Information and Inference: A Journal of the IMA (2023) **12**, 3010–3041
https://doi.org/10.1093/imaiai/iaad047
Advance Access publication on 24 November 2023

The statistical complexity of early-stopped mirror descent[†‡]

Varun Kanade
Department of Computer Science, University of Oxford, Parks Road, Oxford OX1 3QD, UK

Patrick Rebeschini
Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK

and

Tomas Vaškevičius*
Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL), Rte Cantonale, CH-1015 Lausanne, Switzerland
*Corresponding author: tomas.vaskevicius@epfl.ch

[Received on 2 January 2023; revised on 11 September 2023; accepted on 22 October 2023]

Recently there has been a surge of interest in understanding implicit regularization properties of iterative gradient-based optimization algorithms. In this paper, we study the statistical guarantees on the excess risk achieved by early-stopped unconstrained mirror descent algorithms applied to the unregularized empirical risk. We consider the set-up of learning linear models and kernel methods for strongly convex and Lipschitz loss functions while imposing only boundedness conditions on the unknown data-generating mechanism. By completing an inequality that characterizes convexity for the squared loss, we identify an intrinsic link between offset Rademacher complexities and potential-based convergence analysis of mirror descent methods. Our observation immediately yields excess risk guarantees for the path traced by the iterates of mirror descent in terms of offset complexities of certain function classes depending only on the choice of the mirror map, initialization point, step size and the number of iterations. We apply our theory to recover, in a clean and elegant manner via rather short proofs, some of the recent results in the implicit regularization literature while also showing how to improve upon them in some settings.

Keywords: Excess Risk; Regularization; Iterative Regularization; Early Stopping; Rademacher Complexity; Mirror Descent; Fast Rates.

1. Introduction

In a typical supervised statistical learning set-up, we observe a dataset D_n of n input–output pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ sampled i.i.d. from some unknown distribution P. When learning with respect to a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$, the goal is to output a function $\widehat{g}(D_n) : \mathcal{X} \to \mathcal{Y}$ that minimizes the *risk* $R(\widehat{g}(D_n))$ defined as follows for any function $g : \mathcal{X} \to \mathcal{Y}$:

$$R(g) = \mathbb{E}_{(X,Y)\sim P}[\, \ell(g(X), Y)] \,.$$

[†] A prior version of this work appeared at the NeurIPS 2020 conference.

[‡] For the purpose of open access, all the authors have applied the CC BY public copyright licence to any author accepted manuscript version arising from this submission.

A *statistical estimator* is a mapping $\widehat{g} : D_n \mapsto \widehat{g}(D_n)$; we denote the range of this mapping by $\mathcal{G}$. To simplify the notation, we will often write $\widehat{g}$ for $\widehat{g}(D_n)$. Whether $\widehat{g}$ denotes the estimator or the estimator's output $\widehat{g}(D_n)$ will always be clear from the context. Among the most studied statistical estimators is the *empirical risk minimization* (ERM) estimator, which, given a function class $\mathcal{G}$, outputs a function $\widehat{g}_{\mathcal{G}} = \widehat{g}_{\mathcal{G}}(D_n)$ defined as

$$\widehat{g}_{\mathcal{G}} \in \arg\min_{g \in \mathcal{G}} R_n(g), \quad \text{where } R_n(g) := \frac{1}{n}\sum_{i=1}^{n} \ell(g(x_i), y_i), \tag{1.1}$$

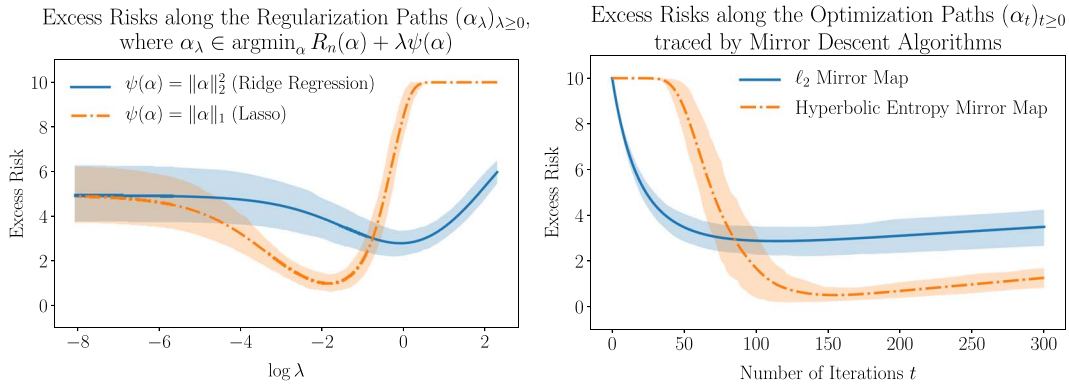
in some cases with a regularization penalty term added to the optimization objective $R_n(g)$, such as ℓ_p norm of the model parameters. We consider the *agnostic* or *distribution-free* setting, i.e. the case where the data-generating distribution P is not constrained to follow a well-specified model or to satisfy some low-noise assumptions. Instead, we only assume that the support of the distribution P is constrained to the set $\mathcal{X} \times \mathcal{Y}$. In the agnostic case, a key performance measure of an estimator $\widehat{g}$ is its *excess risk* with respect to some reference class of functions $\mathcal{F}$ that does not necessarily coincide with the estimator's $\widehat{g}$ range $\mathcal{G}$:

$$\mathcal{E}(\widehat{g}, \mathcal{F}) = R(\widehat{g}) - \inf_{f \in \mathcal{F}} R(f).$$

We remark that the excess risk $\mathcal{E}(\widehat{g}, \mathcal{F})$ is a random variable because $\widehat{g} = \widehat{g}(D_n)$ is a function of the observed random sample D_n. In this paper, we obtain sharp excess risk bounds that hold with high probability for a family of statistical estimators defined as suitably stopped optimization procedures, in a sense explained below.

Traditionally, in learning theory, statistical and computational properties of ERM estimators have been considered separately. From a statistical point of view, localized complexity measures have become a default tool in statistical learning theory and empirical processes theory for controlling the excess risk of ERM algorithms $\widehat{g}_{\mathcal{G}}$ with respect to the function class $\mathcal{G}$ itself, i.e. for controlling $\mathcal{E}(\widehat{g}_{\mathcal{G}}, \mathcal{G})$ [13, 32]. A rich and general theory regarding these complexity measures has been developed and used to provide excess risk bounds in both classification and regression settings, yielding minimax-optimal results in several cases. Such complexity measures depend on combinatorial or geometric parameters of interest, such as the VC-dimension or eigenvalue decay of the kernel matrix and, in particular, they serve as a guiding principle to choose a suitable *explicit regularizer* for a set of candidate models $(\widehat{g}_{\mathcal{G}_\lambda})_{\lambda \in \Lambda}$, where $\lambda \in \Lambda$ is a hyper-parameter that controls the amount of regularization. In practice, some $\lambda^\star \in \Lambda$ is then chosen via some model selection procedure such as cross-validation, aiming to select a model with the smallest risk. From a computational point of view, computing the estimators $(\widehat{g}_{\mathcal{G}_\lambda})_{\lambda \in \Lambda}$ can be done by solving the corresponding optimization problems defined in Equation (1.1), one for each $\lambda \in \Lambda$. An appealing aspect of this approach is that the design and analysis of efficient optimization algorithms, exploiting the geometry of $\mathcal{G}_\lambda$ that arises from the structure of the model as well as the distribution P, can be done independently of the statistical analysis of its performance.

Recent years have also witnessed an increased interest in directly studying the statistical properties of models trained by gradient-based methods, particularly in relation to the notions of *implicit regularization* and *early stopping*. For a family of functions $\mathcal{G} = \{g_\alpha : \alpha \in \mathbb{R}^m\}$ parametrized by a vector α, such methods are fully characterized by the initialization point α_0 and an update rule, which, given α_t and the gradient of the empirical risk at α_t, generates the next iterate α_{t+1}, yielding a set of candidate estimators $(\widehat{g}_{\alpha_t})_{t \geq 0}$. Early stopping has an effect akin to *explicit* regularization discussed



(a) Explicit regularization. (b) Implicit regularization.

FIG. 1. Let $\ell(y, y') = (y - y')^2$ be the quadratic loss and let us consider a distribution P such that $X \sim N(0, I_d)$ and $Y|X = x \sim \langle \alpha', x\rangle + N(0, 5^2)$ for some parameter $\alpha' \in \mathbb{R}^d$. Fix $n = 200, d = 100$ and let α' be a 10-sparse vector with non-zero entries equal to ± 1. Due to the sparsity of α', explicit regularization via ℓ_1 penalization results in a class of models $(\alpha_\lambda)_{\lambda\geq 0}$ that at their minimum achieve significantly lower risk than the class of models generated via ℓ_2 penalization (cf. Fig. 1(a)). Figure 1(b) demonstrates a similar phenomenon from an implicit regularization point of view. Due to the sparsity of α', the choice of a hyperbolic entropy mirror map (cf. Section 4.2) yields an optimization path that at its minimum achieves excess risk nearly an order of magnitude lower than the path generated by the vanilla gradient descent updates. In the plot above, the solid lines denote means over 100 runs, whereas the shaded regions correspond to the 10th and the 90th percentiles.

above, and the *stopping time* $t^\star$ can be chosen in practice via cross-validation, just as in the case of choosing the explicit regularization parameter $\lambda^\star$ corresponding to the best model among $(\widehat{g}_{\mathcal{G}_\lambda})_{\lambda\in\Lambda}$. In modern large-scale machine learning applications, early stopping is often the preferred way to perform model selection, since obtaining a new model is as cheap as performing a step of gradient descent, as opposed to solving a new optimization problem with a different regularization parameter. In Fig. 1, we demonstrate that different choices of optimization algorithms applied to the unregularized empirical risk R_n yield different statistical performance along the optimization path $(\widehat{g}_{\alpha_t})_{t\geq 0}$, in a similar way that a choice of an explicit regularizer affects the statistical performance along the corresponding regularization path. Moreover, in general, early stopping is *crucial* to achieving optimal statistical performance in the same way as selecting an appropriate regularization parameter is crucial for achieving optimal statistical performance for penalized estimators. In particular, the results obtained in this paper cannot be reproduced by restricting the analysis to the statistical estimator obtained at the convergence of the iterative optimization procedure (i.e. taking $t^\star = \infty$). We discuss this point in more detail in Section 3.1 preceding the statement of Theorem 2.

It is by now well understood that changing the update rule that generates the sequence $(\widehat{g}_{\alpha_t})_{t\geq 0}$, e.g. by changing the optimization algorithm or parametrization of the model class, can directly affect both the statistical properties of the iterates $\widehat{g}_{\alpha_t}$ and computational properties, such as an upper-bound on the optimal stopping time $t^\star$. However, most of the literature has focused on the investigation of vanilla gradient descent updates: $\alpha_{t+1} = \alpha_t - \eta\nabla_{\alpha_t} R_n(\widehat{g}_{\alpha_t})$ (cf. Section 2.1). The existing theory does not easily generalize to other update rules corresponding to different problem geometries. A general theory that connects the notion of early stopping for a more general class of update rules with the well-established theory of localized complexities is still missing. More broadly, a general 'language' to reason about the statistical properties of trajectories traced by optimization algorithms applied to the unregularized empirical risk is still lacking.

