# Stochastic linear optimization never overfits with quadratically-bounded losses on general data

**S Deepak Narayanan**
ETH Zurich
dsridharan@ethz.ch

**Mael Macuglia**
ETH Zurich
maelm@ethz.ch

## Abstract

Recently, there has been a number of theoretical works that have focused on analyzing the solutions obtained via first-order optimization methods. These works have focused on providing bounds on the Excess Risk using the iterates obtained via stochastic optimization. The paper of focus by in the current report by Telgarsky [28][1] is a work in this direction where regret-style test error bounds are provided for a variety of stochastic optimization methods. In particular the paper limits its focus to the simpler case of linear predictors, under the assumption that the objective is convex and *quadratically bounded*. Quadratic boundedness is a weaker assumption than Lipschitzness, and still encompasses several losses including the most popular ones such as squared-loss and logistic loss used for regression and classification respectively. The paper provides test error bounds for a variety of settings, and the bounds depend on a single proof technique. We begin by providing a motivation and a thorough literature review, then give a version of the proof sketch, which we believe is more accessible and finally conclude with a discussion on potential extensions and their feasibility.

## 1 Introduction

Modern machine learning revolves around deep networks, owing to their immense success in several tasks, ranging from natural language understanding to perception. Learning the parameters of these networks is carried out using first-order optimization methods. This often encompasses stochastic optimization methods such as stochastic gradient descent [3] or it's adaptive variants such as Adam and AdaGrad [16, 6]. While the empirical success of the solutions obtained via optimization based methods has been remarkable, there has been much desired from a theoretical perspective to better formalize and understand these methods. Given this importance, there has recently been plenty of papers that have focused on understanding stochastic optimization based methods and in particular focused on the generalization behavior of iterates or solutions obtained using these methods. This includes works that range from analyzing stochastic optimization for linear predictors, to those that analyze stochastic optimization for deep networks [14, 25, 22, 32].

Telgarsky [28] is a work in this direction, aimed at providing test error bounds for stochastic *linear* optimization. In this setting, we are concerned with linear predictors, i.e., predictors ($\hat{y}$) of the form $\hat{y} = \mathbf{w}^T x$. The core contribution of [28] is a unified proof technique for obtaining test error bounds. The technique is unified in the sense that the same proof idea goes through for different types of data and for different optimization methods - including Stochastic Mirror Descent (with IID, Markovian and Heavy Tailed Data), Temporal Difference Learning (with Markovian Data), and Batch Mirror Descent (with IID Data). While the general framework of the proof technique remains the same in all of these cases, there are subtle yet important changes made either in terms of the assumptions, or in terms of the tools used for the different cases stated above.

[1] From hereon, whenever we refer to this paper, we cite the paper for brevity.

GML Course Project - 2022.

A couple of key areas in which [28] differs from other prior work are (a) Assumptions: The current work provides test-error bounds for *unconstrained* optimization. This is in stark contrast with multiple prior work, that often assume constraints, or regularizations [21, 12] to control the iterates and obtain bounds. The proof techniques also make weaker assumptions on the objective as opposed to related work [12, 21, 24]. (b) Rates that are locally adapted. The test error bounds provided are with respect to a reference solution ($w_{\text{ref}}$). This is different from related work which use the optimum solution [7, 12, 15, 21]. This a carefully designed notion that is useful in bounding the test error of stochastic optimization methods in a non-asymptotic manner with high probability. Further, this notion of reference solution can help obtain both tighter bounds and address corner cases that previous methods fail to tackle. More interestingly, we can draw comparisons with appropriate choices of $w_{\text{ref}}$ solutions to early stopped solutions. Having introduced the content of the report we now summarize our contributions. Our key contributions in this report are the following:

1. We conduct a thorough literature review, going beyond directly related work and bring together the notions of reference solution, regularized solutions and early stopped solutions.
2. We present the technical proof in the paper in what we believe is a much more approachable and direct manner. We achieve this by first relating the underlying optimization problem and the goal of the paper. We then analyze the bounds in detail, reason about their tightness and crucially highlight how all the different assumptions are put together obtain the test error rates.
3. We compare and contrast two different setups and discuss how the same proof technique uses different tools to achieve the same goal.
4. We finally conclude with an attempt to extend the proof techniques to coordinate descent and argue that unless very strong assumptions are made, such an extension is not feasible.

## 2 Related Work

Central to [28] is a plethora of research works on stochastic convex optimization. In this line of research, the main goal is to minimize a convex function ($F$) over a convex domain ($\mathcal{W}$). However $F$ is unknown and we have access only to a stochastic subgradient oracle, which returns an unbiased estimator of the subgradient at points over the domain. The natural questions to ask are a) How many calls do we need to make to the oracle? and b) What convergence rates do we hope to achieve?

Rakhlin et al [21] analyze this setting for strongly convex and smooth objectives, and prove that Stochastic Gradient Descent (SGD) can achieve optimal rates. However to achieve these rates, they assume bounded subgradient norm. While a common assumption, this is a consequence of either good initialization or by assuming a compact domain, both of which are restrictive and not generally the case in practice. More recently, Harvey et al [12] consider stochastic optimization in the setting where the objective is Lipschitz and strongly convex and provide high probability errors on the final iterate. Similar to [21] they make assumptions that the domain is closed, allowing them to control the subgradient norms. Li and Orabona [17] prove results on nonconvex optimization with adaptive SGD. However they make assumptions on the gradient noise which ensures boundedness of the iterates.

