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 bias of generic optimization methods, including 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 ask the question of whether the global minimum (among the many possible global minima) reached by optimization can be characterized in terms of the potential or norm, and independently of hyperparameter choices such as step size and momentum.

1. Introduction

Implicit bias introduced by optimization algorithms plays a crucial role in the generalization ability of learned models, especially in deep learning as it provides effective regularization 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). Deep neural networks, despite having large capacity, often show remarkably good generalization performance even in the absence of explicit regularization (Zhang et al., 2017) or early stopping (Hoffer et al., 2017). In particular, for large networks where the training objective has many global minima, most of which generalize poorly, optimizing using variants of gradient descent or other local search methods implicitly biases the solutions to special global minima that generalize well.

The implicit bias depends on the choice of optimization algorithm, and changing the optimization algorithm, or even changing associated hyperparameter can change the implicit bias and hence change the effective regularization and thus the generalization properties of the solutions. For example, Wilson et al. (2017) showed that on standard deep learning architectures, common variants of SGD methods for 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.

To rigorously understand what drives generalization in deep neural networks, it is therefore important to understand what the implicit biases are for different algorithms. Can we explicitly relate between the choice of algorithm and the implicit bias? Can we precisely characterize which global minima different optimization algorithms converge to? How does this depend on the loss functions? What other choices — including initialization, step size, momentum, stochasticity, and adaptivity, does the implicit bias depend on?

We already have an understanding of the implicit bias of gradient descent for linear models. In particular, we know that for underconstrained least squares linear regression problems, gradient descent always converges 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 in direction of the hard margin support vector machine solution, and therefore the decision boundary converges to this max margin solution.

Can we extend this characterization to generic 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 manifold, and steepest descent with respect to a generic norm? Can we characterize the implicit bias in terms of the potential or norm?

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 losses with 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 play a significant role, 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 and robust characterization of the implicit bias of the limit direction for generic steepest descent. We also study the adaptive method AdaGrad and optimization over a matrix factorization. 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 sizes 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.

**Property 1 (Losses with a unique finite root).** The loss function \( \ell(f(x); y) \) between a predictor \( f(x) \) and label \( y \) has a unique minimum at a finite value of \( f(x) \). More precisely, for any \( y \) and sequence \( \tilde{y}_t \), if \( \ell(\tilde{y}_t, y) \to \inf_{\tilde{y}} \ell(\tilde{y}, y) \), then \( \tilde{y}_t \to y \), where we assume without loss of generality that \( \inf_{\tilde{y}} \ell(\tilde{y}, y) = 0 \) and the root of \( \ell(\tilde{y}, \tilde{y}) \) is at \( \tilde{y} = y \).

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 \( u(x) = \langle w, x \rangle \) with parameters \( w \in \mathbb{R}^d \) is given by,

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

We are particularly interested in the case when \( N < d \) and the observations are realizable, i.e. \( \min_w L(w) = 0 \). In this case \( L(w) \) is underconstrained and we have multiple global minima denoted by \( G = \{w : L(w) = 0\} \). These multiple global minima are all valid solutions for the empirical optimization problem of minimizing \( L(w) \), but might differ in their behavior on the population objective \( \mathbb{E}\ell(\langle w, x \rangle; y) \). Thus, in order to understand generalization behavior, it is important to understand the properties of the specific global minimum to which different optimization algorithms converge to.

Previous studies on the implicit bias of gradient descent (GD) focused on squared loss, but for underdetermined problems, the implicit global minima of \( L(w) \) is the same for any loss \( \ell \) with a unique finite root (Property 1), and the same arguments can be used for any loss in this family, including e.g., the Huber loss, the truncated squared loss.

2.1 Gradient descent

Gradient descent updates \( w_{t+1} = w(t) - \eta_t \nabla L(w(t)) \) with any step size \( \{\eta_t\} \) that minimizes \( L(w) \) in eq. \( 1 \) always converge to the global minimum that is closest to initialization \( w(0) \) in \( \ell_2 \) distance, \( \arg\min_{w \in G} \|w - w(0)\|_2 \). This can be easily seen as at \( w \), the gradients \( \nabla L(w) = \sum_n \ell'(\langle w, x_n \rangle; y_n)x_n \) are always constrained to a 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 \( N \)-dimensional subspace, there is a unique \( w \) that satisfies the \( N \) linear constraints \( G = \{w : \langle w, x_n \rangle = y_n, \forall n \in [N]\} \).
It is also evident that this implicit bias also holds for gradient descent updates with instance-wise stochastic gradients, where in place of the gradient over the entire dataset $\nabla L(w(t))$, we use stochastic gradients computed from a uniformly random subset of instances $S_t \subseteq [N]$ as defined below:
\[
\nabla \mathcal{L}(w(t)) = \sum_{n \in S_t \subseteq [N]} \nabla w \ell(w(t), x_n); y_n).
\]
(2)

Moreover, when initialized with $w(0) = 0$, the implicit bias characterization also extends to the following generic momentum and acceleration methods:
\[
w(t+1) = w(t) + \beta_t \Delta w(t-1) - \eta \nabla \mathcal{L}(w(t)) + \gamma_t \Delta w(t-1),
\]
(3)

where $\Delta w(t-1) = w(t) - w(t-1)$. This includes Nesterov's acceleration ($\beta_t = \gamma_t$) \cite{Nesterov1983} and Polyak's heavy ball momentum ($\gamma_t = 0$) \cite{Polyak1964}.

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, and further the dependence on initialization is captured through a succinct characterization. Can we get such characterizations also for other optimization algorithms? That is characteristic the bias in terms of the optimization geometry and initialization, but independent of other choices of step size, momentum, and stochasticity.

2.2 Mirror descent

Mirror descent (MD) \cite{Beck2003, Nemirovskii1983} 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 $\psi$ as,
\[
w(t+1) = \arg\min_w \eta \langle w, \nabla \mathcal{L}(w(t)) \rangle + D_\psi(w, w(t)),
\]
(4)

where $D_\psi(w, w') = \psi(w) - \psi(w') - \langle \nabla \psi(w'), w - w' \rangle$ is the Bregman divergence \cite{Bregman1967} w.r.t. $\psi$.

Equivalently, for the unconstrained optimization, the update in eq. (4) can also be written as:
\[
\nabla \psi(w(t+1)) = \nabla \psi(w(t)) - \eta \nabla \mathcal{L}(w(t)).
\]
(5)

For a strongly convex potential $\psi$, the link function $\nabla \psi$ is invertible and hence, the above updates are uniquely defined. $\nabla \psi(w)$ is often referred as the dual variable corresponding to the primal variable $w$.

Examples of strongly convex potentials $\psi$ for mirror descent include, the squared $\ell_2$ norm $\psi(w) = \frac{1}{2} \|w\|^2_2$, 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 $w$, wherein $\psi(w)$ is the entropy potential on the singular values of $w$; general quadratic potentials $\psi(w) = \|w\|_B^2 = w^T D w$ for any positive definite matrix $D$; and the squared $\ell_p$ norm for $p \in (1, 2)$.

From the updates given in 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 $\{x_n \in \mathcal{N} \} + \nabla \psi(w(0))$. The arguments for gradient descent can now be generalized to get the following result.

**Theorem 1.** For any loss $\ell$ with a unique finite root (Property 1), any initialization $w(0)$, and any step size sequence $\{\eta_t\}$, consider the mirror descent updates $w(t)$ given in eq. (4) w.r.t some strongly convex potential $\psi$. If the limit point of the iterates $w_\infty = \lim_{t \to \infty} w(t)$ is a global minimum for $\mathcal{L}$ with $\mathcal{L}(w_\infty) = 0$, then
\[
w_\infty = \arg\min_{w: v_n(w, x_n) = y_n} \mathcal{L}(w).
\]
(6)

In particular, if we start at $w(0) = \arg\min_w \psi(w)$, then we get to $w_\infty = \arg\min_{w \in \mathcal{G}} \psi(w)$, where recall that $\mathcal{G} = \{w: v_n(w, x_n) = y_n\}$ is the set of global minima for $\mathcal{L}(w)$.

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

**Dual momentum:**
\[
\nabla \psi(w(t+1)) = \nabla \psi(w(t)) + \beta_t \Delta z(t-1) - \eta_t \nabla \mathcal{L}(w(t)) + \gamma_t \Delta w(t-1)
\]
(7)

**Primal momentum:**
\[
\nabla \psi(w(t+1)) = \nabla \psi(w(t)) + \beta_t \Delta w(t-1) - \eta_t \nabla \mathcal{L}(w(t)) + \gamma_t \Delta w(t-1)
\]
(8)
where \( \Delta z_{t-1} = w_{t-1} = 0 \), and for \( t \geq 1 \), 
\[
\Delta z_{t-1} = \nabla \psi(w_{t-1}) - \nabla \psi(w_{t-1}) 
\] 
and \( \Delta w_{t-1} = w_{t-1} - w_{t-1} \) are the momentum terms in the primal and dual space, respectively; and \( \{ \beta_t \geq 0, \gamma_t \geq 0 \} \) are the momentum parameters.

If we initialize to \( w_0 = \text{argmin}_w \psi(w) \), then even with dual momentum, the iterates \( \nabla \psi(w_{t}) \) remains in the data manifold, leading to the following extension of Theorem 1.

**Theorem 1a.** Under the conditions in Theorem 1, if initialized at \( w_0 = \text{argmin}_w \psi(w) \), then the mirror descent updates with dual momentum also satisfies (6), i.e. for all \( \{ \eta_t \}, \{ \beta_t \}, \{ \gamma_t \} \), if the \( w_t \) from eq. (7) converges to \( w_\infty \) such that \( \mathcal{L}(w_\infty) = 0 \), then \( w_\infty = \text{argmin}_{w \in \mathcal{G}} \mathcal{D}_\psi(w, w_0) = \text{argmin}_{w \in \mathcal{G}} \psi(w) \).

Furthermore, Theorem 1[1a] also hold when stochastic gradients defined in eq. (2) are used in place of \( \nabla \mathcal{L}(w_t) \) in the mirror descent updates.

For quadratic potentials \( \psi(w) = 1/2 \| w \|^2_2 \), the primal momentum in eq. (8) is equivalent to the dual momentum in eq. (7). For general potentials \( \psi \), iterates from the primal momentum \( \nabla \psi(w_{t}) \) can fall off the data manifold of span\( \{x_n\} \) and the additional components influence the final solution. Thus, the global minima that the iterates \( w_t \) converge to depends on the specific values of the momentum parameters \( \{ \beta_t, \gamma_t \} \) and the step sizes \( \{ \eta_t \} \), as demonstrated in the following example.

**Example 2.** Consider optimizing \( \mathcal{L}(w) \) with \( \{ (x_1 = [1, 2], y_1 = 1) \} \) and \( \ell(u, y) = (u - y)^2 \) using primal momentum updates for MD w.r.t the entropic potential \( \psi(w) = \sum_i w[i] \log w[i] - w[i] [8] \) and initialization \( w_0 = \text{argmin}_w \psi(w) \). Figure 1a shows how different choices of momentum \( \{ \beta_t, \gamma_t \} \) change the limit point \( w_\infty \). In particular, consider a simple case where the primal momentum is used only in the first step, but \( \gamma_t = 0 \) and \( \beta_t = 0 \) for all \( t > 1 \), then

**Proposition 2a.** for any \( \beta_1 > 0 \), there exists \( \{ \eta_t \}, \) such that \( w_t \) from (8) converges to a global minimum, but not to \( \text{argmin}_{w \in \mathcal{G}} \psi(w) \).

### 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,

$$w_{(t+1)} = w_{(t)} - \eta_t H(w_{(t)})^{-1} \nabla L(w_{(t)})$$  \hfill (9)

Natural gradient descent is commonly used when the metric tensor $H(w)$ corresponds to a Bregman divergence w.r.t a strongly convex potential $\psi$. For example, the KL divergence between distributions $P_w$ and $P_{w'}$ parameterized by $w$, corresponds to $\psi$ being the entropy potential over $P_w$. For these cases $H(w) = \nabla^2 \psi(P_w)$.