In this paper, we study a *family* of update rules given by the mirror descent algorithm [15,44]. Mirror descent, which includes vanilla gradient descent as a special case, is increasingly becoming the tool of choice in optimization and machine learning, applied well beyond the traditional settings of convex optimization and online learning. Among the properties that make mirror descent appealing are its ability to exploit non-Euclidean geometries via properly designed mirror maps, the fact that the algorithm admits a general potential-based convergence analysis in terms of Bregman divergences, and its ability to represent a large class of algorithms in a unified and well-developed framework. Our work reveals an inherent connection between the statistical properties of the mirror descent iterates $(\widehat{g}_{\alpha_t})_{t\geq 0}$ and the notion of offset Rademacher complexity [30,36]. Consequently, our work unearths a simple and elegant way to simultaneously analyse upper-bounds on the stopping time $t^\star$, as well as the excess risk $\mathcal{E}(\widehat{g}_{\alpha_t}, \mathcal{F})$ for all $t \leq t^\star$ in terms of the mirror map, the initialization point α_0, the step-size and the function class $\mathcal{F}$.

The rest of the paper is structured as follows.

- In Section 1.1, we introduce the background material on local Rademacher complexities and the family of mirror descent algorithms.
- In Section 1.2, we formulate the assumptions under which we establish the main results of this paper.
- In Section 2, we introduce our proof technique in the simplified setting of the continuous-time mirror descent flow with respect to the quadratic loss. We compare our approach with related work in Section 2.1.
- Section 3 contains our main results. In Section 3.1, we show that early-stopped mirror descent flow satisfies a certain deterministic inequality called *offset condition* (cf. Section 1.1). In turn, it follows that the excess risk of a suitably stopped mirror descent flow can be controlled via offset (local) Rademacher complexity theory, known to yield sharp excess risk bounds in a variety of problem settings. In Section 3.2, we obtain a corresponding result for the discrete-time mirror descent iterates under an additional smoothness assumption on the empirical risk function and strong convexity assumption on the mirror map.
- Example applications of our main results are demonstrated in Section 4.
- Some potential future directions are discussed in Section 5.

1.1 *Background*

We begin this section by explaining the difficulties involved in analysing early-stopped iterative algorithms via the classical notion of localized complexities (Section 1.1.1). We then describe the offset Rademacher complexities, which is a form of localization based on a different mathematical machinery that is more suitable for our setting (Section 1.1.2). Finally, we define the mirror descent updates and outline a short well-known potential-based proof of its convergence (Section 1.1.3).

In what follows, we let $\|g - f\|_n^2 = \frac{1}{n}\sum_{i=1}^n (g(x_i) - f(x_i))^2$ and $\|g - f\|_P^2 = \mathbb{E}[(g(X) - f(X))^2]$ denote the empirical and population ℓ_2 distances between functions g and f, respectively. Further, given a function class $\mathcal{F}$, we denote by $g_{\mathcal{F}} \in \mathcal{F}$ a function that attains risk equal to $\inf_{g\in\mathcal{F}} R(g)$.[1] A table of notation is provided in Appendix A.

[1] If such a function $g_{\mathcal{F}}$ does not exist, we can redefine $g_{\mathcal{F}}$ to be any function in $\mathcal{F}$ such that $R(g_{\mathcal{F}}) \leq \inf_{g\in\mathcal{F}} R(g) + \delta$ for any arbitrarily small $\delta > 0$.

1.1.1 *Local Rademacher Complexities* The classical notion of global Rademacher complexities [10] can only establish the slow rates of order $n^{-1/2}$ on the excess risk (cf. [11], Theorem 2.3)). This observation was one of the primary motivating factors in the development of localized Rademacher complexities [13,32]. Let $\mathcal{G}$ be the range of an estimator $\widehat{g}$. Rather than considering the Rademacher complexity of the whole function class $\mathcal{G}$, localization builds on the idea of computing the Rademacher complexity of the smaller class $\{g \in \mathcal{G} : \|g - g_{\mathcal{G}}\|_n^2 \leq r\}$ for some suitably defined radius r that can be obtained by solving a certain fixed-point equation. More recent work focuses on unbounded and, in particular, heavy-tailed settings [42] as well as extending the scope of localization to study estimators other than ERM, e.g. to study the statistical performance of tournament procedures [40,43]. Crucially, this line of research is rooted in the following two assumptions. First, $(P, \mathcal{G})$ is assumed to satisfy a convexity type assumption known in the literature as the *Bernstein condition* (cf. [11]), which states that for some constant $C > 0$, $R(g) - R(g_{\mathcal{G}}) \geq C\|g - g_{\mathcal{G}}\|_P^2$ for any $g \in \mathcal{G}$. If the class $\mathcal{G}$ is convex and the loss function is quadratic then this condition follows immediately by convexity with $C = 1$ (see [43], Definition 5.2) for more details). The second condition is imposed on the estimator $\widehat{g}$ itself (rather than its range $\mathcal{G}$), which requires that the inequality $R_n(\widehat{g}) \leq R_n(g_{\mathcal{G}})$ holds for all realizations of D_n, a property naturally satisfied by the ERM algorithm over the class $\mathcal{G}$. Our setting, however, does not easily fit into the above assumptions. To see why, note that the sequence $(\widehat{g}_{\alpha_t})_{t \geq 0}$ obtained by some iterative algorithm aimed at minimizing the unconstrained empirical risk is not necessarily explicitly constrained to lie in the class $\mathcal{G}$. Thus by the time the inequality $R_n(\widehat{g}_{\alpha_t}) \leq R_n(g_{\mathcal{G}})$ is satisfied, the iterate $\widehat{g}_{\alpha_t}$ can already be outside the class $\mathcal{G}$, potentially violating the Bernstein condition (cf. Fig. 2) in all cases except when $\mathcal{G}$ is taken to be the union of ranges of $\widehat{g}_{\alpha_t}$ over all $t \geq 0$.

1.1.2 *Offset Rademacher Complexities* When learning with the quadratic loss, a theory of localization based on shifted Rademacher processes was proposed by Liang et al [36] (inspired by prior work in online learning [49]). The use of shifted empirical processes in order to bypass technicalities present in the classical localization arguments dates back at least to [63] and has recently found applications in cross-validation [34], classification [70] and PAC-Bayes bounds [66].

Given an observed data sample $D_n = (x_i, y_i)_{i=1}^n$ let $D_n^x = (x_i)_{i=1}^n$. The *empirical offset Rademacher complexity* is defined as follows.

DEFINITION 1. (Empirical Offset Rademacher Complexity). Let $D_n = (x_i, y_i)_{i=1}^n$ denote an observed data sample and let $\sigma_1, \ldots, \sigma_n$ be a sequence of independent Rademacher random variables (that is, symmetric random variables taking values in $\{-1, +1\}$). For any parameter $\gamma \geq 0$, the offset Rademacher complexity of a function class $\mathcal{G}$ is defined as

$$\mathfrak{R}_{D_n^x}(\mathcal{G}, \gamma) = \mathbb{E}_{\sigma_1, \ldots, \sigma_n}\left[\sup_{g \in \mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^n (\sigma_i g(x_i) - \gamma g(x_i)^2)\right\}\right]. \tag{1.2}$$

Upper-bounds for offset Rademacher complexity of linear functions and kernel classes are demonstrated in Section 1.2. Note that since the terms $-\gamma g(x_i)^2$ are always non-positive, the above notion of complexity is never larger than global Rademacher complexity of the class $\mathcal{G}$, which is recovered with the choice $\gamma = 0$. On the other hand, for any $\gamma > 0$, the quadratic term in the above definition has a localization effect by compensating for the fluctuations in the term involving Rademacher variables (see the discussions in [36], Section 5.2) and ([30], Section 3)). Importantly, the theory of localization via

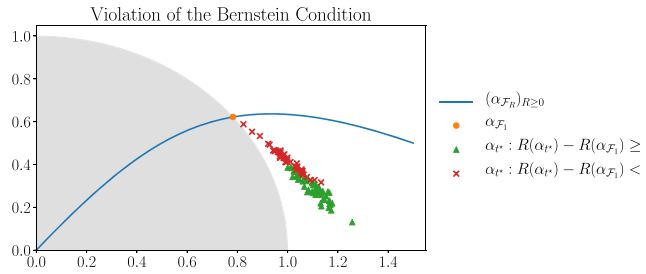


FIG. 2. Let $\ell\ (y, y') = (y - y')^2$ be the quadratic loss; fix $\alpha' = (1.5, 0.5)^\mathsf{T}$, $n = 100$ and consider a distribution P defined as $X \sim N(0, \Sigma)$ and $Y|X = x \sim \langle \alpha', x\rangle + N(0, 0.5^2)$, where $\Sigma = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right]$ For simplicity, we denote linear functions $\langle \alpha, \cdot\rangle$ by the parameter α. For any $R \geq 0$ let $\mathcal{F}_R = \{\alpha : \|\alpha\|_2 \leq R\}$ denote an ℓ_2 ball of radius R and let $\alpha_{\mathcal{F}_R} = \operatorname{argmin}_{\alpha\in\mathcal{F}_R} R(\alpha)$ denote the population risk minimizer in $\mathcal{F}_R$. Let $\widehat{g}$ be some estimator and suppose that we want to upper-bound $\mathcal{E}(\widehat{g}, \mathcal{F}_1)$, where the function class $\mathcal{F}_1$ is denoted by the shaded region in the plot above. The theory of localized Rademacher complexities can be readily used to upper-bound $\mathcal{E}(\widehat{g}, \mathcal{F}_1)$ for proper algorithms (i.e. estimators $\widehat{g} \in \mathcal{F}_1$) that with probability 1 satisfy $R_n(\widehat{g}) \leq R_n(\alpha_{\mathcal{F}_1})$, for instance, if $\widehat{g}$ is an ERM algorithm over the class $\mathcal{F}_1$. The classical theory of localization is, on the other hand, less suitable for analysis of unconstrained iterative algorithms. To see why, consider running mirror descent with the mirror map $\psi(\alpha) = \alpha^\mathsf{T}\Sigma\alpha/2$, the initialization $\alpha_0 = 0$ and the step-size $\eta = 10^{-3}$. Define the stopping time $t^\star = \min\{t \geq 0 : R_n(\alpha_t) \leq R_n(\alpha_{\mathcal{F}_1})\}$ so that our early-stopped estimator is identified with the parameter $\alpha_{t^\star}$. We plot the values of $\alpha_{t^\star}$ (denoted by crosses and triangles) over 100 runs, where the crosses denote instances of $\alpha_{t^\star}$ such that $R(\alpha_{t^\star}) < R(\alpha_{\mathcal{F}_1})$. Such points, in particular, violate the Bernstein condition for all $C > 0$ (i.e. it does not hold that $R(\alpha_{t^\star}) - R(\alpha_{\mathcal{F}_1}) \geq C\|\alpha_{t^\star} - \alpha_{\mathcal{F}_1}\|_P^2$) and hence demonstrate that statistical analysis of early-stopped mirror descent estimators does not easily fit within the classical framework of localized complexity measures. In contrast, all points in $\mathcal{F}_1$ (denoted by the shaded ball) satisfy the Bernstein condition with parameter $C = 1$; thence, bounds on $\mathcal{E}(\widehat{g}, \mathcal{F}_1)$ can be easily obtained via the classical notion of localization whenever $\widehat{g}$ is a proper estimator (i.e. $\widehat{g} \in \mathcal{F}_1$) such that $R_n(\widehat{g}) \leq R_n(\alpha_{\mathcal{F}_1})$ almost surely.

offset complexities replaces the Bernstein condition used in the classical theory of localization by the estimator-dependent *offset condition* defined below.