Other differences in existing literature is that the bounds tend to be often in Expectation and not in high probability [11, 13]. For Temporal Difference Learning for instance, Telgarsky [28] is the first work to provide high probability bounds. Some of the other results that analyze stochastic optimization methods, or it's variants often tailor their proofs to very specific settings making their techniques not generic enough to be applied to different problems. For example much research on logistic regression is tailed to separable data with exponentially tailed losses [25, 26, 15].

Another critical difference between prior works and Telgarsky [28] is in the lack of analysis and bounds for general data. Particularly, most of the analysis restrict themselves to data that can be well controlled . There has not been much results analysing stochastic mirror descent in the context of heavy-tailed data. Telgarksy [28] on the other hand provides high probability test error bounds for this related problem and the most closely related recent work is of that of Vural et al [29], which provides bounds in expectation. There is also related work that tend to modify either the stochastic updates or use procedures such as gradient clipping to control the subgradient norms for heavy tailed data while using stochastic optimization [8, 5, 18].

Another result from Telgarsky[28] is high-probability test error bounds for unconstrained batch mirror descent (and batch gradient descent). Here, data is drawn IID from an underlying density, and the updates are carried out in a batch manner. This setting can be compared with standard gradient descent based results. A line of research discusses the implicit bias or implicit regularization of gradient descent based solutions [26, 14, 19]. These papers demonstrate that gradient descent or gradient flow (gradient descent with infinitesimal learning rate) asymptotically points towards the maximum margin direction when trained with logistic loss, and converges to minimum norm solution for squared loss. In fact, similar results have been shown to hold for interesting problems such as matrix factorization where it has been shown that gradient descent converges to the minimum nuclear norm solution [9].

A recent work by Suggala et al [27] proves that gradient flow paths are pointwise close to the paths obtained by corresponding regularized solutions for strongly convex problems. As a result, they were able to use results from optimization to analyzing regularized solutions and vice versa, and obtain tight excess risk bounds. This helps better understand optimization based solutions' inherent ability to generalize well by relating them to regularized solutions. Ali et al [2] observed that using gradient descent for a large number of iterations is similar to having no $L_2$ regularization and using gradient descent for fewer iterations is similar to having high $L_2$ regularization. Based on this observation, they prove that gradient flow and ridge regression risks are essentially the same with the ridge parameter set to the inverse of number of descent iterations. Here the key takeaway is that running gradient descent for an appropriate number of iterations (not very few, and also not very large) might give us the right balance for generalization – showing that early stopping after a right number of iterations is important. Statistical rates for early stopping were also shown using the same quantities as in regularized regression in [31]. The notion of early stopped solution is also related to the notion of reference solutions; This is central to the proofs in Telgarsky [28] – comparing with good reference solutions might be more useful than comparing with optimal solutions.

Something to keep in mind is that the entire proof in Telgarsky [28] is driven by the so called "implicit bias" term, which pops up in the analysis of Mirror Descent. We will highlight it in the upcoming section, and draw connections to the works mentioned above.

## 3 The Coupling Based Proof Technique

In this section we present the core contribution of Telgarsky [28] – the coupling based proof technique that is used to derived test error bounds for stochastic optimization procedures associated with machine learning problems. We will focus on Stochastic Mirror Descent with IID data and Mirror Descent on full batch data (see Theorem 3.1 and 3.2). One of our aims is to bring clarity about the risk bounds by presenting the proofs in a different and structured manner compared with Telgarsky [28] in the following sense: for all risk bounds provided, we relate the underlying optimization problem, the corresponding machine learning problem and the optimization procedure used. We further discuss and analyze the bounds in detail. We restrict ourselves to these two settings as they highlight the core idea and the differences in dealing with different situations – much of the other results follow the same proof outline.

### 3.1 The Optimization Problem

We assume that we are given a data set $\mathcal{D}_n := \{(x_i, y_i)\}_{i<n}$ where each $(x_i, y_i)$ is drawn IID from the random vector $(X, Y)$ taking values in $(\mathcal{X} \times \mathcal{Y})$. The goal is to predict $y \in \mathcal{Y}$ using the conditional $Y|X$. In other words, given a new realisation $(x, y) \sim (X, Y)$ where we are only able to observe $x$, we aim at making a prediction on $y$. Mathematically the following framework allows us to formalize the latter idea: Find a function $f : \mathcal{X} \to \mathcal{Y}$ such that the random variable $f(X)$ is a good approximation of $Y$. In this report, we consider linear predictors, as in Telgarsky [28] meaning $f$ has the following form $f(x) = \langle w, x \rangle$ where $w \in \Omega$, $\Omega$ being some parameter space. Next, define the notion of loss, risk and excess risk in order to quantify "how good" the predictor $f$ is:

$$\text{Loss} \qquad \ell : (\mathcal{X}, \mathcal{Y}) \times \Omega \to \mathbb{R} \tag{1}$$

$$\text{Risk} \qquad \mathcal{R}(w) = \mathbb{E}_{X,Y} \ell(y, \langle w, x \rangle) \tag{2}$$

$$\text{Excess Risk} \qquad \mathcal{E}(w) = \mathcal{R}(w) - \inf_{v \in \Omega} \mathcal{R}(v) \tag{3}$$

Now we are able to define the stochastic optimization problem

$$\min_{w\in\Omega} \mathcal{R}(w) = \min_{w\in\Omega} \mathbb{E}_{X,Y}\ell(y, \langle w, x\rangle) \tag{4}$$

where the risk $\mathcal{R}$ is assumed to be convex and the loss function $\ell$ is $(C_1, C_2)$ quadratically bounded, and the domain is $\Omega$.