**Connection to mirror descent** When $H(w) = \nabla^2 \psi(w)$ for a strongly convex potential $\psi$, as the step size goes to zero, $\eta \to 0$, the iterates $w_{(t)}$ from natural gradient descent in eq. (9) and mirror descent w.r.t $\psi$ in eq. (4) converge to each other and the common dynamics of $w_{(t)}$ in the limit is given by:

$$\frac{d\nabla \psi(w_{(t)})}{dt} = -\nabla L(w_{(t)}) \iff \frac{dw_{(t)}}{dt} = -\nabla^2 \psi(w_{(t)})^{-1} \nabla L(w_{(t)})$$  \hfill (10)

So as the step sizes are made infinitesimal, the limit point of natural gradient descent $w_\infty = \lim_{t \to \infty} w_{(t)}$ mimics the limit point of mirror descent and hence will be biased towards solutions with minimum divergence to the initialization: as $\eta \to 0$, $w_\infty \to \arg\min_{w \in G} D_\psi(w, w_{(0)})$.

For a quadratic potential $\psi$, where $\psi(w) = \frac{1}{2} \|w\|^2_2$ 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 mirror descent [5] and natural gradient descent [9] updates are exactly equal for all values of $\{\eta_t\}$. Otherwise the updates differ in that [9] is only an approximation of the mirror descent update $\nabla \psi^{-1}(\nabla \psi(w_{(t)}) - \eta_t \nabla L(w_{(t)}))$.

For natural gradient descent with finite step size and non-quadratic potentials $\psi$, 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 easily lead to $w_{(1)}$ for which the dual variable $\nabla \psi(w_{(1)})$ is no longer in the data manifold span$(\{x_n\}) + \nabla \psi(w_{(0)})$, and hence will converge to a different global minimum dependent on the step sizes $\{\eta_t\}$.

**Example 3.** Consider optimizing $L(w)$ with $\{(x_1 = [1, 2], y_1 = 1)\}$ using the natural gradient descent updates in eq. (9) w.r.t. the metric tensor $H(w) = \nabla^2 \psi(w)$ corresponding to the potential $\psi(w) = \sum w[i] \log w[i] - w[i]$ with initialization $w_{(0)} = [1, 1]$. Figure 1b shows that NGD with different step sizes $\eta$ 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 $\eta_1 > 0$, $\lim_{t \to \infty} w_{(t)} = \arg\min_{w \in G} D_\psi(w, w_{(1)}) \neq \arg\min_{w \in G} D_\psi(w, w_{(0)})$.

### 2.4 Steepest Descent

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

$$w_{(t+1)} = w_{(t)} + \eta_t \Delta w_{(t)}$$

where $\Delta w_{(t)} = \arg\min_v \langle \nabla L(w_{(t)}), v \rangle + \frac{1}{2} \|v\|^2$.

(11)

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

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

(12)

Examples of steepest descent include gradient descent, which is steepest descent w.r.t $\ell_2$ norm and coordinate descent which is steepest descent w.r.t $\ell_1$ 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 $w_\infty$ of steepest descent w.r.t an arbitrary norm $\|\cdot\|$ to be the solution closest to initialization in corresponding norm, $\arg\min_{w \in G} \|w - w_{(0)}\|$. This is indeed the case for quadratic norms $\|v\|_D = \sqrt{v^T D v}$ when eq. (11) is equivalent to mirror descent with $\psi(w) = \frac{1}{2} \|w\|_D^2$. Unfortunately, this does not hold for general norms.
We now turn to strictly monotone loss functions. When multiple partial derivatives are maximal, we can choose any one of them, or some convex combination, leading to many possible coordinate descent optimization paths, both with positive bounded step sizes and with infinitesimal step sizes. The connection between optimization paths of coordinate descent and the $\ell_1$ regularization path $\hat{w}(\lambda) = \arg\min_w L(w) + \lambda \|w\|_1$ has been studied by Efron et al. (2004). Using the average of all optimal coordinates in (13) and infinitesimal step sizes is equivalent to forward stage-wise selection, a.k.a. $\epsilon$-boosting (Friedman, 2001). When the $\ell_1$ regularization path $\hat{w}(\lambda)$ 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), and at the limit we get the minimum $\ell_1$ norm solution. But when the regularization path is not monotone, which can and does happen, the paths diverge, and forward stagewise can converge to solutions with sub-optimal $\ell_1$ norms. This matches our understanding that steepest descent w.r.t. a norm, in this case the $\ell_1$ norm, might converge to a solution that is not the minimum norm solution w.r.t. the norm.

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. We noted that this characterization extends for momentum in the dual space as well as to natural gradient descent in the limit of infinitesimal step size $\eta \to 0$. We also saw 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 $\ell$ where the behavior of the implicit bias is fundamentally different, as are the situations when the implicit bias can be characterized. Such losses are defined for 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). $\ell(f(x); y)$ is bounded from below, and $\forall y$, $\ell(f(x); y)$ is strictly monotonically decreasing in $f(x)y$. Without loss of generality, $\forall y$, $\inf_{\hat{y}} \ell(\hat{y}; y) = 0$ and $\ell(\hat{y}; y) \xrightarrow{\gamma \to \infty} 0$.

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

When the training data $\{x_n, y_n\}$ is not linearly separable, the empirical objective $L(w)$ in eq. (1) can have a finite global minimum. However, we are again interested in cases where the empirical objective defined in eq. (1) is ill-posed. For strict monotone losses, if the data set $\{x_n, y_n\}_{n=1}^N$ is linearly separable, the empirical loss $\mathcal{L}(w)$ does not have any finite minimizer and $\mathcal{L}(w) \to 0$ only as $\|w\| \to \infty$. Then for any algorithm such that $\mathcal{L}(w(t)) \to 0$, the iterates diverge to infinity rather than converge and thus, we cannot talk about $\lim_{t \to \infty} w(t)$.

Example 4. Consider minimizing $L(w)$ with $\{(x_1 = [1, 1, 1], y_1 = 1), (x_1 = [1, 2, 0], y_1 = 10)\}$ and $\ell(u, y) = (u - y)^2$ using steepest descent updates w.r.t. the $\ell_{4/3}$ norm. The empirical results for this problem in Figure 1 clearly shows that, even for $\ell_p$ norms where the $\|\cdot\|_p^p$ 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 time limit of $\eta \to 0$, $w(t)$ does not converge to $\arg\min_{w \in G} \|w - w(0)\|_p$.

Coordinate descent Steepest descent w.r.t. the $\ell_1$ norm can be written as coordinate descent, with updates:

$$\Delta w_{(t+1)} \in \text{conv}\{ -\eta \frac{\partial L(w)}{\partial w[j]}, j_t = \arg\max_j |\frac{\partial L(w)}{\partial w[j]}| \}.$$ (13)

When multiple partial derivatives are maximal, we can choose any one of them, or some convex combination, leading to many possible coordinate descent optimization paths, both with positive bounded step sizes and with infinitesimal step sizes. The connection between optimization paths of coordinate descent and the $\ell_1$ regularization path $\hat{w}(\lambda) = \arg\min_w L(w) + \lambda \|w\|_1$ has been studied by Efron et al. (2004). Using the average of all optimal coordinates in (13) and infinitesimal step sizes is equivalent to forward stage-wise selection, a.k.a. $\epsilon$-boosting (Friedman, 2001). When the $\ell_1$ regularization path $\hat{w}(\lambda)$ 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), and at the limit we get the minimum $\ell_1$ norm solution. But when the regularization path is not monotone, which can and does happen, the paths diverge, and forward stagewise can converge to solutions with sub-optimal $\ell_1$ norms. This matches our understanding that steepest descent w.r.t. a norm, in this case the $\ell_1$ norm, might converge to a solution that is not the minimum norm solution w.r.t. the norm.
Instead, we look at the limit direction $\bar{w}_\infty = \lim_{t \to \infty} \frac{w(t)}{\|w(t)\|}$ whenever the limit exists. When the limit exists, we refer to this as convergence in direction. The limit direction entirely specifies the separating hyperplane, or the decision rule, and hence the generalization properties with respect to 0-1 error.

We focus on the exponential loss $\ell(u,y) = \exp(-uy)$. However, all our results can be extended to any loss function with a tight exponential tail, 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 dataset, gradient descent with any initialization and sufficiently small step size converges in direction to maximum margin separator with unit $\ell_2$ norm, i.e., the hard margin support vector classifier,

$$\bar{w}_\infty = \lim_{t \to \infty} \frac{w(t)}{\|w(t)\|} = w^*_n : = \arg\max_{w \neq 0} \min_n \frac{\langle w, x \rangle}{\|w\|_2}.$$

This is a robust characterization of the implicit bias that is independent of all hyperparameters including the initialization. This is fundamentally different from the implicit bias of gradient descent for losses with a unique finite root (Section 2.1) where the characterization was 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 the squared loss, the limit point of steepest descent could depend strongly on the stepsize, and we could not obtain a useful characterization even for infinitesimal step size. In contrast, the following theorem provides a crisp characterization of the limit direction of steepest descent in terms of a maximum margin solution, independent of the the step size (as long as it is small enough) and even the initialization.

**Theorem 5.** For any separable data set $\{x_n,y_n\}_{n=1}^N$, any initial point $w(0)$ and any norm $\|\|$ consider steepest descent updates for linear classification with the exponential loss $\ell(u,y) = \exp(-uy)$ and bounded step sizes that satisfy $\eta_t \leq \max \{ \eta_1, \frac{B}{2\|x_n\|} \}$, where $B : = \max_n \|x_n\|_s$ and $\eta_+ < \infty$ is some maximum step size bound. Then the iterates $w(t)$ satisfy:

$$\lim_{t \to \infty} \min_n \frac{y_n \langle w(t), x_n \rangle}{\|w(t)\|} = \max_w \min_n \frac{y_n \langle w, x_n \rangle}{\|w\|}.$$

In particular, if there is a unique maximum-$\|\|$ margin solution $w^*_n = \arg\max_w \min_n \frac{y_n \langle w, x_n \rangle}{\|w\|}$, then the limit direction converges to it: $\bar{w}_\infty = \lim_{t \to \infty} \frac{w(t)}{\|w(t)\|} = w^*_n$.

A special case of Theorem 5 is for steepest descent w.r.t. the $\ell_1$ 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 $\ell_1$ regularization (Schapire & Freund 2012), and later on AdaBoost with small enough stepsize was shown to converge in direction precisely to the maximum $\ell_1$ margin solution, just as is guaranteed by Theorem 5 (Zhang et al. 2005; Shalev-Shwartz & Singer 2010; Telgarsky 2013). In fact, Telgarsky (2013) generalized the result to a richer variety of exponential–tail loss functions including logistic loss, and a broad class of non-constant step size rules (interestingly, coordinate descent with exact line search might result in step sizes that are too large leading the iterates to converge to a different direction, which is not a max-$\ell_1$-margin direction (Rudin et al. 2004)).
Theorem 5 can be thought of as a generalization of the result 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 $\gamma$, then $\|\nabla \mathcal{L}(w)\|_p \geq \gamma \mathcal{L}(w)$ for all $w$. By using this lower bound on the dual norm of the gradient, with the optimization analysis similar to Telgarsky we are able to show that the loss decreases faster than the increase in the norm of the iterates, establishing convergence to a margin maximizing direction.

In relating the optimization path to the regularization path, it is also interesting to relate Theorem 5 to the fact that for monotone loss functions and $\ell_p$ norms, the $\ell_p$ regularization path $\hat{w}(c) = \arg\min_{w : \|w\|_p < c} \mathcal{L}(w)$ also converges in direction to the maximum margin separator, i.e. $\lim_{c \to \infty} \hat{w}(c) = \text{argmin}_{\|w\|_p \leq p} \mathcal{L}(w) \geq \hat{w}(\infty)$. Although the optimization path and regularization path are not the same, they both converge to the same max-margin separator in the limits of $c \to \infty$ or $t \to \infty$, 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 look at the implicit bias of AdaGrad in this section. To examine this, we focus on the basic (diagonal) AdaGrad:

$$w_{t+1} = w_t - \eta G_t^{-1/2} \nabla \mathcal{L}(w_t),$$

(14)

where $G_t$ is a diagonal matrix such that

$$\forall i : G_{t,ii} = \sum_{u=0}^{t} (\nabla \mathcal{L}(w_u))^2_i.$$

(15)

AdaGrad updates described above corresponds to a pre-conditioned gradient descent, except the pre-conditioning matrix $G_t$ depends on the iteration number. 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 affects the implicit bias. For squared loss, by inspection the updates in eq. (14), we do not expect to get a characterization of the limit point $w_\infty$ 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 (14), we can hypothesize that the robustness to initialization and initial updates depend on whether the matrices $G_t$ diverge or converge: if $G_t$ diverges, then we expect the asymptotic effects to dominate, while if it converges, then the limit direction will indeed depend on the initial conditions. Unfortunately, we can show that, the components of $G_t$ matrix are bounded.

