Characterizing Implicit Bias in Terms of Optimization Geometry

Suriya Gunasekar [1] **Jason Lee** [2] **Daniel Soudry** [3] **Nathan Srebro** [1]

Abstract

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

1. Introduction

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

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

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

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

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

As we shall see, the answer depends on whether the loss function is similar to a squared loss or to a logistic loss.

[1]TTI Chicago, USA [2]USC Los Angeles, USA [3]Technion, Israel. Correspondence to: Suriya Gunasekar <suriya@ttic.edu>, Jason Lee <jasonlee@marshall.usc.edu>, Daniel Soudry <daniel.soudry@gmail.com>, Nathan Srebro <nati@ttic.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

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

In Section 2 we study linear models with loss functions that have unique finite roots. We obtain a robust characterization of the limit point for mirror descent, and discuss how it is independent of step–size and momentum. For natural gradient descent, we show that the step–size does play a role, but get a characterization for infinitesimal step–size. For steepest descent, we show that not only does step–size affects the limit point, but even with infinitesimal step–size, the expected characterization does not hold. The situation is fundamentally different for strictly monotone losses such as the logistic loss (Section 3) where we do get a precise characterization of the limit direction for generic steepest descent. We also study the adaptive gradient descent method (AdaGrad) (Duchi et al., 2011) and optimization over 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–size even for strict monotone losses.

2. Losses with a Unique Finite Root

We first consider learning linear models using losses with a unique finite root, such as the squared loss, where loss function $\ell(f(x), y)$ between a predictor $f(x)$ and label y is minimized at a unique and finite value of $f(x)$.

Property 1 (Losses with a unique finite root). *For any y and sequence $\widehat{y}_t$, $\ell(\widehat{y}_t, y) \to \inf_{\widehat{y}} \ell(\widehat{y}, y) = 0$ if and only if $\widehat{y}_t \to y$. Here we assumed without loss of generality that $\inf_{\widehat{y}} \ell(\widehat{y}, y) = 0$ and the root of $\ell(\widehat{y}, y)$ is at $\widehat{y} = y$.*

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

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

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

2.1. Gradient descent

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

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

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

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

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

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

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

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

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

2.2. Mirror descent

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

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

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

We first look at unconstrained optimization where $\mathcal{W} = \mathbb{R}^d$ and the update in eq. (4) can be equivalently written as:

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

For a strongly convex potential ψ, $\nabla\psi$ is called the link function and is invertible. Hence, the above updates are uniquely defined. Also, w and $\nabla\psi(w)$ are refered as *primal* and *dual* variables, respectively. Examples of strongly convex potentials ψ for mirror descent include, the squared ℓ_2 norm $\psi(w) = 1/2\|w\|_2^2$, which leads to gradient descent; the entropy potential $\psi(w) = \sum_i w[i]\log w[i] - w[i]$; the spectral entropy for matrix valued w, where $\psi(w)$ is the entropy potential on the singular values of w; general quadratic potentials $\psi(w) = 1/2\|w\|_D^2 = 1/2\, w^\top Dw$ for any positive definite matrix D; and the squared ℓ_p norm for $p \in (1, 2]$.

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

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

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

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

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

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

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

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

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

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

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

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

Theorem 1b. *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 converges to* (6)*, i.e., for all $\{\eta_t\}_t, \{\beta_t\}_t, \{\gamma_t\}_t$, if $w_{(t)}$ from eq.* (7) *converges to $w_\infty \in \mathcal{G}$, then $w_\infty = \text{argmin}_{w\in\mathcal{G}}\, \psi(w)$.*

Remark 1. *Theorem 1–1b also hold when stochastic gradients defined in eq.* (2) *are used in place of $\nabla\mathcal{L}(w_{(t)})$.*

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

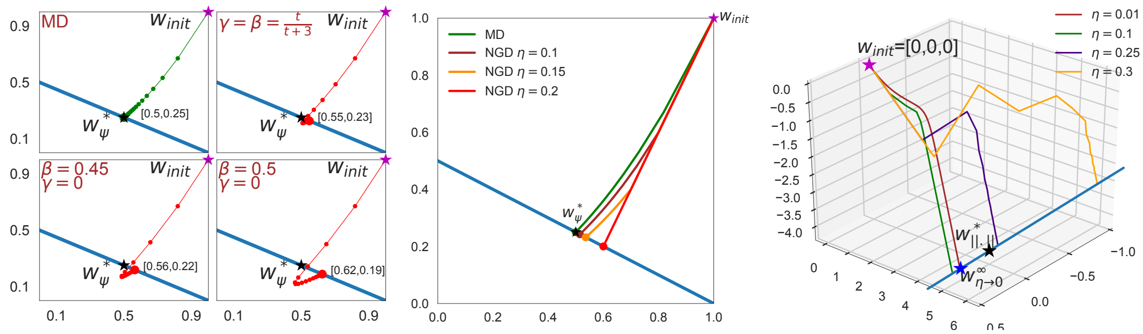
Example 2. *Consider optimizing $\mathcal{L}(w)$ with dataset $\{(x_1 = [1, 2], y_1 = 1)\}$ and squared loss $\ell(u, y) = (u-y)^2$ using primal momentum updates from eq.* (8) *for MD w.r.t. the entropy potential $\psi(w) = \sum_i w[i]\log w[i] - w[i]$ and initialization $w_{(0)} = \text{argmin}_w\,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