**Definition 3.1** (Quadratic Boundedness)**.** A loss function $\ell$ is $(C_1, C_2)$ quadratically bounded (for $C_1, C_2 \geq 0$) if $|\ell'(y, w^T x)| \leq C_1 + C_2(|y| + |w^T x|) \quad \forall y, w^T x$ where $\ell' \in \partial\ell$, the sub-gradient set of $\ell$ at $w$.

Remark: If $\ell$ is $L-$lipschitz, then $\ell$ is $(L, 0)$ quadratically bounded. If $\ell$ is $\mu-$smooth, then $\ell$ is $(|\partial\tilde{\ell}(0)|, \mu)$ quadratically bounded. Here, $\tilde{\ell}(.)$ is an auxilary scalar loss function. We omit the details of this as it is unimportant. The takeaway is that quadratically bounded losses encompasses both Lipschitzness and Smoothness assumptions made in prior work, making it a *much* weaker assumption. In particular, the squared loss is $(0, 1)$ quadratically bounded, and the logistic loss is $(1, 0)$ quadratically bounded. As we will note later, the role of these constants $C_1$ and $C_2$ are crucial in playing a role in the tightness of the bounds.

### 3.2 The Algorithm

Given a Stochastic Optimization Problem of the same form as last section (3.1), the fixed point iterative methods discussed in this report are captured by the general framework of Stochastic Mirror Descent Algorithm (referred to as SMD from hereon). Informally, SMD is a generalization of SGD to non-Euclidean geometry. Given iterates $w_0, w_1, \ldots, w_i$, the SMD update is mathematically defined using the following optimization problem.

$$w_{i+1} := \arg\min_{w\in\Omega} \left\{\langle \eta g_{i+1}, w\rangle + D_\psi(w, w_i)\right\} \tag{5}$$

where $\mathbb{E}g_{i+1}|\{w_j\}_{j<i} \in \partial R(w_i)$, the subgradient set of the objective function with respect to $w$ evaluated at $w_i$ and the step length is $\eta$. The quantity $D_\psi$ defined the Bregman divergence, where the mapping $\psi$ is 1-strongly convex with respect to a norm on $\Omega$.

### 3.3 The Reference Solution

Typically, bounds provided in optimization based proofs compare the performance of the descent iterates with the optimal solution $w^*$ ($w^* := \arg\min_{w\in\Omega} R(w)$) of the optimization problem 3.1. However, comparing with such an optimal solution might not be particularly useful. This could be because the optimal solution is at infinity, or it might be the case that it does not characterize the descent dynamics in an effective manner.

To illustrate this point better, we consider the four point distribution used in Telgarksy [28] and reproduce the results in Figure 1. The two points that are in red have 90 percent probability while the two blue points have 10 percent probability. All the 4 datapoints have label +1. We run 100 parallel runs of SGD for 500 iterations (the trajectories in green). We also run population gradient descent (the trajectory in blue). We run both SGD and population gradient descent with the squared loss and logistic loss as mentioned in Figure 1. The iterates are initialized at the origin.

We can observe that comparing with the population optima in the case of squared loss or the maximum margin solution in the case of logistic loss may not be helpful in characterizing the early phase of training, the circled region in Figure 1. A reference solution (the red arrow in Figure 1) is the solution the descent iterates are compared to, in Telgarksy [28], and we formalize it below.

Mathematically, we require the reference solution to satisfy the following property:

$$\mathcal{E}(w_{\text{ref}}) \leq \frac{1}{\sqrt{t}} D_\psi(w_{\text{ref}}, w_0) \tag{6}$$

where $t$ is the number of iterations of running the first-order stochastic optimization procedure, and $w_0$ is the initial iterate. Intuitively what this says is that if we are willing to run SMD for a large number of iterations, then we require that the excess risk of the reference solution is small, and vice versa. This is reasonable since the longer we run these algorithms, the closer we can hope to get to the

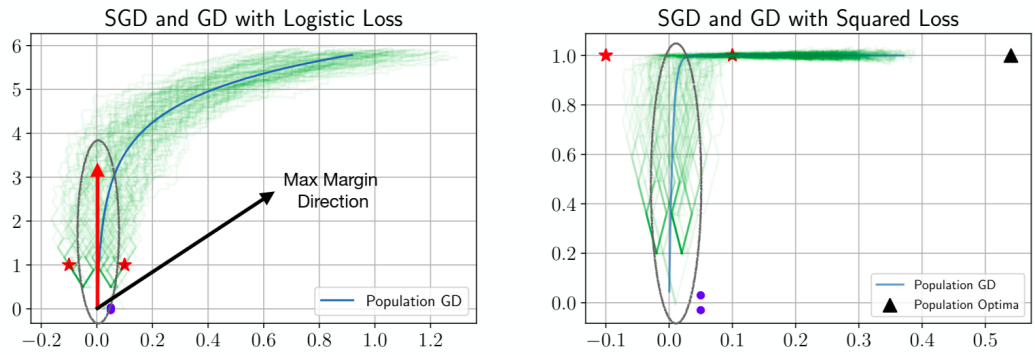


Figure 1: The red arrow depicts a potential $w_{\text{ref}}$. Comparing with $w_{\text{ref}}$ can better characterize the early stage of training. Observe that $w_{\text{ref}}$ can be a low-norm low-risk solution.

optimal solution and we need "better" solutions to characterize this behavior. It is important to note that such a mathematical requirement on the reference solution is a rather cleverly reverse-engineered requirement for the proofs to go through. In practice, we can guarantee the existence of such a $w_{\text{ref}}$, but finding the exact $w_{\text{ref}}$ requires knowledge of the data distribution.