**Lemma 6.** For any training data $\{x_n, y_n\}$, consider the AdaGrad iterates $w_{t}(\ell)$ (eqs. 14, 15) 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 $G_0$, such that $\frac{1}{2} \mathcal{L}(w_0) < 1$, and $\left\| G_0^{-1/4} x_n \right\|_2 \leq 1$, there exists a constant $C < \infty$ such that $\forall i, \forall t : G_{t,ii} \leq C$.

Importantly, since the components of $G_t$ are bounded, the initial conditions $w_0$ have a non-vanishing contribution to the asymptotic behavior of $G_t$ and hence to the limit direction $\hat{w}_\infty = \lim_{t \to \infty} \frac{w_t}{\|w_t\|_p}$, whenever it exists. In other words, the implicit bias of AdaGrad does indeed depend on the initial conditions, including initialization and step size.

### 4. Gradient descent on the factorized parameterization

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

$$\min_W \mathcal{L}(W) = \ell(W, X_n; y_n).$$

(16)
This is the exact same setting as (1) obtained by arranging \( w \) in a matrix form. We can now study another class of optimization algorithms for learning linear models based on matrix factorization where we reparameterize \( W \) as \( W = UV^T \) with unconstrained \( U \in \mathbb{R}^{d \times d} \) and \( V \in \mathbb{R}^{d \times d} \) and we get the following equivalent objective,

\[
\min_{U,V} \mathcal{L}(UV^T) = \sum_{n=1}^{N} \ell(UV^T, X_n ; y_n) 
\]

Note that eq. (17) is exactly the same optimization as eq. (16) with the exact same set of global minima over \( W = UV^T \). 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 updates on \( W \). In particular, gradient descent on \( U, V \) is observed to be biased towards low nuclear norm solutions, which in turns ensures generalization (Srebro et al. 2005) and low rank matrix recovery (Candès & Recht 2009). Since the matrix factorization objective (17) can be viewed as a two-layer neural network with linear activation, understanding the implicit bias then could provide insights into characterizing the implicit bias in more complex neural networks with non-linear activation.

We actually consider a more general formulation of the factorized approach (17) where we optimize \( \mathcal{L}(W) \) over p.s.d. matrices \( W \succ 0 \) using gradient descent over unconstrained symmetric factorization \( W = UV^T \) with \( U \in \mathbb{R}^{d \times d} \):

\[
\min_{U \in \mathbb{R}^{d \times d}} \mathcal{L}(U) = \mathcal{L}(UU^T) = \sum_{n=1}^{N} \ell(UU^T, X_n ; y_n) 
\]

As noted by Gunasekar et al. (2017), the optimization problem (17) over non-p.s.d. factorization \( W = UV^T \) is a special case of the optimization over the p.s.d. factorization (18) for a larger matrix \( \tilde{W} = UU^T \) with data \( \tilde{X}_n \). Specifically, \( \tilde{W} = \begin{bmatrix} A_1 & W \\ W^T & A_2 \end{bmatrix} \) for some p.s.d. matrix variables \( A_1, A_2 \succ 0 \), and \( \tilde{X}_n = \begin{bmatrix} 0 & X_n \\ X_n^T & 0 \end{bmatrix} \).

Let \( U(0) \in \mathbb{R}^{d \times d} \) be any full rank initialization, gradient descent updates in \( U \) are given by,

\[
U(t+1) = U(t) - \eta_t \nabla \mathcal{L}(U(t)),
\]

with corresponding updates in \( W(t) = U(t)U(t)^T \) given by,

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

**Losses with a unique finite root** For squares loss, Gunasekar et al. (2017) noted that the implicit bias of iterates in eq. (20) crucially depended on both the initialization \( U(0) \) as well as the step size \( \eta_t \). 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 \( X_n \) satisfies 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 as zero is a saddle point for eq. (18). 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 (18) 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 linearly separable datasets \( \{X_n, y_n\}_{n=1}^{N} \), and any full rank initialization \( U(0) \), consider the gradient descent iterates \( U(t) \) in eq. (19) and the corresponding sequence of \( W(t) \) in eq. (20) for minimizing \( \mathcal{L}(W) \) with the exponential loss \( \ell(u, y) = \exp(-uy) \). For any sufficiently small step sizes, if \( W(t) \) converges to a global minimum, and additionally the incremental updates \( \Delta W(t) = W(t+1) - W(t) \) and the gradients \( \nabla \mathcal{L}(W(t)) \) converge in direction, i.e. \( \frac{\Delta W(t)}{\|\Delta W(t)\|} \) and \( \frac{\nabla \mathcal{L}(W(t))}{\|\nabla \mathcal{L}(W(t))\|} \) converges, then
We considered the implicit bias of different optimization algorithms for two families of losses — losses we show that even in the strict monotone case, the limit direction with a unique finite root, and strict monotone losses. The implicit bias behavior for these two families is

\( W \) exponentiated gradient descent (Kivinen & Warmuth, 1997), which is MD w.r.t.

extensions to multi-layer linear models with functions with a bit of additional algebra. The preliminary result for matrix factorization also suggest natural believe, the results obtained for univariate estimates in this paper can also be extended for multivariate loss including initialization and initial step sizes.

Again, in contrast to squared loss, is independent of the initialization and step size. Finally, for AdaGrad, robust characterization of the implicit bias as the maximum margin separator with unit nuclear norm. This again, in contrast to squared loss, is independent of initialization and only has mild conditions on the step sizes.

Convergence of \( W_{\infty} \) is necessary for the characterization of implicit bias of limit direction to be relevant, but in Theorem 7 we require a stronger conditions that \( \frac{\Delta W(t)}{\|W(t)\|} \) and \( \frac{\nabla L(W(t))}{\|\nabla L(W(t))\|} \) converge which might not hold in general. Relaxing this condition is of interest for future work.

**Key property:** Let us look at exponential loss \( \ell(u, y) = \exp(-uy) \) when \( W(t) \) converges in direction to, say \( W_{\infty} \) as \( W(t) = W_{\infty}g(t) + \rho(t) \) for some scalar \( g(t) \to \infty \) and \( \rho(t)/g(t) \to 0 \). The gradients are given by \( \nabla L(W(t)) = \sum_{n=1}^{\infty} \exp(-g(t)y_n \langle W_{\infty}, X_n \rangle) \exp(-y_n (\rho(t), X_n))X_n \), which is dominated by positive linear combination of examples \( X_n \) that have the smallest distance to the decision boundary \( \langle W_{\infty}, X_n \rangle \) — which are essentially the support vectors of \( W_{\infty} \). This asymptotic behavior can be used to show optimality of \( W_{\infty} \) to the maximum margin solution subject to nuclear norm constraint in Theorem 7. This idea formalized in the following lemma is of interest beyond the results in this paper.

**Lemma 8.** For almost all linearly separable dataset \( \{x_n, y_n\}_n \), consider any sequence \( w(t) \) that minimizes the empirical objective in (1) with exponential loss, i.e., \( L(w(t)) \to 0 \) for \( \ell(u; y) = \exp(-uy) \). If \( \frac{w(t)}{\|w(t)\|} \) converges, then for every accumulation point \( z_\infty \) of \( \left\{ \frac{-\nabla L(w(t))}{\|\nabla L(w(t))\|} \right\}_\ell \), \( \exists \{\alpha_n \geq 0\}_n \in S \) s.t., \( z_\infty = \sum_{n \in S} \alpha_n x_n \), where \( \bar{w}_\infty := \lim_{t \to \infty} \frac{w(t)}{\|w(t)\|} \) and \( S = \{n : y_n \langle \bar{w}_\infty, x_n \rangle = \min_y y_n \langle \bar{w}_\infty, x_n \rangle\} \) are the indices of the data points with smallest margin to \( \bar{w}_\infty \).

**5. Summary**

We considered the implicit bias of different optimization algorithms for two families of losses — losses with a unique finite root, and strict monotone losses. The implicit bias behavior for these two families is fundamentally different.

In the case of losses with a unique finite root, we have simple characterization of the limit point \( w_\infty = \lim_{t \to \infty} w(t) \) for mirror descent in terms of divergence to initialization, where the divergence is specified by the potential function \( \psi \) which defines the geometry of 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 \( \bar{W}_{\infty} = \lim_{t \to \infty} \frac{W(t)}{\|W(t)\|} \). In particular, for steepest descent with respect to an arbitrary norm, we show that the limit direction converges to the maximum margin separating hyperplane in 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 of the implicit bias as the maximum margin separator with unit nuclear norm. This again, in contrast to squared loss, is independent of the initialization and step size. Finally, for AdaGrad, we show that even in the strict monotone case, the limit direction \( \bar{W}_{\infty} \) does depend on the initial conditions, including initialization and initial step sizes.

Further work is required in order to get a more complete understanding of the implicit bias. First, we believe, the results obtained for univariate estimates in this paper can also be extended for multivariate loss functions with a bit of additional algebra. The preliminary result for matrix factorization also suggest natural extensions to multi-layer linear models with \( W = U_1 U_2 \ldots U_L \), and eventually to non-linear networks.

Second, analysis in this paper applies only to unconstrained optimization and does not immediately extend to optimization with explicit constraints. For example, a common use case of mirror descent is the exponentiated gradient descent (Kivinen & Warmuth, 1997), which is MD w.r.t. \( \psi(w) = \sum_i w[i] \log w[i] - w[i] \).
under an explicit simplex constraint $\sum_i w[i] = 1$. Exponentiated gradient descent, is therefore, not covered by Theorem. Do the implicit bias characterizations in this paper also extend for constrained optimization problems whenever there are multiple global minima $W$ in the feasible set with $L(W) = 0$?

Finally, we would like a more fine grained analysis connecting the iterates $w(t)$ along the optimization path of various algorithms to the estimates along regularization path, $\hat{w}(c) = \arg\min_{R(w) \leq c} L(w)$, where an explicit regularization is added to the optimization objective. In particular, what we show in this paper is that the optimization path and the regularization path meet in their limit points, $t \to \infty$ and $c \to \infty$, respectively. It would be desirable to further understand the relations between the entire optimization and regularization paths.

References

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

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.

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

Gunasekar, Suriya, Woodworth, Blake E, Bhojanapalli, Srinadh, Neyshabur, Behnam, and Srebro, Nati. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6152–6160, 2017.

Hoffer, Elad, 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](http://arxiv.org/abs/1705.08741).

Keshavan, Raghunandan H et al. *Efficient algorithms for collaborative filtering*. PhD thesis, 2012.

Keskar, Nitish Shirish, Mudigere, Dheevatsa, Nocedal, Jorge, Smelyanskiy, Mikhail, and Tang, Ping Tak Peter. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2016.

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

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

Li, Yuanzhi, Ma, Tengyu, and Zhang, Hongyang. Algorithmic regularization in over-parameterized matrix recovery. *arXiv preprint arXiv:1712.09203*, 2017.

Muresan, Marian and Muresan, Marian. *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, Yurii. A method of solving a convex programming problem with convergence rate o (1/k^2). In *Soviet Mathematics Doklady*, 1983.

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

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

Neyshabur, Behnam, Tomioka, Ryota, Salakhutdinov, Ruslan, and Srebro, Nathan. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

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

Ross, Kenneth 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, Cynthia, Daubechies, Ingrid, and Schapire, Robert E. The dynamics of adaboost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5(Dec):1557–1595, 2004.

Schapire, Robert E and Freund, Yoram. *Boosting: Foundations and algorithms*. MIT press, 2012.

Shalev-Shwartz, Shai and Singer, Yoram. 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, Le. Don’t Decay the Learning Rate, Increase the Batch Size. In *ICLR*, 2018.

Soudry, Daniel, Hoffer, Elad, and Srebro, Nathan. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

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

Telgarsky, Matus. 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, Ashia C, Roelofs, Rebecca, Stern, Mitchell, Srebro, Nati, and Recht, Benjamin. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pp. 4151–4161, 2017.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhang, Tong, Yu, Bin, 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 \( \mathcal{P}_X = \text{span}\{\{x_n : n \in [N]\}\} = \{\sum_n \nu_n x_n : \nu_n \in \mathbb{R}\} \) and \( \ell'(u; y) \) be the derivative of \( \ell \) w.r.t \( u \). We have, \( \forall \tilde{w} \in \mathbb{R}^d, \mathcal{L}(\tilde{w}) = \ell'(\langle \tilde{w}, x_n \rangle; y_n) x_n \in \mathcal{P}_X. \) (21)

A.1 Proof of Theorem 1

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

\[
w_\psi^* = \arg\min_w D_\psi(w, w(0)) \text{ s.t., } \forall n, \langle w, x_n \rangle = y_n,
\]

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 (22) are:

Stationarity: \( \nabla \psi(w_\psi^*) - \nabla \psi(w(0)) \in \mathcal{P}_X \), or \( \exists \{\nu_n\}_{n=1}^N \text{ s.t., } \nabla \psi(w_\psi^*) - \nabla \psi(w(0)) = \sum_{n=1}^N \nu_n x_n \)

Primal feasibility: \( \forall n, \langle w_\psi^*, x_n \rangle = y_n \), or \( w_\psi^* \in \mathcal{G} \) (23)

Recall Theorem 1a from Section 2.2.

