems. We explore the question of whether the specific global minimum (among the many possible global minima) reached by an algorithm can be characterized in terms of the potential or norm of the optimization geometry, and independently of hyperparameter choices such as step-size and momentum.

1. Introduction

Implicit bias from the optimization algorithm plays a crucial role in learning deep neural networks as it introduces effective capacity control not directly specified in the objective (Neyshabur et al., 2015b,a; Zhang et al., 2017; Keskar et al., 2016; Wilson et al., 2017; Neyshabur et al., 2017). In overparameterized models where the training objective has many global minima, optimizing using a specific algorithm, such as gradient descent, *implicitly biases* the solutions to some special global minima. The properties of the learned model, including its generalization performance, are thus crucially influenced by the choice of optimization algorithm used. In neural networks especially, characterizing these special global minima for common algorithms such as stochastic gradient descent (SGD) is essential for understanding what the inductive bias of the learned model is and why such large capacity networks often show remarkably good generalization even in the absence of explicit regularization (Zhang et al., 2017) or early stopping (Hoffer et al., 2017).

Implicit bias from optimization depends on the choice of algorithm, and changing the algorithm, or even changing associated hyperparameter can change the implicit bias. For example, Wilson et al. (2017) showed that for some standard deep learning architectures, variants of SGD algorithm with different choices of momentum and adaptive gradient updates (AdaGrad and Adam) exhibit different biases and thus have different generalization performance; Keskar et al. (2016), Hoffer et al. (2017) and Smith (2018) study how the size of the mini-batches used in SGD influences generalization; and Neyshabur et al. (2015a) compare the bias of path-SGD (steepest descent with respect to a scale invariant path-norm) to standard SGD.

It is therefore important to explicitly relate different optimization algorithms to their implicit biases. Can we precisely characterize which global minima different algorithms converge to? How does this depend on the loss function? What other choices including initialization, step-size, momentum, stochasticity, and adaptivity, does the implicit bias depend on? In this paper, we provide answers to some of these questions for simple linear regression and classification models. While neural networks are certainly more complicated than these simple linear models, the results here provide a segue into understanding such biases for more complex models.

For linear models, we already have an understanding of the implicit bias of gradient descent. For underdetermined least squares objective, gradient descent can be shown to converge to the minimum Euclidean norm solution. Recently, Soudry et al. (2017) studied gradient descent for linear logistic regression. The logistic loss is fundamentally different from the squared loss in that the loss function has no attainable global minima. Gradient descent iterates therefore diverge (the norm goes to infinity), but Soudry et al. showed that they diverge in the direction of the hard margin support vector machine solution, and therefore the decision boundary converges to this maximum margin separator.

Can we extend such characterization to other optimization methods that work under different (non-Euclidean) geometries such as mirror descent with respect to some potential, natural gradient descent with respect to a Riemannian metric, and steepest descent with respect to a generic norm? Can we relate the implicit bias to these geometries?

As we shall see, the answer depends on whether the loss function is similar to a squared loss or to a logistic loss. This difference is captured by two family of losses: (a) loss functions that have a unique finite root, like the squared loss and (b) strictly monotone loss functions where the infimum is unattainable, like the logistic loss. For losses with a unique finite root, we study the *limit point* of the optimization iterates, $w_\infty = \lim_{t\to\infty} w_{(t)}$. For monotone losses, we study the *limit direction* $\bar{w}_\infty = \lim_{t\to\infty} \frac{w_{(t)}}{\|w_{(t)}\|}$.

In Section 2 we study linear models with loss functions that have unique finite roots. We obtain a robust characterization of the limit point for mirror descent, and discuss how it is independent of step-size and momentum. For natural gradient descent, we show that the step-size does play a role, but get a characterization for infinitesimal step-size. For steepest descent, we show that not only does step-size affects the limit point, but even with infinitesimal step-size, the expected characterization does not hold. The situation is fundamentally different for strictly monotone losses such as the logistic loss (Section 3) where we do get a precise characterization of the limit direction for generic steepest descent. We also study the adaptive gradient descent method (AdaGrad) Duchi et al. (2011) (Section 3.3) and optimization over matrix factorization (Section 4). Recent studies considered the bias of such methods for least squares problems (Wilson et al., 2017; Gunasekar et al., 2017), and here we study these algorithms for monotone loss functions, obtaining a more robust characterization for matrix factorization problems, while concluding that the implicit bias of AdaGrad depends on initial conditions including step-size even for strict monotone losses.

2. Losses with a Unique Finite Root

We first consider learning linear models using losses with a unique finite root, such as the squared loss, where the loss $\ell(\widehat{y}, y)$ between a prediction $\widehat{y}$ and label y is minimized at a unique and finite value of $\widehat{y}$. We assume without loss of generality, that $\min_{\widehat{y}} \ell(\widehat{y}, y) = 0$ and the unique minimizer is $\widehat{y} = y$.

Property 1 (Losses with a unique finite root). *For any y, a sequence $\{\widehat{y}_t\}_{t=1}^\infty$ minimizes $\ell(.,y)$, i.e., $\ell(\widehat{y}_t, y) \overset{t\to\infty}{\longrightarrow} \inf_{\widehat{y}} \ell(\widehat{y}, y) = 0$ if and only if $\widehat{y}_t \overset{t\to\infty}{\longrightarrow} y$.*

Denote the training dataset $\{(x_n, y_n) : n = 1, 2, \ldots, N\}$ with features $x_n \in \mathbb{R}^d$ and labels $y_n \in \mathbb{R}$. The empirical loss (or risk) minimizer of a linear model $f(x) = \langle w, x\rangle$ with parameters $w \in \mathbb{R}^d$ is given by,

$$\min_w \mathcal{L}(w) := \sum_{n=1}^{N} \ell(\langle w, x_n\rangle, y_n). \tag{1}$$

We are particularly interested in the case where $N < d$ and the observations are realizable, i.e., $\min_w \mathcal{L}(w) = 0$. Under these conditions, the optimization problem in eq. (1) is underdetermined and has multiple global minima denoted by $\mathcal{G} = \{w : \mathcal{L}(w) = 0\} = \{w : \forall n,\ \langle w, x_n\rangle = y_n\}$. Note that the set of global minima $\mathcal{G}$ is the same for any loss ℓ with unique finite root (Property 1), including, e.g., the Huber loss, the truncated squared loss. Which specific global minima $w \in \mathcal{G}$ do different optimization algorithms reach when minimizing the empirical loss objective $\mathcal{L}(w)$?

2.1 Gradient descent

Consider gradient descent updates for minimizing $\mathcal{L}(w)$ with step-size sequence $\{\eta_t\}_t$ and initialization $w_{(0)}$,

$$w_{(t+1)} = w_{(t)} - \eta_t \nabla \mathcal{L}(w_{(t)}).$$

If $w_{(t)}$ minimizes the empirical loss in eq. (1), then the iterates converge to the unique global minimum that is closest to initialization $w_{(0)}$ in ℓ_2 distance, i.e., $w_{(t)} \to \operatorname{argmin}_{w \in \mathcal{G}} \|w - w_{(0)}\|_2$. This can be easily seen as for any w, the gradients $\nabla \mathcal{L}(w) = \sum_n \ell'(\langle w, x_n \rangle, y_n) x_n$ are always constrained to the fixed subspace spanned by the data $\{x_n\}_n$, and thus the iterates $w_{(t)}$ are confined to the low dimensional affine manifold $w_{(0)} + \text{span}(\{x_n\}_n)$. Within this low dimensional manifold, there is a unique global minimizer w that satisfies the linear constraints in $\mathcal{G} = \{w : \langle w, x_n \rangle = y_n, \forall n \in [N]\}$.

The same argument also extends for updates with instance-wise stochastic gradients, where we use a stochastic estimate $\widetilde{\nabla} \mathcal{L}(w_{(t)})$ of the full gradient $\nabla \mathcal{L}(w_{(t)})$ computed from a random subset of instances $S_t \subseteq [N]$,

$$\widetilde{\nabla} \mathcal{L}(w_{(t)}) = \sum\nolimits_{n \in S_t \subset [n]} \nabla_w \ell(\langle w_{(t)}, x_{n_t} \rangle, y_{n_t}). \tag{2}$$

Moreover, when initialized with $w_{(0)} = 0$, the implicit bias characterization also extends to the following generic momentum and acceleration based updates,

$$w_{(t+1)} = w_{(t)} + \beta_t \Delta w_{(t-1)} - \eta_t \nabla \mathcal{L}\big(w_{(t)} + \gamma_t \Delta w_{(t-1)}\big), \tag{3}$$

where $\Delta w_{(t-1)} = w_{(t)} - w_{(t-1)}$. This includes Nesterov's acceleration ($\beta_t = \gamma_t$) (Nesterov, 1983) and Polyak's heavy ball momentum ($\gamma_t = 0$) (Polyak, 1964).

For losses with a unique finite root, the implicit bias of gradient descent therefore depends only on the initialization and not on the step-size or momentum or mini-batch size. Can we get such succinct characterization for other optimization algorithms? That is, characterize the bias in terms of the optimization geometry and initialization, but independent of choices of step-sizes, momentum, and stochasticity.

2.2 Mirror descent

Mirror descent (MD) (Beck & Teboulle, 2003; Nemirovskii & Yudin, 1983) was introduced as a generalization of gradient descent for optimization over geometries beyond the Euclidean geometry of gradient descent. In particular, mirror descent updates are defined for any strongly convex and differentiable potential ψ as

$$w_{(t+1)} = \operatorname*{argmin}_{w \in \mathcal{W}} \eta_t \left\langle w, \nabla \mathcal{L}(w_{(t)}) \right\rangle + D_\psi(w, w_{(t)}), \tag{4}$$

where $D_\psi(w, w') = \psi(w) - \psi(w') - \langle \nabla \psi(w'), w - w' \rangle$ is the *Bregman divergence* (Bregman, 1967) w.r.t. ψ, and $\mathcal{W}$ is some constraint set for parameters w.

We first look at unconstrained optimization where $\mathcal{W} = \mathbb{R}^d$ and the update in eq. (4) is equivalent to

$$\nabla \psi(w_{(t+1)}) = \nabla \psi(w_{(t)}) - \eta_t \nabla \mathcal{L}(w_{(t)}). \tag{5}$$

For a strongly convex potential ψ, $\nabla \psi$ is called the link function and is invertible. Hence, the above updates are uniquely defined. Also, w and $\nabla \psi(w)$ are referred as *primal* and *dual* variables, respectively.

Examples of potentials ψ for mirror descent include the squared ℓ_2 norm $\psi(w) = 1/2\|w\|_2^2$, which leads to gradient descent; the entropy potential $\psi(w) = \sum_i w[i] \log w[i] - w[i]$; the spectral entropy for matrix valued w, where $\psi(w)$ is the entropy potential on the singular values of w; general quadratic potentials $\psi(w) = 1/2\|w\|_D^2 = 1/2\, w^\top D w$ for any positive definite matrix D; and the squared ℓ_p norms for $p \in (1, 2]$.

From eq. (5), we see that rather than the primal iterates $w_{(t)}$, it is the dual iterates $\nabla \psi(w_{(t)})$ that are constrained to the low dimensional data manifold $\nabla \psi(w_{(0)}) + \text{span}(\{x_n\}_{n \in [N]})$. The arguments for gradient descent can now be generalized to get the following result.

Theorem 1. *For any loss ℓ with a unique finite root (Property 1), any realizable dataset $\{x_n, y_n\}_{n=1}^N$, and any strongly convex potential ψ, consider the mirror descent iterates $w_{(t)}$ from eq.* (5) *for minimizing the empirical loss $\mathcal{L}(w)$ in eq.* (1)*. For all initializations $w_{(0)}$, if the step-size sequence $\{\eta_t\}_t$ is chosen such that the limit point of the iterates $w_\infty = \lim_{t\to\infty} w_{(t)}$ is a global minimizer of $\mathcal{L}$, i.e., $\mathcal{L}(w_\infty) = 0$, then w_∞ is given by*

$$w_\infty = \underset{w:\forall n, \langle w, x_n\rangle = y_n}{\operatorname{argmin}} D_\psi(w, w_{(0)}). \tag{6}$$

In particular, if we start at $w_{(0)} = \operatorname{argmin}_w \psi(w)$ (so that $\nabla\psi(w_{(0)}) = 0$), then we get to $w_\infty = \operatorname{argmin}_{w\in\mathcal{G}} \psi(w)$, where recall that $\mathcal{G} = \{w : \forall n, \langle w, x_n\rangle = y_n\}$ is the set of global minima for $\mathcal{L}(w)$.

The analysis of Theorem 1 can also be extended for special cases of constrained mirror descent (eq. (4)) when $\mathcal{L}(w)$ is minimized over realizable affine equality constraints.

Theorem 1a. *Under the conditions of Theorem 1, consider constrained mirror descent updates $w_{(t)}$ from eq.* (4) *with realizable affine equality constraints, that is $\mathcal{W} = \{w : Gw = h\}$ for some $G \in \mathbb{R}^{d'\times d}$ and $h \in \mathbb{R}^{d'}$ and additionally, $\exists w \in \mathcal{W}$ with $\mathcal{L}(w) = 0$. For all initializations $w_{(0)}$, if the step-size sequence $\{\eta_t\}_t$ is chosen to asymptotically minimize $\mathcal{L}$, i.e., $\mathcal{L}(w_\infty) = 0$, then $w_\infty = \operatorname{argmin}_{w\in\mathcal{G}\cap\mathcal{W}} D_\psi(w, w_{(0)})$.*

For example, in exponentiated gradient descent (Kivinen & Warmuth, 1997), which is mirror descent w.r.t $\psi(w) = \sum_i w[i]\log w[i] - w[i]$, under the explicit simplex constraint $\mathcal{W} = \{w : \sum_i w[i] = 1\}$, Theorem 1a shows that using uniform initialization $w_{(0)} = \frac{1}{d}\mathbf{1}$, mirror descent will return the the maximum entropy solution $w_\infty = \operatorname{argmin}_{w\in\mathcal{G}\cap\mathcal{W}} \sum_i w[i]\log w[i]$.

Let us now consider momentum for mirror descent. There are two possible generalizations of the gradient descent momentum in eq. (3): adding momentum either to primal variables $w_{(t)}$, or to dual variables $\nabla\psi(w_{(t)})$,

$$\text{Dual momentum:} \qquad \nabla\psi(w_{(t+1)}) = \nabla\psi(w_{(t)}) + \beta_t \Delta z_{(t-1)} - \eta_t \nabla\mathcal{L}\left(w_{(t)} + \gamma_t \Delta w_{(t-1)}\right) \tag{7}$$

$$\text{Primal momentum:} \qquad \nabla\psi(w_{(t+1)}) = \nabla\psi\big(w_{(t)} + \beta_t \Delta w_{(t-1)}\big) - \eta_t \nabla\mathcal{L}\left(w_{(t)} + \gamma_t \Delta w_{(t-1)}\right) \tag{8}$$

where $\Delta z_{(-1)} = \Delta w_{(-1)} = 0$, and for $t \geq 1$, $\Delta z_{(t-1)} = \nabla\psi(w_{(t)}) - \nabla\psi(w_{(t-1)})$ and $\Delta w_{(t-1)} = w_{(t)} - w_{(t-1)}$ are the momentum terms in the primal and dual space, respectively; and $\{\beta_t \geq 0, \gamma_t \geq 0\}_t$ are the momentum parameters.

If we initialize at $w_{(0)} = \operatorname{argmin}_w \psi(w)$, then even with dual momentum $\nabla\psi(w_{(t)})$ continues to remain in the data manifold. This leads to the following extension of Theorem 1.