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

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,

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



(a) Mirror descent primal momentum (Example 2)

(b) Natural gradient descent (Example 3)

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

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

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

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

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

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

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

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

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

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

2.4. Steepest Descent

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

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

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

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

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

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

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

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

$$\Delta w_{(t+1)} \in \mathrm{conv}\Big\{-\eta_t \frac{\partial \mathcal{L}(w)}{\partial w[j_t]} e_{j_t} : j_t = \underset{j}{\mathrm{argmax}} \Big|\frac{\partial \mathcal{L}(w)}{\partial w[j]}\Big|\Big\},$$

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

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

2.5. Summary for losses with a unique finite root

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

3. Strictly Monotone Losses

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

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

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

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

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

3.1. Gradient descent

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

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

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

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

3.2. Steepest Descent

Recall that for the squared loss, the limit point of steepest descent could depend strongly on the step–size, and we were unable 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 as a maximum margin solution, independent of step–size (as long as it is small enough) and initialization. Let $\|.\|_\star$ denote the dual norm of $\|.\|$.

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

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

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

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

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

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

3.3. Adaptive Gradient Descent (AdaGrad)

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

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

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

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

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

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

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

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

4. Gradient descent on the factorized parameterization

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

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

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

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

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

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

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

Specifically, both the objective as well as gradient descent updates of eq. (16) are equivalent to the problem in eq. (17) with larger data matrices $\widetilde{X}_n = \begin{bmatrix} 0 & X_n \\ X_n^\top & 0 \end{bmatrix}$ and loss optimized over larger p.s.d. matrix $\widetilde{W} = \widetilde{U}\widetilde{U}^\top = \begin{bmatrix} A_1 & W \\ W^\top & A_2 \end{bmatrix}$.

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

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

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

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

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

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

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

Theorem 7. *For almost all datasets $\{X_n, y_n\}_{n=1}^N$ separable by a p.s.d. linear classifier, consider the gradient descent iterates $U_{(t)}$ in eq.* (18) *for minimizing $\bar{\mathcal{L}}(U)$ with the exponential loss $\ell(u, y) = \exp(-uy)$ and the corresponding sequence of linear predictors $W_{(t)}$ in eq.* (19). *For any full rank initialization $U_{(0)}$ and any sufficiently small step–size sequences $\{\eta_t\}$ such that $\eta_t < \infty$ is smaller than the local Lipschitz at $W_{(t)}$, if $W_{(t)}$ asymptotically minimizes $\mathcal{L}$, i.e., $\mathcal{L}(W_{(t)}) \to 0$, and additionally the incremental updates $W_{(t+1)} - W_{(t)}$ and the gradients $\nabla\mathcal{L}(W_{(t)})$ converge in direction, then the limit direction $\bar{W}_\infty = \lim\limits_{t\to\infty} \frac{W_{(t)}}{\|W_{(t)}\|_*}$ exists, and is given by the maximum margin separator with unit nuclear norm $\|.\|_*$,*

$$\bar{W}_\infty = \underset{W \succcurlyeq 0}{\operatorname{argmax}} \min_n y_n \langle W, X_n \rangle \text{ s.t., } \|W\|_* \leq 1.$$

Here we note that convergence of $W_{(t)}$ in direction is necessary for the characterization of implicit bias to be relevant, but in Theorem 7 we require stronger conditions that the incremental updates $W_{(t+1)} - W_{(t)}$ and the gradients $\nabla\mathcal{L}(W_{(t)})$ converge in direction. Relaxing this condition is of interest for future work.

Key property Let us look at exponential loss when $W_{(t)}$ converges in direction to, say $\bar{W}_\infty$ as $W_{(t)} = \bar{W}_\infty g(t) + \rho(t)$ for some scalar $g(t) \to \infty$ and $\frac{\rho(t)}{g(t)} \to 0$. Consequently, the gradients $\nabla\mathcal{L}(W_{(t)}) = \sum_n \mathrm{e}^{-g(t) y_n \langle \bar{W}_\infty, X_n \rangle} e^{-y_n \langle \rho(t), X_n \rangle} y_n X_n$ will asymptotically be dominated by linear combinations of examples X_n that have the smallest distance to the decision boundary, i.e., the support vectors of $\bar{W}_\infty$. This behavior can be used to show optimality of $\bar{W}_\infty$ to the maximum margin solution subject to nuclear norm constraint in Theorem 7.

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

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

5. Summary

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

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

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

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

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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