**Theorem 1** For any loss \( \ell \) with a unique finite root (Property 1), any initialization \( w(0) \), and any step size sequence \( \{\eta_t\} \), consider the mirror descent updates \( w(t) \) given in eq. (4) w.r.t some strongly convex potential \( \psi \). If the limit point of the iterates \( w_\infty = \lim_{t \to \infty} w(t) \) is a global minimum for \( \mathcal{L} \), i.e., \( \mathcal{L}(w_\infty) = 0 \), then \( w_\infty = \arg\min_{w} D_\psi(w, w(0)). \)

**Theorem 1a** Under the conditions in Theorem 1, if initialized at \( w(0) = \arg\min_w \psi(w) \), then the mirror descent updates with dual momentum also satisfies (6), i.e., for all \( \{\eta_t\}, \{\beta_t\}, \{\gamma_t\} \), if the \( w(t) \) from mirror descent with dual momentum (eq. (7)) converges to \( w_\infty \) such that \( \mathcal{L}(w_\infty) = 0 \), then \( w_\infty = \arg\min_{w} D_\psi(w, w(0)) = \arg\min_{w} \psi(w) \).

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

\[
\forall t, \nabla \psi(w(t)) - \nabla \psi(w(0)) = \sum_{t'<t} \nabla \psi(w(t'+1)) - \nabla \psi(w(t')) = \sum_{t'<t} -\eta_{t'} \nabla \mathcal{L}(w(t')) \in \mathcal{P}_X,
\]

where the last inclusion follows as \( \forall t', -\eta_{t'} \nabla \mathcal{L}(w(t')) \in \mathcal{P}_X \) from (21).

Thus, \( \forall t, w(t) \) from (5) always satisfies the stationarity condition of (23). 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 (23). Combining the above arguments, we have that if \( \mathcal{L}(w_\infty) = 0 \), then \( w_\infty = \arg\min_{w} \psi(w) \).

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

\[
\nabla \psi(w(t+1)) = \sum_{t' \leq t} \tilde{\beta}_{t'} \nabla \psi(w(t')) + \tilde{\gamma}_{t'} \nabla \mathcal{L}(\tilde{w}_{t'}). \tag{25}
\]

**Claim:** If \( \nabla \psi(w(0)) = 0 \), then for all updates of the form (25) satisfies \( \psi(w(t)) \in \mathcal{P}_X \)—this can be easily proved by induction:(a) for \( t = 0, \nabla \psi(w(0)) = 0 \in \mathcal{P}_X \); (b) let \( \forall t' \leq t, \nabla \psi(w(t')) \in \mathcal{P}_X \), (c) then using the inductive assumption and (21), we have \( \nabla \psi(w(t+1)) = \sum_{t' \leq t} \tilde{\beta}_{t'} \nabla \psi(w(t')) + \tilde{\gamma}_{t'} \nabla \mathcal{L}(\tilde{w}_{t'}) \in \mathcal{P}_X \).

Dual momentum in (7) is a special case of (25) with appropriate choice of \( \tilde{\beta}_{t'}, \tilde{\gamma}_{t'} \in \mathbb{R}, \), and \( \tilde{w}_{t'} \in \mathbb{R}^d \).
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 \(\ell(u, y) = (u - y)^2\).

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 \(\mathcal{P}_X = \text{span}(x_1) = \{z : 2z[1] - z[2] = 0\}\).

Initialization \(w(0) = [1, 1] \) satisfies \(\nabla \psi(w(0)) = 0\), and hence \(\psi(w(0)) = \arg\min_w \psi(w)\)

1. **Proof of Proposition 2a.** We use primal momentum with \(\beta > 0\) only in the first step, and \(\forall t \geq 2\), \(\beta_t = \gamma_t = 0\).

   Thus, for \(t > 2\), the updates follow the path of standard MD initialized at \(\nabla \psi(w(2))\) and thus, for appropriate choice of \(\{\eta_t\}\), we have \(w_{\infty} = \lim_{t \to \infty} w(t) \in \mathcal{G}\)

   Also, from (24), \(w_\infty\) satisfies \(\nabla \psi(w_\infty) - \nabla \psi(w(2)) \in \mathcal{P}_X \Rightarrow \nabla \psi(w_\infty) \in \nabla \psi(w(2)) + \mathcal{P}_X\).

From stationarity condition in (23), \(w_\infty = w^*_\psi \arg\min_{w \in \mathcal{G}} \psi(w)\) if and only if \(\nabla \psi(w(2)) \in \mathcal{P}_X\).

   We show that this is not the case for any \(\beta > 0\) and any \(\gamma \geq 0\). Working through the steps in (8), we have:

   \(\Delta w_{(-1)} = 0\), \(\nabla \psi(w(0)) = 0\) and \(\nabla \psi(w) = \log w\), thus,

   - \(\nabla \psi(w(1)) = r_0 x_1 \Rightarrow w(1) = \exp(r_0 x_1)\), and
   - \(\nabla \psi(w(2)) = \nabla \psi((1 + \beta_1) w(1)) + r_1 x_1 = \log (1 + \beta_1) + r_0 x_1 + r_1 x_1 \in \log (1 + \beta_1) + \mathcal{P}_X \not\in \mathcal{P}_X\),

   where \(r_0 = \eta_0 (y_1 - \langle w(0), x_1 \rangle)\) and \(r_1 = \eta_1 (y_1 - \langle w(1), x_1 \rangle)\).

2. **Proof of Proposition 2a.** The arguments are similar to the proof of Proposition 2a. In Example 3, we again use a finite \(\eta_t > 0\) to get \(w(1)\) and then follow the NGD using infinitesimal \(\eta_t\) 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. (24), the NGD updates for this example converges to a global minimum \(w_{\infty} = \lim_{t \to \infty} w(t) \in \mathcal{G}\), that satisfies \(\nabla \psi(w_{\infty}) - \nabla \psi(w(1)) \in \mathcal{P}_X \Rightarrow \nabla \psi(w_{\infty}) \in \nabla \psi(w(1)) + \mathcal{P}_X\).

   From stationarity condition in (23), \(w_\infty = w^*_\psi \arg\min_{w \in \mathcal{G}} \psi(w)\) if and only if \(\nabla \psi(w(1)) \in \mathcal{P}_X\).

   For natural gradient descent, \(w(1) = w(0) - \eta_t \nabla^2 \psi(w(0)) \nabla L(w(0)) = [1 + \eta_1 r_0, 1 + 2\eta_1 r_0]\), where \(r_0 = \eta_0 (y_1 - \langle w(0), x_1 \rangle)\). Thus, we have \(\nabla \psi(w(1)) \in \mathcal{P}_X \iff 2 \nabla \psi(w(1))[1] - \nabla \psi(w(1))[2] = 0 \iff 2 \log (w(1)[1]) - \log (w(1)[2]) = 0 \iff \log ((1 + \eta_t^2 r_0^2) \not= 0\).

   For any \(\eta_t\) such that \(\eta_t^2 r_0^2 \not= 0\), we get a contradiction.

Appendix B. Steepest descent for strictly monotone losses

We prove Theorem 5 in this section. The following lemma is a standard result in convex analysis.

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

\[
\max_w -f^*(-Aw) - g^*(-w) \leq \min_r f(r) + g(A^T r).
\]

Let \(X \in \mathbb{R}^{N \times d}\) be the data matrix with \(x_n\) along the rows of \(X\). Without loss of generality \(y_n = 1\), as for linear models \(y_n\) can be absorbed into \(x_n\). Define the \(\|\cdot\|_\gamma\) maximum margin as,

\[
\gamma = \max_{\|w\| = 1} \frac{\langle w, x_n \rangle}{\|w\|} = \max_{\|w\| = 1} e_n^T X w,
\]

where \(e_n\) is the \(n\)th standard basis in \(\mathbb{R}^N\).

Our primary technical novelty is the following duality lemma that generalizes [Shalev-Shwartz & Singer 2010] from \(\ell_1\) norm to any norm. We wish to show that \(\|\nabla L(w)\|_* \geq \gamma L(w)\) for all \(w\), where \(\|\cdot\|_*\) is the dual norm of \(\|\cdot\|\).
Define \( r_n(w) = \exp(-w^\top x_n) \). For succinctness, we often refer \( r(w) \in \mathbb{R}^N \) without the dependence on \( w \) as \( r \). By noting that \( L(w) = |r|_1 \) and \( \nabla L(w) = X^\top r, \|\nabla L(w)\|_* \geq \gamma L(w) \) can be restated as \( \|X^\top r\|_1 \geq \gamma \). Since we require this for all \( w, r_n \geq 0 \), and since norms are homogeneous, this is equivalent to \( \min_{r \in \Delta N - 1} \|X^\top r\|_* \geq \gamma, \) 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:

\[
\min_{r \in \Delta N - 1} \|X^\top r\|_* \geq \max_{\|w\| = 1} \min_n e_n^\top Xw = \gamma. \tag{28}
\]

This implies, for exponential loss \( \ell(u, y) = \exp(-uy) \), the iterates \( w(t) \) from the steepest descent path satisfy, \( \forall t, \|\nabla L(w(t))\|_* \geq \gamma L(w(t)) \).

**Proof.** Let \( f(r) = 1_{r \in \Delta N - 1} \) and \( g(z) = \|z\|_* \), where \( 1_X \) is the indicator function which is 0 if \( E \) is satisfied and \( \infty \) otherwise. Thus,

\[
\min_{r \in \Delta N - 1} \|X^\top r\|_* = \min_r f(r) + g(X^\top r). \tag{29}
\]

The conjugates are \( f^*(w) = \max_{z \in \Delta N - 1} \langle w, z \rangle = \max_n w_n, \) and \( g^*(w) = 1_{\|w\| \leq 1} \). The LHS of Lemma 9 is

\[
\max_w (-f^*(-Xw) - g^*(-w)) = \max_w (-\max_n e_n^\top Xw - 1_{\|w\| \leq 1})
\]

\[
= \max -\max_n e_n^\top Xw = \max_{\|w\| \leq 1} \min_n e_n^\top Xw \overset{(a)}{=} \gamma, \tag{30}
\]

where \((a)\) follows from definition of maximum \(-\)margin in eq. (27). By weak duality (Lemma 9) on eqs. (29) and (30), we have shown that \( \min_{r \in \Delta N - 1} \|X^\top r\|_* \geq \gamma \Rightarrow \forall r, \|X^\top r\|_* \geq \gamma \|r\|_1. \)

Hence, recalling that for exponential loss \( r_n(w) = \exp(-w^\top x_n), L(w) = \|r\|_1 \) and \( \nabla L(w) = X^\top r, \) we have \( \forall t, \|\nabla L(w(t))\|_* \geq \gamma L(w(t)) \).