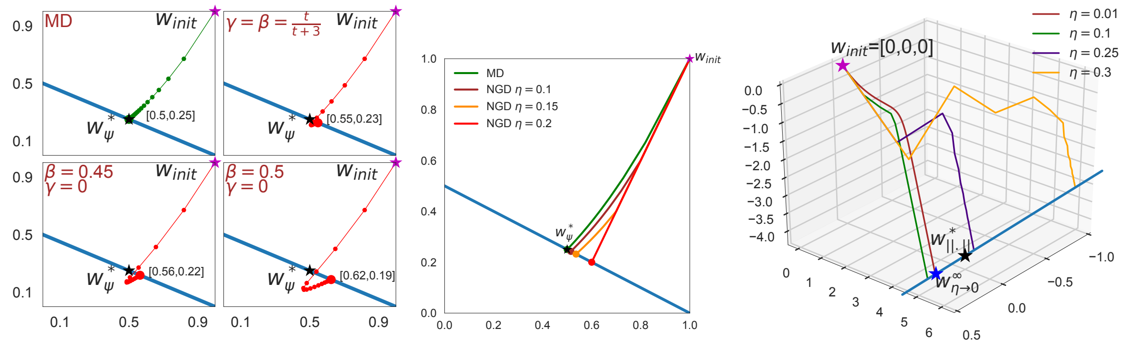
Theorem 1b. *Under the conditions in Theorem 1, if initialized at $w_{(0)} = \operatorname{argmin}_w \psi(w)$, then the mirror descent updates with dual momentum also converge to* (6)*, i.e., for all $\{\eta_t\}_t, \{\beta_t\}_t, \{\gamma_t\}_t$, if $w_{(t)}$ from eq.* (7) *converges to $w_\infty \in \mathcal{G}$, then $w_\infty = \operatorname{argmin}_{w\in\mathcal{G}} \psi(w)$.*

Remark 1. *Following the same arguments, we can show that Theorem 1–1b also hold when instancewise stochastic gradients defined in eq.* (2) *are used in place of $\nabla\mathcal{L}(w_{(t)})$.*

Let us now look at primal momentum. For general potentials ψ, the dual iterates $\nabla\psi(w_{(t)})$ from the primal momentum can fall off the data manifold and the additional components influence the final solution. Thus, the specific global minimum that the iterates $w_{(t)}$ converge to will depend on the values of momentum parameters $\{\beta_t, \gamma_t\}_t$ and step-sizes $\{\eta_t\}_t$ as demonstrated in the following example.

Example 2. *Consider optimizing $\mathcal{L}(w)$ with dataset $\{(x_1 = [1, 2], y_1 = 1)\}$ and squared loss $\ell(u, y) = (u - y)^2$ using primal momentum updates from eq.* (8) *for MD w.r.t. the entropy potential $\psi(w) = \sum_i w[i]\log w[i] - w[i]$. For initialization $w_{(0)} = \operatorname{argmin}_w \psi(w)$, Figure 1a shows how different choices of momentum $\{\beta_t, \gamma_t\}$ change the limit point w_∞. Additionally, we show the following:*

Proposition 2a. *In Example 2, consider the case where primal momentum is used only in the first step, but $\gamma_t = 0$ and $\beta_t = 0$ for all $t \geq 2$. For any $\beta_1 > 0$, there exists $\{\eta_t\}_t$, such that $w_{(t)}$ from* (8) *converges to a global minimum, but not to $\operatorname{argmin}_{w\in\mathcal{G}} \psi(w)$.*



(a) Mirror descent primal momentum (Example 2)

(b) Natural gradient descent (Example 3)

(c) Steepest descent w.r.t $\|.\|_{4/3}$ (Example 4)

Figure 1: Dependence of implicit bias on step-size and momentum: In (a)–(c), the blue line denotes the set $\mathcal{G}$ of global minima for the respective examples. In (a) and (b), ψ is the entropy potential and all algorithms are initialized with $w_{(0)} = [1, 1]$ so that $\psi(w_{(0)}) = \operatorname{argmin}_w \psi(w)$. $w^*_\psi = \operatorname{argmin}_{\psi \in \mathcal{G}} \psi(w)$ denotes the minimum potential global minima we expect to converge to. (a) **Mirror descent with primal momentum (Example 2):** the global minimum that eq. (8) converges to depends on the momentum parameters—the sub-plots contain the trajectories of eq. (8) for different choices of $\beta_t = \beta$ and $\gamma_t = \gamma$. (b) **Natural gradient descent (Example 3):** for different step-sizes $\eta_t = \eta$, eq. (9) converges to different global minima. Here, η was chosen to be small enough to ensure $w_{(t)} \in \operatorname{dom}(\psi)$. (c) **Steepest descent w.r.t** $\|.\|_{4/3}$ **(Example 4):** the global minimum to which eq. (11) converges to depends on η. Here $w_{(0)} = [0, 0, 0]$, $w^*_{\|.\|} = \operatorname{argmin}_{\psi \in \mathcal{G}} \|w\|_{4/3}$ denotes the minimum norm global minimum, and $w^\infty_{\eta\to 0}$ denotes the solution of infinitesimal SD with $\eta \to 0$. Note that even as $\eta \to 0$, the expected characterization does not hold, i.e., $w^\infty_{\eta\to 0} \neq w^*_{\|.\|}$.

2.3 Natural gradient descent

Natural gradient descent (NGD) was introduced by Amari (1998) as a modification of gradient descent, wherein the updates are chosen to be the steepest descent direction w.r.t a Riemannian metric tensor H that maps w to a positive definite local metric $H(w)$. The updates are given by,

$$w_{(t+1)} = w_{(t)} - \eta_t H\big(w_{(t)}\big)^{-1} \nabla \mathcal{L}(w_{(t)}). \tag{9}$$

In many instances, the metric tensor H is specified by the Hessian $\nabla^2\psi$ of a strongly convex potential ψ. For example, when the metric over the Riemannian manifold is the KL divergence between distributions P_w and $P_{w'}$ parameterized by w, the metric tensor is given by $H(w) = \nabla^2\psi(P_w)$, where the potential ψ is the entropy potential over P_w.

Connection to mirror descent When $H(w) = \nabla\psi^2(w)$ for a strongly convex potential ψ, as the step-size η goes to zero, the iterates $w_{(t)}$ from natural gradient descent in eq. (9) and mirror descent w.r.t ψ in eq. (4) converge to each other, and the common dynamics in the limit is given by,

$$\frac{\mathrm{d}\nabla\psi(w_{(t)})}{\mathrm{d}t} = -\nabla\mathcal{L}(w_{(t)}) \implies \frac{\mathrm{d}w_{(t)}}{\mathrm{d}t} = -\nabla^2\psi(w_{(t)})^{-1}\nabla\mathcal{L}(w_{(t)}). \tag{10}$$

Thus, as the step-sizes are made infinitesimal, the limit point of natural gradient descent $w_\infty = \lim_{t\to\infty} w_{(t)}$ is also the limit point of mirror descent and hence will be biased towards solutions with minimum divergence to the initialization, i.e., as $\eta \to 0$, $w_\infty = \operatorname{argmin}_{w\in\mathcal{G}} D_\psi(w, w_{(0)})$.

For general step-sizes $\{\eta_t\}$, if the potential ψ is quadratic, $\psi(w) = {}^1\!/_2\|w\|^2_D$ for some positive definite D, we get linear link functions $\nabla\psi(w) = Dw$ and constant metric tensors $\nabla^2\psi(w) = H(w) = D$, and the natural gradient descent updates (9) are the same as the mirror descent (5). Otherwise the updates in eq. (9) is only an approximation of the mirror descent update $\nabla\psi^{-1}(\nabla\psi(w_{(t)}) - \eta_t\nabla\mathcal{L}(w_{(t)}))$.

For natural gradient descent with finite step-size and non-quadratic potentials ψ, the characterization in eq. (6) generally does not hold. We can see this as for any initialization $w_{(0)}$, a finite $\eta_1 > 0$ will lead to $w_{(1)}$ for which the dual variable $\nabla\psi(w_{(1)})$ is no longer in the data manifold $\text{span}(\{x_n\}) + \nabla\psi(w_{(0)})$, and hence will converge to a different global minimum dependent on the step-sizes $\{\eta_t\}_t$.

Example 3. *Consider optimizing $\mathcal{L}(w)$ with squared loss over dataset $\{(x_1 = [1,2], y_1 = 1)\}$ using the natural gradient descent w.r.t. the metric tensor given by $H(w) = \nabla^2\psi(w)$, where $\psi(w) = \sum_i w[i]\log w[i] - w[i]$, and initialization $w_{(0)} = [1,1]$. Figure 1b shows that NGD with different step-sizes η converges to different global minima. For a simple analytical example: take one finite step $\eta_1 > 0$ and then follow the continuous time path in eq.* (10)*.*

Proposition 3a. *For almost all $\eta_1 > 0$, $\lim_{t\to\infty} w_{(t)} = \operatorname{argmin}_{w\in\mathcal{G}} D_\psi(w, w_{(1)}) \neq \operatorname{argmin}_{w\in\mathcal{G}} D_\psi(w, w_{(0)})$.*

2.4 Steepest Descent

Gradient descent is also a special case of steepest descent (SD) w.r.t a generic norm $\|.\|$ (Boyd & Vandenberghe, 2004) with updates given by,

$$w_{(t+1)} = w_{(t)} + \eta_t \Delta w_{(t)}, \text{ where } \Delta w_{(t)} = \operatorname*{argmin}_v \left\langle \nabla\mathcal{L}(w_{(t)}), v\right\rangle + \frac{1}{2}\|v\|^2. \tag{11}$$

The optimality of $\Delta w_{(t)}$ in eq. (11) requires $-\nabla\mathcal{L}(w_{(t)}) \in \partial\|\Delta w_{(t)}\|^2$, which is equivalent to,

$$\langle \Delta w_{(t)}, -\nabla\mathcal{L}(w_{(t)})\rangle = \|\Delta w_{(t)}\|^2 = \|\nabla\mathcal{L}(w_{(t)})\|_\star^2. \tag{12}$$

Examples of steepest descent include gradient descent, which is steepest descent w.r.t ℓ_2 norm and coordinate descent, which is steepest descent w.r.t ℓ_1 norm. In general, the update $\Delta w_{(t)}$ in eq. (11) is not uniquely defined and there could be multiple direction $\Delta w_{(t)}$ that minimize eq. (11). In such cases, any minimizer of eq. (11) is a valid steepest descent update and satisfies eq. (12).

Generalizing gradient descent, we might expect the limit point w_∞ of steepest descent w.r.t an arbitrary norm $\|.\|$ to be the solution closest to initialization in corresponding norm, $\operatorname{argmin}_{w\in\mathcal{G}} \|w - w_{(0)}\|$. This is indeed the case for quadratic norms $\|v\|_D = \sqrt{v^\top Dv}$ when eq. 11 is equivalent to mirror descent with $\psi(w) = 1/2\|w\|_D^2$. Unfortunately, this does not hold for general norms.

Example 4. *Consider minimizing $\mathcal{L}(w)$ with dataset $\{(x_1 = [1,1,1], y_1 = 1), (x_1 = [1,2,0], y_1 = 10)\}$ and loss $\ell(u,y) = (u-y)^2$ using steepest descent updates w.r.t. the $\ell_{4/3}$ norm. The empirical results for this problem in Figure 1c clearly show that even for ℓ_p norms where the $\|.\|_p^2$ is smooth and strongly convex, the corresponding steepest descent converges to a global minimum that depends on the step-size. Further, even in the continuous step-size limit of $\eta \to 0$, $w_{(t)}$ does not converge to $\operatorname{argmin}_{w\in\mathcal{G}} \|w - w_{(0)}\|$.*

Coordinate descent Steepest descent w.r.t. the ℓ_1 norm is called the coordinate descent, with updates:

$$\Delta w_{(t+1)} \in \text{conv}\left\{-\eta_t \frac{\partial\mathcal{L}(w)}{\partial w[j_t]} e_{j_t} : j_t = \operatorname*{argmax}_j \left|\frac{\partial\mathcal{L}(w)}{\partial w[j]}\right|\right\},$$

where $\text{conv}(S)$ denotes the convex hull of the set S, and $\{e_j\}$ are the standard basis, i.e., when multiple partial derivatives are maximal, we can choose any convex combination of the maximizing coordinates, leading to many possible coordinate descent optimization paths.

The connection between optimization paths of coordinate descent and the ℓ_1 *regularization path* given by, $\widehat{w}(\lambda) = \operatorname{argmin}_w \mathcal{L}(w) + \lambda\|w\|_1$, has been studied by Efron et al. (2004). The specific coordinate descent path where updates are along the average of all optimal coordinates and the step-sizes are infinitesimal is equivalent to forward stage-wise selection, a.k.a. ϵ-boosting (Friedman, 2001). When the ℓ_1 regularization path $\widehat{w}(\lambda)$ is monotone in each of the coordinates, it is identical to this stage-wise selection path, i.e., to a coordinate descent optimization path (and also to the related LARS path) (Efron et al., 2004). In this case, at the limit of $\lambda \to 0$ and $t \to \infty$, the optimization and regularization paths, both converge to the minimum ℓ_1 norm solution. However, when the regularization path $\widehat{w}(\lambda)$ is not monotone, which can and does happen, the optimization and regularization paths diverge, and forward stage-wise selection can converge to solutions with sub-optimal ℓ_1 norm. This matches our understanding that steepest descent w.r.t. a norm $\|.\|$, in this case the ℓ_1 norm might converge to a solution that is *not* always the minimum $\|.\|$ norm solution.

2.5 Summary for losses with a unique finite root

For losses with a unique finite root, we characterized the implicit bias of generic mirror descent algorithm in terms of the potential function and initialization. This characterization extends for momentum in the dual space as well as to natural gradient descent in the limit of infinitesimal step-size. We also saw that the characterization breaks for mirror descent with primal momentum and natural gradient descent with finite step-sizes. Moreover, for steepest descent with general norms, we were unable to get a useful characterization even in the infinitesimal step size limit. In the following section, we will see that for strictly monotone losses, we *can* get a characterization also for steepest descent.

3. Strictly Monotone Losses

We now turn to strictly monotone loss functions ℓ where the behavior of the implicit bias is fundamentally different, and as are the situations when the implicit bias can be characterized. Such losses are common in classification problems where $y = \{-1, 1\}$ and $\ell(f(x), y)$ is typically a continuous surrogate of the 0-1 loss. Examples of such losses include logistic loss, exponential loss, and probit loss.

Property 2 (Strict monotone losses). *$\ell(\widehat{y}, y)$ is bounded from below, and $\forall y$, $\ell(\widehat{y}, y)$ is strictly monotonically decreasing in $\widehat{y}$. Without loss of generality, $\forall y$, $\inf_{\widehat{y}} \ell(\widehat{y}, y) = 0$ and $\ell(\widehat{y}, y) \overset{\widehat{y}y \to \infty}{\longrightarrow} 0$.*

We look at classification models that fit the training data $\{x_n, y_n\}_n$ with linear decision boundaries $f(x) = \langle w, x \rangle$ with decision rule given by $\widehat{y}(x) = \text{sign}(f(x))$. In many instances of the proofs, we also assume without loss of generality that $y_n = 1$ for all n, since for linear models, the sign of y_n can equivalently be absorbed into x_n.

We again look at unregularized empirical risk minimization objective of the form in eq. (1), but now with strictly monotone losses. When the training data $\{x_n, y_n\}_n$ is not linearly separable, the empirical objective $\mathcal{L}(w)$ can have a finite global minimum. However, if the dataset is linearly separable, i.e., $\exists w : \forall n, y_n \langle w, x_n \rangle > 0$, the empirical loss $\mathcal{L}(w)$ is again ill-posed, and moreover $\mathcal{L}(w)$ does not have any finite minimizer, i.e, $\mathcal{L}(w) \to 0$ only as $\|w\| \to \infty$. Thus, for any sequence $\{w_{(t)}\}_{t=0}^{\infty}$, if $\mathcal{L}(w_{(t)}) \to 0$, then $w_{(t)}$ necessarily diverges to infinity rather than converge, and hence we cannot talk about $\lim_{t\to\infty} w_{(t)}$. Instead, we look at the limit direction $\bar{w}_\infty = \lim\limits_{t\to\infty} \frac{w_{(t)}}{\|w_{(t)}\|}$ whenever the limit exists. We refer to existence of this limit as convergence in direction. Note that, the limit direction fully specifies the decision rule of the classifier that we care about.

We focus on the exponential loss $\ell(u, y) = \exp(-uy)$. However, our results can be extended to loss functions with tight exponential tails, including logistic and sigmoid losses, along the lines of Soudry et al. (2017) and Telgarsky (2013).

3.1 Gradient descent

Soudry et al. (2017) showed that for almost all linearly separable datasets, gradient descent with *any initialization and any bounded step-size* converges in direction to maximum margin separator with unit ℓ_2 norm, i.e., the hard margin support vector machine classifier,

$$\bar{w}_\infty = \lim_{t\to\infty} \frac{w_{(t)}}{\|w_{(t)}\|_2} = w^*_{\|.\|_2} := \underset{\|w_{(t)}\|_2 \leq 1}{\operatorname{argmax}} \min_n y_n \langle w, x_n \rangle .$$

This characterization of the implicit bias is independent of both the step-size as well as the initialization. We already see a fundamentally difference from the implicit bias of gradient descent for losses with a unique finite root (Section 2.1) where the characterization depended on the initialization.