To make this statement more concrete, consider the following regularized optimization problem

$$u_{\text{ref}}(\lambda) := \arg\min \left\{ \mathcal{R}(u) + \frac{\lambda}{2} D_\psi(u, w_0) : u \in \mathbb{R}^d \right\}$$

It can be shown that for any $\lambda > 0$, $\mathcal{E}(u_{\text{ref}}(\lambda)) \leq \lambda D_\psi(u_{\text{ref}}(\lambda), w_0)$. In particular, the natural choice of $\lambda = \frac{1}{\sqrt{t}}$ gives us the required bound on the reference solution – i.e; $w_{\text{ref}} = u_{\text{ref}}(\frac{1}{\sqrt{t}})$ is one way to obtain a $w_{\text{ref}}$ with the required mathematical property.

More importantly, the fact that the reference solution can be viewed as a solution to a regularized problem can help us to draw connections with early stopped solutions. Recall from Section 2 where we discussed the work on Ali et al [2] in the context of full batch gradient descent. We noted there that the risk incurred by that of gradient descent iterates follows that of the ridge regularized iterates very closely. This connection gives an alternate viewpoint about the reference solutions – they can be viewed as early-stopped solutions. This connection can help us view the bounds in a different light – How good are the iterates of a stochastic optimization procedure as compared to an optimally regularized solution?

In summary, the reference solution can (a) Help characterize the early stage training dynamics better; (b) Can be more representative of the learning dynamics in the case that the optimal solution is very far away or at infinity and (c) Can even be a better solution to compare to, especially from a generalization perspective as it can be viewed as a solution to a regularized problem on the population risk!

Having motivated the notion of reference solutions and having drawn connections to early stopping, we will now discuss the proof sketch.

### 3.4 Proof Sketch with SMD

Assuming an unconstrained stochastic optimization problem of the form 3.1 where $\Omega = \mathbb{R}^d$, the proof technique consists of the following ideas.

1. Couple the unconstrained SMD iterates $\{w_i\}_{i<t}$ with constrained iterates $\{v_i\}_{i<t}$. Here the latter are constrained as they live within a ball centered at $w_{\text{ref}}$ of radius $B_w$.
2. Demonstrate that the unconstrained iterates $\{w_i\}_{i<t}$ remain with high-probability within the ball of radius $B_w$ via an implicitly biased analysis of SMD, and choosing $B_w$ appropriately.
3. Use concentration inequalities together with implicitly biased SMD analysis to provide high probability bounds on the objective function (i.e the population risk).

### 3.4.1 Coupling unconstrained iterates with constrained iterates

As has been noted in Section 2, we need some way of controlling the iterates in order to obtain reasonable bounds. In Telgarsky [28] this is achieved using concentration inequalities. Specifically, the idea is to argue that the unconstrained iterates of the optimization procedure are the same as the constrained set of coupled iterates, with high-probability. We can then apply concentration inequalities on the unconstrained iterates.

Therefore, define the projected iterates $\{v_i\}_{i<n}$ coupled to the unconstrained iterates $\{w_i\}_{i<n}$ in the following way: $v_0 = w_0$; $v_{i+1}$ is coupled to $w_{i+1}$ as it shares the same randomness and is defined as follows

$$w_{i+1} := \arg\min_{w\in\mathbb{R}^d} \left\{\langle \eta g_{i+1}, w\rangle + \mathcal{D}_\psi(w, w_i)\right\} \tag{7}$$

$$v'_{i+1} := \arg\min_{v\in\mathbb{R}^d} \left\{\langle \eta h_{i+1}, v\rangle + \mathcal{D}_\psi(v, v_i)\right\}$$

$$v_{i+1} = \Pi_S\left(v'_{i+1}\right) \tag{8}$$

where $g_{i+1} \in \partial_w \ell(y_{i+1}, x_{i+1}^T w_i)$; $h_{i+1} \in \partial_v \ell(y_{i+1}, x_{i+1}^T v_i)$ and $S := \left\{v \in \mathbb{R}^d : \|v - w_{\text{ref}}\| \le B_w\right\}$

The main idea in Telgarsky [28] is to choose the radius $B_w$ large enough such that the unconstrained iterates $\{w_i\}_{i<n}$ lie inside the ball with high probability. Therefore, the projection step is never invoked and results in the constrained and unconstrained iterates being the same.

### 3.4.2 Implicitly Biased Analysis of MD

Assume a mirror map $\psi$ and $w_{\text{ref}}$, $v_0 \in S \cap \text{dom}(\psi)$ to be given, where $S$ is a closed convex set. Further assume an optimization problem of the form 3.1. Running the SMD algorithm (3.2) with $\Omega = S$ for $t$ iterations provides the set of iterates $\{w_i\}_{i<t}$. By conventional Mirror Descent analysis, for convex $\ell$, it can be shown that the following inequality holds:

$$\underbrace{D_\psi(w_{\text{ref}}, w_t)}_{\text{"Implicit Bias"}} \le D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{i<t} \left[\ell_{i+1}(w_{\text{ref}}) - \ell_{i+1}(w_i)\right] + \sum_{i<t} \frac{\eta^2}{2} \|g_{i+1}\|_*^2 \tag{9}$$

In the above inequality, $\ell_{i+1}(w) := \ell(y_{i+1}, x_{i+1}^T w_i)$, and $g_{i+1}$ is the same defined in Section 3.2. In standard mirror descent analyses the term on left hand side (LHS), marked as "implicit bias" above, is often ignored as it is a non-negative quantity and lower bounded by zero [4, 7]. However, observe that the LHS is precisely the "distance" between iterate $w_t$ and the center of the ball discussed in the previous section. To ensure that this iterate stays within the ball defined earlier, we need to ensure that the right hand side is less than or equal to $aB_w^2, a \le 1$.

$$D_\psi(w_{\text{ref}}, w_t) \le D_\psi(w_{\text{ref}}, w_0) + \underbrace{\eta \sum_{i<t} \left[\ell_{i+1}(w_{\text{ref}}) - \ell_{i+1}(w_i)\right]}_{\text{Concentration}} + \sum_{i<t} \frac{\eta^2}{2} \|g_{i+1}\|_*^2 \tag{10}$$

Using the property that $\ell$ in quadratically bounded, and with regularity assumptions on $\ell_i(w_{\text{ref}}), (|\ell_i(w_{\text{ref}})| \le C \text{ for } C \ge 0)$, we can in fact use standard concentration inequalities like Azuma-Hoeffding [30] to bound the RHS above.