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

\[
w(t+1) = w(t) + \eta_t \Delta w(t), \text{ where } \Delta w(t) \text{ s.t.,}
\]

\[
\langle \Delta w(t), -\nabla L(w(t)) \rangle = \|\Delta w(t)\|^2 = \|\nabla L(w(t))\|_*^2. \tag{31}
\]

**Lemma 11.** For exponential loss \( \ell(u, y) = \exp(-uy) \), consider the steepest descent iterates \( w(t) \) for minimizing \( L(w(t)) \), with any initialization \( w(0) \) and any finite step size \( \eta_t \) that leads to a strictly decreasing sequence \( L(w(t)) \) and satisfies \( 0 < \eta_t \leq \max \{ \eta_+, \frac{4}{B^2 L(w(t))} \} \), where \( B = \max_n \|x_n\|_\ast \). Then the following holds:

(A) \( \sum_{t=0}^{\infty} \eta_t \|\nabla L(w(t))\|^2_* \leq \infty, \) and hence \( \|\nabla L(w(t))\|_* \to 0. \)

(B) Iterates \( w(t) \) converge to a global minima \( L(w(t)) \to 0, \) and hence \( \forall n \langle w(t), x_n \rangle \to \infty. \)

(C) \( \sum_{t=0}^{\infty} \|\nabla L(w(t))\|_* = \infty. \)

**Proof.** 1. **Proof of (A):** We have that \( \|x_n\|_\ast \leq B \) for all \( n \). Recall that \( r_n(w) = \exp(-\langle w, x_n \rangle), L(w) = \sum_n r_n(w), \) and \( \nabla L(w) = \sum_n r_n(w)x_n. \) Thus, for all \( u \), we have

\[
\sum_n r_n(w)(x_n^\top v)^2 \leq \sum_n r_n(w)\|x_n\|_\ast \|v\|^2 \leq L(w)B^2 \|v\|^2. \tag{32}
\]
Using Taylors reminder theorem for the convex loss $\mathcal{L}$, we have

$$\mathcal{L}(w(t+1)) \leq \mathcal{L}(w(t)) + \eta_t \langle \nabla \mathcal{L}(w(t)), \Delta w(t) \rangle + \sup_{\beta \in (0,1)} \frac{\eta_t^2}{2} \Delta w(t)^\top \nabla^2 \mathcal{L} \left( w(t) + \beta \eta_t \Delta w(t) \right) \Delta w(t)$$

where (a) follows from eq. (32); (b) follows as $\eta \Delta w(t)$ is a descent step and along with convexity of $\mathcal{L}(w)$ we have $\sup_{\beta \in (0,1)} \mathcal{L} \left( w(t) + \beta \eta \Delta w(t) \right) = \mathcal{L}(w(t))$; and (c) follows as $\eta_t \leq \frac{2}{B^2 \mathcal{L}(w(t))}$ from the assumption, and $\|\Delta w(t)\| = \|\nabla \mathcal{L}(w(t))\|$, from eq. [31].

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

$$\forall t, \sum_{u=0}^{t} \eta_t \|\nabla \mathcal{L}(w(u))\|_\ast^2 \leq 2 \sum_{u=0}^{t} \mathcal{L}(w(u)) - \mathcal{L}(w(t+1)) = 2 \left( \mathcal{L}(w(0)) - \mathcal{L}(w(t+1)) \right) < \infty. \quad (34)$$

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 $\int_{0}^{t} \|\nabla \mathcal{L}(w(t))\|_\ast^2 < \infty$.

Thus, we have $\lim_{t \to \infty} \|\nabla \mathcal{L}(w(t))\|_\ast = 0$ — 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 any linearly separable data), then $\forall t < \infty, v^\top \nabla \mathcal{L}(w(t)) = \sum_{n \in [N]} \exp(-\langle w(t), x_n \rangle) x_n^\top v > 0$.

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

Finally, using triangle inequality,

$$\infty = \lim_{t \to \infty} \|w(t)\| \leq \|w(0)\| + \eta \sum_{t=0}^{\infty} \|\Delta w(t)\| = \|w(0)\| + \sum_{t=0}^{\infty} \eta_t \|\nabla \mathcal{L}(w(t))\|_\ast, \quad (35)$$

where we used $\|\Delta w(t)\| = \|\nabla \mathcal{L}(w(t))\|_\ast$ from [31]. This gives us $\sum_{t=0}^{\infty} \eta_t \|\nabla \mathcal{L}(w(t))\|_\ast = \infty$ in (C).

**B.1 Remaining steps in the proof of Theorem [5]**

The steepest descent updates in eq. (31) can be equivalently written as:

$$w(t+1) = w(t) - \eta_t \gamma_t \mathcal{P}(t), \quad \text{where}$$

$$\gamma_t \triangleq \|\nabla \mathcal{L}(w(t))\|_\ast, \quad \text{and} \quad \mathcal{P}(t) \text{ s.t.,}$$

$$\langle \mathcal{P}(t), \nabla \mathcal{L}(w(t)) \rangle = \|\nabla \mathcal{L}(w(t))\|_\ast, \quad \|\mathcal{P}(t)\| = 1.$$
From eq. 33 using $\gamma_t = \|\nabla \mathcal{L}(w(t))\|_*$, we have that

$$\mathcal{L}(w(t+1)) \leq \mathcal{L}(w(t)) - \eta_t \gamma_t^2 + \frac{B^2 \mathcal{L}(w(t)) \eta_t^2}{2} = \mathcal{L}(w(t)) \left[ 1 - \frac{\eta_t \gamma_t^2}{\mathcal{L}(w(t))} + \frac{\eta_t^2 B^2 \gamma_t^2}{2} \right]$$

\((a)\) \hspace{1cm} \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)

\((b)\) \hspace{1cm} \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),

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

**Step 1: Lower bound the unnormalized margin:** From eq. (36), we have,

$$\max_{n \in [N]} \langle w(t+1), x_n \rangle \leq \mathcal{L}(w(t+1)) \leq \mathcal{L}(w(0)) \exp \left( - \sum_{u \leq t} \frac{\eta_u \gamma_u^2}{\mathcal{L}(w(u))} + \sum_{u \leq t} \frac{\eta_u^2 B^2 \gamma_u^2}{2} \right). \quad (37)$$

By applying $- \log$,

$$\min_{n \in [N]} \langle w(t+1), x_n \rangle \geq \sum_{u \leq t} \frac{\eta_u \gamma_u^2}{\mathcal{L}(w(u))} - \sum_{u \leq t} \frac{\eta_u^2 B^2 \gamma_u^2}{2} - \log \mathcal{L}(w(0)). \quad (38)$$

**Step 2: Upper bound $\|w(t+1)\|$:** Using $\|\Delta w(u)\| = \|\nabla \mathcal{L}(w(u))\|_* = \gamma_u$, we have,

$$\|w(t+1)\| \leq \|w(0)\| + \sum_{u \leq t} \eta_u \|\Delta w(u)\| \leq \|w(0)\| + \sum_{u \leq t} \eta_u \gamma_u. \quad (39)$$

Combining eqs. (38) and (39), $\forall n \in [N]$, we have that

$$\frac{\langle w(t+1), x_n \rangle}{\|w(t+1)\|} \geq \sum_{u \leq t} \frac{\eta_u \gamma_u^2}{\mathcal{L}(w(u))} - \sum_{u \leq t} \frac{\eta_u^2 B^2 \gamma_u^2}{2} - \log \mathcal{L}(w(0)). \quad (40)$$

$$:= (I) + (II) + (III) \quad (41)$$

From Lemma 11 we have $\sum_{u \leq t} \eta_u \gamma_u \to \infty$ and $\|w(t)\| \to \infty$. Additionally,

(I) using $\gamma_u \geq \gamma \mathcal{L}(w(t))$ from the duality Lemma 10, we have $\sum_{u \leq t} \frac{\eta_u \gamma_u^2}{\mathcal{L}(w(u))} \geq \gamma \sum_{u \leq t} \eta_u \gamma_u$, and thus,

$$\frac{\sum_{u \leq t} \eta_u \gamma_u^2}{\sum_{u \leq t} \eta_u \gamma_u + \|w(0)\|} \geq \gamma \sum_{u \leq t} \eta_u \gamma_u \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). Thus, $\frac{\sum_{u \leq t} \eta_u^2 B^2 \gamma_u^2}{\||w(0)\||} \to 0.$

(III) $\log \mathcal{L}(w(0)) \to 0.$

Combining the above in (41), we have

$$\lim_{t \to \infty} \frac{w_{t+1}^T x_n}{\|w_{t+1}\|} \geq \gamma := \max_w \frac{w^T x_n}{\|w\|. \quad (42)$$
Appendix C. Adagrad

**Lemma 12.** Let $\mathcal{L}(w) = \sum_{n=1}^{N} \exp(-w^\top x_n)$, $\|\cdot\|_t$ be some $w_{(t)}$-dependent norm, and $\|\cdot\|_{t,*}$ be its dual, and assume that and that $\forall t : \|x_n\|_{t,*} \leq 1$. We examine the steepest descent (SD) Sequence:

$$w_{(t+1)} = w_{(t)} - \eta \beta_t p_{(t)}$$

where

$$p_{(t)}^\top \nabla \mathcal{L}(w_{(t)}) = \|\nabla \mathcal{L}(w_{(t)})\|_{t,*} \triangleq \beta_t ; \|p_{(t)}\|_t = 1.$$ 

Then, for any $w_{(0)}$ such that $\sum_{u=0}^{\infty} \beta_{u}^2 < \infty$ and therefore $\lim_{t \to \infty} \|\beta_t\| = 0$.

**Proof.** First we note that

$$p_{(t)}^\top \nabla^2 \mathcal{L}(w_{(t)}) p_{(t)} = \sum_{n=1}^{N} \exp(-w^\top x_n) (x_n^\top p_{(t)})^2 \leq \sum_{n=1}^{N} \exp(-w^\top x_n) = \mathcal{L}(w)$$

since $\|x_n\|_{t,*} \leq 1$ and $\|p_{(t)}\|_t = 1$. Therefore,

$$\max_{r \in (0,1)} \nabla^2 \mathcal{L}(w_{(t)}) - r \eta \beta_t p_{(t)} \leq \max_{r \in (0,1)} \mathcal{L}(w_{(t)}) - r \eta \beta_t p_{(t)} \leq \mathcal{L}(w_{(t)}),$$

since $-r \eta \beta_t p_{(t)}$ is a descent direction of $w_{(t)}$.

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

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

To calculate the last term, we first note that

$$p_{(t)}^\top \nabla^2 \mathcal{L}(w_{(t)}) p_{(t)} = \sum_{n=1}^{N} \exp(-w^\top x_n) (x_n^\top p_{(t)})^2 \leq \sum_{n=1}^{N} \exp(-w^\top x_n) = \mathcal{L}(w)$$

since $\|x_n\|_{t,*} \leq 1$ and $\|p_{(t)}\|_t = 1$. Therefore,

$$\max_{r \in (0,1)} \nabla^2 \mathcal{L}(w_{(t)}) - r \eta \beta_t p_{(t)} \leq \max_{r \in (0,1)} \mathcal{L}(w_{(t)}) - r \eta \beta_t p_{(t)} \leq \mathcal{L}(w_{(t)}),$$

since $-r \eta \beta_t p_{(t)}$ is a descent direction of $w_{(t)}$. Substituting this into eq. 44 we find

$$\mathcal{L}(w_{(t+1)}) \leq \mathcal{L}(w_{(t)}) - \eta \beta_t \nabla \mathcal{L}(w_{(t)})^\top p_{(t)} + \frac{1}{2} \beta_t^2 \mathcal{L}(w_{(t)})$$

$$= \mathcal{L}(w_{(t)}) - \eta \left(1 - \frac{\eta}{2} \mathcal{L}(w_{(0)}) \right) \beta_t^2.$$

Summing over the last equation, we obtain

$$\mathcal{L}(w_{(t)}) \leq \mathcal{L}(w_{(0)}) - \eta \left(1 - \frac{\eta}{2} \mathcal{L}(w_{(0)}) \right) \sum_{u=1}^{t} \beta_u^2.$$ 

Therefore,

$$\sum_{u=1}^{t} \beta_u^2 \leq \frac{\mathcal{L}(w_{(0)}) - \mathcal{L}(w_{(t)})}{\eta \left(1 - \frac{\eta}{2} \mathcal{L}(w_{(0)}) \right)}.$$ 

Recall we assumed $\frac{\eta}{2} \mathcal{L}(w_{(0)}) < 1$, and that $\mathcal{L}(w_{(t)}) \geq 0$. Therefore, the right-hand side is a bounded positive number. This entails that $\sum_{u=0}^{\infty} \beta_u^2 < \infty$ and therefore $\lim_{t \to \infty} \beta_t = 0$. \qed
Lemma 13. Let $\mathcal{L}(w) = \sum_{n=1}^{N} \exp(-w^T x_n)$. We examine the AdaGrad

$$w_{t+1} = w_t - \eta G_t^{-1/2} \nabla \mathcal{L}(w_t)$$

(45)

where $G_t$ is a diagonal matrix such that

$$\forall i : G_{t,ii} = \sum_{u=0}^{t} (\nabla \mathcal{L}(w_u))^2 .$$

Then, for any $w(0)$ such that $\frac{2}{d} \mathcal{L}(w(0)) < 1$, and if $\|G_{(0)}^{-1/2} x_n\|_2 \leq 1$, $\exists C < \infty$ such that

$$\forall i, \forall t : G_{(t),ii} < C.$$

\textbf{Proof.} First, we note that AdaGrad is a special case of the steepest descent algorithm as in Lemma 12 with respect to $\|v\|_i = \|G_t^{1/2} v\|_2$. Here the dual norm $\|v\|_{t,*} = \|G_t^{-1/2} v\|_2$. Since $G_{(t),ii}$ is monotonically decreasing for all $t$, $\|G_{(t)}^{-1/2} x_n\|_2 \leq \|G_{(0)}^{-1/2} x_n\|_2 \leq 1$, and so we can apply Lemma 12. This implies that

$$\infty > \sum_{t=0}^{\infty} \|G_{(t)}^{-1/2} \nabla \mathcal{L}(w_t)\|_2^2$$

$$= \sum_{i=1}^{d} \sum_{t=0}^{\infty} (\nabla \mathcal{L}(w_t))^2 \left[ \sum_{u=0}^{t} (\nabla \mathcal{L}(w_u))^2 \right]^{-1/2}$$

$$\geq \sum_{i=1}^{d} \sum_{t=0}^{\infty} (\nabla \mathcal{L}(w_t))^2 \left[ \sum_{u=0}^{\infty} (\nabla \mathcal{L}(w_u))^2 \right]^{-1/2}$$

$$= \sum_{i=1}^{d} \left( \sum_{t=0}^{\infty} (\nabla \mathcal{L}(w_t))^2 \right)$$

This implies that

$$\forall i : \sum_{t=0}^{\infty} (\nabla \mathcal{L}(w_t))_i^2 < \infty ,$$

\hfill \Box

Appendix D. Gradient descent on factorized parameterization

We first prove the Lemma 8 and Lemma D.2 that hold for any general linear model (1) with exponential tail strictly monotone losses $\ell(u, y)$. These results are not specific to the matrix factorization setup in Section 4.