Can we similarly characterize the implicit bias of different algorithms establishing $w_{(t)}$ converges in direction and calculating $\bar{w}_\infty$? Can we do this even when we *could not* characterize the limit point $w_\infty = \lim_{t\to\infty} w_{(t)}$ for losses with unique finite roots? As we will see in the following section, we can indeed answer these questions for steepest descent w.r.t arbitrary norms.

3.2 Steepest Descent

Recall that for squared loss, the limit point of steepest descent depends on the step-size, and we were unable obtain a useful characterization even for infinitesimal step-size and zero initialization. In contrast, for exponential loss, the following theorem provides a crisp characterization of the limit direction of steepest descent as a maximum margin solution, independent of step-size (as long as it is small enough) and initialization. Let $\|.\|_\star$ denote the dual norm of $\|.\|$.

Theorem 5. *For any separable dataset $\{x_n, y_n\}_{n=1}^N$ and any norm $\|\cdot\|$, consider the steepest descent updates from eq.* (12) *for minimizing $\mathcal{L}(w)$ in eq.* (1) *with the exponential loss $\ell(u, y) = \exp(-uy)$. For all initializations $w_{(0)}$, and all bounded step-sizes satisfying $\eta_t \leq \max\{\eta_+, \frac{1}{B^2 \mathcal{L}(w_{(t)})}\}$, where $B := \max_n \|x_n\|_\star$ and $\eta_+ < \infty$ is any finite upper bound, the iterates $w_{(t)}$ satisfy the following,*

$$\lim_{t\to\infty} \min_n \frac{y_n \left\langle w_{(t)}, x_n \right\rangle}{\|w_{(t)}\|} = \max_{w:\|w\|\leq 1} \min_n y_n \langle w, x_n \rangle.$$

In particular, if there is a unique maximum-$\|.\|$ margin solution $w^\star_{\|.\|} = \operatorname{argmax}_{w:\|w\|\leq 1} \min_n y_n \langle w, x_n \rangle$, then the limit direction is given by $\bar{w}_\infty = \lim\limits_{t\to\infty} \frac{w_{(t)}}{\|w_{(t)}\|} = w^\star_{\|.\|}$.

A special case of Theorem 5 is for steepest descent w.r.t. the ℓ_1 norm, which as we already saw corresponds to coordinate descent. More specifically, coordinate descent on the exponential loss can be thought of as an alternative presentation of AdaBoost (Schapire & Freund, 2012), where each coordinate represents the output of one "weak learner". Indeed, initially mysterious generalization properties of boosting have been understood in terms of implicit ℓ_1 regularization (Schapire & Freund, 2012), and later on AdaBoost with small enough step-size was shown to converge in direction precisely to the maximum ℓ_1 margin solution (Zhang et al., 2005; Shalev-Shwartz & Singer, 2010; Telgarsky, 2013), just as guaranteed by Theorem 5. In fact, Telgarsky (2013) generalized the result to a richer variety of exponential tailed loss functions including logistic loss, and a broad class of non-constant step-size rules. Interestingly, coordinate descent with exact line search can result in infinite step-sizes, leading the iterates to converge in a different direction that is not a max-ℓ_1-margin direction (Rudin et al., 2004), hence the maximum step-size bound in Theorem 5.

Theorem 5 is a generalization of the result of Telgarsky to steepest descent with respect to other norms, and our proof follows the same strategy as Telgarsky. We first prove a generalization of the duality result of Shalev-Shwartz & Singer (2010): if there is a unit norm linear separator that achieves margin γ, then $\|\nabla\mathcal{L}(w)\|_\star \geq \gamma\mathcal{L}(w)$ for all w. By using this lower bound on the dual norm of the gradient, we are able to show that the loss decreases faster than the increase in the norm of the iterates, establishing convergence in a margin maximizing direction.

In relating the optimization path to the regularization path, it is also relevant to relate Theorem 5 to the result by Rosset et al. (2004) that for monotone loss functions and ℓ_p norms, the ℓ_p regularization path $\widehat{w}(c) = \operatorname{argmin}_{w:\|w\|_p\leq c}\mathcal{L}(w_{(t)})$ also converges in direction to the maximum margin separator, i.e., $\lim_{c\to\infty}\widehat{w}(c) = w^\star_{\|\cdot\|_p}$. Although the optimization path and regularization path are not the same, they both converge to the same max-margin separator in the limits of $c\to\infty$ and $t\to\infty$, for the regularization path and steepest descent optimization path, respectively.

3.3 Adaptive Gradient Descent (AdaGrad)

Adaptive gradient methods, such as AdaGrad (Duchi et al., 2011) or Adam (Kingma & Adam, 2015) are very popular for neural network training. We now look at the implicit bias of the basic (diagonal) AdaGrad.

$$w_{(t+1)} = w_{(t)} - \eta\mathbf{G}_{(t)}^{-1/2}\nabla\mathcal{L}\left(w_{(t)}\right), \tag{13}$$

where $\mathbf{G}_{(t)}\in\mathbb{R}^{d\times d}$ is a diagonal matrix such that,

$$\forall i:\ \mathbf{G}_{(t)}[i,i] = \sum_{u=0}^{t}\left(\nabla\mathcal{L}\left(w_{(u)}\right)[i]\right)^2 . \tag{14}$$

AdaGrad updates described above correspond to a pre-conditioned gradient descent, where the pre-conditioning matrix $\mathbf{G}_{(t)}$ adapts across iterations. It was observed by Wilson et al. (2017) that for neural networks with squared loss, adaptive methods tend to degrade generalization performance in comparison to non-adaptive methods (e.g., SGD with momentum), even when both methods are used to train the network until convergence to a global minimum of training loss. This suggests that adaptivity does indeed affect the implicit bias. For squared loss, by inspection the updates in eq. (13), we do not expect to get a characterization of the limit point w_∞ that is independent of the step-sizes.

However, we might hope that, like for steepest descent, the situation might be different for strictly monotone losses, where the asymptotic behavior could potentially nullify the initial conditions. Examining the updates in eq. (13), we can see that the robustness to initialization and initial updates depend on whether the matrices $\mathbf{G}_{(t)}$ diverge or converge: if $\mathbf{G}_{(t)}$ diverges, then we expect the asymptotic effects to dominate, but if it is bounded, then the limit direction will depend on the initial conditions.

Unfortunately, the following theorem shows that, the components of $\mathbf{G}_{(t)}$ matrix are bounded, and hence even for strict monotone losses, the initial conditions $w_{(0)}, \mathbf{G}_{(0)}$ and step-size η will have a non-vanishing contribution to the asymptotic behavior of $\mathbf{G}_{(t)}$ and hence to the limit direction $\bar{w}_\infty = \lim\limits_{t\to\infty}\frac{w_{(t)}}{\|w_{(t)}\|}$, whenever it exists. In other words, the implicit bias of AdaGrad does indeed depend on initialization and step-size.

Theorem 6. *For any linearly separable training data $\{x_n, y_n\}_{n=1}^N$, consider the AdaGrad iterates $w_{(t)}$ from eq.* (13) *for minimizing $\mathcal{L}(w)$ with exponential loss $\ell(u,y) = \exp(-uy)$. For any fixed and bounded step-size $\eta < \infty$, and any initialization of $w_{(0)}$ and $\mathbf{G}_{(0)}$, such that $\frac{\eta}{2}\mathcal{L}\left(w_{(0)}\right) < 1$, and $\left\|\mathbf{G}_{(0)}^{-1/4}x_n\right\|_2 \leq 1$, $\forall i, \forall t:\ \mathbf{G}_{(t)}[i,i] < \infty$.*

4. Gradient descent on the factorized parameterization

Consider the empirical risk minimization in eq. (1) for matrix valued $X_n \in \mathbb{R}^{d\times d}$, $W \in \mathbb{R}^{d\times d}$

$$\min_{W} \mathcal{L}(W) = \ell(\langle W, X_n\rangle, y_n). \tag{15}$$

This is the exact same setting as eq. (1) obtained by arranging w and x_n as matrices. We can now study another class of algorithms for learning linear models based on matrix factorization, where we reparameterize W as $W = UV^\top$ with *unconstrained* $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{d\times d}$ to get the following equivalent objective,

$$\min_{U,V} \mathcal{L}(UV^\top) = \sum_{n=1}^{N} \ell(\langle UV^\top, X_n\rangle, y_n). \tag{16}$$

Note that although non-convex, eq. (16) is equivalent to eq. (15) with the exact same set of global minima over $W = UV^\top$. Gunasekar et al. (2017) studied this problem for squared loss $\ell(u, y) = (u-y)^2$ and noted that gradient descent on the factorization yields radically different implicit bias compared to gradient descent on W. In particular, gradient descent on U, V is often observed to be biased towards low nuclear norm solutions, which in turns ensures generalization (Srebro et al., 2005) and low rank matrix recovery (Recht et al., 2010; Candes & Recht, 2009). Since the matrix factorization objective in eq. (16) can be viewed as a two-layer neural network with linear activation, understanding the implicit bias here could provide direct insights into characterizing the implicit bias in more complex neural networks with non-linear activations.

Gunasekar et al. (2017) noted that, the optimization problem in eq. (16) over factorization $W = UV^\top$ can be cast as a special case of optimization over p.s.d. matrices with unconstrained symmetric factorization $W = UU^\top$:

$$\min_{U\in\mathbb{R}^{d\times d}} \bar{\mathcal{L}}(U) = \mathcal{L}(UU^\top) = \sum_{n=1}^{N} \ell\left(\langle UU^\top, X_n\rangle, y_n\right). \tag{17}$$

Specifically, in terms of both the objective as well as gradient descent updates, a problem instance of eq. (16) is equivalent to a problem instance of eq. (17) with larger data matrices $\widetilde{X}_n = \begin{bmatrix} 0 & X_n \\ X_n^\top & 0 \end{bmatrix}$ and loss optimized over larger p.s.d. matrix of the form $\widetilde{U}\widetilde{U}^\top = \begin{bmatrix} A_1 & W \\ W^\top & A_2 \end{bmatrix}$, where $W = UV^\top$ corresponds to the optimization variables in the original problem instance of eq. (16) and A_1 and A_2 some p.s.d matrices that are irrelevant for the objective.

Henceforth, we will also consider the symmetric matrix factorization in (17). Let $U_{(0)} \in \mathbb{R}^{d\times d}$ be any full rank initialization, gradient descent updates in U are given by,

$$U_{(t+1)} = U_{(t)} - \eta_t \nabla\bar{\mathcal{L}}(U_{(t)}), \tag{18}$$

with corresponding updates in $W_{(t)} = U_{(t)}U_{(t)}{}^\top$ given by,

$$W_{(t+1)} = W_{(t)} - \eta_t\left[\nabla\mathcal{L}(W_{(t)})W_{(t)} + W_{(t)}\nabla\mathcal{L}(W_{(t)})\right] + \eta_t^2\nabla\mathcal{L}(W_{(t)})W_{(t)}\nabla\mathcal{L}(W_{(t)}) \tag{19}$$

Losses with a unique finite root For squared loss, Gunasekar et al. (2017) showed that the implicit bias of iterates in eq. (19) crucially depended on both the initialization $U_{(0)}$ as well as the step-size η. Gunasekar et al. conjectured, and provided theoretical and empirical evidence that gradient descent on the factorization converges to the minimum nuclear norm global minimum, but only if the initialization is infinitesimally close to zero and the step-sizes are infinitesimally small. Li et al. (2017), later proved the conjecture under additional assumption that the measurements X_n satisfy certain *restricted isometry property (RIP)*.

In the case of squared loss, it is evident that for finite step-sizes and finite initialization, the implicit bias towards the minimum nuclear norm global minima is not exact. In practice, not only do we need $\eta > 0$, but we also cannot initialize very close to zero since zero is a saddle point for eq. (17). The natural question motivated by the results in Section 3 is: for strictly monotone losses, can we get a characterization of the implicit bias of gradient descent for the factorized objective in eq. (17) that is more robust to initialization and step-size?

Strict monotone losses In the following theorem, we again see that the characterization of the implicit bias of gradient descent for factorized objective is more robust in the case of strict monotone losses.

Theorem 7. *For almost all datasets $\{X_n, y_n\}_{n=1}^N$ separable by a p.s.d. linear classifier, consider the gradient descent iterates $U_{(t)}$ in eq.* (18) *for minimizing $\bar{\mathcal{L}}(U)$ with the exponential loss $\ell(u,y) = \exp(-uy)$ and the corresponding sequence of linear predictors $W_{(t)}$ in eq.* (19). *For any full rank initialization $U_{(0)}$ and any finite step-size sequence $\{\eta_t\}_t$, if $W_{(t)}$ asymptotically minimizes $\mathcal{L}$, i.e., $\mathcal{L}(W_{(t)}) \to 0$, and additionally the updates $U_{(t)}$ and the gradients $\nabla\mathcal{L}(W_{(t)})$ converge in direction, then the limit direction $\bar{U}_\infty = \lim_{t\to\infty} \frac{U_{(t)}}{\|U_{(t)}\|_*}$ is a scaling of a first order stationary point (f.o.s.p) of the following non-convex optimization problem*

$$\bar{U}_\infty \propto \textit{f.o.s.p.} \min_{U \in \mathbb{R}^{d\times d}} \|U\|_2^2 \quad \textit{s.t.,} \quad \forall n, y_n \left\langle UU^\top, X_n \right\rangle \geq 1. \tag{20}$$

Remark 2. *Any global minimum U^* of eq.* (20) *corresponds to predictor W^* that minimizes the nuclear norm $\|.\|_*$ of linear p.s.d. classifier with margin constraints,*

$$W^* = \operatorname*{argmin}_{W \succcurlyeq 0} \|W\|_* \textit{ s.t., } \forall n, y_n \langle W, X_n\rangle \geq 1. \tag{21}$$

Additionally, in the absence of rank constraints on U, all second order stationary points of eq. (20) *are global minima for the problem. More general, we expect a stronger result that $\bar{W}_\infty = \bar{U}_\infty \bar{U}_\infty^\top$, which is also the limit direction of $W_{(t)}$, is a minimizer of eq.* (21). *Showing a stronger result that $W_{(t)}$ indeed converges in direction to W^* is of interest for future work.*

Here we note that convergence of $U_{(t)}$ in direction is necessary for the characterization of implicit bias to be relevant, but in Theorem 7, we require stronger conditions that the gradients $\nabla\mathcal{L}(W_{(t)})$ also converge in direction. Relaxing this condition is of interest for future work.

Key property Let us look at exponential loss when $W_{(t)}$ converges in direction to, say $\bar{W}_\infty$. Then $\bar{W}_\infty$ can be expressed as $W_{(t)} = \bar{W}_\infty g(t) + \rho(t)$ for some scalar $g(t) \to \infty$ and $\frac{\rho(t)}{g(t)} \to 0$. Consequently, the gradients $\nabla\mathcal{L}(W_{(t)}) = \sum_n \mathrm{e}^{-g(t) y_n \langle W_\infty, X_n\rangle} e^{-y_n\langle\rho(t), X_n\rangle}\, y_n X_n$ will asymptotically be dominated by linear combinations of examples X_n that have the smallest distance to the decision boundary, i.e., the support vectors of $\bar{W}_\infty$. This behavior can be used to show optimality of $\bar{U}_\infty$ such that $\bar{W}_\infty = \bar{U}_\infty \bar{U}_\infty^\top$ to the first order stationary points of the maximum margin problem in eq. 20.

This idea formalized in the following lemma, which is of interest beyond the results in this paper.