DEFINITION 2. (Offset Condition). A triplet $(P, \mathcal{F}, \widehat{g})$ satisfies the offset condition with parameters $\varepsilon, \gamma \geq 0$, if for $D_n \sim P^n$, with probability 1, we have $R_n(\widehat{g}) - R_n(g_\mathcal{F}) + \gamma\|\widehat{g} - g_\mathcal{F}\|_n^2 \leq \varepsilon$.

The above condition with $\varepsilon = 0$ was introduced in [36] where it was called the *geometric inequality* and shown to hold for ERM estimators over convex classes $\mathcal{F}$ as well as the two-step star estimator [7] over general classes for finite aggregation. A key advantage offered by the theory of offset complexities is that the range of $\widehat{g}$ need not be a subset of $\mathcal{F}$, as long as the offset condition is satisfied. This allows us to consider very general estimators $\widehat{g}$, possibly with non-convex ranges $\mathcal{G}$. In this respect, our work can be seen as showing that early-stopped mirror descent satisfies the offset condition defined above. Once an estimator is shown to satisfy the offset condition, its excess risk $\mathcal{E}(\widehat{g}, \mathcal{F})$ can be controlled in terms of the expected offset complexity $\mathbb{E}_{D_n}\left[\mathfrak{R}_{D_n^x}(\mathcal{G} - g_\mathcal{F}, \gamma)\right]$. In particular, ([36], Theorem 3) provides such guarantees for the *expected* excess risk. Recently, it was shown by Kanade et al. [30] that the expected offset Rademacher complexity provides excess risk guarantees that hold with *high probability*. Before we state this result, let us introduce some additional notation. For a function class $\mathcal{F}$, denote its star hull around zero by $\operatorname{star}(\mathcal{F}) = \{\lambda f : f \in \mathcal{F}, \lambda \in [0, 1]\}$; also, for a function class $\mathcal{F}$ and any function g, let $\mathcal{F} - g = \{f - g : f \in \mathcal{F}\}$.

THEOREM 1. (Theorem 3.3 in [30]). Let P be a data-generating distribution supported on $\mathcal{X} \times [-b, b]$ (that is, $\mathcal{Y} = [-b, b]$), where $b > 0$. Let $\mathcal{F}$ be a class of reference functions mapping $\mathcal{X}$ to $[-b, b]$. Suppose that the following two conditions hold:

1. There exists $C_b > 0$ such that for any $y \in [-b, b]$ the loss function $\ell(\cdot, y)$ is C_b-Lipschitz;
2. The estimator $\widehat{g}$ satisfies the offset condition (see Definition 2) with parameters $\varepsilon, \gamma \geq 0$. Moreover, the estimator's $\widehat{g}$ range is a function class $\mathcal{G}$ mapping $\mathcal{X}$ to $[-b, b]$.

Then, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$, we have

$$\mathcal{E}(\widehat{g}, \mathcal{F}) \leq c_1 C_b' \mathbb{E}_{D_n}\Big[\mathfrak{R}_{D_n^x}(\mathrm{star}(\mathcal{G} - g_{\mathcal{F}}), (C_b')^{-1}\gamma)\Big] + c_2 \frac{\gamma^{-1}(C_b')^2 \log(1/\delta)}{n} + \varepsilon,$$

where $c_1, c_2 > 0$ are some universal constants and $C_b' = C_b + \gamma b$.

The generality of the above result allows us to improve upon the existing bounds in the early stopping literature even for the vanilla gradient descent updates (cf. Section 2.1). Indeed, observe that the above bound does not impose any restrictions on the data-generating distribution P other than boundedness. On the other hand, concerning random design excess risk bounds treated in this paper, the existing works [51,64] connecting early-stopped gradient descent iterates to the notion of local Rademacher (or Gaussian) complexity rely on a well-specified model assumption in their analysis.

1.1.3 *Mirror Descent* The key object characterizing the geometry of the mirror descent algorithm is the *mirror map* ψ, a strictly convex and differentiable function mapping some open set $\mathcal{D} \subseteq \mathbb{R}^m$ to $\mathbb{R}$ whose gradient is surjective, i.e. $\{\nabla\psi(\alpha) \mid \alpha \in \mathcal{D}\} = \mathbb{R}^m$. By slightly abusing notation, we use $R_n(\alpha) := R_n(g_\alpha)$ to denote the empirical risk of g_α. When optimizing the empirical risk $R_n(\alpha)$, the mirror descent updates in continuous and discrete time are given, respectively, by

$$\frac{d}{dt}\alpha_t = -\Big(\nabla^2\psi(\alpha_t)\Big)^{-1}\nabla R_n(\alpha_t) \quad \text{and} \quad \nabla\psi(\alpha_{t+1}) = \nabla\psi(\alpha_t) - \eta\nabla R_n(\alpha_t), \tag{1.3}$$

where $\eta > 0$ is the step-size. We remark that the choice $\psi(\alpha) = \frac{1}{2}\|\alpha\|_2^2$ reduces the above updates to gradient descent. A key notion in the analysis of mirror descent algorithms is the *Bregman divergence*, defined as $D_\psi(\alpha', \alpha) = \psi(\alpha') - \psi(\alpha) - \langle\nabla\psi(\alpha), \alpha' - \alpha\rangle$ for all α', α in the domain of ψ. By convexity of ψ, the Bregman divergence D_ψ is non-negative and enters the analysis of mirror descent algorithms through the following elementary equality:

$$-\frac{d}{dt}D_\psi(\alpha', \alpha_t) = \langle -\nabla R_n(\alpha_t), \alpha' - \alpha_t\rangle. \tag{1.4}$$

Let $\bar{\alpha}_t = \frac{1}{t}\int_0^t \alpha_t dt$. In the optimization literature, the above equation can be used to establish that $R_n(\bar{\alpha}_t)$ can get arbitrarily close to $R_n(\alpha')$ from above, for any reference point α'. In particular, by convexity of R_n, we have $\langle -\nabla R_n(\alpha_t), \alpha' - \alpha_t\rangle \geq R_n(\alpha_t) - R_n(\alpha')$ and so

$$\frac{1}{t}D_\psi(\alpha', \alpha_0) \geq \frac{1}{t}\int_0^t -\frac{d}{ds}D_\psi(\alpha', \alpha_s)ds \geq \frac{1}{t}\int_0^t (R_n(\alpha_s) - R_n(\alpha'))ds \geq R_n(\bar{\alpha}_t) - R_n(\alpha'), \tag{1.5}$$

where the last line follows by convexity of R_n. Remarkably, the above proof works independently of the choice of the mirror map ψ, establishing convergence for a *family* of algorithms in a unified framework. For more information we refer the interested reader to the surveys by Bubeck [18] and Bansal and Gupta

[9]. The latter survey focuses entirely on such potential-based proofs in a variety of settings, including acceleration.

1.2 *Assumptions*

In this section, we formulate the assumptions on the loss function and the representation of prediction functions needed to establish our main results presented in Section 3.

The first assumption requires that the loss function ℓ is strongly convex and differentiable, in the sense formulated below.

Assumption 1. (Strong Convexity and Differentiability). Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be the loss function. For any $y' \in \mathcal{Y}$ let $\ell_{y'} : \mathcal{Y} \to [0, \infty)$ denote the function $y \mapsto \ell(y, y')$. We assume that the following two conditions are satisfied.

1. For any $y' \in \mathcal{Y}$, the function $\ell_{y'}$ is differentiable.
2. There exists $\gamma > 0$ such that for any $y' \in \mathcal{Y}$ the function $\ell_{y'}$ is γ-strongly convex, in the sense that for any $y_1, y_2 \in \mathcal{Y}$ the following inequality holds:

$$\ell_{y'}(y_1) \geq \ell_{y'}(y_2) + \ell'_{y'}(y_2)(y_1 - y_2) + \tfrac{\gamma}{2}(y_1 - y_2)^2.$$

A classical example of a loss function satisfying the above condition is the quadratic loss $\ell(y, y') = (y - y')^2$, which is 2-strongly convex. Observe that the above condition is much weaker than assuming that the empirical risk function $\alpha \mapsto R_n(g_\alpha)$ is strongly convex.

The second assumption concerns the representation of functions. We will only consider parametric classes of linear functions in the following sense. We identify the parameter system by the set $\mathbb{R}^m$ for some natural number m and denote the parameters by $\alpha \in \mathbb{R}^m$. Each vector $\alpha \in \mathbb{R}^m$ identifies a linear function f_α satisfying the conditions described below.

Assumption 2. (Function Class Representation). Let $D_n = (x_i, y_i)_{i=1}^n$ denote the observed data sample. We assume that the learning algorithm has access to a (possibly data dependent) matrix $Z = Z(D_n) \in \mathbb{R}^{n \times m}$, such that for any $\alpha \in \mathbb{R}^m$, the corresponding function f_α satisfies $f_\alpha(x_i) = (Z\alpha)_i$ for any $i = 1, \ldots, n$.

Let us provide two example settings commonly studied in iterative regularization literature that admit the conditions specified in Assumption 2.

Example 1. (Linear Regression). Given a data sample $D_n = (x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^d$, let $m = d$ and let the i-th row of $Z \in \mathbb{R}^{n \times d}$ be given by the vector x_i^T. For any $\alpha \in \mathbb{R}^d$, let $g_\alpha(\cdot) = \langle \alpha, \cdot \rangle$ be the linear function identified by α. This is the setting of simulations performed in Figs 1 and 2, for different choices of mirror maps.

Letting $\mathcal{G} = \{g_\alpha : \alpha \in \mathbb{R}^d\}$ be the set of all d-dimensional linear functions, its (empirical) offset Rademacher complexity $\mathfrak{R}_n(\mathcal{G}, \gamma)$ can be upper bounded as follows via a direct computation:

$$\begin{aligned}\mathfrak{R}_n(\mathcal{G}, \gamma) &= \mathbb{E}_{\sigma_1, \ldots, \sigma_n}\left[\sup_{g \in \mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^n (\sigma_i g(x_i) - \gamma g(x_i)^2)\right\}\right] \\ &= \frac{1}{n}\mathbb{E}_{\sigma_1, \ldots, \sigma_n}\left[\sup_{\alpha \in \mathbb{R}^d}\left\{\langle \sigma, Z\alpha \rangle - \gamma \alpha^\mathsf{T} Z^\mathsf{T} Z \alpha\right\}\right]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\frac{1}{4\gamma}\left(\sum_{i=1}^{n}\sigma_i Z_i\right)^{\mathsf{T}}(Z^{\mathsf{T}}Z)^{\dagger}\left(\sum_{i=1}^{n}\sigma_i Z_i\right)\right]\\
&= \frac{1}{4\gamma n}\sum_{i=1}^{n} Z_i^{\mathsf{T}}(Z^{\mathsf{T}}Z)^{\dagger}Z_i\\
&= \frac{1}{4\gamma n}\,\mathrm{trace}\left(\sum_{i=1}^{n} Z_i^{\mathsf{T}}(Z^{\mathsf{T}}Z)^{\dagger}Z_i\right)\\
&= \frac{1}{4\gamma n}\,\mathrm{trace}\left((Z^{\mathsf{T}}Z)^{\dagger}(Z^{\mathsf{T}}Z)\right)\\
&\leq \frac{1}{4\gamma n}\mathrm{rank}(Z)\\
&\leq \frac{d}{4\gamma n},
\end{aligned}
$$

where $(Z^{\mathsf{T}}Z)^{\dagger}$ denotes the Moore–Penrose inverse of $Z^{\mathsf{T}}Z$. We remark that the global Rademacher complexity corresponding to $\gamma = 0$ is infinite in the above example, while it leads to the 'slow rate' of order $1/\sqrt{n}$ under the additional boundedness condition $\|\alpha\|_2 \leq 1$ instead of the 'fast rate' d/n obtained above.