We repeat the process for each iterate $w_i, i \le t$ as we want that $\{w_i\}_{i\le t} = \{v_i\}_{i\le t}$. The crucial thing is that once we have that all the unconstrained iterates essentially live inside the ball mentioned above, we can first swap out the $w_t$ for the constrained iterates $v_t$ (with high probability). We can then apply concentration inequalities to obtain regret-style test-error bounds. We now present two specific cases that are of particular interest and discuss the differences in the assumptions and proofs carried out by Telgarsky [28].

## 3.5 Realizable Case of Stochastic Mirror Descent with IID Data

### 3.5.1 Theorem Analysis and Discussion

**Theorem 3.1.** *Let $\ell$ be convex, and $(C_1, C_2)$ quadratically bounded. Let $t$ be the number of iterations. Suppose $((x_i, y_i))_{i\le t}$ are IID samples, and let $\max\{\|x_i\|_* |y_i|\} \le 1$, with probability 1. Assume*

*$\rho$ self-boundedness and $|\ell_i(w_{ref})| \leq C_4, \forall i$, with probability 1. Further in the realizable case, $\mathcal{R}(w_{ref}) \leq \rho D_\psi(w_{ref}, w_0)/t$. Given $w_0, w_{ref}$, for $\eta \leq \frac{1}{2\rho}$, with probability at least $1 - 2t\delta, \forall i \leq t$, the following holds.*

$$\frac{8}{3i\eta} D_\psi(w_{ref}, w_i) + \frac{1}{i}\sum_{j<i} \mathcal{R}(w_j) - \frac{4}{\eta}\mathcal{R}(w_{ref}) \leq \frac{2B^2}{i\eta}$$

*where* $B := \sqrt{1 + C_1 + C_2(1 + \|w_{ref}\|)} B_w; B_w := \max\left\{1, 4\sqrt{D_\psi(w_{ref}, w_0)}, \sqrt{\frac{64C_4}{\rho}\ln\frac{1}{\delta}}\right\}$

In this realizable case we have $\mathcal{E}(w_{\text{ref}}) = \mathcal{R}(w_{\text{ref}})$; This is not a very strong assumption, as in many cases we do have $\inf_v \mathcal{R}(v) = 0$. For example, with separable data in classification settings. Also, our mathematical requirement on the reference solution has changed to $\mathcal{O}(\frac{1}{t})$ v/s $\mathcal{O}(\frac{1}{\sqrt{t}})$[2]. As noted earlier, we are always guaranteed the existence of such solutions. Finally, $\rho$ self-boundedness is this condition: $\ell'(w) \leq 2\rho\ell(w)$, where $\ell' \in \partial\ell$. This property is true for squared and logistic losses. It is important for obtaining good bounds as it helps replace $\ell'$ with $\ell$ upto scaling in inequality 9 on the RHS. This helps us apply concentration bounds that help introduce the Risk terms into the equation. Otherwise, we are stuck with scalar derivative terms, that even under concentration don't give us useful quantities. We make this more concrete mathematically later.

Further we can analyze the constant $B$. Observe that in very high dimensions the norm of $\|w_{\text{ref}}\|$ might be a problematic quantity to deal with. The "distance" between the reference solution and the initial point might also be problematic if it is the dominating term in $B_w$. Observe that this dependency on the norm can be rid off, if $C_2$ is 0. This implies that the bounds are inherently tighter for the logistic loss, instead of the squared loss[3], and that turns out to be a consequence of *Lipschitzness* of the Logistic Loss. Also important to note that the bounds are potentially tighter than those in multiple related work [12, 21], that carry bounds depending on the diameter of the compact set of their constrained optimization problems. The diameter could be much larger than $B^2$ in the proof above. Also note that the above setting is like that of the online setting, where we get a new datapoint at every iteration and we provide test-error bounds for each iteration.

### 3.5.2 Proof Overview

Assuming $t$ iterations of SMD 3.2, we have the following inequality restated from above

$$D_\psi(w_{\text{ref}}, w_t) \leq D_\psi(w_{\text{ref}}, w_0) + \eta\sum_{i<t}\left[\ell_{i+1}(w_{\text{ref}}) - \ell_{i+1}(w_i)\right] + \sum_{i<t}\frac{\eta^2}{2}\|g_{i+1}\|_*^2 \tag{11}$$

To demonstrate that unconstrained iterates as the same as the constrained iterates, consider the RHS in the equation above. We have, from data regularity, self boundedness and the bound on $\eta$,

$$\sum_{i<t}\frac{\eta^2}{2}\|g_{i+1}\|_*^2 = \sum_{i<t}\frac{\eta^2}{2}\ell'(w_i)^2\|x_{j+1}y_{j+1}\|_*^2 \leq \sum_{i<t}\eta^2\rho\ell_{i+1}(w_i)$$