Lemma 8. *For almost all linearly separable datasets $\{x_n, y_n\}_{n=1}^N$, consider any sequence $w_{(t)}$ that minimizes $\mathcal{L}(w)$ in eq.* (1) *with exponential loss, i.e., $\mathcal{L}(w_{(t)}) \to 0$. If $\frac{w_{(t)}}{\|w_{(t)}\|}$ converges, then for every accumulation point z_∞ of $\left\{\frac{-\nabla\mathcal{L}(w_{(t)})}{\|\nabla\mathcal{L}(w_{(t)})\|}\right\}_t$, $\exists\{\alpha_n \geq 0\}_{n\in S}$ s.t., $z_\infty = \sum_{n\in S} \alpha_n y_n x_n$, where $\bar{w}_\infty = \lim_{t\to\infty} \frac{w_{(t)}}{\|w_{(t)}\|}$ and $S = \{n : y_n \langle \bar{w}_\infty, x_n\rangle = \min_n y_n \langle \bar{w}_\infty, x_n\rangle\}$ are the indices of the data points with smallest margin to $\bar{w}_\infty$.*

5. Summary

We studied the implicit bias of different optimization algorithms for two families of losses, losses with a unique finite root and strict monotone losses, where the biases are fundamentally different. In the case of losses with a unique finite root, we have a simple characterization of the limit point $w_\infty = \lim_{t\to\infty} w_{(t)}$ for mirror descent. But for this family of losses, such a succinct characterization does not extend to steepest descent with respect to general norms. On the other hand, for strict monotone losses, we noticed that the initial updates of the algorithm, including initialization and initial step-sizes are nullified when we analyze the asymptotic limit direction $\bar{w}_\infty = \lim_{t\to\infty} \frac{w_{(t)}}{\|w_{(t)}\|}$. We show that for steepest descent, the limit direction is a maximum margin separator within the unit ball of the corresponding norm. We also looked at other optimization algorithms for strictly monotone losses. For matrix factorization, we again get a more robust characterization that relates the limit direction to the maximum margin separator with unit nuclear norm. This again, in contrast to squared loss Gunasekar et al. (2017), is independent of the initialization and step-size. However, for AdaGrad, we show that even for strict monotone losses, the limit direction $\bar{w}_\infty$ could depend on the initial conditions.

In our results, we characterize the implicit bias for linear models as minimum norm (potential) or maximum margin solutions. These are indeed very special among all the solutions that fit the training data, and in particular, their generalization performance can in turn be understood from standard analyses Bartlett & Mendelson (2003).

Going forward, for more complicated non-linear models, especially neural networks, further work is required in order to get a more complete understanding of the implicit bias. The preliminary result for matrix factorization provides us tools to attempt extensions to multi-layer linear models, and eventually to non-linear networks. Even for linear models, the question of what is the implicit bias is when $\mathcal{L}(w)$ is optimized with explicitly constraints $w \in \mathcal{W}$ is an open problem. We believe similar characterizations can be obtained when there are multiple feasible solutions with $\mathcal{L}(w) = 0$. We also believe, the results for single outputs considered in this paper can also be extended for multi-output loss functions.

Finally, we would like a more fine grained analysis connecting the iterates $w_{(t)}$ along the optimization path of various algorithms to the regularization path, $\widehat{w}(c) = \operatorname{argmin}_{\mathcal{R}(w) \leq c} \mathcal{L}(w)$, where an explicit regularization is added to the optimization objective. In particular, our positive characterizations show that the optimization and regularization paths meet at the limit of $t \to \infty$ and $c \to \infty$, respectively. It would be desirable to further understand the relations between the entire optimization and regularization paths, which will help us understand the non-asymptotic effects from early stopping.

Acknowledgments

The authors are grateful to M.S. Nacson, Y. Carmon, and the anonymous ICML reviewers for helpful comments on the manuscript. The research was supported in part by NSF IIS award 1302662. The work of DS was supported by the Taub Foundation.

References

Amari, S. I. Natural gradient works efficiently in learning. *Neural computation*, 1998.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2003.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 2003.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 1967.

Candes, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics*, 2004.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 2001.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6152–6160, 2017.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *NIPS*, pp. 1–13, may 2017. URL `http://arxiv.org/abs/1705.08741`.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2016.

Kingma, D. and Adam, J. B. Adam: A method for stochastic optimisation. In *International Conference for Learning Representations*, volume 6, 2015.

Kivinen, J. and Warmuth, M. K. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 1997.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix recovery. *arXiv preprint arXiv:1712.09203*, 2017.

Muresan, M. and Muresan, M. *A concrete approach to classical analysis*, volume 14. Springer, 2009.

Nemirovskii, A. and Yudin, D. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

Nesterov, Y. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, 1983.

Neyshabur, B., Salakhutdinov, R. R., and Srebro, N. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2015a.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*, 2015b.

Neyshabur, B., Tomioka, R., Salakhutdinov, R., and Srebro, N. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 1964.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 2010.

Ross, K. A. *Elementary analysis*. Springer, 1980.

Rosset, S., Zhu, J., and Hastie, T. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 2004.

Rudin, C., Daubechies, I., and Schapire, R. E. The dynamics of adaboost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5(Dec):1557–1595, 2004.

Schapire, R. E. and Freund, Y. *Boosting: Foundations and algorithms*. MIT press, 2012.

Shalev-Shwartz, S. and Singer, Y. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. *Machine learning*, 80(2-3):141–163, 2010.

Smith, Kindermans, L. Don't Decay the Learning Rate, Increase the Batch Size. In *ICLR*, 2018.

Soudry, D., Hoffer, E., and Srebro, N. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Srebro, N., Alon, N., and Jaakkola, T. S. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems*, pp. 1321–1328, 2005.

Telgarsky, M. Margins, shrinkage and boosting. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pp. II–307. JMLR. org, 2013.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pp. 4151–4161, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhang, T., Yu, B., et al. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

Appendix A. Losses with a unique finite root

Let $\mathcal{P}_{\mathcal{X}} = \text{span}(\{x_n : n \in [N]\}) = \{\sum_n \nu_n x_n : \nu_n \in \mathbb{R}\}$. Let $\ell'(u, y)$ be the derivative of ℓ w.r.t first operand u, then we can see that, for any ℓ,

$$\forall w \in \mathbb{R}^d, \nabla\mathcal{L}(w) = \sum_{n=1}^{N} \ell'(\langle w, x_n \rangle, y_n))\, x_n \in \mathcal{P}_{\mathcal{X}}. \tag{22}$$

A.1 Proof of Theorem 1-1b

For a strongly convex potential ψ, denote the global optimum with minimum Bregman divergence $D_\psi(., w_{(0)})$ to the initialization $w_{(0)}$ as

$$w_\psi^* = \operatorname*{argmin}_w D_\psi(w, w_{(0)}) \text{ s.t.}, \forall n,\ \langle w, x_n \rangle = y_n, \tag{23}$$

where recall that $D_\psi(w, w_{(0)}) = \psi(w) - \psi(w_{(0)}) - \langle \nabla\psi(w_{(0)}), w - w_{(0)} \rangle$.

The KKT optimality conditions for (23) are as follows,

$$\begin{aligned}
&\textbf{Stationarity:} && \nabla\psi(w_\psi^*) - \nabla\psi(w_{(0)}) \in \mathcal{P}_{\mathcal{X}},\ \text{or } \exists\{\nu_n\}_{n=1}^N \text{ s.t., } \nabla\psi(w_\psi^*) - \nabla\psi(w_{(0)}) = \sum_{n=1}^{N} \nu_n x_n \\
&\textbf{Primal feasibility:} && \forall n, \langle w_\psi^*, x_n \rangle = y_n,\ \text{or } w_\psi^* \in \mathcal{G}
\end{aligned} \tag{24}$$

Recall Theorem 1–1b from Section 2.2.

Theorem 1. *For any loss ℓ with a unique finite root (Property 1), any realizable dataset $\{x_n, y_n\}_{n=1}^N$, and any strongly convex potential ψ, consider the mirror descent iterates $w_{(t)}$ from eq.* (5) *for minimizing the empirical loss $\mathcal{L}(w)$ in eq.* (1)*. For all initializations $w_{(0)}$, if the step-size sequence $\{\eta_t\}_t$ is chosen such that the limit point of the iterates $w_\infty = \lim_{t\to\infty} w_{(t)}$ is a global minimizer of $\mathcal{L}$, i.e., $\mathcal{L}(w_\infty) = 0$, then w_∞ is given by*

$$w_\infty = \operatorname*{argmin}_{w:\forall n, \langle w, x_n \rangle = y_n} D_\psi(w, w_{(0)}). \tag{6}$$

Theorem 1a. *Under the conditions of Theorem 1, consider constrained mirror descent updates $w_{(t)}$ from eq.* (4) *with realizable affine equality constraints, that is $\mathcal{W} = \{w : Gw = h\}$ for some $G \in \mathbb{R}^{d' \times d}$ and $h \in \mathbb{R}^{d'}$ and additionally, $\exists w \in \mathcal{W}$ with $\mathcal{L}(w) = 0$. For all initializations $w_{(0)}$, if the step-size sequence $\{\eta_t\}_t$ is chosen to asymptotically minimize $\mathcal{L}$, i.e., $\mathcal{L}(w_\infty) = 0$, then $w_\infty = \operatorname{argmin}_{w \in \mathcal{G} \cap \mathcal{W}} D_\psi(w, w_{(0)})$.*

Theorem 1b. *Under the conditions in Theorem 1, if initialized at $w_{(0)} = \operatorname{argmin}_w \psi(w)$, then the mirror descent updates with dual momentum also converge to* (6)*, i.e., for all $\{\eta_t\}_t, \{\beta_t\}_t, \{\gamma_t\}_t$, if $w_{(t)}$ from eq.* (7) *converges to $w_\infty \in \mathcal{G}$, then $w_\infty = \operatorname{argmin}_{w \in \mathcal{G}} \psi(w)$.*

Remark 1. *Following the same arguments, we can show that Theorem 1–1b also hold when instancewise stochastic gradients defined in eq.* (2) *are used in place of $\nabla\mathcal{L}(w_{(t)})$.*

Proof. (a) **Generic mirror descent: Theorem 1** Recall the updates of mirror descent: $\nabla\psi(w_{(t+1)}) - \nabla\psi(w_{(t)}) = -\eta_t \nabla\mathcal{L}(w_{(t)})$ Using telescoping sum, we have,

$$\forall t,\ \nabla\psi(w_{(t)}) - \nabla\psi(w_{(0)}) = \sum_{t'<t} \nabla\psi(w_{(t'+1)}) - \nabla\psi(w_{(t')}) = \sum_{t'<t} -\eta_{t'} \nabla\mathcal{L}(w_{(t')}) \in \mathcal{P}_{\mathcal{X}}, \tag{25}$$

where the last inclusion follows as $\forall t', -\eta_{t'} \nabla\mathcal{L}(w_{(t')}) \in \mathcal{P}_{\mathcal{X}}$ from (22).

Thus, for all t, $w_{(t)}$ from mirror descent updates in eq. (5) always satisfy the stationarity condition of eq. (24). Additionally, if $w_{(t)}$ converges to a global minimum, then $w_\infty = \lim_{t\to\infty} w_{(t)} \in \mathcal{G} = \{w : \forall n, \langle w, x_n \rangle = y_n\}$ also satisfies the primal feasibility condition in eq. (24). Combining the above arguments, we have that if $\mathcal{L}(w_\infty) = 0$, then $w_\infty = \operatorname{argmin}_{w\in\mathcal{G}} D_\psi(w, w_{(0)})$.

(b) **Realizable affine equality constraints: Theorem 1a** For $\mathcal{W} = \{w : Gw = h\}$ for some $G \in \mathbb{R}^{d'\times d}$ and $h \in \mathbb{R}^{d'}$ with a realizable feasible solution $w \in \mathcal{W}$ satisfying $\forall n, \langle x_n, w\rangle = y_n$, the claim is that $w_{(t)}$ from eq. (4) converges to

$$w^*_{\psi,\mathcal{W}} = \operatorname*{argmin}_w D_\psi(w, w_{(0)}) \text{ s.t.}, \forall n,\ \langle w, x_n\rangle = y_n \text{ and } w \in \mathcal{W} = \{w : Gw = h\}. \tag{26}$$

The KKT optimality conditions for (26) are as follows,

$$\begin{aligned} &\textbf{Stationarity:} && \exists\{\nu_n\}_{n=1}^N, \{\mu_j\}_{j=1}^{d'} \text{ s.t., } \nabla\psi(w^*_{\psi,\mathcal{W}}) - \nabla\psi(w_{(0)}) = \sum_{n=1}^N \nu_n x_n + \sum_{j=1}^{d'} \mu_j g_j \\ &\textbf{Primal feasibility:} && \forall n, \left\langle w^*_{\psi,\mathcal{W}}, x_n\right\rangle = y_n, \text{ and } Gw^*_{\psi,\mathcal{W}} = h \end{aligned} \tag{27}$$

To show that the limit point of mirror descent updates in eq. (4) satisfy the above KKT conditions, we first note that the updates are equivalently computed as follows,

$$\begin{aligned} w_{(t+1)} &= \operatorname*{argmin}_{w:Gw=h} \eta_t \left\langle w, \nabla\mathcal{L}(w_{(t)})\right\rangle + D_\psi(w, w_{(t)}) \\ &= \operatorname*{argmin}_{w:Gw=h} D_\psi(w, \nabla\psi^{-1}\left(\nabla\psi(w_{(t)}) - \eta_t\nabla\mathcal{L}(w_{(t)})\right)). \end{aligned} \tag{28}$$

Let g_j for $j = 1, 2, \ldots, d'$ denote the rows of G. From the optimality conditions of eq. (28), we get that

$$\exists\{\mu_j\}_{j=1}^{d'}, \quad \text{s.t.} \quad \nabla\psi(w_{(t+1)}) = \nabla\psi(w_{(t)}) - \eta_t\mathcal{L}(w_{(t)}) + \sum_{j=1}^{d'} \mu_j g_j \text{ and } Gw_{(t+1)} = h. \tag{29}$$

Again, primal feasibility is satisfied whenever $w_{(t)} \to \mathcal{G}$ since mirror descent iterates are always feasible points $w_{(t)} \in \mathcal{W}$. The stationarity condition follows from using eq. (29) with same arguments of the unconstrained case.

(c) **Dual momentum:** For *any* $\widetilde{\beta}_{t'}, \widetilde{\gamma}_{t'} \in \mathbb{R}$ and $\widetilde{w}_{(t')} \in \mathbb{R}^d$, consider a general update of the form

$$\nabla\psi(w_{(t+1)}) = \sum_{t'\le t} \widetilde{\beta}_{t'}\nabla\psi(w_{(t')}) + \widetilde{\gamma}_{t'}\nabla\mathcal{L}(\widetilde{w}_{(t')}). \tag{30}$$

Claim: If $\nabla\psi(w_{(0)}) = 0$, then for all updates of the form (30) satisfies $\nabla\psi(w_{(t)}) \in \mathcal{P}_\mathcal{X}$
—this can be easily proved by induction:(a) for $t = 0$, $\nabla\psi(w_{(0)}) = 0 \in \mathcal{P}_\mathcal{X}$; (b) let $\forall t' \le t$, $\nabla\psi(w_{(t')}) \in \mathcal{P}_\mathcal{X}$, (c) then using the inductive assumption and eq. (22), we have $\nabla\psi(w_{(t+1)}) = \sum_{t'\le t} \widetilde{\beta}_{t'}\nabla\psi(w_{(t')}) + \widetilde{\gamma}_{t'}\nabla\mathcal{L}(\widetilde{w}_{(t')}) \in \mathcal{P}_\mathcal{X}$.

Dual momentum in eq. (7) is a special case of eq. (30) with appropriate choice of $\widetilde{\beta}_{t'}, \widetilde{\gamma}_{t'} \in \mathbb{R}$, and $\widetilde{w}_{t'} \in \mathbb{R}^d$.

(d) **Instancewise stochastic gradient descent (Remark 1):** In the above arguments, the only property of gradient $\nabla\mathcal{L}$ that we used is that $\forall w \in \mathbb{R}^d, \nabla\mathcal{L}(w) = \sum_{n=1}^N \ell'(\langle w, x_n\rangle, y_n))\, x_n \in \mathcal{P}_\mathcal{X}$ (eq. (22)). This property also holds for instancewise stochastic gradients as defined in eq. (2), i.e., $\nabla\widetilde{\mathcal{L}}(w) = \sum_{n\in S_t} \ell'(\langle w, x_n\rangle, y_n))\, x_n$, hence all the results follow.

□

A.2 Proofs of propositions in Section 2

A.2.1 Primal momentum and natural gradient descent

Recall the optimization problem in Examples 2–3: $\{(x_1 = [1, 2], y_1 = 1)\}$, and $\ell(u, y) = (u - y)^2$. We have $\mathcal{P}_{\mathcal{X}} = \text{span}(x_1) = \{z : 2z[1] - z[2] = 0\}$.

For entropy potential $\psi(w) = \sum_i w[i] \log w[i] - w[i]$, we have $\nabla\psi(w) = \log w$ (where the $\log$ is taken elementwise), and initialization $w_{(0)} = [1, 1]$ satisfies $\nabla\psi(w_{(0)}) = 0$ which is the optimality condition for $\min_w \psi(w)$.