We now turn to the second example that admits the conditions of Assumption 2, namely, the setting of non-parametric regression in reproducing kernel Hilbert spaces (RKHS), frequently considered in iterative regularization literature (e.g. [14, 51, 64, 67]).

EXAMPLE 2. (Kernel Regression). Let $D_n = (x_i, y_i)_{i=1}^n$ be an observed data sample, where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ for some abstract space $\mathcal{X}$ and $\mathcal{Y} \subseteq \mathbb{R}$. Let $k : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a Mercer kernel which for any $x, y \in \mathcal{X}$ satisfies $\langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}} = k(x,y)$ and consequently induces a reproducing kernel Hilbert space $\mathcal{H}$ equipped with norm $\|\cdot\|_{\mathcal{H}}$. Then, conditionally on the observed sample D_n, denote by $K \in R^{n\times n}$ a matrix such that $K_{ij} = k(x_i, x_j)$. To each $\alpha \in \mathbb{R}^n$, we associate a function $g_\alpha \in \mathcal{H}$ defined as $g_\alpha = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$. Thus, for any $i = 1, \ldots, n$, we have $g_\alpha(x_i) = (K\alpha)_i$ and hence we may set $m = n$ and $Z = K$. We refer the interested reader to the book by Scholkopf and Smola [56] for more background on reproducing kernel Hilbert spaces.

Let $\mathcal{H}_1 = \{h \in \mathcal{H} : \|g\|_{\mathcal{H}} \leq 1\}$. We will show how to bound the (empirical) offset Rademacher complexity $\mathfrak{R}_n(\mathcal{H}_1, \gamma)$. Keeping the data sample D_n fixed, by the Representer theorem it is enough to consider functions of the form $\{g_\alpha = \sum_{i=1}^n \alpha_i k(x_i, \cdot) : \alpha^{\mathsf{T}}K\alpha \leq 1\} \subseteq \mathcal{H}_1$. Hence, repeating the calculations carried out in Example 1 while taking the additional constraint $\alpha^{\mathsf{T}}K\alpha \leq 1$ into account, for any $\gamma \in (0, 1]$ we have

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{H}_1, \gamma) &= \frac{1}{n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{\alpha\in\mathbb{R}^n:\alpha^{\mathsf{T}}K\alpha\leq 1}\left\{\langle\sigma, K\alpha\rangle - \gamma\alpha^{\mathsf{T}}K^2\alpha\right\}\right]\\
&= \frac{1}{\gamma n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{\alpha\in\mathbb{R}^n:\alpha^{\mathsf{T}}K\alpha\leq 1}\left\{\langle\sigma, K(\gamma\alpha)\rangle - (\gamma\alpha)^{\mathsf{T}}K^2(\gamma\alpha)\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{\gamma n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{\alpha\in\mathbb{R}^n:\alpha^{\mathsf{T}}K\alpha\leq 1}\left\{\langle\sigma,K\alpha\rangle-\alpha^{\mathsf{T}}K^2\alpha\right\}\right]\\
&= \frac{1}{\gamma n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{\alpha\in\mathbb{R}^n}\inf_{\lambda>0}\left\{\langle\sigma,K\alpha\rangle-\alpha^{\mathsf{T}}K^2\alpha-\lambda(\alpha^{T}K\alpha-1)\right\}\right]\\
&\leq \frac{1}{\gamma n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\inf_{\lambda>0}\left\{\lambda+\sup_{\alpha\in\mathbb{R}^n}\left\{\langle\sigma,K\alpha\rangle-\alpha^{\mathsf{T}}(K^2+\lambda K)\alpha\right\}\right\}\right]\\
&\leq \frac{1}{\gamma n}\inf_{\lambda>0}\left\{\lambda+\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{\alpha\in\mathbb{R}^n}\left\{\langle\sigma,K\alpha\rangle-\alpha^{\mathsf{T}}(K^2+\lambda K)\alpha\right\}\right]\right\}\\
&= \frac{1}{\gamma n}\inf_{\lambda>0}\left\{\lambda+\frac{1}{4}\operatorname{trace}\left((K^2+\lambda K)^{\dagger}K^2\right)\right\}.
\end{aligned}
$$

We will now rewrite the above bound in a more familiar form resulting from the classical local Rademacher complexity bounds for kernel classes. Letting $\mu_1 \geq \mu_2 \cdots \geq \mu_n \geq 0$ denote the eigenvalues of the kernel matrix K and choosing $\lambda > 0$ that balances the λ term with the trace term, we obtain

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{H}_1,\gamma) &\leq \frac{1}{\gamma n}\cdot 2\inf\left\{\lambda>0:\lambda\geq\frac{1}{4}\sum_{i=1}^{n}\frac{\mu_i}{\mu_i+\lambda}\right\}\\
&= \frac{1}{\gamma n}\cdot 2\inf\left\{\lambda>0:\lambda^2\geq\frac{1}{4}\sum_{i=1}^{n}\frac{\lambda\mu_i}{\mu_i+\lambda}\right\}\\
&= \frac{1}{\gamma n}\cdot 2\inf\left\{\lambda>0:\lambda\geq\frac{1}{2}\sqrt{\sum_{i=1}^{n}\frac{\lambda\mu_i}{\mu_i+\lambda}}\right\}\\
&\leq \frac{1}{\gamma n}\cdot 2\inf\left\{\lambda>0:\lambda\geq\frac{1}{2}\sqrt{\sum_{i=1}^{n}\min\{\mu_i,\lambda\}}\right\}.
\end{aligned}
$$

It can now be verified that the obtained upper-bound on the offset Rademacher complexity is the same as the corresponding bounds obtainable via the classical localization theory[2] ; see, for example, ([62], Corollary 14.5). In particular, the order of magnitude of $\mathfrak{R}_n(\mathcal{H}_1,\gamma)$ is determined by the rate of eigenvalue decay of the kernel matrix K, which can be readily computed from the given data sample. Consequences of the above bound for various kernel classes are discussed in ([62], Section 13.4.2).

2. Summary of Techniques and Main Results

We develop a general theory for learning linear models (including kernel machines) with strongly convex and Lipschitz loss functions, which shows how the optimization trajectory of *unconstrained*

[2] Indeed, the offset Rademacher complexity can be upper bounded via the classical notion of local Rademacher complexity, as shown in ([30], Lemma 3.5).

mirror descent applied to minimize the unregularized empirical risk is *inherently* connected to excess risk guarantees offered via the offset Rademacher complexity theory. Unlike in most prior work on early stopping, the notion of statistical complexity appears naturally from intrinsic properties of mirror descent applied to the unregularized empirical risk, without invoking lower level arguments related to concentration to the *fictitious* population version of the algorithm. Furthermore, our theory leads to an explicit characterization of stopping times from the point of view of both optimization and statistics, which directly yields excess risk bounds and allows us to re-derive previously established results, and some new results, in a much simpler fashion.

As discussed in Section 1.1, early-stopped unconstrained iterative algorithms do not easily fit within the mathematical framework of classical localization techniques, partially explaining the scarcity of results connecting localized complexity measures with such algorithms. Offset Rademacher complexities, on the other hand, open up another avenue for establishing such connections via the design of update rules tailored to satisfy the offset condition (cf. Definition 2). In the view of the preconditions of Theorem 1, instead of optimizing the empirical risk R_n, a natural approach to consider is an application of some iterative optimization algorithm to directly minimize the term appearing in the definition of the offset condition: $\tilde{R}_n^{\alpha',\gamma}(\alpha) = R_n(\alpha) - R_n(\alpha') + \gamma\|g_\alpha - g_{\alpha'}\|_n^2$. For any $\gamma > 0$, the gradient $\nabla_\alpha \tilde{R}_n^{\alpha',\gamma}(\alpha)$ depends on the unknown reference point α' and hence cannot be computed in practice. Remarkably, we show that the mirror descent updates applied to the empirical loss R_n simultaneously *implicitly* minimizes $\tilde{R}_n^{\alpha',1}$ for *all* reference points α' up to a certain stopping time (which depends on α') while also *staying inside a certain Bregman 'ball'* centred at α' up to the corresponding stopping time. While mirror descent was developed within the framework of convex optimization, it has also found applications in a wide range of problems including bandits [1], online learning [28], the k-server problem [19] and metrical task systems [20]. In this respect, our work can be seen as an exposition of yet another example where mirror descent naturally solves a problem outside of its originally intended scope.

In this paper, we show that the excess risk of early-stopped mirror descent iterates can be controlled using the offset Rademacher complexity theory. In order to do so, we need to show that the pre-conditions of Theorem 1 hold. In particular, we need to show that suitably early-stopped mirror descent iterates satisfy the offset condition, and also, we need to guarantee that the early-stopped iterates lie in some bounded set, the offset Rademacher complexity of which will give the resulting excess risk bound.

To show the key idea of the proof technique that we present in this paper, for simplicity, let us temporarily fix the loss function $\ell(y, y') = (y - y')^2$ to be the quadratic loss. In addition, for the sake of the exposition, we will only consider the continuous-time mirror descent flow in this section, deferring the analysis of discrete-time updates to Section 3.2.

The key insight behind our main result is the following identity, linking the potential-based analysis of mirror descent (cf. Section 1.1) to the statistical guarantees derived from offset complexities via the offset condition (cf. Definition 2).