\subsection*{D.1 Convergence of $-\nabla \mathcal{L}(w_t)$ in direction}

Recall Lemma 8.

\textbf{Lemma 8.} For almost all linearly separable dataset $\{x_n, y_n\}_n$, consider any sequence $w_t$ that minimizes the empirical objective in (1) with exponential loss, i.e., $\mathcal{L}(w_t) \to 0$. If $\frac{\nabla \mathcal{L}(w_t)}{\|w_t\|}$ converges, then for every accumulation point $z_\infty$ of $\left\{ \frac{-\nabla \mathcal{L}(w_t)}{\|\nabla \mathcal{L}(w_t)\|} \right\}$, $\exists \{\alpha_n \geq 0 : n \in S\}$ s.t., $z_\infty = \sum_{n \in S} \alpha_n x_n$, 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 $\bar{w}_\infty$. 19
Here for almost all \{x_n\} means that with probability 1 over \{x_n\} are drawn 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 – as the sign of y can be absorbed into \(x\), i.e. \(x_n \leftarrow y_n x_n\).

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 \(w(t)\) is in \(G = \{w : Xw > 0, \|w\| \to \infty\}\).

Also, 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 \(\rho(t)/g(t)\).

From these conditions, we also have \(\forall n, X\bar{w}_\infty > 0\). We introduce some additional notation:

- Let \(\gamma = \min_n \langle x_n, w_\infty \rangle = \min_n e_n^T X\bar{w}_\infty > 0\) the margin of \(\bar{w}_\infty\), where \(e_n\) are standard basis in \(\mathbb{R}^N\).
- Denote by \(S := \{n : \langle x_n, \bar{w}_\infty \rangle = \gamma\}\) the indices of support vectors of \(\bar{w}_\infty\).
- \(\tilde{\gamma} = \min_{n \in S} \langle x_n, w \rangle > \gamma\) is the second smallest margin of \(\langle x_n, w \rangle\).
- \(\tilde{\gamma}_n = \langle x_n, \bar{w}_\infty \rangle \geq \gamma\).
- \(\alpha_n(t) = \exp(-\langle \rho(t), x_n \rangle)\), \(\alpha \in \mathbb{R}^N\) be a vector of \(\alpha_n\) stacked. For any \(J \subseteq [N]\), similar to \(X_J\), let \(\alpha_J \in \mathbb{R}^{|J|}\) be a sub-vector with components corresponding to the indices in \(J\).
- \(B = \max_n \|x_n\|_2\).

Since \(\rho(t)/g(t) \to 0, \forall \epsilon_1, \epsilon_2, \exists \epsilon_1, \epsilon_2\) such that

\[
\forall t > \epsilon_1, \min_n -\langle \rho(t), x_n \rangle \geq -\|\rho(t)\|_2 B \geq -\epsilon_1 \gamma g(t),
\]
\[
\forall t > \epsilon_2, \max_n \langle \rho(t), x_n \rangle \leq \|\rho(t)\|_2 B \leq \epsilon_2 \tilde{\gamma} g(t)
\]

\[\tag{46}\]

The following claim is useful:

Claim 1. For almost \{x_n\} we have, \(|S| < d\) and \(\sigma_S(X_S) > 0\), where \(\sigma_k(A)\) is the kth singular value of \(A\).

Proof. Since, \(S = \{n : \langle \bar{w}_\infty, x_n \rangle = \gamma\}\), we have \(X_S\bar{w}_\infty = \gamma 1_S \in \mathbb{R}^{|S|}\).

For any fixed subset \(J\) if \(|J| > d\), then with probability 1,

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

\[\tag{47}\]

This is because if \(X\) is random and from a continuous distribution, and the rank deficient column span of \(X_J\) will miss any fixed vector \(v\) (that is independent of \(X\) with probability 1. Since \(1_S \in \text{span}(X_S)\), then for almost all \(X\), \(|S| \leq d\) and \(\sigma_S(X_S) > 0\).

Exponential loss: We prove the lemma for exponential loss \(\ell(u, y) = \exp(-uy)\) and extensions to general exponential tail (Property 2) can be followed by essentially following the steps in [Soudry et al., 2017]. For exponential loss, the gradients are given by

\[-\nabla \mathcal{L}(w(t)) = \sum_{n \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^T x_n) x_n + \sum_{n \in S^c} \exp(-\tilde{\gamma}_n g(t)) \exp(-\rho(t)^T x_n) x_n
\]
\[= I(t) + II(t),\]

\[\tag{48}\]

where \(I(t) = \sum_{n \in S} \exp(-\gamma g(t)) \exp(-\rho(t)^T x_n) x_n\) and \(II(t) = \sum_{n \in S^c} \exp(-\tilde{\gamma}_n g(t)) \exp(-\rho(t)^T x_n) x_n\).

We will show that \(\lim_{t \to \infty} \frac{\|II(t)\|}{\|I(t)\|} = 0\). Recall that \(\alpha(t)\) 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 \(I(t)\): For large enough \(t > \epsilon_1\), we have

\[\|I(t)\|_2 = \exp(-\gamma g(t)) \|X_S\alpha_S(t)\|_2 \geq \exp(-\gamma g(t)) \sigma_S(X_S)\|\alpha_S(t)\|_2 \geq \exp(-\gamma g(t)) \sigma_S(X_S) \min_n \alpha_n(t)
\]
\[\geq \sigma_S(X_S) \exp((-1 + \epsilon_1)\gamma g(t)) := C_1 \exp((-1 + \epsilon_1)\gamma g(t)),\]

\[\tag{49}\]
where (a) follows from the definition of \( \alpha_n = \exp(-\langle \rho(t), x_n \rangle) \) and (46), and \( C_1 > 0 \) is a constant independent of \( t \).

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

\[
\|II(t)\|_2 = \sum_{n \in S^e} \exp(-\bar{\gamma}_n g(t)) \exp(-\rho(t)^\top x_n) x_n \leq N \max_n \exp(-\bar{\gamma}_n g(t)) \alpha_n \|x_n\|_2 \leq \exp(-\bar{\gamma} g(t)) BN \max_n \alpha_n^{(a)} \leq BN \exp(-(1 - \epsilon_2)\bar{\gamma} g(t)) \coloneqq C_2 \exp(-(1 - \epsilon_2)\bar{\gamma} g(t)),
\]

where (a) uses \( \forall n \not\in S, \bar{\gamma}_n \geq \bar{\gamma} \) and (b) follows from the definition of \( \alpha_n = \exp(-\langle \rho(t), x_n \rangle) \) and (46), and \( C_2 > 0 \) is a constant independent of \( t \).

**Remaining steps in the proof:** By combining (49) and (50) using \( \epsilon_1 = (\bar{\gamma} - \gamma)/4\gamma \) and \( \epsilon_2 = (\bar{\gamma} - \gamma)/4\gamma \) and an appropriate constant \( C > 0 \), we have for any norm \( \|\cdot\| \)

\[
\frac{\|II(t)\|}{\|I(t)\|} \leq C \exp\left(-\frac{1}{2}(\bar{\gamma} - \gamma) g(t)\right) \xrightarrow{(a)} 0,
\]

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

Finally, \( -\nabla N(w(t)) = \frac{I(t)}{\|I(t)\|} + \frac{II(t)}{\|II(t)\|} \).

Since \( \frac{\|II(t)\|}{\|I(t)\| + \|II(t)\|} \xrightarrow{t \to \infty} 0 \), and \( I(t) \propto \sum_{n \in S} \alpha_n(t)x_n \) for \( \alpha_n(t) > 0 \), then every limit point of \( -\nabla N(w(t)) \xrightarrow{t \to \infty} \sum_{n \in S} \alpha_n x_n \) for some \( \alpha_n > 0 \). This finishes the proof for exponential loss.

\qed

**D.2 Convergence of \( w(t) \) in direction**

**Lemma 14** (t;dr \( w(t) \) converges in direction if \( \Delta w(t) \) converges in direction). *Assume \( \|w(t)\| \to \infty \) and \( \|\Delta w(t)\| \to \infty \), \( \forall t < \infty \). If \( \frac{\Delta w(t)}{\|\Delta w(t)\|} \to \tilde{\omega}_\infty \), then \( \frac{w(t)}{\|w(t)\|} \to \tilde{\omega}_\infty \) under (a) any bounded discrete step-size update of \( w(t+1) = w(t) + \eta_t \Delta w(t) \) for \( 0 < \eta_t < \infty \), or (b) the continuous time dynamics of \( \eta_t \to 0 \) with \( \frac{d}{dt} w(t) = \Delta w(t) \).

**Proof. Discrete Update:** Let \( 0 < \eta_- \leq \eta_t \leq \eta_+ < \infty \). Since \( \frac{\Delta w(t)}{\|\Delta w(t)\|} \to \tilde{\omega}_\infty \), we can write \( \Delta w(t) = \tilde{\omega}_\infty h(t) + \xi(t) \) where \( h(t) = \|\Delta w(t)\| \) and \( \frac{\xi(t)}{h(t)} \to 0 \).

Define

\[
g(t) := \sum_{u \leq t} \eta_u h(u), \text{ and } \rho(t) := w(t) - \tilde{\omega}_\infty g(t) = \sum_{u \leq t} \eta_u \xi(u) + w(0).
\]

In order to prove the lemma, we need to show that \( g(t) \to \infty \) and \( \frac{\rho(t)}{g(t)} \to 0 \). We have the following,

1. As \( \|w(t)\| \to \infty \), we have \( \|w(t)\| \leq \|w(0)\| + \sum_{u < \xi} \eta_u \|\Delta w(t)\| = \|w(0)\| + g(t) \xrightarrow{t \to \infty} \infty \).
2. Also, \( g(t) = \sum_{u < \xi} h(t) \) is a strictly monotonically increasing as \( \forall t < \infty, h(t) > 0 \).
3. Finally,

\[
\lim_{t \to \infty} \frac{\rho(t+1) - \rho(t)}{g(t+1) - g(t)} = \lim_{t \to \infty} \frac{\xi(t)}{\eta_t h(t)} \xrightarrow{(a)} 0,
\]

where in (a) we use that \( \eta_t \in [\eta_-, \eta_+] \) and thus \( 0 = \lim_{t \to \infty} \frac{\xi(t)}{\eta_+ h(t)} \leq \lim_{t \to \infty} \frac{\xi(t)}{\eta_+ h(t)} \leq \lim_{t \to \infty} \frac{\xi(t)}{\eta_- h(t)} = 0 \).

Summarizing, we have \( g(t) \) strictly monotone and unbounded and \( \lim_{t \to \infty} \frac{\rho(t+1) - \rho(t)}{g(t+1) - g(t)} = 0 \). Thus, by using the Stolz-Cesaro (Theorem 20), we have \( \lim_{t \to \infty} \frac{\rho(t)}{g(t)} = 0 \).