Thanks to self boundedness, we were able to swap out the derivative with the loss, thereby allowing us to use standard concentration inequalities (that have a martingale structure), on the loss, to obtain risk bounds. Therefore we have

$$\begin{aligned}
D_\psi(w_{\text{ref}}, w_t) &\leq D_\psi(w_{\text{ref}}, w_0) + \eta\sum_{i<t}\left[\ell_{i+1}(w_{\text{ref}}) - (1-\eta\rho)\ell_{i+1}(w_i)\right] \\
&\leq D_\psi(w_{\text{ref}}, w_0) + \eta\sum_{i<t}\left[\ell_{i+1}(w_{\text{ref}})\right] (\text{Since } \ell \geq 0) \\
&\leq D_\psi(w_{\text{ref}}, w_0) + \underbrace{\frac{5i\eta}{4}\mathcal{R}(w_{\text{ref}}) + 4C_4\ln\left(\frac{1}{\delta}\right)}_{\text{By Martingale Concentration}} \\
&\leq \underbrace{D_\psi(w_{\text{ref}}, w_0) + \frac{5D_\psi(w_{\text{ref}}, w_0)}{8} + 4C_4\ln\left(\frac{1}{\delta}\right)}_{\text{Need this to be } \leq B_w^2\,\forall t}
\end{aligned}$$

[2]As noted earlier these conditions are rather reverse engineerined to obtain good rates.

[3]Recall that Logistic Loss is (1, 0) - Quadratically Bounded and Squared Loss - (0, 1) Quadratically Bounded

Note that in the penultimate step Telgarsky [28] uses a variant of Freedman's inequality, first introduced by Agarwal et al [1]. The step from the penultimate term to the last term highlights the reverse engineering done to obtain the bound on $\mathcal{R}(w_{\text{ref}})$. Note that choosing $\mathcal{R}(w_{\text{ref}}) \leq \mathcal{O}(1/t^{\alpha})$, $0 < \alpha < 1$ would mean the diameter of the ball would then have a dependence on the number of iterations, something that we don't desire. Now the value of $B_w$ is clear from the above bound, which was also mentioned in $B$ appearing in Theorem 3.1. Crucially it depends only on the *known* quantities. Note that we need the above concentration for all $i < t$. That requires a union bound and therefore with probability at least $1 - t\delta$, we have the unconstrained iterates and constrained iterates being the same.

After plugging in $\eta$, and rewriting the inequality above, we have,

$$\begin{aligned} D_{\psi}(w_{\text{ref}}, w_t) &\leq D_{\psi}(w_{\text{ref}}, w_0) + \frac{\eta}{2} \sum_{i<t} \left[2\ell_{i+1}(w_{\text{ref}}) - \ell_{i+1}(w_i)\right] \\ &= D_{\psi}(w_{\text{ref}}, w_0) + \underbrace{\frac{\eta}{2} \sum_{i<t} \left[2\ell_{i+1}(w_{\text{ref}}) - \ell_{i+1}(v_i)\right]}_{\text{Concentration Here}} \\ &\leq D_{\psi}(w_{\text{ref}}, w_0) + \eta \sum_{i<t} \left[\frac{5}{4}\mathcal{R}(w_{\text{ref}}) - \frac{3}{8}\mathcal{R}(v_i)\right] + \mathcal{O}\left(\ln(1/\delta)\right) \end{aligned}$$

simplifying which we get the bound stated in Theorem 3.1. Here we were able to provide a concentration, thanks to quadratic boundedness and the equality above which is true with high probability from the previous step. For all $t$ on the LHS, we again do a union bound for the result.

### 3.6 The Case of Mirror Descent on full batch data

#### 3.6.1 A Different Stochastic Optimization Problem

In this section the underlying minimization problem is a slight variation of the optimization problem from section 3.1. In practice, the underlying probability measure appearing in the risk equation (2) is unknown, i.e $\mathbb{P}_{X,Y}$ is an unknown quantity in the following equation

$$\min_{w \in \Omega} \mathcal{R}(w) = \min_{w \in \Omega} \int \ell(y, \langle w, x \rangle) d\mathbb{P}_{X,Y}$$

In order to have a computable quantity, given a data set $\mathcal{D}_n := \{(x_i, y_i)\}_{i<n}$, the unknown probability measure $\mathbb{P}_{X,Y}$ is replace by its empirical counterpart yielding the following empirical minimization problem (ERM)

$$\min_{w \in \Omega} \hat{\mathcal{R}}(w) = \min_{w \in \Omega} \frac{1}{n} \sum_{i \in [n]} \ell(y_i, \langle w, x_i \rangle) \tag{12}$$

#### 3.6.2 Theorem Analysis and Discussion

In the case of Mirror Descent with Batch Data, this change of the optimization problem necessitates a change in the approaches used for the same proof technique to go through. In particular, Telgarsky [28] utilizes Rademacher bounds instead of the double concentration we exploited in the previous sketch.

**Theorem 3.2.** *Suppose $\ell$ is convex and $(C_1, C_2)$ quadratically bounded. Suppose $((x_i, y_i))_{i \leq n}$ are drawn IID and let the regularity conditions on the features and labels from Theorem 3.1 hold true. Consider the Mirror Descent Algorithm from Section 3.2 run for $t$ iterations. Suppose we have $t \leq n$. Let $w_{ref}$, $w_0$ be given. Replace $g_{i+1}$ by $\partial_w \hat{\mathcal{R}}(w_i)$ in Equation 3.2, where $\hat{\mathcal{R}}(w_i) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, w_i^T x_i)$. Let the assumption on the reference solution hold true from Equation 6. Then with probability at least 1 - $4\delta$, $\forall i \leq t$*

$$\frac{1}{i\eta} D_{\psi}(w_{ref}, w_i) + \frac{1}{i} \sum_{j<i} \mathcal{R}(w_j) - \mathcal{R}(w_{ref}) \leq \frac{B_w^2}{8i\eta}$$

Here $\eta \leq \mathcal{O}(1/\sqrt{t})$ and $B_w = \max\left\{1, \mathbf{1}[C_2 \geq 0]\|w_{\text{ref}}\|, 4\sqrt{D_{\psi}(w_{\text{ref}}, w_0)}\right\}$.