1. **Proof of Proposition 2a:** we use primal momentum with $\beta_1 > 0$ only in the first step, and $\forall t \geq 2$, $\beta_t = \gamma_t = 0$. We get the following initial updates

Since $\forall t \geq 2$, $\beta_t = \gamma_t = 0$, we first note that for $t > 2$, the updates merely follow the path of standard MD initialized at $\nabla\psi(w_{(2)})$ for a convex loss function. This implies the following:

- for appropriate choice of $\{\eta_t\}_{t\geq 2}$ (given by convergence analysis of mirror descent for convex functions), we can get $w_\infty = \lim_{t\to\infty} w_{(t)} \in \mathcal{G}$, and
- from eq. (25), w_∞ satisfies $\nabla\psi(w_\infty) - \nabla\psi(w_{(2)}) \in \mathcal{P}_{\mathcal{X}} \Rightarrow \nabla\psi(w_\infty) \in \nabla\psi(w_{(2)}) + \mathcal{P}_{\mathcal{X}}$.

Since w_∞ satisfies primal feasibility, from stationarity condition in eq. (24), we have

$$w_\infty = w^*_\psi = \operatorname*{argmin}_{w\in\mathcal{G}} \psi(w) \textit{ if and only if } \nabla\psi(w_{(2)}) \in \mathcal{P}_{\mathcal{X}}.$$

We show that this is not the case for any $\beta_1 > 0$ and any $\gamma_1 \geq 0$. Recall that $\Delta w_{(-1)} = 0$, $\nabla\psi(w_{(0)}) = 0$ and $\nabla\psi(w) = \log w$. Working through the steps in eq. (8), for scalars $r_0 = \eta_0(y_1 - \langle w_{(0)}, x_1\rangle)$ and $\widetilde{r}_1 = \eta_t(y_1 - \langle w_{(1)} + \gamma_1\Delta w_{(0)}, x_1\rangle)$, and any $\beta_1 > 0$, we have:

- $\nabla\psi(w_{(1)}) = r_0 x_1 \implies w_{(1)} = \exp(r_0 x_1)$, and
- $\nabla\psi(w_{(2)}) = \nabla\psi((1 + \beta_1)w_{(1)}) + \widetilde{r}_1 x_1 = \log(1 + \beta_1) + r_0 x_1 + \widetilde{r}_1 x_1 \in \log(1 + \beta_1) + \mathcal{P}_{\mathcal{X}} \notin \mathcal{P}_{\mathcal{X}}$.□

2. **Proof of Proposition 3a:** The arguments are similar to the proof of Proposition 2a. In Example 3, we again use a finite $\eta_1 > 0$ to get $w_{(1)}$ and then follow the NGD using infinitesimal η initialized at $w_{(1)}$.

We know that for infinitesimal step-size, the NGD path starting at $w_{(1)}$ follows the corresponding infinitesimal MD path on a convex problem and hence from eq. (25), the NGD updates for this example converges to a global minimum $w_\infty = \lim_{t\to\infty} w_{(t)} \in \mathcal{G}$, that satisfies $\nabla\psi(w_\infty) - \nabla\psi(w_{(1)}) \in \mathcal{P}_{\mathcal{X}} \Rightarrow \nabla\psi(w_\infty) \in \nabla\psi(w_{(1)}) + \mathcal{P}_{\mathcal{X}}$.

From stationarity condition in (24), $w_\infty = w^*_\psi = \operatorname{argmin}_{w\in\mathcal{G}} \psi(w)$ if and only if $\nabla\psi(w_{(1)}) \in \mathcal{P}_{\mathcal{X}}$.

For natural gradient descent, $w_{(1)} = w_{(0)} - \eta_1 \nabla^2\psi(w_{(0)})^{-1}\nabla\mathcal{L}(w_{(0)}) = [1 + \eta_1 r_0, 1 + 2\eta_1 r_0]$, where $r_0 = \eta_0(y_1 - \langle w_{(0)}, x_1\rangle)$. We then have $\nabla\psi(w_{(1)}) \in \mathcal{P}_{\mathcal{X}} \Leftrightarrow 2\nabla\psi(w_{(1)})[1] - \nabla\psi(w_{(1)})[2] = 0 \Leftrightarrow 2\log(w_{(1)}[1]) - \log(w_{(1)}[2]) = 0 \Leftrightarrow \log\left(1 + \frac{\eta_1^2 r_0^2}{1+2\eta_1 r_0}\right) = 0$.

For any η_1 such that $\frac{\eta_1^2 r_0^2}{1+2\eta_1 r_0} \neq 0$, we get a contradiction. □

Appendix B. Steepest descent for strictly monotone losses

We prove Theorem 5 in this section.

Theorem 5. *For any separable dataset $\{x_n, y_n\}_{n=1}^N$ and any norm $\|\cdot\|$, consider the steepest descent updates from eq.* (12) *for minimizing $\mathcal{L}(w)$ in eq.* (1) *with the exponential loss $\ell(u, y) = \exp(-uy)$. For all* *initializations $w_{(0)}$, and all bounded step-sizes satisfying $\eta_t \leq \max\{\eta_+, \frac{1}{B^2\mathcal{L}(w_{(t)})}\}$, where $B := \max_n \|x_n\|_\star$ and $\eta_+ < \infty$ is any finite upper bound, the iterates $w_{(t)}$ satisfy the following,*

$$\lim_{t\to\infty} \min_n \frac{y_n \left\langle w_{(t)}, x_n \right\rangle}{\|w_{(t)}\|} = \max_{w:\|w\|\leq 1} \min_n y_n \left\langle w, x_n \right\rangle.$$

In particular, if there is a unique maximum-$\|.\|$ margin solution $w^\star_{\|.\|} = \operatorname{argmax}_{w:\|w\|\leq 1} \min_n y_n \langle w, x_n\rangle$, then the limit direction is given by $\bar{w}_\infty = \lim\limits_{t\to\infty} \frac{w_{(t)}}{\|w_{(t)}\|} = w^\star_{\|.\|}$.

The proof is divided into three subsections

1. Generalized duality lemma: we show that for all norms and all w, $\|\nabla\mathcal{L}(w)\|_\star \geq \gamma\mathcal{L}(w)$.
2. Properties of $\nabla\mathcal{L}(w_{(t)})$ and $\mathcal{L}(w_{(t)})$ for steepest descent: we prove two lemmata that show some useful properties of $\nabla\mathcal{L}(w_{(t)})$ and $\mathcal{L}(w_{(t)})$.
3. Remaining steps in the proof: putting together above lemmata to prove Theorem 5.

B.1 Generalized duality lemma: for all norms and all w, $\|\nabla\mathcal{L}(w)\|_\star \geq \gamma\mathcal{L}(w)$

The following lemma is a standard result in convex analysis.

Lemma 9 (Fenchel Duality)**.** *Let $A \in \mathbb{R}^{m\times n}$, and $f : \mathbb{R}^m \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}$ be two closed convex functions and $f^\star, g^\star$ be their Fenchel conjugate functions, respectively. Then,*

$$\max_{w\in\mathbb{R}^n} -f^\star(Aw) - g^\star(-w) \leq \min_{r\in\mathbb{R}^m} f(r) + g(A^\top r). \tag{31}$$

Let $X \in \mathbb{R}^{N\times d}$ be the data matrix with x_n along the rows of X. Without loss of generality, let $y_n = 1$, as for linear models y_n can be absorbed into x_n. Let e_n denote the n^{th} standard basis in $\mathbb{R}^N$.

We define the $\|\cdot\|$- maximum margin as,

$$\gamma = \max_{w\neq 0} \min_{n\in[N]} \frac{\langle w, x_n\rangle}{\|w\|} = \max_{\|w\|\leq 1} \min_{n\in[N]} e_n^\top X w. \tag{32}$$

Our primary technical novelty is the following duality lemma that generalizes similar result in Telgarsky (2013) for ℓ_1 norm to general norms: we want to show that $\|\nabla\mathcal{L}(w)\|_\star \geq \gamma\mathcal{L}(w)$ for all w, where recall that $\|.\|_\star$ is the dual norm of $\|.\|$.

Define $r_n(w) = \exp\!\left(-w^\top x_n\right) \geq 0$ and let $r(w) = [r_n(w)]_{n=1}^N \in \mathbb{R}^N$. Note that $\mathcal{L}(w) = \|r(w)\|_1$ and $\nabla\mathcal{L}(w) = X^\top r(w)$. We can now restate $\|\nabla\mathcal{L}(w)\|_\star \geq \gamma\mathcal{L}(w)$ as $\frac{\|X^\top r(w)\|_\star}{\|r(w)\|_1} \geq \gamma$.

In the following lemma, we show this holds for any $r_n(w) \geq 0$. Since norms are homogeneous, this is equivalent to $\min_{r\in\Delta_{N-1}} \left\|X^\top r\right\|_\star \geq \gamma$, where $\Delta_{N-1} = \{v \in \mathbb{R}^N : v \geq 0, \|v\|_1 = 1\}$ is the N-dimensional probability simplex.

Lemma 10. *For any norm $\|.\|$, the following duality holds:*

$$\min_{r\in\Delta_{N-1}} \left\|X^\top r\right\|_\star \geq \max_{\|w\|\leq 1} \min_{n\in[N]} e_n^\top X w = \gamma. \tag{33}$$

This implies, for exponential loss $\ell(u, y) = \exp(-uy)$, the following holds

$$\forall w,\ \|\nabla\mathcal{L}(w)\|_\star \geq \gamma\mathcal{L}(w). \tag{34}$$

Proof. Let $\mathbf{1}_E$ denote the indicator function which takes value 0 if E is satisfied and ∞ otherwise.

Define $f(r) = \mathbf{1}_{r\in\Delta_{N-1}}$ and $g(z) = \|z\|_\star$, so that

$$\min_{r\in\Delta_{N-1}} \left\|X^\top r\right\|_* = \min_{r\in\mathbb{R}^N} f(r) + g(X^\top r). \tag{35}$$

The conjugates are $f^*(y) = \max_{r\in\Delta_{N-1}} \langle y, r\rangle = \max_{n=1}^N \langle y, e_n\rangle$, and $g^*(w) = \mathbf{1}_{\|w\|\leq 1}$. The LHS of Lemma 9 is

$$\begin{aligned}\max_w\big(-f^*(Xw) - g^*(-w)\big) &= \max_w \big(-\max_n e_n^\top Xw - \mathbf{1}_{\|w\|\leq 1}\big)\\ &= \max_{\|w\|\leq 1}\min_n e_n^\top X(-w) \overset{(a)}{=} \max_{\|w\|\leq 1}\min_n e_n^\top Xw \overset{(b)}{=} \gamma,\end{aligned} \tag{36}$$

where (a) follows from central symmetry of $\{w : \|w\| \leq 1\}$, and (b) from definition of maximum $\|.\|$-margin in eq. (32).

Using weak duality (Lemma 9) on eqs. (35) and (36), we have $\forall r$, $\|X^\top r\|_\star \geq \gamma\|r\|_1$. Finally, recalling that for exponential loss $r_n(w) = \exp\big(-w^\top x_n\big)$, $\mathcal{L}(w) = \|r(w)\|_1$ and $\nabla\mathcal{L}(w) = X^\top r(w)$, we have $\forall w$, $\|\nabla\mathcal{L}(w)\|_\star \geq \gamma\mathcal{L}(w)$. $\square$

B.2 Properties of $\nabla\mathcal{L}(w_{(t)})$ and $\mathcal{L}(w_{(t)})$ for steepest descent

Recall the steepest descent updates in eqs. (11) and (12) :

$$\begin{aligned}&w_{(t+1)} = w_{(t)} + \eta_t\Delta w_{(t)}, \text{ where } \Delta w_{(t)} \text{ satisfies}\\ &\langle\Delta w_{(t)}, -\nabla\mathcal{L}(w_{(t)})\rangle = \|\Delta w_{(t)}\|^2 = \|\nabla\mathcal{L}(w_{(t)})\|_\star^2.\end{aligned} \tag{37}$$

Lemma 11. *For exponential loss $\ell(u, y) = \exp(-uy)$, consider the steepest descent iterates $w_{(t)}$ for minimizing $\mathcal{L}(w_{(t)})$, with any initialization $w_{(0)}$ and any finite step-size η_t that leads to a strictly decreasing sequence $\mathcal{L}(w_{(t)})$ and satisfies $0 < \eta_t \leq \max\{\eta_+, \frac{1}{B^2\mathcal{L}(w_{(t)})}\}$, where $B = \max_n \|x_n\|_\star$. Then the following holds:*

(A) $\sum_{t=0}^\infty \eta_t\|\nabla\mathcal{L}(w_{(t)})\|_\star^2 \leq \infty$, and hence $\|\nabla\mathcal{L}(w_{(t)})\|_\star \to 0$.
(B) Iterates $w_{(t)}$ converge to a global minima $\mathcal{L}(w_{(t)}) \to 0$, and hence $\forall n$ $\big\langle w_{(t)}, x_n\big\rangle \to \infty$.
(C) $\sum_{t=0}^\infty \eta_t\|\nabla\mathcal{L}(w_{(t)})\|_\star = \infty$.

Proof. 1. **Proof of (A):** We have that $\|x_n\|_\star \leq B$ for all n. Recall that $r_n(w) = \exp(-\langle w, x_n\rangle) \geq 0$, $\mathcal{L}(w) = \sum_n r_n(w)$, and $\nabla\mathcal{L}(w) = \sum_n r_n(w)x_n$. Thus, for all v, we have

$$v^\top\nabla^2\mathcal{L}(w)v = \sum_n r_n(w)(x_n^\top v)^2 \leq \sum_n r_n(w)\|x_n\|_\star^2\|v\|^2 \leq \mathcal{L}(w)B^2\|v\|^2. \tag{38}$$

Using Taylor's reminder theorem for the convex loss $\mathcal{L}$, we have

$$\begin{aligned}\mathcal{L}(w_{(t+1)}) &\leq \mathcal{L}(w_{(t)}) + \eta_t\left\langle\nabla\mathcal{L}(w_{(t)}), \Delta w_{(t)}\right\rangle + \sup_{\beta\in(0,1)}\frac{\eta_t^2}{2}\Delta w_{(t)}{}^\top\nabla^2\mathcal{L}\left(w_{(t)} + \beta\eta_t\Delta w_{(t)}\right)\Delta w_{(t)}\\ &\overset{(a)}{\leq} \mathcal{L}(w_{(t)}) - \eta_t\left\|\nabla\mathcal{L}(w_{(t)})\right\|_\star^2 + \frac{\eta_t^2B^2}{2}\sup_{\beta\in(0,1)}\mathcal{L}\left(w_{(t)} + \beta\eta_t\Delta w_{(t)}\right)\left\|\Delta w_{(t)}\right\|^2\\ &\overset{(b)}{\leq} \mathcal{L}(w_{(t)}) - \eta_t\left\|\nabla\mathcal{L}(w_{(t)})\right\|_\star^2 + \frac{\eta_t^2B^2}{2}\mathcal{L}(w_{(t)})\left\|\Delta w_{(t)}\right\|^2\\ &\overset{(c)}{\leq} \mathcal{L}(w_{(t)}) - \frac{\eta_t}{2}\left\|\nabla\mathcal{L}(w_{(t)})\right\|_\star^2,\end{aligned} \tag{39}$$

where (a) follows from eq. (38) and from the condition on update direction in eq. (37); (b) follows as $\eta_t \Delta w_{(t)}$ is a descent step and along with convexity of $\mathcal{L}(w)$ we have $\sup_{\beta\in(0,1)} \mathcal{L}\left(w_{(t)} + \beta\eta_t\Delta w_{(t)}\right) \leq \mathcal{L}(w_{(t)})$; and (c) follows as $\eta_t \leq \frac{1}{B^2\mathcal{L}(w_{(t)})}$ from the assumption and also using $\|\Delta w_{(t)}\| = \|\nabla\mathcal{L}(w_{(t)})\|_\star$ from eq. 37.

Thus, $\mathcal{L}(w_{(t)}) - \mathcal{L}(w_{(t+1)}) \geq \frac{\eta_t}{2}\left\|\nabla\mathcal{L}(w_{(t)})\right\|_\star^2$, which implies

$$\forall t,\ \sum_{u=0}^{t} \eta_u \left\|\nabla\mathcal{L}(w_{(u)})\right\|_\star^2 \leq 2\sum_{u=0}^{t} \mathcal{L}(w_{(u)}) - \mathcal{L}(w_{(u+1)}) = 2\left(\mathcal{L}(w_{(0)}) - \mathcal{L}(w_{(t+1)})\right) < \infty\,. \tag{40}$$

where the final inequality follows as $\mathcal{L}(w_{(0)}) < \infty$ and $\mathcal{L}(w_{(t)}) \geq 0\ \forall t$.