LEMMA 1. Let $\ell(y, y') = (y-y')^2$ be the quadratic loss and suppose that the function class representation assumption (Assumption 2) holds. Then, for any $\alpha, \alpha' \in \mathbb{R}^m$, the following holds:

$$\left\langle -\nabla R_n(\alpha), \alpha' - \alpha \right\rangle = R_n(\alpha) - R_n(\alpha') + \|g_\alpha - g_{\alpha'}\|_n^2.$$

Proof. By Assumption 2, there exists some $Z \in \mathbb{R}^{n\times m}$ such that for any parameter $\alpha \in \mathbb{R}^m$ we have $g_\alpha(x_i) = (Z\alpha)_i$. Hence, we can express the empirical loss function as $R_n(\alpha) = \frac{1}{n}\|Z\alpha - y\|_2^2$, where

$y \in \mathbb{R}^n$ is a vector with the i^{th} entry equal to y_i. Hence, for any $\alpha, \alpha' \in \mathbb{R}^m$ we have

$$\begin{aligned}
\left\langle -\nabla R_n(\alpha), \alpha' - \alpha \right\rangle &= \frac{2}{n} \left\langle -Z^{\mathsf{T}}(Z\alpha - y), \alpha' - \alpha \right\rangle \\
&= \frac{2}{n} \left\langle -(Z\alpha - Z\alpha' + Z\alpha' - y), Z(\alpha' - \alpha) \right\rangle \\
&= \frac{2}{n} \left\| Z\alpha - Z\alpha' \right\|_2^2 - \frac{1}{n} \cdot 2 \left\langle Z\alpha' - y, Z(\alpha' - \alpha) \right\rangle \\
&= \frac{2}{n} \left\| Z\alpha - Z\alpha' \right\|_2^2 - \frac{1}{n} \cdot \left(\left\| Z\alpha' - y \right\|_2^2 + \left\| Z(\alpha - \alpha') \right\|_2^2 - \| Z\alpha - y \|_2^2 \right) \\
&= \frac{2}{n} \left\| Z\alpha - Z\alpha' \right\|_2^2 - \frac{1}{n} \cdot \left(n R_n(\alpha') + \left\| Z(\alpha - \alpha') \right\|_2^2 - n R_n(\alpha) \right) \\
&= R_n(\alpha) - R_n(\alpha') + \left\| g_\alpha - g_{\alpha'} \right\|_n^2,
\end{aligned}$$

where the fourth line follows by applying the equality $2\langle a, b \rangle = \|a\|_2^2 + \|b\|_2^2 - \|a - b\|_2^2$, which holds for any vectors $a, b \in \mathbb{R}^m$. □

To appreciate the interest in the above lemma, we shall now revisit the potential-based proof of mirror descent presented in Equation (1.5) in Section 1.1. This time, instead of using the convexity of R_n which gives $\left\langle -\nabla R_n(\alpha_t), \alpha' - \alpha_t \right\rangle \geq R_n(\alpha_t) - R_n(\alpha')$, we directly plug in the identity given in Lemma 1 into Equation (1.4) which yields the following *equality*:

$$-\frac{d}{dt} D_\psi(\alpha', \alpha_t) = R_n(\alpha_t) - R_n(\alpha') + \| g_{\alpha_t} - g_{\alpha'} \|_n^2.$$

The above equation shows that while $R_n(\alpha_t) - R_n(\alpha') + \| g_{\alpha_t} - g_{\alpha'} \|_n^2 > 0$, the iterates of mirror descent stay within the Bregman ball $\{\alpha \in \mathbb{R}^m : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', \alpha_0)\}$. At the same time, the integration argument used in Equation (1.5) establishes that the term $R_n(\alpha_t) - R_n(\alpha') + \| g_{\alpha_t} - g_{\alpha'} \|_n^2$ eventually gets arbitrarily close to 0, and thus the early-stopped mirror descent iterates satisfy the offset condition (cf. Definition 2). For a visual demonstration of the above proof sketch see Fig. 3. We provide full details of this argument in the proof of Theorem 2 as well as a discrete-time version in Theorem 3.

Summary of contributions:

1. Our work extends the scope of offset Rademacher complexities to a family of early-stopped mirror descent methods. Additionally, we extend the scope of mirror descent to be used as a computationally efficient statistical device in an i.i.d. batch statistical learning setting.
2. Our main results, in a short and transparent way, yield bounds on the excess risk of the iterates of (both continuous-time and discrete-time) mirror descent using offset Rademacher complexities. In contrast to prior work, our arguments require no direct use of low-level mathematical techniques such as symmetrization, peeling or concentration to the population version of the algorithm.
3. In Section 4, we demonstrate some selected applications of our main results and comment on the connections to the related work therein.

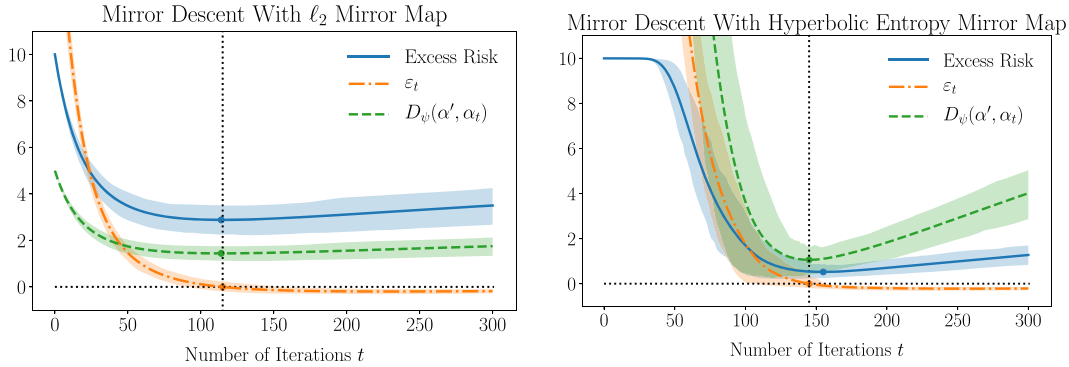


FIG. 3. Consider the setting of Fig. 1 and let $\varepsilon_t = R_n(\alpha_t) - R_n(\alpha') + \left\| g_{\alpha_t} - g_{\alpha'} \right\|_n^2$. The above plots illustrate the following two points. First, there exists a stopping time time $t^\star$ such that $\varepsilon_{t^\star} \approx 0$ (denoted by the vertical dotted line). Hence, the triplet $(P, \{g_{\alpha'}\}, g_{\alpha_{t^\star}})$ satisfies the offset condition (cf. Definition 2) with parameters $(c = 1, \varepsilon \approx 0)$. Second, while $\varepsilon_t \geq 0$, the Bregman divergence $D_\psi(\alpha', \alpha_t)$ denoted by the green line is non-increasing. It follows that the estimator $g_{\alpha_{t^\star}}$ is constrained to lie in the set $\{g_\alpha : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', \alpha_0)\}$, the offset complexity of which can be used to upper bound the excess risk of interest. Crucially, this type of analysis does not directly rely on the particular form taken by the mirror descent update rules, which bypasses the limitations present in prior work (cf. Section 2.1) and allows us to provide excess risk guarantees for a family of mirror descent algorithms. In the plot above, the solid lines denote means over 100 runs, the dots denote the minimum of each solid line and the shaded regions correspond to the 10^{th} and the 90^{th} percentiles.

2.1 *Comparison with Related Work*

The idea of iterative regularization has a long history. Early ideas can be traced back to the stochastic approximation arguments of Robbins and Monro [54]. Even more closely related are the ideas put forth by Louis Landweber [33], yielding one of the regularization schemes in the theory of inverse problems; see the book by [23] for further details and a more extensive background from the inverse problems point of view. In the Statistics literature, the first work to analyse early-stopped gradient descent in connection to minimax optimality appears to be due to Bühlmann and Yu [21], formulated in the context of L_2-boosting algorithms. Regarding early stopping regularization for boosting algorithms, see also the works [12,16,29,68]. However, from the practical perspective, early stopping regularization was used long before, for example, in neural network training [48].

Closer to the setting investigated in this paper, statistical and computational properties of unconstrained *gradient descent* updates have been a subject of intense study over the past two decades, with most of the existing results focusing on the quadratic loss in reproducing kernel Hilbert spaces (RKHS) [14,17,21,51,67], while for general loss functions, see [39,64]. It shall be noted that, in contrast to our work, some of the works investigating early stopping focus on attaining bounds in the $\|\cdot\|_n$ or in the $\|\cdot\|_P$ norms. The quality measure $\|\cdot\|_n$ assumes that the design is non-random. At the same time, the quality measure $\|\cdot\|_P$ is different from the excess risk considered in this paper; as discussed in ([57], Section 1), bounds obtained in $\|\cdot\|_P$ norm do not, in general, imply bounds on the excess risk. In addition, the analysis in the above-cited works [14,21,51,67] is closely tied to the ℓ_2 geometry of the gradient updates and to the quadratic loss function. Such a set-up admits closed-form expressions for the early-stopped gradient descent iterates in terms of relatively simple linear operators acting on the observed labels. Spectral properties of these linear operators are then analysed as a function of the number of iterations, which can be solved for a stopping time via some form of bias-variance decomposition. Our work, in

contrast, enables simultaneously studying a *family* of update rules, characterized by different choices of the mirror map, in a unified framework without relying on the access to closed-form expressions of the iterates.

One of the primary contributions of our work is the connection between mirror descent iterates and localized complexity measures. To the best of our knowledge, there are only two prior works making connections of a similar nature, albeit only in the setting of Euclidean gradient descent updates, that is, with the choice of the mirror map $\psi(\alpha) = \|\alpha\|_2^2/2$ [51,64]. Such connections are observed in an algebraic fashion in the former work, while localized complexities appear more naturally in [64], via the analysis of the range of estimators defined by gradient descent iterates up to the stopping time. In this respect, the work in [64] is the closest to ours. In Theorem 4, we show how a straightforward application of our main results immediately recovers results similar to the ones obtained in [51,64] and defer an extended discussion of similarities and differences to Section 4.1.

Beyond the Euclidean set-up, interest in understanding the generalization properties of neural networks has sparked research into *implicit* regularization properties of various factorized models. In the context of neural networks, the authors of [6,25,26,35,65] show that iterates of gradient descent applied to factorized matrix models are implicitly biased towards some sparsity-inducing structure such as low-rankness or low nuclear norm. Such results, however, hold under certain limit statements, such as vanishing initialization or step-size, the number of iterations going to infinity or no noise in the problem. In the setting of linear regression, matrix factorization models reduce to vector Hadamard product factorizations, where early-stopped gradient descent was shown to yield minimax optimal rates for sparse recovery with the analysis vitally relying on the restricted isometry property [59,69]. In Theorem 5, we demonstrate a simple analysis of such updates within our framework *without any assumptions* on the design matrix other than bounded columns, yielding a (up to a log factor) minimax optimal algorithm for in-sample linear prediction under ℓ_1 norm constraints.

Implicit regularization properties of mirror descent have recently attracted a considerable amount of attention; however, most results in this area either focus on optimization guarantees that do not provide any direct link to statistical guarantees on out-of-sample prediction [8,27], or establish a connection to statistics via some forms of explicit regularization [58]. Specifically, it is shown in ([58], Section 3) that the *continuous-time* mirror descent flow satisfies excess risk guarantees similar to guarantees obtainable via a regularization path of an explicitly penalized procedure. However, this analysis is based on a strong convexity assumption on the *empirical risk function*, an assumption not present in our work (see the discussion following Assumption 1). Without strong convexity of the empirical loss function, ([58], Section 4) show that mirror descent iterates can diverge from a regularization path of a corresponding explicitly regularized procedure. This does not cause issues in our analysis because our excess risk guarantees, obtained for suitably early-stopped mirror descent iterates, do not rely on pointwise closeness to some specific explicit regularization scheme. In Proposition 6, we show how the analysis of such problems naturally fit into our framework and defer an extended discussion and comparison with the other related work [2] to Section 4.3. Yet other papers have used early stopping to solvers applied directly to appropriately constrained problems and regularization-promoting structures encoded directly into the loss function [41].