**Continuous time dynamics** Here we have \( \dot{w}(t) \coloneqq \frac{d}{dt} w(t) = \Delta w(t) = \tilde{\omega}_\infty h(t) + \xi(t) \).
Define \( g(t) := \int_{u=0}^{t} h(u) \, du \) and \( \rho(t) := w(t) - \bar{w}_\infty g(t) = \int_{u=0}^{t} \bar{w}_\infty h(u) \) with \( g(t) := \frac{dg(t)}{dt} = h(t) \) and \( \rho(t) := \frac{d\rho(t)}{dt} = \bar{w}_\infty \) and \( \lim_{t \to \infty} \frac{\rho(t) - \bar{w}_\infty}{g(t)} = 0 \). In order to prove the lemma, we need to show that \( g(t) \to \infty \) and \( \lim_{t \to \infty} \frac{\rho(t)}{g(t)} \to 0 \).

Since \( \|w(t)\| \to \infty \), \( \|w(t)\| \leq \int_{t=0}^{t} \|\Delta w(t)\| \, dt = g(t) \to \infty \).

Thus, we have \( g(t) \to \infty \), \( \forall t < \infty \), \( g(t) = h(t) > 0 \), and \( \lim_{t \to \infty} \frac{\rho(t)}{g(t)} = 0 \). Thus, using L’Hopital’s Rule (Theorem 19), we have \( \lim_{t \to \infty} \frac{\rho(t)}{g(t)} = 0 \). \( \square \)

D.3 Proof of Theorem 7

Recall Theorem 7:

**Theorem 7** For almost all linearly separable dataset \( \{X_n, y_n\}_{n=1}^{N} \), any full rank initialization \( U(0) \), consider the gradient descent iterates \( U(t) \) in eq. (19) and the corresponding sequence of \( W(t) \) in eq. (20) for minimizing \( L(W) \) with the exponential loss \( f(u, y) = \exp(-uy) \). For any fixed and sufficiently small step size \( \eta_n = \eta \), if \( W(t) \) converges to a global minimum, and additionally the incremental updates \( \Delta W(t) = W(t+1) - W(t) \) and the gradients \( \nabla L(W(t)) \) converge in direction, i.e. \( \frac{\Delta W(t)}{\|\Delta W(t)\|} \) and \( \frac{\nabla L(W(t))}{\|\nabla L(W(t))\|} \) converges, then

(a) the limit direction of \( W(t) \) exists \( \bar{W}_\infty = \lim_{t \to \infty} \frac{W(t)}{\|W(t)\|} \) and \( \lim_{t \to \infty} \frac{\Delta W(t)}{\|\Delta W(t)\|} = \frac{\Delta W(t)}{\|\Delta W(t)\|} \).

(b) the limit direction \( \bar{W}_\infty \) is the maximum margin separator with bounded nuclear norm \( \|\cdot\|_* \), given by

\[
\bar{W}_\infty = \arg\max_{W \geq 0} \min_{y_n} \frac{y_n(W, X_n)}{\|W\|_*}.
\]

**Proof.** From the assumption that \( \frac{\Delta W(t)}{\|\Delta W(t)\|} \) converges, let \( \bar{W}_\infty = \lim_{t \to \infty} \frac{W(t)}{\|W(t)\|} \). Lemma D.2 shows the first part of the theorem that \( W(t) \) normalized by the nuclear norm converges \( \frac{W(t)}{\|W(t)\|} \to \bar{W}_\infty \). Also, since \( W(t) \) minimizes a strictly monotone loss, we have that \( \|W(t)\|_* \to \infty \) and \( \forall n, y_n \langle \bar{W}_\infty, X_n \rangle > 0 \).

Let \( \gamma := \min_{n} y_n \langle \bar{W}_\infty, X_n \rangle \), in order to show (b), we equivalently show that \( \bar{W}_\infty := \bar{W}_\infty / \gamma \) is the solution to the following nuclear norm constrained maximum margin solution:

\[
W^* = \arg\min_{W \geq 0} \|W\|_*, \text{ s.t. } \forall n, y_n \langle W, X_n \rangle \geq 1.
\]

The KKT optimality conditions of (54) is given by

**Stationarity:**

\[
W^* = \sum_{n} \alpha_n X_n W^*,
\]

**Complementary slackness:**

\[
\alpha_n = 0, \forall i \notin S := \{i \in [n] : y_n \langle W^*, X_n \rangle = 1\},
\]

**Dual feasibility:**

\[
\alpha \geq 0, \text{ and } I - \sum_{n} \alpha_n X_n \geq 0,
\]

**Primal feasibility:**

\[
y_n \langle W^*, X_n \rangle \geq 1, W^* \geq 0.
\]

**Primal feasibility** We already get primal feasibility for \( \bar{W}_\infty := \bar{W}_\infty / \gamma \) as it has unit margin by the scaling, and from (20), \( \forall t, W(t) = U(t)U(t)^\dagger \) \( \geq 0 \), and hence it must converge in direction to a p.s.d. matrix \( \bar{W}_\infty \geq 0 \).

**Dual feasibility and complementary slackness:** Let \( S = \{n : y_n \langle \bar{W}_\infty, X_n \rangle = \gamma\} = \{n : y_n \langle \bar{W}_\infty, X_n \rangle = 1\} \), and \( \lambda_{\max}(\cdot) \) denote the maximum eigenvalue of a symmetric matrix \( Z \).

For a p.s.d. \( \bar{W}_\infty \), we have \( -\nabla L(W(t)) = \sum_{n} \exp(-y_n \langle W(t), X_n \rangle) \langle W(t), X_n \rangle \geq \gamma > 0 \). This implies, \( \lambda_{\max}(-\nabla L(W(t))) > 0 \). From the assumption in Theorem 7, we have that \( -\nabla L(W(t)) \to 0 \), thus denote \( \lambda_{\max}(-\nabla L(W(t))) \to 0 \).
Using Lemma\(^8\) we have \(Z_\infty = \sum_{n \in S} \alpha_n X_n\) for some \(\alpha_n \geq 0\) (with \(\alpha_n = 0, \forall n \notin S\)). We propose \(\{\alpha_n\}\) as our candidate dual certificate for \(\hat{W}_\infty\) which satisfies complementary slackness by definition. Further, as \(\lambda_{\max}(Z_\infty) = 1\), we also have \(I - Z_\infty = I - \sum \alpha_n X_n \geq 0\).

**Stationarity:** This is the main condition to verify: \(\hat{W}_\infty = Z_\infty \hat{W}_\infty\), or equivalently \(\hat{W}_\infty = Z_\infty \hat{W}_\infty\).

From Lemma\(^2\) and Lemma\(^8\), we have the following:

1. From assumption that \(\|\Delta W(t)\|_* \rightarrow W_\infty\), we define the following (note that unlike eq. (52), we have absorbed \(\eta\) into definition of \(\Delta W(t)\) here):

   \[
   \Delta W(t) = W(t+1) - W(t) = \hat{W}_\infty h(t) + \xi(t) \quad \text{s.t.,} \quad h(t) = \|\Delta W(t)\|_*, \quad \frac{\xi(t)}{h(t)} \rightarrow 0, \|\hat{W}_\infty\|_* = 1, \quad (59)
   \]

2. From the construction in eq. (52) in proof of Lemma\(^2\)

   \[
   W(t) = \hat{W}_\infty g(t) + \rho(t), \quad \text{where} \quad g(t) = \sum \limits_u h(u), \quad \rho(t) = \sum \limits_u \xi(u) \quad \text{and} \quad \frac{\rho(t)}{g(t)} \rightarrow 0. \quad (60)
   \]

3. Since \(\mathcal{L}(W(t)) \rightarrow 0, \nabla \mathcal{L}(W(t)) \rightarrow 0\). Thus, using Lemma\(^8\) we have \(Z_\infty = \sum_{n \in S} \alpha_n\), such that

   \[
   -\nabla \mathcal{L}(W(t)) = Z_\infty p(t) + \zeta(t), \quad \text{where} \quad \frac{\zeta(t)}{p(t)} \rightarrow 0 \quad \text{and} \quad p(t) = \lambda_{\max}(\nabla \mathcal{L}(W(t))) \rightarrow 0. \quad (61)
   \]

4. \(\nabla \mathcal{L}(W(t)) W(t) \nabla \mathcal{L}(W(t)) = p(t) g(t) \delta_1(t) \quad \text{where} \quad \delta_1(t) := p(t) Z_\infty \hat{W}_\infty Z_\infty + Z_\infty \hat{W}_\infty \frac{\zeta(t)}{p(t)} + Z_\infty \frac{\rho(t)}{g(t)} Z_\infty + Z_\infty \rho(t) \frac{\zeta(t)}{g(t)} + \frac{\zeta(t)}{p(t)} \hat{W}_\infty Z_\infty + \frac{\zeta(t)}{p(t)} \hat{W}_\infty Z_\infty + \frac{\zeta(t)}{p(t)} \hat{W}_\infty + \frac{\zeta(t)}{p(t)} \hat{W}_\infty \rightarrow 0 \quad (62)
   \]

Using \(W(t)\) and \(-\nabla \mathcal{L}(W(t))\) from (60) and (61), respectively, for the updates \(\Delta W(t)\) from eq. (20), we have

\[
\Delta W(t) = -\eta \frac{\nabla \mathcal{L}(W(t))}{\|\nabla \mathcal{L}(W(t))\|_*} \rightarrow W_\infty \quad \text{with} \quad W_\infty \succeq 0 \quad \text{and} \quad h(t) = \|\Delta W(t)\|_*, \quad \text{we have,}
   \]

\[
1 = \|\hat{W}_\infty\|_* = \langle \hat{W}_\infty, I \rangle = \lim_{t \rightarrow \infty} \frac{\Delta W(t)}{\|\Delta W(t)\|_*}, I \rangle = \lim_{t \rightarrow \infty} \frac{2\eta p(t) g(t)}{h(t)} \langle Z_\infty \hat{W}_\infty, I \rangle. \quad (63)
   \]

\[
\lim_{t \rightarrow \infty} \frac{2\eta p(t) g(t)}{h(t)} = \frac{1}{\langle Z_\infty, \hat{W}_\infty \rangle} := D \geq 1, \quad (64)
   \]

\[
\hat{W}_\infty = \frac{\eta p(t) g(t)}{\Delta g(t)} [Z_\infty \hat{W}_\infty + \hat{W}_\infty Z_\infty + \delta(t)] = \frac{D}{2} (Z_\infty \hat{W}_\infty + \hat{W}_\infty Z_\infty) \quad (65)
   \]

where constant \(D\) defined above is independent of \(t\), and in (a) we use \(\langle \hat{W}_\infty, I - Z_\infty \rangle \geq 0 \) as \(\hat{W}_\infty, I - Z_\infty \succeq 0\), and hence \(\langle \hat{W}_\infty, Z_\infty \rangle \leq \langle \hat{W}_\infty, I \rangle = 1\).

**Claim 2.** If \(D = 1\), then the stationarity condition in (55) holds.

**Proof.** If \(D = 1\), from eq. (63) \(\langle \hat{W}_\infty, I \rangle = \langle \hat{W}_\infty, Z_\infty \rangle \implies \text{trace} (\hat{W}_\infty (I - Z_\infty)) = 0 \implies \hat{W}_\infty = W_\infty Z_\infty\), where the last implication follows as both \(\hat{W}_\infty\) and \((I - Z_\infty)\) are p.s.d.

\[ \square \]
Showing \( D = 1 \)  
Let \( Z_\infty = \sum_i \lambda_i z_i z_i^T \) be the eigenvalue decomposition of \( Z_\infty \) with \( \lambda_1 = \lambda_{\text{max}} (Z_\infty) = 1 \). 
Let \( \mu_i(t) = \langle W(t), z_i z_i^T \rangle \geq 0 \), \( \bar{\mu}_i(t) = \frac{\mu_i(t)}{\delta_i(t)} \), and \( \bar{\mu}_i^\infty = \lim_{t \to \infty} \bar{\mu}_i(t) = \langle \bar{W}_\infty, z_i z_i^T \rangle \).

For any \( W \), we have \( \langle Z_\infty W, z_i z_i^T \rangle = \lambda_1 \langle W, z_i z_i^T \rangle \) from the eigenvalue decomposition of \( Z_\infty \).

**Claim 3.** For all \( i \), \( \bar{\mu}_i^\infty > 0 \Rightarrow \lambda_i = 1/D. \)

**Proof.** From \( (55) \) we have, \( \mu_i^\infty = \langle \bar{W}_\infty, z_i z_i^T \rangle = D \langle Z_\infty \bar{W}_\infty, z_i z_i^T \rangle = D \lambda_i \mu_i^\infty \Rightarrow D = \lambda_i^{-1}. \)

In particular, from above proposition, if \( \exists i : \lambda_i = 1 \) and \( z_i^\top \bar{W}_\infty z_i > 0 \), then \( D = 1 \) and thus, \( \bar{W}_\infty = Z_\infty \bar{W}_\infty \).