In the continuous time limit of $\eta \to 0$, (A) is equivalently expressed as $\int_0^t \|\nabla\mathcal{L}(w_{(t)})\|_\star^2 < \infty$. Thus, we have $\lim_{t\to\infty}\|\nabla\mathcal{L}(w_{(t)})\|_\star = 0$—both for any finite $\eta_t > 0$ as well as in the continuous time limit of $\eta \to 0$.

2. **Proof of** (B) **and** (C) **:** Consider any $v \in \mathbb{R}^d$ that linearly separates the data, i.e., $\forall n, \langle v, x_n\rangle > 0$ (such a v always exists for linearly separable data), then we have

$$\forall t < \infty,\ v^\top\nabla\mathcal{L}(w_{(t)}) = \sum_{n\in[N]} \exp\big(-\left\langle w_{(t)}, x_n\right\rangle\big)x_n^\top v > 0.$$

Since $\lim_{t\to\infty} v^\top\nabla\mathcal{L}(w_{(t)}) = 0$, it must be that $\forall n,\ \exp\big(-\left\langle w_{(t)}, x_n\right\rangle\big) \to 0$, and thus $\left\|w_{(t)}\right\| \to \infty$.

Using triangle inequality, we have

$$\infty = \lim_{t\to\infty}\left\|w_{(t)}\right\| \leq \|w_{(0)}\| + \sum_{t=0}^{\infty}\eta_t\|\Delta w_{(t)}\| = \|w_{(0)}\| + \sum_{t=0}^{\infty}\eta_t\|\nabla\mathcal{L}(w_{(t)})\|_\star, \tag{41}$$

where we used $\|\Delta w_{(t)}\| = \|\nabla\mathcal{L}(w_{(t)})\|_\star$ from (37). This gives us $\sum_{t=0}^\infty \eta_t\|\nabla\mathcal{L}(w_{(t)})\|_\star = \infty$ in (C). □

We next show that under the conditions of Theorem 5, $\mathcal{L}(w_{(t)})$ forms a decreasing sequence, and hence satisfies the assumption in Lemma 11.

Lemma 12. *If step-sizes η_t satisfy $\eta_t = \frac{c_t}{B^2\mathcal{L}(w_t)}$ for $c_t \leq \sqrt{2}$, then $\mathcal{L}(w_{(t+1)}) \leq \mathcal{L}(w_{(t)})$.*

Proof. From the Taylor expansion of $\mathcal{L}\left(w\right)$ in eq. (39), we have

$$\begin{aligned}\mathcal{L}(w_{(t+1)}) &\leq \mathcal{L}(w_{(t)}) - \eta_t\|\nabla\mathcal{L}(w_{(t)})\|_\star^2 + \frac{\eta_t^2B^2}{2}\|\nabla\mathcal{L}(w_{(t)})\|_\star^2 \sup_{\beta\in(0,1)}\mathcal{L}(w_{(t)} + \beta\eta_t\Delta w_{(t)})\\ &\overset{(a)}{\leq} \mathcal{L}(w_{(t)}) - \eta_t\|\nabla\mathcal{L}(w_{(t)})\|_\star^2\left(1 - \frac{\eta_tB^2}{2}\max\left(\mathcal{L}(w_{(t)}), \mathcal{L}(w_{(t+1)})\right)\right),\end{aligned} \tag{42}$$

where (a) follows from convexity of $\mathcal{L}$.

We want to show that $\mathcal{L}(w_{(t+1)}) \leq \mathcal{L}(w_{(t)})$. Let us assume the contrary that $\mathcal{L}(w_{(t+1)}) > \mathcal{L}(w_{(t)})$.

From eq. (42), we have

$$\begin{aligned}\mathcal{L}(w_{(t)}) \overset{(a)}{<} \mathcal{L}(w_{(t+1)}) \leq \mathcal{L}(w_{(t)}) - \eta_t\|\nabla\mathcal{L}(w_{(t)})\|_\star^2\left(1 - \frac{\eta_tB^2\mathcal{L}(w_{t+1})}{2}\right)\\ \overset{(b)}{\Longrightarrow}\left(1 - \frac{\eta_tB^2\mathcal{L}(w_{t+1})}{2}\right) \leq 0,\end{aligned} \tag{43}$$

where (a) follows from the contradictory assumption and (b) follows as $\eta_t\|\nabla\mathcal{L}(w_{(t)})\|_\star^2 \geq 0$.

Following up from eq. (43), we have

$$
\begin{aligned}
\mathcal{L}(w_{(t+1)}) &\leq \mathcal{L}(w_{(t)}) + \eta_t \|\nabla\mathcal{L}(w_{(t)})\|_\star^2 \left(\frac{\eta_t B^2 \mathcal{L}(w_{t+1})}{2} - 1\right) \\
&\overset{(a)}{\leq} \mathcal{L}(w_{(t)}) + \eta_t B^2 \mathcal{L}(w_{(t)})^2 \left(\frac{\eta_t B^2 \mathcal{L}(w_{t+1})}{2} - 1\right) \\
&\overset{(b)}{\leq} \mathcal{L}(w_{(t)}) - c_t \mathcal{L}(w_t) + \frac{c_t^2 \mathcal{L}(w_{t+1})}{2} \\
\implies \mathcal{L}(w_{(t+1)}) &\leq \frac{1-c_t}{1-0.5c_t^2} \mathcal{L}(w_{(t)}) \overset{(c)}{\leq} \mathcal{L}(w_{(t)}).
\end{aligned}
$$

where in (a) we used $\|\nabla\mathcal{L}(w_t)\|_\star = \left\|\sum_n \exp\left(w_t^T x_n\right) x_n\right\|_\star \leq B\mathcal{L}(w_t)$ from triangle inequality and $\left(\frac{\eta_t B^2 \mathcal{L}(w_{t+1})}{2} - 1\right) \geq 0$ from eq. (43), (b) follows from using $\eta_t \leq \frac{c_t}{B^2 \mathcal{L}(w_t)}$ for some $0 < c_t \leq \sqrt{2}$, and (c) follows as for $0 < c_t \leq \sqrt{2}$, $\frac{1-c_t}{1-0.5c_t^2} \leq 1$. This shows $\mathcal{L}(w_{t+1}) \leq \mathcal{L}(w_t)$ which is a contradiction. $\square$

B.3 Remaining steps in the proof of Theorem 5

The steepest descent updates in eq. (37) can be equivalently written as:

$$
\begin{aligned}
&w_{(t+1)} = w_{(t)} - \eta_t \gamma_t p_{(t)}, \text{ where} \\
&\gamma_t \triangleq \left\|\nabla\mathcal{L}(w_{(t)})\right\|_\star, \text{ and } p_{(t)} \triangleq \frac{\Delta w_{(t)}}{\left\|\nabla\mathcal{L}(w_{(t)})\right\|_\star}, \text{ which satisfies} \\
&\left\langle p_{(t)}, \nabla\mathcal{L}(w_{(t)})\right\rangle = \left\|\nabla\mathcal{L}(w_{(t)})\right\|_\star, \left\|p_{(t)}\right\| = 1.
\end{aligned}
\tag{44}
$$

From eq . 39, using $\gamma_t = \|\nabla\mathcal{L}(w_{(t)})\|_\star = \|\Delta w_{(t)}\|$, we have that

$$
\begin{aligned}
\mathcal{L}(w_{(t+1)}) &\leq \mathcal{L}(w_{(t)}) - \eta_t\gamma_t^2 + \frac{\eta_t^2 B^2 \mathcal{L}(w_{(t)})\gamma_t^2}{2} = \mathcal{L}(w_{(t)}) \left[1 - \frac{\eta_t \gamma_t^2}{\mathcal{L}(w_{(t)})} + \frac{\eta_t^2 B^2 \gamma_t^2}{2}\right] \\
&\overset{(a)}{\leq} \mathcal{L}(w_{(t)}) \exp\left(-\frac{\eta_t \gamma_t^2}{\mathcal{L}(w_{(t)})} + \frac{\eta_t^2 B^2 \gamma_t^2}{2}\right) \\
&\overset{(b)}{\leq} \mathcal{L}(w_{(0)}) \exp\left(-\sum_{u\leq t} \frac{\eta_u \gamma_u^2}{\mathcal{L}(w_{(u)})} + \sum_{u \leq t} \frac{\eta_u^2 B^2 \gamma_u^2}{2}\right),
\end{aligned}
\tag{45}
$$

where we get (a) by using $(1+x) \leq \exp(x)$, and (b) using recursion.

Step 1: Lower bound the unnormalized margin: From eq. (45), we have,

$$
\max_{n\in[N]} \exp\left(-\left\langle w_{(t+1)}, x_n\right\rangle\right) \leq \mathcal{L}(w_{(t+1)}) \leq \mathcal{L}(w_{(0)}) \exp\left(-\sum_{u\leq t} \frac{\eta_u \gamma_u^2}{\mathcal{L}(w_{(u)})} + \sum_{u\leq t} \frac{\eta_u^2 B^2 \gamma_u^2}{2}\right). \tag{46}
$$

By applying $-\log$,

$$
\min_{n\in[N]} \left\langle w_{(t+1)}, x_n\right\rangle \geq \sum_{u\leq t} \frac{\eta_u \gamma_u^2}{\mathcal{L}(w_{(u)})} - \sum_{u\leq t} \frac{\eta_u^2 B^2 \gamma_u^2}{2} - \log \mathcal{L}(w_{(0)}). \tag{47}
$$

Step 2: Upper bound $\left\|\boldsymbol{w_{(t+1)}}\right\|$: Using $\|\Delta w_{(u)}\| = \|\nabla\mathcal{L}(w_{(u)})\|_\star = \gamma_u$, we have,

$$\left\|w_{(t+1)}\right\| \leq \left\|w_{(0)}\right\| + \sum_{u\leq t}\eta_u\left\|\Delta w_{(u)}\right\| \leq \|w_{(0)}\| + \sum_{u\leq t}\eta_u\gamma_u. \tag{48}$$

Step 3: Lower bound on normalized margin: Combining eqs. (47) and (48)$\forall n \in [N]$, we have that

$$\begin{aligned}\frac{\left\langle w_{(t+1)}, x_n\right\rangle}{\left\|w_{(t+1)}\right\|} &\geq \frac{\sum_{u\leq t}\frac{\eta_u\gamma_u^2}{\mathcal{L}(w_{(u)})}}{\sum_{u\leq t}\eta_u\gamma_u + \|w_{(0)}\|} - \left(\frac{\sum_{u\leq t}\frac{\eta_u^2 B^2\gamma_u^2}{2} + \log\mathcal{L}(w_{(0)})}{\|w_{(t+1)}\|}\right). &(49)\\ &:= (I) + (II). &(50)\end{aligned}$$

We look at the two terms separately,

(I) From the duality Lemma 10, we have $\gamma_u = \|\nabla\mathcal{L}(w_{(u)})\|_\star \geq \gamma\mathcal{L}(w_{(u)})$. Hence, $\sum_{u\leq t}\frac{\eta_u\gamma_u^2}{\mathcal{L}(w_{(u)})} \geq \gamma\sum_{u\leq t}\eta_u\gamma_u$ and further using $\sum_{u\leq t}\eta_u\gamma_u \to \infty$ from Lemma 11, we have

$$\frac{\sum_{u\leq t}\frac{\eta_u\gamma_u^2}{\mathcal{L}(w_{(u)})}}{\sum_{u\leq t}\eta_u\gamma_u + \|w_{(0)}\|} \geq \gamma\frac{\sum_{u\leq t}\eta_u\gamma_u}{\sum_{u\leq t}\eta_u\gamma_u + \|w_{(0)}\|} \to \gamma$$

(II) For any bounded $\eta \leq \eta_+$, $\sum_{u\leq t}\frac{\eta_u^2 B^2\gamma_u^2}{2} \leq \frac{\eta_+ B^2}{2}\sum_{u\leq t}\eta_u\gamma_u^2 < \infty$ (from Lemma 11).

Along with using $\|w_{(t)}\| \to \infty$ from Lemma 11, we get $\frac{\sum_{u\leq t}\frac{\eta_u^2 B^2\gamma_u^2}{2} + \log\mathcal{L}(w_{(0)})}{\|w_{(t+1)}\|} \to 0$.

Using the above bounds in (50), we get $\lim_{t\to\infty}\frac{w_{(t+1)}^\top x_n}{\left\|w_{(t+1)}\right\|} \geq \gamma := \max_w \frac{w^\top x_n}{\|w\|}$ □

Appendix C. Adagrad

Lemma 13. *Let $\mathcal{L}(w) = \sum_{n=1}^N \exp\left(-w^\top x_n\right)$, $\|\cdot\|_t$ be some $w_{(t)}$-dependent norm, and $\|\cdot\|_{t,*}$ be its dual, and assume that and $\forall t:\ \|x_n\|_{t,*} \leq 1$. We examine the following adaptive steepest descent update sequence w.r.t adaptive norm $\|\cdot\|_t$:*

$$w_{(t+1)} = w_{(t)} - \eta\gamma_t p_{(t)}, \tag{51}$$

where $\left\|\nabla\mathcal{L}\left(w_{(t)}\right)\right\|_{t,*} \triangleq \gamma_t$ and $p_{(t)}$ is the normalized update satisfying $\left\|p_{(t)}\right\|_t = 1$ and $p_{(t)}^\top\nabla\mathcal{L}\left(w_{(t)}\right) = \left\|\nabla\mathcal{L}\left(w_{(t)}\right)\right\|_{t,*}$.

For these adaptive steepest descent updates, for any initialization, $w_{(0)}$ such that $\eta\mathcal{L}\left(w_{(0)}\right) < 1$, if $w_{(t)}$ minimizes $\mathcal{L}$, i.e., $\mathcal{L}(w_{(t)}) \to 0$, then we have $\sum_{u=0}^\infty \gamma_t^2 < \infty$.

Proof. First we note that since $\|x_n\|_{t,*} \leq 1$ and $\left\|p_{(t)}\right\|_t = 1$.

$$p_{(t)}^\top\nabla^2\mathcal{L}(w)\,p_{(t)} = \sum_{n=1}^N \exp\left(-w^\top x_n\right)\left(x_n^\top p_{(t)}\right)^2 \leq \sum_{n=1}^N \exp\left(-w^\top x_n\right) = \mathcal{L}(w). \tag{52}$$

Additionally, following the arguments of Lemma 12, we can show that $-\eta\gamma_t p_{(t)}$ for $\eta \leq \frac{1}{\mathcal{L}(w_{(0)})}$ is a descent direction, hence from convexity of $\mathcal{L}$, we have

$$\max_{r\in(0,1)} p_{(t)}^\top\nabla^2\mathcal{L}\left(w_{(t)} - r\eta\gamma_t p_{(t)}\right)p_{(t)} \leq \max_{r\in(0,1)}\mathcal{L}\left(w_{(t)} - r\eta\gamma_t p_{(t)}\right) \leq \mathcal{L}\left(w_{(t)}\right). \tag{53}$$

From the Taylor expansion of $\mathcal{L}(w)$

$$\mathcal{L}\left(w_{(t+1)}\right) \leq \mathcal{L}\left(w_{(t)}\right) - \eta\gamma_t \nabla\mathcal{L}\left(w_{(t)}\right)^\top p_{(t)} + \frac{1}{2}\eta^2\gamma_t^2 \max_{r\in(0,1)} p_{(t)}^\top \nabla^2\mathcal{L}\left(w_{(t)} - r\eta\gamma_t p_{(t)}\right) p_{(t)}. \tag{54}$$

Substituting eq. (52) and (53) into eq. 54, we find

$$\begin{aligned}
\mathcal{L}\left(w_{(t+1)}\right) &\leq \mathcal{L}\left(w_{(t)}\right) - \eta\gamma_t \nabla\mathcal{L}\left(w_{(t)}\right)^\top p_{(t)} + \frac{1}{2}\eta^2\gamma_t^2\mathcal{L}\left(w_{(t)}\right) \\
&= \mathcal{L}\left(w_{(t)}\right) - \eta\left(1 - \frac{\eta}{2}\mathcal{L}\left(w_{(t)}\right)\right)\gamma_t^2 \overset{(a)}{\leq} \mathcal{L}\left(w_{(t))}\right) - \frac{\eta}{2}\gamma_t^2,
\end{aligned}$$

where (a) follows from assumption that $\eta \leq \frac{1}{\mathcal{L}(w_{(0)})} \leq \frac{1}{\mathcal{L}(w_{(t)})}$.