Recent work has also focused on providing statistical guarantees for iterates generated via gradient descent updates in stochastic [3,38,45,55], accelerated [22,47] and distributed settings [37,52,53]. These works provide statistical guarantees without establishing connections to localized complexity measures; we anticipate such connections to be studied within our framework in future work, for a family of mirror descent algorithms.

3. Main Results

The main results of this paper establish that early-stopped mirror descent iterates satisfy the preconditions needed to obtain excess risk bounds via the offset Rademacher complexity theory; see the statement of Theorem 1. In particular, we show that suitably stopped mirror descent iterates satisfy the offset condition (see Definition 2) while remaining in a certain bounded set, the offset complexity of which yields an excess risk upper-bound.

In the argument sketched in Section 2, we have restricted our attention to the simplified case of the quadratic loss that allowed us to prove the *equality* (cf. Lemma 1) $\langle -\nabla R_n(\alpha), \alpha' - \alpha \rangle = R_n(\alpha) - R_n(\alpha') + \|g_\alpha - g_{\alpha'}\|_n^2$. However, the argument presented in Section 2 only relied on having a *lower-bound* on $\langle -\nabla R_n(\alpha), \alpha' - \alpha \rangle$. Such a lower-bound follows directly from the strong convexity assumption on the loss function (Assumption 1), as we show in the following lemma.

LEMMA 2. Suppose that the loss function ℓ satisfies the γ-strong convexity assumption (Assumption 1) and suppose that the function class representation assumption (Assumption 2) holds. Then, the following inequality holds for any $\alpha' \in \mathbb{R}^m$:

$$\langle -\nabla R_n(\alpha), \alpha' - \alpha \rangle \geq R_n(\alpha) - R_n(\alpha') + \frac{\gamma}{2}\|g_\alpha - g_{\alpha'}\|_n^2.$$

Proof. Recall that $(x_i, y_i)_{i=1}^n$ denotes the observed data sample and for any $y' \in \mathbb{R}$ denote by $\ell_{y'}$ the function $y \mapsto \ell(y, y')$. By Assumption 1, the following holds for any $i = 1, \ldots, n$:

$$\ell_{y_i}(g_{\alpha'}(x_i)) \geq \ell_{y_i}(g_\alpha(x_i)) + \ell'_{y_i}(g_\alpha(x_i))(g_{\alpha'}(x_i) - g_\alpha(x_i)) + \frac{\gamma}{2}(g_\alpha(x_i) - g_{\alpha'}(x_i))^2.$$

Summing the above equation for $i = 1, \ldots, n$ and dividing both sides by n yields

$$R_n(\alpha') \geq R_n(\alpha) + \frac{1}{n}\sum_{i=1}^n \ell'_{y_i}(g_\alpha(x_i))(g_{\alpha'}(x_i) - g_\alpha(x_i)) + \frac{\gamma}{2}\left\|g_\alpha - g_{\alpha'}\right\|_n^2. \tag{3.1}$$

Finally, by Assumption 2, we have $g_\alpha(x_i) = (Z\alpha)_i = z_i^\top \alpha$, where $z_i \in \mathbb{R}^m$ is the i-th row of the matrix $Z \in \mathbb{R}^{n \times m}$. Hence, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \ell'_{y_i}(g_\alpha(x_i))(g_{\alpha'}(x_i) - g_\alpha(x_i)) &= \frac{1}{n}\sum_{i=1}^n \ell'_{y_i}(z_i^\top \alpha)(z_i^\top \alpha' - z_i^\top \alpha) \\
&= \frac{1}{n}\sum_{i=1}^n \left(z_i \ell'_{y_i}(z_i^\top \alpha)\right)^\top (\alpha' - \alpha) \\
&= \frac{1}{n}\sum_{i=1}^n \left(\nabla_\alpha \ell_{y_i}(g_\alpha(x_i))\right)^\top (\alpha' - \alpha) \\
&= \langle \nabla R_n(\alpha), \alpha' - \alpha \rangle.
\end{aligned}$$

Plugging the above identity into the inequality (3.1) and rearranging completes the proof. □

We will now state and prove our main results. The rest of this section is split into two parts. First, we prove a continuous-time result (Section 3.1) already sketched in Section 2. Next, in Section 3.2, we prove a discrete-time result under additional assumptions on smoothness of the empirical risk function and strong convexity of the mirror map.

3.1 *Continuous-Time Version of the Main Result*

In this section, we state and prove a continuous-time version of our main theorem, which demonstrates the key ideas behind our approach in the simplest setting. The first part of the theorem shows that the iterates of mirror descent stay within a certain Bregman ball up to the prescribed stopping time $t^\star$. The second part of the theorem immediately establishes that when the parametrization given by $\alpha \in \mathbb{R}^m$ is independent of the data[3], the early-stopped estimator $g_{\alpha_{t^\star}}$ satisfies the offset condition (cf. Definition 2) with parameters $\gamma/2$ (depending on the strong convexity of the loss function ℓ) and any $\varepsilon > 0$ (depending on the amount of prescribed computational resources). For the applications we consider, we choose ε to match the complexity measure of interest and recover the statistical-computational trade-offs consistent with the previous results in the literature. In particular, $t^\star = O(D_\psi(\alpha', \alpha_0)/\varepsilon)$, so that achieving higher statistical accuracy requires more computational power; we also note that the dependence of $t^\star$ on the unknown radius $D_\psi(\alpha', \alpha_0)$ is unavoidable purely from an optimization point of view. Finally, we remark that early-stopping is, in general, necessary to transform the results of the below theorem into sharp statistical guarantees. Indeed, in Section 4.2, we demonstrate an application to sparse linear prediction problem where early-stopped mirror descent iterates satisfy an excess risk bound with logarithmic dependence on the ambient dimension, whereas at convergence (i.e. $t^\star = \infty$), this is no longer true. To see this, consider the setting of Section 4.2 and take $d = n$ and $Z = \sqrt{n} I_n$,where I_n is the $n \times n$ identity matrix; then, at $t^\star = \infty$ we obtain the (unique) ordinary least squares solution with excess risk of constant order (i.e. a trivial guarantee).

THEOREM 2. Suppose that the γ-strong convexity assumption (Assumption 1) and the function representation assumption (Assumption 2) hold. Let $\alpha_0 \in \mathbb{R}^m$ be the initialization point and $\psi : \mathbb{R}^m \to \mathbb{R}$ be a mirror map (cf. Section 1.1.3). Consider the continuous-time mirror descent dynamics given by

$$\frac{d}{dt}\alpha_t = -(\nabla^2 \psi(\alpha_t))^{-1} \nabla R_n(\alpha_t).$$

Then, for any chosen reference point α' and any $\varepsilon > 0$, there exists a stopping time $t^\star = t^\star(D_n, \psi, \alpha_0, \alpha') \le 2D_\psi(\alpha', \alpha_0)/\varepsilon$ such that:

1. For all $0 \le t \le t^\star$, $g_{\alpha_t} \in \mathcal{G}(\psi, \alpha_0, \alpha') = \{g_\alpha \in \mathbb{R}^m : D_\psi(\alpha', \alpha) \le D_\psi(\alpha', \alpha_0)\}$. In particular, up to the stopping time $t^\star$, the mirror descent iterates remain in the set $\mathcal{G}(\psi, \alpha_0, \alpha')$.
2. At the stopping time $t^\star$, we have $R_n(\alpha_{t^\star}) - R_n(\alpha') + \frac{\gamma}{2}\|g_{\alpha_{t^\star}} - g_{\alpha'}\|_n^2 \le \varepsilon$. In particular, at the stopping time $t^\star$, the estimator $g_{\alpha_{t^\star}}$ satisfies the offset condition (cf. Definition 2) with parameters ε and $\gamma/2$.

[3] When the parametrization is data-dependent, such as in the setting of kernel methods, our main theorems also establish that the early-stopped mirror descent iterates satisfy the offset condition (cf. Definition 2). We analyse a concrete example and provide full details in Theorem 4.

Proof. Using Lemma 2 instead of Lemma 1, we may repeat the argument sketched in Section 2. To simplify the notation, let $\delta_t = R_n(\alpha_t) - R_n(\alpha')$ and $r_t = \frac{\gamma}{2}\|g_{\alpha_t} - g_{\alpha'}\|_n^2$. As discussed in Section 1.1.3, the continuous-time mirror descent iterates satisfy the following identity:

$$-\frac{d}{dt}D_\psi(\alpha', \alpha_t) = \left\langle -\nabla R_n(\alpha_t), \alpha' - \alpha_t \right\rangle.$$

Combining the above equation with Lemma 2 we obtain the following bound on the continuous-time change of Bregman divergence

$$-\frac{d}{dt}D_\psi(\alpha', \alpha_t) \geq r_t + \delta_t.$$

Let $T = 2D_\psi(\alpha', \alpha_0)/\varepsilon$. Integrating both sides of the above inequality we obtain

$$D_\psi(\alpha', \alpha_0) - D_\psi(\alpha', \alpha_T) = \int_0^T -\frac{d}{dt}D_\psi(\alpha', \alpha_t)dt \geq \int_0^T (r_t + \delta_t)dt$$

$$\implies \inf_{0\leq t\leq T}\left\{r_t + \delta_t\right\} \leq \frac{1}{T}\int_0^T (r_t + \delta_t)dt \leq \frac{D_\psi(\alpha', \alpha_0)}{T} = \frac{\varepsilon}{2}.$$

It follows that the following infimum is well defined:

$$t^\star = \inf\{0 \leq t \leq T \mid r_t + \delta_t \leq \varepsilon\}.$$

Hence, $r_{t^\star} + \delta_{t^\star} \leq \varepsilon$, which proves the second assertion of this theorem. To prove the first assertion, observe that for all $0 \leq t \leq t^\star$ we have

$$D_\psi(\alpha', \alpha_0) - D_\psi(\alpha', \alpha_t) \geq \int_0^t (r_t + \delta_t)dt \geq t\varepsilon \geq 0.$$

The above inequality implies that $D_\psi(\alpha', \alpha_t) \leq D_\psi(\alpha', \alpha_0)$, which concludes our proof. □

3.2 *Discrete-Time Version of the Main Result*

In the following theorem, we prove a discrete-time counterpart to the continuous-time theorem proved in the previous section. We will show a variant of a discrete-time result under smoothness of the empirical loss function R_n and under strong convexity of the mirror map; such assumptions are natural from the optimization point of view (see, e.g. the monograph by Bubeck [18]).

Let $\|\cdot\|$ denote any norm. We say that R_n is β-smooth with respect to $\|\cdot\|$ if $R_n(\alpha') \leq R_n(\alpha) + \langle \nabla R_n(\alpha), \alpha' - \alpha\rangle + \frac{\beta}{2}\left\|\alpha - \alpha'\right\|^2$ for any α, α' in the domain of R_n. We also say that the mirror map ψ is ρ-strongly convex with respect to $\|\cdot\|$ if for any α, α' we have $D_\psi(\alpha', \alpha) \geq \frac{\rho}{2}\left\|\alpha' - \alpha\right\|^2$.

With the definition of smoothness and strong convexity with respect to general norms in place, we are now ready to state the discrete time theorem.