**Assume the contrary** that

\[
\lim_{t \to \infty} \bar{\mu}_1(t) \rightarrow 0 \text{ and } \exists k \text{ s.t., } \lambda_k < 1 \text{ and } \lim_{t \to \infty} \bar{\mu}_k(t) = 1/D \Rightarrow \lim_{t \to \infty} \frac{\bar{\mu}_1(t)}{\bar{\mu}_k(t)} = 0. \quad (66)
\]

We will show that this is not possible:

**Step I: Expression for \( \mu_i(t+1) \):** From \( (62) \), we have \( \Delta W(t) = \eta p(t) [Z_\infty W(t) + W(t) Z_\infty + g(t) \vec{\delta}(t)] \) for \( \delta(t) = (p(t) g(t))^{-1} (\zeta(t) W(t) + W(t) \zeta(t)) \rightarrow 0 \). Thus,

\[
\Delta \mu_i(t) := \mu_i(t+1) - \mu_i(t) = 2 \eta p(t) \lambda_i \langle W(t), z_i z_i^T \rangle + \eta p(t) g(t) \langle \vec{\delta}(t), z_i z_i^T \rangle.
\]

Defining \( \tilde{\delta}_i(t) = \frac{1}{2} \langle \vec{\delta}(t), z_i z_i^T \rangle \rightarrow 0 \) and using \( \mu_i(t) = g(t) \tilde{\mu}_i(t) \), we have \( \forall t \)

\[
\mu_i(t+1) = g(t) \left[ (1 + 2 \eta p(t) \lambda_i) \tilde{\mu}_i + 2 \eta p(t) \tilde{\delta}_i(t) \right].
\]

**Step II: Bound on \( \frac{\mu_i(t+1)}{\mu_k(t+1)} \):** Defining \( \kappa(t) := \frac{\bar{\mu}_i(t)}{\bar{\mu}_k(t)} = \frac{\mu_i(t)}{\mu_k(t)} \), we have the following:

\[
\kappa(t+1) = \frac{\mu_i(t+1)}{\mu_k(t+1)} = \frac{(1 + 2 \eta p(t) \lambda_i) \bar{\mu}_i}{(1 + 2 \eta p(t) \lambda_k) \bar{\mu}_k + 2 \eta p(t) \tilde{\delta}_k(t)} \leq \frac{(1 + 2 \eta p(t) \lambda_i) \bar{\mu}_i}{(1 + 2 \eta p(t) \lambda_k) \bar{\mu}_k + 2 \eta p(t) \tilde{\delta}_k(t)} \quad (a)
\]

\[
= \tau(t) \kappa(t) - \tilde{\delta}(t), \quad (b)
\]

where \( (a) \) follows by dividing the numerator and denominator by \( \bar{\mu}_k(t) > 0 \).

**Step III: Show \( \kappa(t) \to \infty \):** The following propositions are proved in Section \( \text{D.4} \).

**Proposition 15.** \( \exists \bar{c} > 0, t_0 \text{ s.t., } \forall t > t_0, \tau(t) \geq 1 + 2 \eta p(t) \bar{c} \). In particular, \( \bar{c} = \frac{\lambda_1 - \lambda_k}{2(\lambda_1 + \lambda_k)} > 0 \).

**Proposition 16.** For any \( t_0 \), we have \( \sum_{u=t_0}^t \bar{\delta}(u) \to 0 \), and further \( \sum_{u=t_0}^t 2 \eta p(u) \to \infty \).

Thus, extending eq. \( (67) \) we have \( \forall t > t_0 \) and \( \bar{c} = \frac{\lambda_1 - \lambda_k}{2(\lambda_1 + \lambda_k)} > 0 \),

\[
\kappa(t+1) = \tau(t) \kappa(t) - \tilde{\delta}(t) \geq \prod_{u=t_0}^t (1 + 2 \eta p(u) \bar{c}) \kappa(t_0) - \sum_{u=t_0}^t \tilde{\delta}(t) \geq \left( 1 + \sum_{u=t_0}^t 2 \eta p(u) \bar{c} \right) \kappa(t_0) - \sum_{u=t_0}^t \tilde{\delta}(t) \quad (68)
\]

\[
= \kappa(t_0) + \left( \sum_{u=t_0}^t 2 \eta p(u) \right) \left[ \bar{c} \kappa(t_0) - \frac{\sum_{u=t_0}^t \tilde{\delta}(u)}{\sum_{u=t_0}^t 2 \eta p(u)} \right],
\]

where \( (a) \) follows from iterating over \( t \) and Proposition 15 and \( (b) \) by expanding the product and ignoring the higher order positive terms as \( 2 \eta p(t) \to 0 \).
Since we start with a full rank $U(0)$ and hence a full rank $W(0)$, and $\{X_n\}$ are in general position for almost all datasets, with small enough $\eta$ (smaller than the inverse of local Lipschitz), for any finite $t_0$, the iterates $U(t) = U(t-1) - \eta \nabla \mathbb{L}(W(t-1))U(t-1)$ remains full rank and hence $\kappa(t_0) > 0$, and thus, $\kappa(t_0)\bar{c} > 0$.

Also, from Proposition 16, we have $\sum_{u=t_0}^{t} \delta(u) \rightarrow 0$ and hence for large enough $t$, $\bar{c}\kappa(t_0) - \sum_{u=t_0}^{t} \delta(u) > 0$.

Combining this with $\sum_{u=t_0}^{t} 2\eta p(u) \rightarrow \infty$, we have $\kappa(t) \rightarrow \infty$, thus contradicting 66. □

D.4 Proof of Proposition 15 and Proposition 16

Proposition 15. $\exists \bar{c} > 0, t_0$ s.t., $\forall t > t_0$, $\tau(t) \geq 1 + 2\eta p(t)\bar{c}$. In particular, $\bar{c} = \frac{\lambda_1 - \lambda_2}{2(\lambda_1 + \lambda_2)} > 0$.

Proof. We have $p(t) \rightarrow 0$ (as $\nabla \mathbb{L}(W(0)) = 0$), and $\delta_k(t) \rightarrow 0$, $\bar{\mu}_k(t) \rightarrow \bar{\mu}_k^\infty > 0$. We then pick $\epsilon = \frac{1 - \lambda_1(1 + \lambda_2)}{1 + \epsilon} = \frac{1 - \lambda_2}{2(1 + \epsilon)} > 0$.

For $t > t_0$, $2\eta p(t) \leq 1$ and $\frac{\delta_k(t)}{\bar{\mu}_k(t)} \leq \epsilon$, where we have

$$0 < \epsilon = \frac{1 - \lambda_k - \bar{c} (1 + \lambda_k)}{1 + \epsilon} = \frac{1 - \bar{c}}{1 + 2\eta p(t)\bar{c} - \lambda_k} \quad (69)$$

Now using $\frac{\delta_k(t)}{\bar{\mu}_k(t)} \leq \epsilon$ and $\lambda_1 = 1$, we have the following for $t \geq t_0$

$$\tau(t) = \frac{1 + 2\eta p(t)\lambda_1}{1 + 2\eta p(t)\left(\lambda_k + \frac{\delta_k(t)}{\bar{\mu}_k(t)}\right)} \geq \frac{1 + 2\eta p(t)}{1 + 2\eta p(t)(\lambda_k + \epsilon)} \quad (70)$$

where $(a)$ follows from eq. (69). □

Proposition 16. For any $t_0$, we have $\sum_{u=t_0}^{t} 2\eta p(u) \rightarrow 0$.

Proof. First show $\sum_{u=t_0}^{t} 2\eta p(u) \rightarrow \infty$: Recall that $g(t) = \sum_{u=t} h(u)$, where $h(u) = \|\Delta W(t)\|_*$ > 0 (from eq. (60)-(59)). We then have

$$0 < \log g(t+1) - \log g(t) = \log \left(\frac{g(t+1)}{g(t)}\right) \leq \frac{g(t+1) - g(t)}{g(t)} = \frac{\frac{\Delta g(u)}{g(u)} \sum_{u=t}^{t+1} \frac{h(u)}{g(u)}}{g(t)}$$

where in $(a)$ we used $\log(x) \leq x - 1$. Summing over $t$, we have $\sum_{u=t_0}^{t} \frac{\Delta g(u)}{g(u)} \sum_{u=t_0}^{t+1} \frac{h(u)}{g(u)} \rightarrow \infty$.

Also, recall from eq. (64) that $\frac{2\eta p(t)}{h(t)/g(t)} \rightarrow D$.

Now using $a_t = \sum_{u=t_0}^{t} 2\eta p(u)$ and monotonically increasing divergent sequence $b_t = \sum_{u=t_0}^{t} \frac{h(u)}{g(u)} \rightarrow \infty$ in Stolz-Cesaro theorem (Theorem 20), we get

$$\lim_{t \rightarrow \infty} \frac{\sum_{u=t_0}^{t} 2\eta p(u)}{\sum_{u=t_0}^{t} \frac{h(u)}{g(u)}} = \lim_{t \rightarrow \infty} \frac{a_t}{b_t} = \lim_{t \rightarrow \infty} \frac{a_t - a_{t-1}}{b_t - b_{t-1}} = \lim_{t \rightarrow \infty} \frac{2\eta p(t)g(t)}{h(t)} = D.$$

Hence, $\lim_{t \rightarrow \infty} \sum_{u=t_0}^{t} 2\eta p(u) = D \lim_{t \rightarrow \infty} \sum_{u=t_0}^{t} 2\eta p(u) = D \lim_{t \rightarrow \infty} \sum_{u=t_0}^{t} \frac{\Delta g(u)}{g(u)} = \infty$.

Bound $\sum_{u=t_0}^{t} \delta(t)$: We have from definition $0 < \delta(t) = \frac{2\eta p(t)\delta_1(t)}{1 + 2\eta p(t)\lambda_k + 2\eta p(t)\bar{\mu}_k(t)} \leq 2\eta p(t) \frac{\delta_1(t)}{\bar{\mu}_k(t)} \rightarrow 0$.

Since $\delta_1(t) \rightarrow 0$ and $\bar{\mu}_k(t) \rightarrow \bar{\mu}_k^\infty > 0$, we also have $\frac{\delta_1(t)}{\eta p(t)} \rightarrow 0$. 25
Again using Stolz-Cesaro theorem with $c_t = \sum_{u=t_0}^t \delta(u)$ and $d_t = \sum_{u=t_0}^t 2\eta p(u) \to \infty$, we have

$$\lim_{t \to \infty} \frac{\sum_{u=t_0}^t \delta(u)}{\sum_{u=t_0}^t 2\eta p(u)} = \lim_{t \to \infty} \frac{c_t - c_{t-1}}{d_t - d_{t-1}} = \lim_{t \to \infty} \frac{\delta(t)}{2\eta p(t)} = 0. \quad (72)$$

\[\square\]

Appendix E. Preliminaries

**Lemma 17** (Sub-differentials of norms). For a generic norm $\|v\|$ for $v \in \mathcal{V}$, recall the dual norm $\|y\|_* = \sup_{\|v\| \leq 1} \langle y, v \rangle$. The sub-differential of a norm $\|v\|$ 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. $\partial \|v\| = \{ y : \|y\|_* = 1, \text{ and } \langle y, v \rangle = \|v\| \}$ Conversely, $\forall y \in \partial \|v\|$, from the definition, we have $\forall \Delta, \|v\| + \|\Delta\| \geq \|v + \Delta\| \geq \|v\| + \langle y, \Delta \rangle \implies \|y\|_* = \sup_{\|\Delta\| \neq 0} \frac{\langle y, \Delta \rangle}{\|\Delta\|} \leq \|v\|$.

2. From above result, $y \in \partial \|v\| \iff \|y\|_* = \|v\|$ and $\langle y, v \rangle = \|v\| \iff \|y\|_* = \sup_{\|v\| \leq 1} \langle y, v \rangle \Rightarrow \langle y, v \rangle = \sup_{\|v\| \leq 1} \langle y, v \rangle$.

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

\[\square\]

**Lemma 18** (Limit points of a compact sets). If $\{a_t\}_{t=1}^\infty$ is a sequence contained in a compact set $a_1 \in C$, 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 19** (L-Hôpital’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 $\lim_{x \to s} f(x) g(x) = L$ exists. If either (a) $\lim_{x \to s} f(x) = \lim_{x \to s} g(x) = 0$, or (b) $\lim_{x \to s} \frac{f(x)}{g(x)}$ exists and is equal to $L$

**Theorem 20** (Stolz–Cesaro theorem, proof in Theorem 1.23 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 $L$. 

26