Summing over the last equation, we get that $\frac{\eta}{2}\sum_{u=1}^{t}\gamma_u^2 \leq \mathcal{L}\left(w_{(0)}\right) - \mathcal{L}\left(w_{(t)}\right) < \infty$. □

Recall the AdaGrad update $w_{(t+1)} = w_{(t)} - \eta\mathbf{G}_{(t)}^{-1/2}\nabla\mathcal{L}\left(w_{(t)}\right)$, where $\mathbf{G}_{(t)}$ is a diagonal matrix such that

$$\forall i: \ \mathbf{G}_{(t)}[i,i] = \sum_{u=0}^{t}\left(\nabla\mathcal{L}\left(w_{(u)}\right)[i]\right)^2 .$$

We now prove the Theorem 6. Recall the statement,

Theorem 6. *For any linearly separable training data $\{x_n, y_n\}_{n=1}^N$, consider the AdaGrad iterates $w_{(t)}$ from eq.* (13) *for minimizing $\mathcal{L}(w)$ with exponential loss $\ell(u,y) = \exp(-uy)$. For any fixed and bounded step-size $\eta < \infty$, and any initialization of $w_{(0)}$ and $\mathbf{G}_{(0)}$, such that $\frac{\eta}{2}\mathcal{L}\left(w_{(0)}\right) < 1$, and $\left\|\mathbf{G}_{(0)}^{-1/4}x_n\right\|_2 \leq 1$, $\forall i, \forall t: \ \mathbf{G}_{(t)}[i,i] < \infty$.*

Proof. First, we note that AdaGrad is a special case of the adaptive steepest descent algorithm described in Lemma 13 with respect the norm $\|v\|_t = \left\|\mathbf{G}_{(t)}^{1/2}v\right\|_2$. Here the dual norm $\|v\|_{t,*} = \left\|\mathbf{G}_{(t)}^{-1/2}v\right\|_2$.

Also from the definition of $\mathbf{G}_{(t)}$, we have that $\mathbf{G}_{(t)}^{-1}[i,i]$ is monotonically decreasing for all t, and thus $\left\|\mathbf{G}_{(t)}^{-1/2}x_n\right\|_2 \leq \left\|\mathbf{G}_{(0)}^{-1/2}x_n\right\|_2 \leq 1$, and so we can apply Lemma 13. This implies that

$$\begin{aligned}
\infty > \sum_{t=0}^{\infty}\left\|\nabla\mathcal{L}(w_{(t)})\right\|_{t,*}^2 &= \sum_{t=0}^{\infty}\left\|\mathbf{G}_{(t)}^{-1/2}\nabla\mathcal{L}\left(w_{(t)}\right)\right\|_2^2 \\
&= \sum_{i=1}^{d}\sum_{t=0}^{\infty}\left(\nabla\mathcal{L}\left(w_{(t)}\right)[i]\right)^2\left[\sum_{u=0}^{t}\left(\nabla\mathcal{L}\left(w_{(u)}\right)[i]\right)^2\right]^{-1/2} \\
&\geq \sum_{i=1}^{d}\sum_{t=0}^{\infty}\left(\nabla\mathcal{L}\left(w_{(t)}\right)[i]\right)^2\left[\sum_{u=0}^{\infty}\left(\nabla\mathcal{L}\left(w_{(u)}\right)[i]\right)^2\right]^{-1/2} \\
&= \sum_{i=1}^{d}\sqrt{\sum_{t=0}^{\infty}\left(\nabla\mathcal{L}\left(w_{(t)}\right)[i]\right)^2}
\end{aligned}$$

This implies that

$$\forall i, \forall t: \ \mathbf{G}_{(t)}[i,i] = \sum_{u=0}^{t}\left(\nabla\mathcal{L}\left(w_{(u)}\right)[i]\right)^2 \leq \sum_{t=0}^{\infty}\left(\nabla\mathcal{L}\left(w_{(t)}\right)[i]\right)^2 < \infty\,,$$

□

Appendix D. Gradient descent on factorized parameterization

We first prove the Lemma 8 on convergence of $-\nabla\mathcal{L}(w_{(t)})$. This lemma holds for any general linear model (1) with exponential loss $\ell(u,y)=\exp(-uy)$ and are of interest beyond the matrix factorization setup in Section 4.

D.1 Convergence of $-\nabla\mathcal{L}(w_{(t)})$

Lemma 8. *For almost all linearly separable datasets $\{x_n,y_n\}_{n=1}^N$, consider any sequence $w_{(t)}$ that minimizes $\mathcal{L}(w)$ in eq. (1) with exponential loss, i.e., $\mathcal{L}(w_{(t)})\to 0$. If $\frac{w_{(t)}}{\|w_{(t)}\|}$ converges, then for every accumulation point z_∞ of $\left\{\frac{-\nabla\mathcal{L}(w_{(t)})}{\|\nabla\mathcal{L}(w_{(t)})\|}\right\}_t$, $\exists\{\alpha_n\geq 0\}_{n\in S}$ s.t., $z_\infty=\sum\limits_{n\in S}\alpha_n y_n x_n$, where $\bar{w}_\infty=\lim\limits_{t\to\infty}\frac{w_{(t)}}{\|w_{(t)}\|}$ and $S=\{n: y_n\left<\bar{w}_\infty,x_n\right>=\min_n y_n\left<\bar{w}_\infty,x_n\right>\}$ are the indices of the data points with smallest margin to $\bar{w}_\infty$.*

Here *for almost all* $\{x_n,y_n\}$ means that with probability 1 over the dataset $\{x_n,y_n\}$ such that the signed features y_nx_n are drawn independently from a distribution that is absolutely continuous w.r.t the d dimensional Lebesgue measure.

Proof. Without loss of generality assume $\forall n, y_n=1$, else the sign of y can be absorbed into x as $x_n\leftarrow y_nx_n$.

Let $X\in\mathbb{R}^{N\times d}$ denote the data matrix with $x_n\in\mathbb{R}^d$ along the rows of X. Also, for any $J\subseteq[N]$, $X_J\in\mathbb{R}^{|J|\times d}$ denotes the submatrix of X with only the rows corresponding to indices in J.

We have that $\lim_{t\to\infty}\mathcal{L}(w_{(t)})=0$ for strictly monotone loss over separable data, this implies asymptotically $w_{(t)}$ satisfies $Xw_{(t)}>0, \|w_{(t)}\|\to\infty$.

Since $w_{(t)}$ converges in direction to $\bar{w}_\infty$, we can write $w_{(t)}=g(t)\bar{w}_\infty+\rho_{(t)}$ for a scalar $g(t)=\|w_{(t)}\|\to\infty$ and vector $\rho(t)\in\mathbb{R}^d$ such that $\frac{\rho(t)}{g(t)}\to 0$. Additionally, this implies $\forall n$, $X\bar{w}_\infty>0$.

We introduce some additional notation:

- Denote the asymptotic margin of x_n as $\bar{\gamma}_n:=\left<x_n,\bar{w}_\infty\right>$. Additionally, we define the following:
  - Let $\gamma=\min_n\left<x_n,\bar{w}_\infty\right>=\min_n e_n^\top X\bar{w}_\infty>0$ denote the smallest margin, where $e_n\in\mathbb{R}^N$ are standard basis.
  - Let $S:=\{n:\left<x_n,\bar{w}_\infty\right>=\gamma\}$ denote the indices of support vectors of $\bar{w}_\infty$.
  - Denote the second smallest margin of $\bar{w}_\infty$ as $\bar{\gamma}:=\min_{n\notin S}\left<x_n,\bar{w}_\infty\right>>\gamma$.
- Define $\alpha_n(t):=\exp(-\left<\rho(t),x_n\right>)$ and let $\alpha(t)\in\mathbb{R}^N$ be a vector of $\alpha_n(t)$ stacked. For any $J\subset[N]$ and $\alpha\in\mathbb{R}^N$, similar to X_J, let $\alpha_J\in\mathbb{R}^{|J|}$ be the sub-vector with components corresponding to the indices in J
- $B=\max_n\|x_n\|_2$,

Since $\|\rho(t)\|/g(t)\to 0$ and $\gamma,\bar{\gamma}>0$, we have $\forall\epsilon_1,\epsilon_2>0$, $\exists t_{\epsilon_1},t_{\epsilon_2}$ such that

$$\begin{aligned}&\forall t>t_{\epsilon_1},\ \forall n,\quad \left<\rho(t),x_n\right>\leq\|\rho(t)\|_2B\leq\epsilon_1\gamma g(t),\ \text{and}\\&\forall t>t_{\epsilon_2},\ \forall n,\quad \left<\rho(t),x_n\right>\geq-\|\rho(t)\|_2B\geq-\epsilon_2\bar{\gamma}g(t)\end{aligned}\tag{55}$$

The first prove the following claim:

Claim 1. *For almost all $\{x_n\}$, $|S|<d$ and $\sigma_{|S|}(X_S)>0$, where $\sigma_k(A)$ is the k^{th} singular value of A.*

Proof. Since, $S=\{n:\left<\bar{w}_\infty,x_n\right>=\gamma\}$, we have $X_S\bar{w}_\infty=\gamma 1_S\in\mathbb{R}^{|S|}$.

If X is randomly drawn from a continuous distribution, for any fixed subset J if $|J| > d$, the column span of X_J is rank deficient and will miss any fixed vector v that is independent of X with probability 1. Thus,

$$\mathbb{R}^{|J|} \ni 1_J \notin \text{colspan}(X_J), \text{ for almost all } X_J \in \mathbb{R}^{|J|\times d}. \tag{56}$$

Since we always have $1_S \in \text{colspan}(X_S)$, this implies for almost all X, $|S| \leq d$ and $\sigma_{|S|}(X_S) > 0$. $\square$

Exponential loss: For exponential loss, the gradient at $w_{(t)}$ is given by

$$\begin{aligned}-\nabla\mathcal{L}(w_{(t)}) &= \sum_{n\in S}\exp(-\gamma g(t))\exp\big(-\rho(t)^\top x_n\big)x_n + \sum_{n\in S^c}\exp(-\bar{\gamma}_n g(t))\exp\big(-\rho(t)^\top x_n\big)x_n \\ &:= I(t) + II(t),\end{aligned} \tag{57}$$

where $I(t) = \sum_{n\in S}\exp(-\gamma g(t))\exp\big(-\rho(t)^\top x_n\big)x_n$ and $II(t) = \sum_{n\notin S}\exp(-\bar{\gamma}_n g(t))\exp\big(-\rho(t)^\top x_n\big)x_n$.

To prove the lemma, we need to show that the gradient are dominated by the positive span of support vectors. Towards this goal, we will now show that $\lim\limits_{t\to\infty}\frac{\|II(t)\|}{\|I(t)\|} = 0$.

Recall that $\alpha(t) = [\alpha_n(t)]_n$ is defined as $\alpha_n(t) = \exp(-\langle\rho(t), x_n\rangle)$ and $\alpha_S(t) \in \mathbb{R}^{|S|}$ is a subvector restricted to indices in S. The following are true for any $\epsilon_1, \epsilon_2 > 0$.

Step 1. *Lower bound on $I(t)$:* There exists t_{ϵ_1} such that for all $t > t_{\epsilon_1}$, we have

$$\begin{aligned}\|I\|_2 &= \exp(-\gamma g(t))\|X_S\alpha_S(t)\|_2 \geq \exp(-\gamma g(t))\sigma_{|S|}(X_S)\|\alpha_S(t)\|_2 \\ &\geq \exp(-\gamma g(t))\sigma_{|S|}(X_S)\max_{n\in S}\alpha_n(t) \\ &\overset{(a)}{\geq} \sigma_{|S|}(X_S)\exp(-(1+\epsilon_1)\gamma g(t)) := C_1\exp(-(1+\epsilon_1)\gamma g(t)),\end{aligned} \tag{58}$$

where (a) follows from (55), from which we get $\alpha_n(t) = \exp(-\langle\rho(t), x_n\rangle) \geq \exp(-\epsilon_1\gamma g(t))$, and $C_1 = \sigma_{|S|}(X_S) > 0$ is a constant independent of t.

Step 2. *Upper bound on $II(t)$:* Again, for large enough $t > t_{\epsilon_2}$, we have

$$\begin{aligned}\|II(t)\|_2 &= \sum_{n\notin S}\exp(-\bar{\gamma}_n g(t))\exp\big(-\rho(t)^\top x_n\big)x_n \leq N\max_n\exp(-\bar{\gamma}_n g(t))\alpha_n\|x_n\|_2 \\ &\overset{(a)}{\leq} \exp(-\bar{\gamma}g(t))BN\max_n\alpha_n \\ &\overset{(b)}{\leq} BN\exp(-(1-\epsilon_2)\bar{\gamma}g(t)) := C_2\exp(-(1-\epsilon_2)\bar{\gamma}g(t)),\end{aligned} \tag{59}$$

where (a) uses $\forall n\notin S, \bar{\gamma}_n \geq \bar{\gamma}$ (recall that $\bar{\gamma}$ is the second smallest margin to $\bar{w}_\infty$) and (b) follows from (55), using $\alpha_n = \exp(-\langle\rho(t), x_n\rangle) \leq \exp(\epsilon_2\bar{\gamma}g(t))$, and $C_2 = BN > 0$ is again a constant independent of t.

Step 3. *Remaining steps in the proof:* By combining (58) and (59) using $\epsilon_1 = {(\bar{\gamma}-\gamma)}/{4\gamma}$ and $\epsilon_2 = {(\bar{\gamma}-\gamma)}/{4\bar{\gamma}}$ and an appropriate constant $C > 0$, we have for any norm $\|.\|$

$$\frac{\|II(t)\|}{\|I(t)\|} \leq C\exp\bigg(-\frac{1}{2}(\bar{\gamma}-\gamma)g(t)\bigg) \overset{(a)}{\to} 0, \tag{60}$$

where (a) follows from $\bar{\gamma} > \gamma$ and $g(t) = \|w_{(t)}\| \to \infty$.

Finally, note that $-\frac{\nabla\mathcal{L}(w_{(t)})}{\|\nabla\mathcal{L}(w_{(t)})\|} = \frac{I(t)}{\|I(t)+II(t)\|} + \frac{II(t)}{\|I(t)+II(t)\|}$. Since $\Big\|\frac{II(t)}{\|I(t)+II(t)\|}\Big\| \leq \frac{\|II(t)\|/\|I(t)\|}{1-\|II(t)\|/\|I(t)\|} \overset{t\to\infty}{\to} 0$, and $I(t) \propto \sum_{n\in S}\alpha_n(t)x_n$ for $\alpha_n(t) > 0$, we have shown that every limit point of $-\frac{\nabla\mathcal{L}(w_{(t)})}{\|\nabla\mathcal{L}(w_{(t)})\|} \to \sum_{n\in S}\alpha_n x_n$ for some $\alpha_n > 0$.

Recall that in the beginnning of the proof we made a change of variable that $x_n \leftarrow y_n x_n$. Reversing this change of variable finishes the proof for exponential loss. $\square$

D.2 Proof of Theorem 7

Theorem 7. *For almost all datasets $\{X_n, y_n\}_{n=1}^N$ separable by a p.s.d. linear classifier, consider the gradient descent iterates $U_{(t)}$ in eq.* (18) *for minimizing $\bar{\mathcal{L}}(U)$ with the exponential loss $\ell(u,y) = \exp(-uy)$ and the corresponding sequence of linear predictors $W_{(t)}$ in eq.* (19)*. For any full rank initialization $U_{(0)}$ and any finite step-size sequence $\{\eta_t\}_t$, if $W_{(t)}$ asymptotically minimizes $\mathcal{L}$, i.e., $\mathcal{L}(W_{(t)}) \to 0$, and additionally the updates $U_{(t)}$ and the gradients $\nabla\mathcal{L}(W_{(t)})$ converge in direction, then the limit direction $\bar{U}_\infty = \lim\limits_{t\to\infty} \frac{U_{(t)}}{\|U_{(t)}\|_*}$ is a scaling of a first order stationary point (f.o.s.p) of the following non-convex optimization problem*

$$\bar{U}_\infty \propto \textit{f.o.s.p.} \min_{U \in \mathbb{R}^{d\times d}} \|U\|_2^2 \quad \textit{s.t.,} \quad \forall n, y_n \left\langle UU^\top, X_n \right\rangle \geq 1. \tag{20}$$

Proof. In this proof, $\|.\|_F$, $\|.\|_*$, and $\|.\|_{\text{op}}$ denote the Frobenious norm, nuclear norm, and operator norm, respectively.