THEOREM 3. Suppose that the γ-strong convexity assumption (Assumption 1) and the function representation assumption (Assumption 2) hold. Additionally, suppose that the empirical risk function R_n is

β-smooth and the mirror map ψ (cf. Section 1.1.3) is ρ-strongly convex with respect to some norm $\|\cdot\|$. Let $\alpha_0 \in \mathbb{R}^m$ be the initialization point and let $0 < \eta \leq \frac{\rho}{\beta}$ be the step size. Consider the discrete-time mirror descent updates given by

$$\nabla\psi(\alpha_{t+1}) = \nabla\psi(\alpha_t) - \eta\nabla R_n(\alpha_t).$$

Then, for any chosen reference point α' and any $\varepsilon > 0$, there exists a stopping time $t^\star = t^\star(D_n, \psi, \alpha_0, \alpha', \eta) \leq (D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha'))/(\eta\varepsilon)$ such that:

1. For all $0 \leq t \leq t^\star$, $g_{\alpha_t} \in \mathcal{G}(\psi, \alpha_0, \alpha', \eta) = \{g_\alpha : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha')\}$. In particular, up to the stopping time $t^\star$, the mirror descent iterates remain in the set $\mathcal{G}(\psi, \alpha_0, \alpha', \eta)$.
2. At the stopping time $t^\star$, we have $R_n(\alpha_{t^\star}) - R_n(\alpha') + \frac{\gamma}{2}\|g_{\alpha_{t^\star}} - g_{\alpha'}\|_n^2 \leq \varepsilon$. In particular, at the stopping time $t^\star$, the estimator $g_{\alpha_{t^\star}}$ satisfies the offset condition (cf. Definition 2) with parameters ε and $\gamma/2$.

Before providing the proof, we briefly comment on the above theorem. First, the step-size condition $\eta \leq \rho/\beta$ and the number of iterations $O(1/\varepsilon)$ needed to reach a desired level of accuracy are identical to the guarantees proved in purely convex optimization settings (cf. Theorem 4.4 in [18]). On the other hand, comparing Theorems 2 and 3, in the discrete setting we pay a price of $\eta R_n(\alpha')$ in the radius of the Bregman ball where our early-stopped estimator lies. This is consistent with prior work in the early stopping literature, where such an expansion of the radius dependent on the 'noise level' (when measured by $R(\alpha')$, which for the population risk minimizer α' in a well-specified least-squares regression model corresponds to the variance of the additive response-variable noise) propagates into the resulting bounds (cf. definition of C in Theorem 1 in [64]). Our work, on the other hand, allows for a more fine-grained control of statistical-computational trade-offs via a selection of a small enough step-size η.

We now introduce two lemmas supporting the proof of Theorem 3. The first lemma is a well-known generalization of the Euclidean identity $\|a\|_2^2 + \|b\|_2^2 = \|a-b\|_2^2 + 2\langle a, b\rangle$, which holds for any Bregman divergence D_ψ induced by any mirror map ψ.

LEMMA 3. For any mirror map ψ and any points x, y, z in the domain of ψ we have

$$D_\psi(z, x) - D_\psi(z, y) = \langle\nabla\psi(x) - \nabla\psi(y), x - z\rangle - D_\psi(x, y).$$

Proof. The identity follows by the definition of Bregman divergence. □

The second lemma proves a discrete-time counterpart to the identity given in Equation (1.4), which combined with Lemma 2 states that $-\frac{d}{dt}D_\psi(\alpha', \alpha_t) = \langle -\nabla R_n(\alpha_t), \alpha' - \alpha_t\rangle \geq \delta_t + r_t$. We remind our reader that $\delta_t = R_n(\alpha_t) - R_n(\alpha')$ and $r_t = \frac{\gamma}{2}\|g_{\alpha_t} - g_{\alpha'}\|_n^2$.

LEMMA 4. Consider the setting of Theorem 3. Then, the discrete-time mirror descent iterates $(\alpha_t)_{t\geq 0}$ satisfy the following inequality for any reference point α' and any time $t \geq 0$:

$$D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t),$$

where $\delta_{t+1} = R_n(\alpha_{t+1}) - R_n(\alpha')$ and $r_t = \|g_{\alpha_t} - g_{\alpha'}\|_n^2$.

Proof. Combining Lemma 3 with the definition of discrete-time mirror descent updates (cf. Equation (1.3)) we have

$$\begin{aligned}
&D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \\
&= \left\langle \nabla\psi(\alpha_t) - \nabla\psi(\alpha_{t+1}), \alpha_t - \alpha' \right\rangle - D_\psi(\alpha_t, \alpha_{t+1}) \\
&= \left\langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha' \right\rangle - \left(\psi(\alpha_t) - \psi(\alpha_{t+1}) - \left\langle \nabla\psi(\alpha_{t+1}), \alpha_t - \alpha_{t+1} \right\rangle\right) \\
&= \left\langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha' \right\rangle - \left(-D_\psi(\alpha_{t+1}, \alpha_t) + \left\langle \nabla\psi(\alpha_t) - \nabla\psi(\alpha_{t+1}), \alpha_t - \alpha_{t+1} \right\rangle\right) \\
&= \left\langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha' \right\rangle - \left(-D_\psi(\alpha_{t+1}, \alpha_t) + \left\langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha_{t+1} \right\rangle\right) \\
&= \left\langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha' \right\rangle + D_\psi(\alpha_{t+1}, \alpha_t) + \left\langle -\eta\nabla R_n(\alpha_t), \alpha_t - \alpha_{t+1} \right\rangle. \qquad (3.2)
\end{aligned}$$

By the ρ-strong convexity of the mirror map ψ, the second term in Equation (3.2) can be lower bounded as $D_\psi(\alpha_{t+1}, \alpha_t) \geq \frac{\rho}{2}\left\|\alpha_{t+1} - \alpha_t\right\|^2$. The last term in Equation (3.2) can be lower bounded using the β-smoothness condition of the empirical risk function R_n, which yields $\langle -\nabla R_n(\alpha_t), \alpha_t - \alpha_{t+1}\rangle = \langle \nabla R_n(\alpha_t), \alpha_{t+1} - \alpha_t\rangle \geq R_n(\alpha_{t+1}) - R_n(\alpha_t) - \frac{\beta}{2}\|\alpha_{t+1} - \alpha_t\|^2$. We can hence continue from Equation (3.2) as follows:

$$\begin{aligned}
&D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \\
&= \left\langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha' \right\rangle + D_\psi(\alpha_{t+1}, \alpha_t) + \left\langle -\eta\nabla R_n(\alpha_t), \alpha_t - \alpha_{t+1} \right\rangle \\
&\geq \left\langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha' \right\rangle + \frac{\rho}{2}\left\|\alpha_{t+1} - \alpha_t\right\|^2 + \eta\left(R_n(\alpha_{t+1}) - R_n(\alpha_t) - \frac{\beta}{2}\|\alpha_{t+1} - \alpha_t\|^2.\right) \\
&= \left\langle \eta\nabla R_n(\alpha_t), \alpha_t - \alpha' \right\rangle + \left(\frac{\rho - \eta\beta}{2}\right)\left\|\alpha_{t+1} - \alpha t\right\|^2 + \eta(\delta_{t+1} - \delta_t).
\end{aligned}$$

Since $\eta \leq \rho/\beta$, the second term is lower bounded by 0. Also, by Lemma 2, the first term can be lower bounded as follows: $\eta\langle -\nabla R_n(\alpha_t), \alpha' - \alpha_t\rangle \geq \eta(\delta_t + r_t)$. Combining these two observations with the last equation above we obtain

$$D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t),$$

which completes our proof. □

With Lemma 4 at hand, we can prove Theorem 3 following along the same steps used to prove Theorem 2, albeit with the continuous-time equation $-\frac{d}{dt}D_\psi(\alpha', \alpha_t) = \delta_t + r_t$ replaced with its discrete-time counterpart $D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t)$. In the discrete-time equation, δ_t is replaced with δ_{t+1}, which results in the expansion of the radius of the Bregman ball in which the mirror descent iterates lie before the prescribed stopping time (cf. the discussion following the statement of Theorem 3 above).

Proof of Theorem 3. By Lemma 4 we have $D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t)$. Let $T = \left\lceil \frac{D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha')}{\eta \varepsilon} \right\rceil$. Summing both sides of the above equation for $t = 0, \ldots, T$ we obtain

$$D_\psi(\alpha', \alpha_0) - D_\psi(\alpha', \alpha_T) \geq \eta r_0 + \sum_{t=1}^{T} \eta(r_t + \delta_t) + \eta \delta_{T+1}$$

$$\Longrightarrow \min_{t=1,\ldots,T} \{\delta_t + r_t\} \leq \frac{\sum_{t=1}^{T} r_t + \delta_t}{T} \leq \frac{D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha')}{\eta T} \leq \varepsilon,$$

where in the last inequality we have used the definition of T and facts that $D_\psi(\alpha', \alpha_T) \geq 0$, $r_0 \geq 0$ and $\delta_{T+1} \geq -R_n(\alpha')$.

It follows that the following minimum is well defined: $t^\star = \min\{t = 0, \ldots, T \mid r_t + \delta_t \leq \varepsilon\}$. Hence, $r_{t^\star} + \delta_{t^\star} \leq \varepsilon$, which proves the second assertion of the theorem. To prove the first assertion of the theorem, note that for any $1 \leq t \leq t^\star$ by telescoping the equation $D_\psi(\alpha', \alpha_t) - D_\psi(\alpha', \alpha_{t+1}) \geq \eta(\delta_{t+1} + r_t)$ from 0 to $t - 1$ we obtain

$$D_\psi(\alpha', \alpha_0) - D_\psi(\alpha', \alpha_t) \geq \eta r_0 + \sum_{t=1}^{t-1} \eta(r_t + \delta_t) + \eta \delta_t$$

$$\Longrightarrow D_\psi(\alpha', \alpha_t) \leq D_\psi(\alpha', \alpha_0) - \sum_{t=1}^{t-1} \eta(r_t + \delta_t) - \eta \delta_t \leq D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha'),$$

where in the last line we have used the facts that $\delta_t + r_t > \varepsilon > 0$ and $-\delta_t \leq R_n(\alpha')$. □

4. Selected Applications of the Main Results

In this section, we discuss three selected applications of our main theorems.

Most of the results on early stopping in prior literature are shown for vanilla gradient descent updates in the non-parametric regression setting over reproducing kernel Hilbert spaces (cf. Section 2.1). In such settings, the parametrization α depends on the observed data. Theorem 4 that we present in Section 4.1 demonstrates that such data-dependent parametrizations easily fit within our framework. Additionally, we obtain results that in some ways improve upon related work, e.g. we obtain bounds on excess risk with no assumptions on the distribution P other than boundedness of its support.

In Section 4.2, we consider a problem of bounding the in-sample linear prediction error under the quadratic loss and under the ℓ_1 constraints on the optimal predictor. Such a setting has recently attracted a lot of attention in implicit regularization literature, specifically, when the design matrix is assumed to satisfy regularity conditions such as the restricted isometry property [59,69]. In Theorem 5, we obtain an up to logarithmic factors minimax-optimal bound in the setting where the design matrix does not satisfy the restricted isometry condition, but instead, its columns are bounded in ℓ_2 norm.