Note here that the learning rate now depends on the number of iterations and the number of iterations has to be smaller than the number of datapoints. The former and latter and tightly intertwined. If we don't assume the latter, the radius of the ball becomes a function of $n$ and $t$, which would make the proof very limited in scope. To fix these issues, $\eta$ has to be a function of $t$ and $t$ has to be less than $n$. Further, as remarked earlier we still get the benefits of a Lipschitz objective in obtaining tighter bounds. Moreover, this section is the one that is most closely related to early stopping [2] and it's connections to regularized solutions. Recall that Mirror Descent on Batch Data is a generalization of standard Gradient Descent.

### 3.6.3 Proof Overview

Given a reference solution $w_{\text{ref}}$, applying Mirror Descent analysis results (3.4.2) to the optimization problem (12) yields the following

$$D_\psi(w_{\text{ref}}, w_i) \leq D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{j<i} \left[\hat{\mathcal{R}}(w_{\text{ref}}) - \hat{\mathcal{R}}(w_i)\right] + \eta^2 \sum_{j<i} \|\nabla \hat{R}(w_j)\|_*^2 \tag{13}$$

where it is assumed that the unconstrained iterates $\{w_j\}_{j<i}$ from MD procedure are coupled with the constrained iterates $\{v_j\}_{j<i}$ in the same strong sense as in the proof sketch 3.4.1 over a convex set $S := \left\{v \in \mathbb{R}^d : \|v - w_{\text{ref}}\| \leq B_w\right\}$. For now, let us assume $\{w_j\}_{j<i} = \{v_j\}_{j<i}$. We prove this after deriving the high probability bounds on the risk $\mathcal{R}$. Therefore we relate the empirical risk with $\mathcal{R}$ using Rademacher Generalization Bounds on $S$. This, along with quadratic boundedness of the loss gives us, with high probability,

$$\sum_{j<i} \left[\hat{\mathcal{R}}(w_{\text{ref}}) - \hat{\mathcal{R}}(v_j)\right] \leq \sum_{j<i} \left[\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)\right] + \mathcal{O}\left(\frac{t}{\sqrt{n}} C_3\right) \tag{14}$$

where $C_3 := (C_1 + C_2(2B_w + \|w_{\text{ref}}\|)$ Further, quadratic boundedness together with data regularity assumptions yields the following upper bound on the norm of the empirical risk:

$$\|\partial \hat{\mathcal{R}}(v_j)\|_* \leq C_3 \tag{15}$$

Plugging equations (14) and (15) into (13), we have,

$$D_\psi(w_{\text{ref}}, w_i) \leq D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{j<i} \left[\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)\right] + \eta^2 t C_3^2 + + \mathcal{O}\left(\eta \frac{t}{\sqrt{n}} C_3\right) \tag{16}$$

$$\leq \mathcal{O}\left(B_w^2\right) + \eta \sum_{j<i} \left[\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)\right] \tag{17}$$

This together with the learning rate $\eta \leq \mathcal{O}(1/\sqrt{t})$ yields the required bounds. Observe that the choice of $\eta$, and the condition $t \leq n$ are intertwined and critical: with another choice of $\eta$, we will obtain *vacuous* bounds as the bounds will grow in looseness as $t$ increases!

It remains to show that the assumption $\{w_j\}_{j<i} = \{v_j\}_{j<i}$ is true. We will prove this inductively. The base case is true by definition. Using the reference solution condition $\mathcal{E}(w_{\text{ref}}) \leq D_\psi(w_{\text{ref}}, w_0)/\sqrt{t}$ and equation (17), and the non-negativity of $\mathcal{R}$, and assuming $\{w_j\}_{j<i} = \{v_j\}_{j<i}$ we have for $w_i$, from the equation above that,

$$D_\psi(w_{\text{ref}}, w_i) \leq \mathcal{O}\left(B_w^2\right) + \underbrace{\eta \sum_{j<i} \left[\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)\right]}_{\mathcal{O}(B_w^2)} \leq a B_w^2 \tag{18}$$

where $a \leq 1$ for appropriate choice of constants, which proves $w_i = v_i$.

In summary, the most important takeaway from the two proof overviews is the different approaches that the proof technique introduced by Telgarsky [28] takes because of the different optimization problems that they solve. While the overarching idea still remains the same, the tools to achieve them are vastly different – Rademacher bounds versus a martingale inequality. The central idea in both proofs is to not discard the implicit bias term and find a ***sufficiently*** large radius so that the unconstrained iterates can be controlled. The choices of $\eta$ and the assumption on the reference

solution are all intertwined: They are essentially reverse-engineered in a clever way to make the proofs go through and have bounds that don't become arbitrarily loose with regard to number of iterations. We do not discuss the details of the other results presented in [28], as we believe we have covered the essential idea and discussed in detail the different aspects of the proof technique that makes it work.