From the assumption of theorem, we have that $U_{(t)}$ converges in direction. Let $\bar{U}_\infty = \lim\limits_{t\to\infty} \frac{U_{(t)}}{\|U_{(t)}\|_F}$. Noting that for $W_{(t)} = U_{(t)}{U_{(t)}}^\top$, $\|W_{(t)}\|_* = \|U_{(t)}\|_F^2$, we have that $\lim\limits_{t\to\infty} \frac{W_{(t)}}{\|W_{(t)}\|_*} = \lim\limits_{t\to\infty} \frac{U_{(t)}}{\|U_{(t)}\|_F} \frac{U_{(t)}}{\|U_{(t)}\|_F}^\top = \bar{U}_\infty \bar{U}_\infty^\top$. Denote $\bar{W}_\infty = \lim\limits_{t\to\infty} \frac{W_{(t)}}{\|W_{(t)}\|_*} = \bar{U}_\infty \bar{U}_\infty^\top$.

Since $W_{(t)}$ minimizes a strictly monotone loss, we have that $\|W_{(t)}\|_* \to \infty$ and $\forall n, y_n \left\langle \bar{W}_\infty, X_n \right\rangle > 0$. Let $\gamma = \min_n y_n \left\langle \bar{W}_\infty, X_n \right\rangle$ denote the margin of $\bar{W}_\infty$ and $S = \{n : y_n \left\langle \bar{W}_\infty, X_n \right\rangle = \gamma\}$ denote the indices of the support vectors of $\bar{W}_\infty$.

In order to prove the theorem, we can can equivalently show that a positive scaling of $\bar{U}_\infty$ given by $\bar{\bar{U}}_\infty = \bar{U}_\infty / \sqrt{\gamma}$ is the first order stationary point of eq. (20).

In the remainder of the proof we show that $\bar{\bar{U}}_\infty$ satisfies the following KKT optimality conditions of (20):

To show:

$$y_n \left\langle \bar{\bar{U}}_\infty \bar{\bar{U}}_\infty^\top, X_n \right\rangle \geq 1 \text{ and } \exists \alpha \geq 0 \text{ s.t.,} \qquad \textbf{(primal and dual feasibility)} \tag{61}$$

$$\forall i \notin S : \alpha_n = 0 \text{ and} \qquad \textbf{(complementary slackeness)} \tag{62}$$

$$\bar{\bar{U}}_\infty = \sum_n \alpha_n y_n X_n \bar{\bar{U}}_\infty. \qquad \textbf{(stationarity)} \tag{63}$$

Primal feasibility This holds by definition since $\bar{\bar{U}}_\infty \bar{\bar{U}}_\infty^\top = \bar{W}_\infty / \gamma$ has unit margin by the scaling.

Dual feasibility and complementary slackness Denote $Z_{(t)} = -\nabla\mathcal{L}(W_{(t)}) = \sum_n \exp\bigl(-y_n \left\langle W_{(t)}, X_n \right\rangle\bigr) y_n X_n$. From the assumptions in the theorem, we have that $Z_{(t)}$ converge in direction. Let $\bar{Z}_\infty = \lim\limits_{t\to\infty} \frac{Z_{(t)}}{\|Z_{(t)}\|_{\text{op}}}$. In addition, we also assume that $\mathcal{L}(W_{(t)}) \to 0$ and that $U_{(t)}$ convergence in direction, which in turn implies convergence in direction of $W_{(t)} = U_{(t)}{U_{(t)}}^\top$. Thus, from Lemma 8, we have $\bar{Z}_\infty = \sum_{n\in S} \alpha_n y_n X_n$ for some $\{\alpha_n\}_{n=1}^N$ such that $\alpha_n \geq 0$ and $\alpha_n = 0$ for all $n \notin S$. We propose this $\{\alpha_n\}_{n=1}^N$ as our candidate dual certificate, which satisfies both dual feasibility and complementary slackness.

Stationarity: To prove the theorem, we now need to show that: $\bar{\bar{U}}_\infty = D\bar{Z}_\infty \bar{\bar{U}}_\infty$, for some positive scalar D, or equivalently that $\bar{U}_\infty = D\bar{Z}_\infty \bar{U}_\infty$. This forms the main part of the proof.

Using the assuptions in the theorem, we have that $U_{(t)}$ and $Z_{(t)}$ converges in direction, we introducing the following notation to conveniently represent these quantities.

1. Since $\frac{U_{(t)}}{\|U_{(t)}\|_F} \to \bar{U}_\infty$, we define $g(t)$ and $\rho_{(t)}$ satisfying the following,

$$U_{(t)} = \bar{U}_\infty g(t) + \rho_{(t)} \;\text{ s.t., }\; g(t) := \|U_{(t)}\|_F \to \infty \text{ and } \frac{\rho_{(t)}}{g(t)} \to 0. \tag{64}$$

2. For exponential loss, $\mathcal{L}(W_{(t)}) \to 0$ implies $Z_{(t)} = -\nabla\mathcal{L}(W_{(t)}) \to 0$. Thus, using the previously introduced notation $\bar{Z}_\infty = \lim_{t\to\infty} \frac{Z_{(t)}}{\|Z_{(t)}\|_{\text{op}}}$, we define $p(t)$ and $\zeta_{(t)}$ as follows

$$Z_{(t)} = -\nabla\mathcal{L}(W_{(t)}) = \bar{Z}_\infty p(t) + \zeta_{(t)} \text{ s.t., } p(t) := \left\|Z_{(t)}\right\|_{\text{op}} \to 0 \text{ and } \frac{\zeta_{(t)}}{p(t)} \to 0. \tag{65}$$

To show stationarity, we need to show that $\bar{U}_\infty = D\bar{Z}_\infty\bar{U}_\infty$, which requires that the columns of $\bar{W}_\infty$ are spanned subset of eigenvectors of $\bar{Z}_\infty$ that correspond to the same eigen value.

Let $\Delta U_{(t)} = U_{(t+1)} - U_{(t)}$. Substituting expressions of $U_{(t)}$ and $Z_{(t)}$ from (64) and (65), respectively, for the updates $\Delta U_{(t)}$ from eq. (18), we have

$$\begin{aligned}\Delta U_{(t)} = \eta_t Z_{(t)} U_{(t)} &= \eta_t p(t) g(t)\left[\bar{Z}_\infty\bar{U}_\infty + \bar{Z}_\infty\frac{\rho_{(t)}}{g(t)} + \frac{\zeta_{(t)}}{p(t)}\bar{U}_\infty\right]\\ &\overset{(a)}{=} \eta_t p(t) g(t)[\bar{Z}_\infty\bar{U}_\infty + \delta_{(t)}]\end{aligned} \tag{66}$$

where in (a) we collect all the diminishing terms into $\delta_{(t)} = \bar{Z}_\infty\frac{\rho_{(t)}}{g(t)} + \frac{\zeta_{(t)}}{p(t)}\bar{U}_\infty \to 0$ as from eqs. (64)–(65), we have $\frac{\rho_{(t)}}{g(t)}, \frac{\zeta_{(t)}}{p(t)} \to 0$ and $\bar{Z}_\infty$ and $\bar{W}_\infty$ are finite quanitities independent of t.

Summing over t, we have that

$$U_{(t)} - U_{(0)} = \bar{Z}_\infty\bar{U}_\infty\sum_{u<t}\eta_u p(u)g(u) + \sum_{u<t}\delta_{(u)}\eta_u p(u) g(u) \tag{67}$$

Claim 2. $\left\|\bar{Z}_\infty\bar{U}_\infty\right\| > 0$ *and* $\sum_{u<t}\eta_u p(u)g(u) \to \infty$.

Proof. First, recall that for the limit direction $\bar{W}_\infty = \bar{U}_\infty\bar{U}_\infty^\top$, $\min_n y_n\left\langle\bar{W}_\infty, X_n\right\rangle = \gamma > 0$ and $\bar{Z}_\infty = \sum_{n\in S}\alpha_n y_n X_n$ for $\alpha_n \geq 0$. Thus, $\left\langle Z_\infty, \bar{W}_\infty\right\rangle = \left\langle\bar{Z}_\infty\bar{U}_\infty, \bar{U}_\infty\right\rangle = \sum_{n\in S}\alpha_n\gamma > 0$ for $\bar{U}_\infty \neq 0$, and hence $\left\|\bar{Z}_\infty\bar{U}_\infty\right\| > 0$.

Secondly, since $\delta_{(t)} \to 0$ in eq. (67), $\exists t_0$ such that $\forall t > t_0$, $\left\|\delta_{(t)}\right\| \leq 1$ and since all the incremental updates to gradient descent are finite, we have that $\sup_t\left\|\delta_{(t)}\right\| < \infty$. Additionally, since $p(t) = \left\|Z_{(t)}\right\|_{\text{op}}$ and $g(u) = \left\|U_{(t)}\right\|_F$ are positive, we have that $b_t = \sum_{u<t}\eta_u p(u)g(u)$ is monotonic increasing, thus if $\limsup_{t\to\infty} b_t = \infty$ then $\lim_{t\to\infty} b_t = \infty$. On contrary, if $\limsup_{t\to\infty} b_t = C < \infty$, then we have from eq. (67), $\left\|U_{(t)}\right\| \leq \left\|U_{(0)}\right\| + \left\|\bar{Z}_\infty\bar{U}_\infty\right\|C + \left(\sup_t\left\|\delta_{(t)}\right\|\right)C < \infty$ which is a contradiction to $\left\|U_{(t)}\right\| \to \infty$. $\square$

From the above claim, we have that the sequence $b_t = \sum_{u<t}\eta_u p(u)g(u)$ is monotonic increasing and diverging. Thus, for $a_t = \sum_{u<t}\delta_{(u)}\eta_u p(u)g(u)$, using Stolz-Cesaro theorem (Theorem 17), we have that

$$\begin{aligned}\lim_{t\to\infty}\frac{a_t}{b_t} = \lim_{t\to\infty}\frac{\sum_{u<t}\delta_{(u)}\eta_u p(u)g(u)}{\sum_{u<t}\eta_u p(u)g(u)} = \lim_{t\to\infty}\frac{a_{t+1}-a_t}{b_{t+1}-b_t} = \lim_{t\to\infty}\delta_{(t)} = 0.\\ \implies \text{for } \widetilde{\delta}_{(t)} \to 0, \text{ we have } \sum_{u<t}\delta_{(u)}\eta_u p(u)g(u) = \widetilde{\delta}_{(t)}\sum_{u<t}\eta_u p(u)g(u),\end{aligned} \tag{68}$$

Subtituting eq. (68) in eq. (67), we have

$$U_{(t)} \stackrel{(a)}{=} \left[\bar{Z}_\infty \bar{U}_\infty + \delta'_{(t)}\right] \left[\sum_{u<t} \eta_u p(u) g(u)\right] \tag{69}$$

$$\implies \frac{U_{(t)}}{\|U_{(t)}\|} = \frac{\bar{Z}_\infty \bar{U}_\infty + \delta'_{(t)}}{\left\|\bar{Z}_\infty \bar{U}_\infty + \delta'_{(t)}\right\|_F} \xrightarrow{(a)} \frac{\bar{Z}_\infty \bar{U}_\infty}{\|\bar{Z}_\infty \bar{U}_\infty\|} \tag{70}$$

$$\implies \bar{U}_\infty = \lim_{t\to\infty} \frac{U_{(t)}}{\|U_{(t)}\|} = \frac{1}{\|\bar{Z}_\infty \bar{U}_\infty\|} \bar{Z}_\infty \bar{U}_\infty, \tag{71}$$

where in (a) we absorbed all the diminishing terms into $\delta'_{(t)} = \widetilde{\delta}_{(t)} + U_{(0)}/\sum_{u<t} \eta_u p(u) g(u) \to 0$ and (b) follows since $\bar{Z}_\infty \bar{U}_\infty \neq 0$ and hence dominates $\widetilde{\delta}_{(t)}$.

We have thus shown that $\bar{U}_\infty = D\bar{Z}_\infty \bar{U}_\infty$ for $D = \frac{1}{\|\bar{Z}_\infty \bar{U}_\infty\|}$ which completes the proof of the theorem. ◻

Appendix E. Preliminaries

Lemma 14 (Sub-differentials of norms)**.** *For a generic norm $\|v\|$ for $v \in \mathcal{V}$, recall the dual norm $\|y\|_\star = \sup_{\|v\|\leq 1} \langle y, v\rangle$. The sub-differential of a norm $\|.\|$ at v is defined as $\partial\|v\| = \{y : \forall \Delta \in \mathcal{V},\ \|v + \Delta\| \geq \|v\| + \langle y, \Delta\rangle\}$.*

We have the following results on the properties on the sub-differentials are readily established:

1. $\partial\|v\| = \{y : \|y\|_\star = 1,$ and $\langle y, v\rangle = \|v\|\}$

2. $y \in \partial\|v\|^2$ if and only if $v \in \partial\|v\|^2$

3. if there exists $v_1, v_2 \in \mathcal{V}$ and $g \in \mathcal{V}^\star$ such that $g \in \partial\|v_1\|$ and $g \in \partial\|v_2\|$, then $forall \alpha, \beta > 0$, $g \in \partial\|\alpha v_1 + \beta v_2\|$.

Proof. 1. It can be easily verified that $\{y : \|y\|_\star = 1, \text{and } \langle y, v\rangle = \|v\|\} \subseteq \partial\|v\|$ Conversely, $\forall y \in \partial\|v\|$, from the definition, we have $\forall \Delta$, $\|v\| + \|\Delta\| \geq \|v + \Delta\| \geq \|v\| + \langle y, \Delta\rangle \implies \|y\|_\star = \sup_{\Delta\neq 0}\langle \frac{\Delta}{\|\Delta\|}, y\rangle \leq 1$. Using $\|y\|_\star \leq 1$ along with $\Delta = -v$, we have $\langle y, v\rangle \geq \|v\| = \sup_{\|y\|_\star\leq 1} \langle v, y\rangle \Rightarrow \langle y, v\rangle = \sup_{\|y\|_\star\leq 1} \langle v, y\rangle \|v\|$, which by homogeneity of norms implies $\|y\|_\star = 1$.

2. From above result, $y \in \partial\frac{1}{2}\|v\|^2 \Leftrightarrow \|y\|_* = \|v\|$ and $\langle y, v\rangle = \|v\|^2 = \|y\|_\star^2 \Leftrightarrow v \in \partial\|y\|_\star^2$.

3. $g \in \partial\|v_1\| \cap \partial\|v_2\|$ implies $\|g\|_\star = 1$, $\|v_1\| = \langle g, v_1\rangle$, and $\|v_2\| = \langle g, v_2\rangle$. Using triangle inequality, $\|\alpha v_1 + \beta v_2\| \leq \alpha\|v_1\| + \beta\|v_2\| = \langle g, \alpha v_1 + \beta v_2\rangle \leq \sup_{\|y\|_\star\leq 1} \langle y, \alpha v_1 + \beta v_2\rangle = \|\alpha v_1 + \beta v_2\| \implies \|\alpha v_1 + \beta v_2\| = \langle g, \alpha v_1 + \beta v_2\rangle$.

◻

Lemma 15 (Limit points of a compact sets)**.** *If $\{a_t\}_{t=1}^\infty$ is a sequence contained in a compact set $a_t \in C$, then there exists at least one limit point of $\{a_t\}$ in C. That is, $\exists a^\infty \in C$ and a subsequence $\{a_{t_k}\}_{k=1}^\infty$, such that $\lim_{k\to\infty} a_{t_k} = a^\infty$.*

Theorem 16 (L-Hopital's Rule, proof in Theorem 30.2 of Ross (1980))**.** *Let $s \in \mathbb{R} \cup \{-\infty, \infty\}$, and $f(x)$ and $g(x)$ be continuous and differentiable functions such that $\lim_{x\to s} \frac{f'(x)}{g'(x)} = L$ exists. If either (a) $\lim_{x\to s} f(x) = \lim_{x\to s} g(x) = 0$, or (b) $\lim_{x\to s} |g(x)| = \infty$, then $\lim_{x\to s} \frac{f(x)}{g(x)}$ exists and is equal to L*

Theorem 17 (Stolz–Cesaro theorem, proof in Theorem 1.22 of Muresan & Muresan (2009))**.** *Assume that $\{a_k\}_{k=1}^\infty$ and $\{b_k\}_{k=1}^\infty$ are two sequences of real numbers such that $\{b_k\}_{k=1}^\infty$ is strictly monotonic and* *diverging (i.e., monotonic increasing with $b_k \to \infty$ or monotonic decreasing with $b_k \to -\infty$). Additionally, if $\lim_{k\to\infty} \frac{a_{k+1}-a_k}{b_{k+1}-b_k} = L$ exists, then $\lim_{k\to\infty} \frac{a_k}{b_k}$ exists and is equal to L.*