# 4 The Case of Coordinate Descent

In this section, we explore if these proof techniques can be used for the case of Coordinate Descent. In coordinate descent, at each iteration, one of the coordinates is updated. There are plenty of ways to pick which coordinate to update such as cyclical, Gauss-Southwell rule and uniform selection [23, 20, 10]. Regardless of which approach is used, we find that the initial step of controlling the unconstrained iterates via coupling is not straightforward unless very strong (and impractical) assumptions are made. Formally, assume that we are interested in minimizing the empirical risk, as in the batch mirror descent case. Further let us assume that the objective is convex and $L$-smooth (we only needed convexity in the previous proofs[4]). Denoting the $i^{\text{th}}$ coordinate of the gradient as $\nabla \ell(w)_i$, the update rule in generic coordinate descent is as follows (we assume some coordinate $i_t \in [d]$ is used at step $t$):

$$w_{t+1} = w_t - \eta \nabla \ell(w_t)_{i_t} \mathbf{e}_{i_t}$$

We need to relate it to a reference solution and to obtain an inequality similar to the implicitly biased ones we used. To achieve that, we subtract $w_{\text{ref}}$ from both sides and square, obtaining

$$\|w_{t+1} - w_{\text{ref}}\|^2 = \|w_t - w_{\text{ref}}\|^2 + \eta^2 \nabla \ell(w_t)_{i_t}^2 - 2\eta \langle w_t - w_{\text{ref}}, \nabla \ell(w_t)_{i_t} \mathbf{e}_{i_t} \rangle$$

Taking summation from $t = 0$ to $T - 1$ both sides, we have

$$\|w_T - w_{\text{ref}}\|^2 \leq \|w_0 - w_{\text{ref}}\|^2 + \eta^2 \sum_t \nabla \ell(w_{i_t})^2 - 2\eta \sum_t \langle w_t - w_{\text{ref}}, \nabla \ell(w_t)_{i_t} \mathbf{e}_{i_t} \rangle$$

In the case of realizable SMD, we could control $\sum_t \nabla \ell(w_t)_{i_t}^2$ using self boundedness. In the case of Batch Mirror Descent, we could control using quadratic boundedness, Rademacher bounds and the properties of the ball. In this case, unless we assume that $\ell$ is Lipschitz or equivalent, there is no direct way to control the term. Now the more problematic term is $\sum_t \langle w_t - w_{\text{ref}}, \nabla \ell(w_t)_{i_t} \mathbf{e}_{i_t} \rangle$. In standard mirror descent analysis, we invoke convexity of the objective to relate this term (or its generalization) to the objective values at the current iterate and at the reference solution. More precisely, from first order definition of convexity we have

$$\ell(w_{\text{ref}}) \geq \ell(w_t) + \langle \nabla \ell(w_t), w_{\text{ref}} - w_t \rangle$$

We therefore need to demonstrate that $\langle \nabla \ell(w_t), w_{\text{ref}} - w_t \rangle \geq \langle \nabla \ell(w_t)_{i_t} \mathbf{e}_{i_t}, w_{\text{ref}} - w_t \rangle$, which is the same as showing that

$$\sum_{j \neq i_t} \nabla \ell(w_t)_j (w_{\text{ref}} - w_t)_j \geq 0$$

It is unclear if we can demonstrate this generally and hence an extension to coordinate descent seems unlikely. Now since we can no longer control use convexity in the form we need, it is unclear if an extension is obvious.

# 5 Summary and Conclusion

In this report, we discussed the proof technique of Telgarsky [28] in detail and presented connections to various interesting problems by conducting a detailed literature review. We then presented the proofs in a more approachable manner and finally concluded with an analysis of a potential failure case in Coordinate Descent.

[4] Coordinate Descent may fail to converge to global optima for non-smooth convex functions but smoothness based guarantees are available. For e.g. here.

## References


[1] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Bejing, China, 22–24 Jun 2014. PMLR.

[2] Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1370–1378. PMLR, 2019.

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD.

[4] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[5] Damek Davis and Dmitriy Drusvyatskiy. High probability guarantees for stochastic convex optimization. In *Conference on Learning Theory*, pages 1411–1427. PMLR, 2020.

[6] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12(null):2121–2159, jul 2011.

[7] John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

[8] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

[9] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization, 2017.

[10] Mert Gurbuzbalaban, Asuman Ozdaglar, Pablo A Parrilo, and Nuri Vanli. When cyclic coordinate descent outperforms randomized coordinate descent. *Advances in Neural Information Processing Systems*, 30, 2017.

[11] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[12] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[13] Yuzheng Hu, Ziwei Ji, and Matus Telgarsky. Actor-critic is implicitly biased towards high entropy optimal policies. *arXiv preprint arXiv:2110.11280*, 2021.

[14] Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2109–2136. PMLR, 09–12 Jul 2020.

[15] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression, 2018.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum, 2020.

[18] Zhipeng Lou, Wanrong Zhu, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23:1–22, 2022.

[19] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[20] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641. PMLR, 2015.

[21] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[22] Daniel A Roberts. Sgd implicitly regularizes generalization error. *arXiv preprint arXiv:2104.04874*, 2021.

[23] Ankan Saha and Ambuj Tewari. On the finite time convergence of cyclic coordinate descent methods, 2010.

[24] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2, page 5, 2009.

[25] Ohad Shamir. Gradient methods never overfit on separable data. *J. Mach. Learn. Res.*, 22:85–1, 2021.

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

[27] Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[28] Matus Telgarsky. Stochastic linear optimization never overfits with quadratically-bounded losses on general data. *arXiv preprint arXiv:2202.06915*, 2022.

[29] Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65–102. PMLR, 2022.

[30] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[31] Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[32] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24280–24314. PMLR, 17–23 Jul 2022.