Some recent works [2,58] investigated the connections between *continuous-time* optimization paths traced by gradient and mirror descent algorithms, and regularization paths of suitably regularized problems. Via Proposition 6 proved in Section 4.3, we demonstrate that such questions can also be addressed within our framework.

Crucially, various different questions recently studied in the related literature naturally fit within the framework developed in our paper, which provides a simple and unified way to approach such problems. Moreover, in all of the three considered examples, we prove results that in some aspects improve upon the prior work spanning several different sub-areas in the early stopping literature.

4.1 *Early Stopping for Non-Parametric Regression*

In this section, we consider the kernel regression setting described in Example 2 in Section 1.2.

Let P be any distribution supported on $\mathcal{X} \times [-M, M]$ for some constant $M > 0$. Let $k : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a Mercer kernel which induces a Hilbert space of functions $\mathcal{H}$ equipped with norm $\|\cdot\|_{\mathcal{H}}$. Assume that $\sup_{x \in \mathcal{X}} k(x, x) \leq L$ for some constant $L > 0$ and, conditionally on the observed data sample $D_n = (x_i, y_i)_{i=1}^n$, denote by $K \in \mathbb{R}^{n \times n}$ a matrix such that $K_{ij} = k(x_i, x_j)$.

In the theorem below, we consider the discrete-time mirror descent updates defined as

$$\alpha_0 = 0, \qquad \alpha_{t+1} = \alpha_t - \frac{\eta}{n}(K\alpha_t - y). \tag{4.1}$$

The above updates correspond to mirror descent updates with the mirror map $\psi(\alpha) = \alpha^\mathsf{T} K \alpha$.[4] Observe that the mirror map ψ is 2-strongly convex with respect to the norm $\|\cdot\|_K$ defined by $\|\alpha\|_K^2 = \psi(\alpha) = \alpha^\mathsf{T} K \alpha$. To each $\alpha \in \mathbb{R}^n$, we associate a $g_\alpha \in \mathcal{H}$ defined as $g_\alpha = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$. For any α, α', the squared distance between the functions $g_{\alpha'}$ and g_α with respect to the RKHS norm $\|\cdot\|_{\mathcal{H}}$ is given by the Bregman divergence $D_\psi(\alpha', \alpha)$:

$$\|g_{\alpha'} - g_\alpha\|_{\mathcal{H}}^2 = (\alpha' - \alpha)^\mathsf{T} K (\alpha' - \alpha) = \|\alpha' - \alpha\|_K^2 = D_\psi(\alpha', \alpha).$$

We make the following assumptions on the loss function ℓ:

1. The loss function ℓ satisfies the γ-strong convexity assumption (Assumption 1);
2. For any $b > 0$ and any $y' \in [-b, b]$, the function $\ell(\cdot, y') : [-b, b] \to [0, \infty)$ is C_b-Lipschitz;
3. The empirical risk function $R_n(\alpha)$ is $\beta = \beta(D_n)$-smooth with respect to the $\|\cdot\|_K$ norm.

For example, if $\ell(y, y') = (y - y')^2$ is the quadratic loss function, then the strong convexity parameter satisfies $\gamma = 2$, the Lipschitzness parameter satisfies $C_b = 4b$, and the smoothness parameter satisfies $\beta = 2\lambda_{\max}(K/n)$, where $\lambda_{\max}(K/n)$ denotes the maximum eigenvalue of the normalized (data-dependent) kernel matrix K/n.

Since our parameter system is data-dependent (both K and the parametrization given by α depend on the observed data points), there is, in general, no single $\alpha' \in \mathbb{R}^n$ such that $g' = g_{\alpha'}$ for all realization of D_n, where $g' \in \mathcal{H}$ is some arbitrary reference function of interest. Hence, Theorem 3 does not immediately establish that early-stopped mirror descent iterates satisfy the offset condition. Our proof that we present below demonstrates how a data-dependent parameter system can be analysed within our

[4] Typically, the map $\nabla\psi : \mathbb{R}^n \to \mathbb{R}^n$ is required to be surjective to ensure that the elements of the dual space can always be mapped back to the primal space, which is not necessarily the case with the choice of the mirror map $\psi(\alpha) = \alpha^\mathsf{T} K \alpha$. However, note that pre-multiplying both sides of Equation (4.1) by $2K$, for any $t \geq 0$ it holds that $\nabla\psi(\alpha_{t+1}) = \nabla\psi(\alpha_t) - \eta \nabla R_n(\alpha_t)$ and hence the updates defined in (4.1) are mirror descent updates.

framework. The key idea is to find $\alpha'(D_n) \in \mathbb{R}^n$, one for each dataset D_n, such that $g_{\alpha'(D_n)}$ is 'close enough' to a reference function of interest g'.

THEOREM 4. Consider the set-up described above and consider the discrete-time mirror descent updates (4.1) with any step size $\eta \in (0, 2/\beta)$. Fix any $R > 0$ and let $\mathcal{F}_R = \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \leq R\}$. Then, there exists a data-dependent stopping time $t^\star \leq (R')^2/(\eta \mathbb{E}_{D_n}\Big[\mathfrak{R}_{D_n^x}(\mathcal{F}_{R'}, C_b^{-1}\gamma/4)\Big])$ such that for any $\delta \in (0,1)$, with probability at least $1-\delta$ it holds that

$$\mathcal{E}(g_{\alpha_{t^\star}}, \mathcal{F}_R) \leq c_1 C_b \mathbb{E}_{D_n}\Big[\mathfrak{R}_{D_n^x}(\mathcal{F}_{R'}, C_b^{-1}\gamma/4)\Big] + c_2 \frac{\gamma^{-1} C_b^2 \log(1/\delta)}{n},$$

where $R' > 0$ satisfies $(R')^2 = 10R^2 + 2\sup_{y\in[-M,M]} \ell(0,y)$, $b = \max\{M, L(R'+R)\}$, and $c_1, c_2 > 0$ are universal constants.

Before presenting the proof, we compare the above theorem with the related works connecting early stopping and localized complexity measures [51,64] in the setting of RKHS. First, the work [51] considers the quadratic loss, while the work [64] considers general loss functions under strong convexity, smoothness and Lipschitzness conditions, in close similarity to the setting considered above. Second, the works [51,64] considered vanilla gradient descent updates; in contrast, the above theorem follows as a corollary of Theorem 3 that treats a general family of mirror descent algorithms. The flexibility of Theorem 3 comes from the fact that we use different mathematical machinery to obtain excess risk bounds; namely, we rely on localization via offset Rademacher complexities, which allows us to consider a more general class of algorithms for the reasons outlined in Section 1.1. Consequently, and in contrast to the results obtained in [51,64], our main results can also be applied to provide statistical guarantees *along the whole optimization path* (cf. Proposition 6)). Third, concerning the random design setting considered in this work, the bounds in [51,64] were proved under a well-specified model and i.i.d. noise assumptions, neither of which is present in Theorem 4 considered in this section. Finally, the differences aside, Theorem 4 is similar to the results obtained in [51,64]. In particular, we recover similar conditions on the step size and provide almost identical statistical and computational guarantees. We refer to ([62], Chapters 13 and 14) for further discussions concerning the statistical optimality of localized complexity measures for non-parametric regression.

Proof of Theorem 4. By the Representer theorem, there exists $\alpha' = \alpha'(D_n)$, such that $g_{\alpha'(D_n)} \in \operatorname{argmin}_{g\in\mathcal{F}_R} R_n(g)$ and hence, by convexity of $\mathcal{F}_R$ and $\frac{\gamma}{2}$-strong convexity of ℓ, the triplet $(P, \mathcal{F}_R, g_{\alpha'(D_n)})$ satisfies the offset condition with parameters $\varepsilon = 0$ and $\gamma/2$. Consequently, with probability one we have

$$R_n(\alpha'(D_n)) - R_n(g_{\mathcal{F}_R}) + \frac{\gamma}{2}\|g_{\alpha'(D_n)} - g_{\mathcal{F}_R}\|_n^2 \leq 0. \tag{4.2}$$

Since $\alpha_0 = 0$, we have $D_\psi(\alpha'(D_n), \alpha_0) = \|\alpha'(D_n)\|_K^2 = \|g_{\alpha'(D_n)}\|_{\mathcal{H}}^2 \leq R^2$.

Also, by the fact that $g_{\alpha'(D_n)}$ is an empirical risk minimizer, we have $R_n(\alpha'(D_n)) \leq R_n(0) \leq \sup_{y\in[-M,M]} \ell(0,y) = \ell_{\max}$. Hence, applying Theorem 3 with $\alpha' = \alpha'(D_n)$, for any $\varepsilon > 0$, there exists a data-dependent stopping time $t^\star$ satisfying

$$t^\star \leq (D_\psi(\alpha', \alpha_0) + \eta R_n(\alpha'))/(\eta\varepsilon) \leq (R^2 + \eta\ell_{\max})/(\eta\varepsilon) \leq (R^2 + \ell_{\max})/(\eta\varepsilon),$$

such that the following two inequalities hold with probability 1:

$$R_n(\alpha_{t^\star}) - R_n(\alpha'(D_n)) + \frac{\gamma}{2}\|g_{\alpha_{t^\star}} - g_{\alpha'(D_n)}\|_n^2 \le \varepsilon,$$
$$D_\psi(\alpha'(D_n), \alpha_{t^\star}) \le D_\psi(\alpha'(D_n), \alpha_0) + \ell_{\max}. \tag{4.3}$$

Combining the first inequality above with Equation (4.2), the following holds with probability 1:

$$\begin{aligned} &R_n(\alpha_{t^\star}) - R_n(g_{\mathcal{F}_R}) + \frac{\gamma}{4}\|g_{\alpha_{t^\star}} - g_{\mathcal{F}_R}\|_n^2 \\ &\le R_n(\alpha_{t^\star}) - R_n(\alpha'(D_n)) + R_n(\alpha'(D_n)) - R_n(g_{\mathcal{F}_R}) \\ &\quad + \frac{\gamma}{2}\|g_{\alpha_{t^\star}} - g_{\alpha'(D_n)}\|_n^2 + \frac{\gamma}{2}\|g_{\alpha'(D_n)} - g_{\mathcal{F}_R}\|_n^2 \\ &\le \varepsilon. \end{aligned}$$

Thus, the triplet $(P, \mathcal{F}_R, g_{\alpha_{t^\star}})$ satisfies the offset condition with parameters $(\varepsilon, \gamma/4)$. In addition, by Equation (4.3) we have

$$\begin{aligned} \left\|g_{\alpha_{t^\star}} - g_{\mathcal{F}_R}\right\|_{\mathcal{H}}^2 &\le 2\left\|g_{\alpha_{t^\star}} - g_{\alpha'(D_n)}\right\|_{\mathcal{H}}^2 + 2\left\|g_{\alpha'(D_n)} - g_{\mathcal{F}_R}\right\|_{\mathcal{H}}^2 \\ &\le 2D_\psi(\alpha'(D_n), \alpha_{t^\star}) + 8R^2 \\ &\le \underbrace{10R^2 + 2\ell_{\max}}_{\text{denote by}(R')^2}. \end{aligned}$$

Hence $g_{\alpha_{t^\star}} - g_{\mathcal{F}_R} \in \mathcal{F}_{R'}$. Finally, by the assumption $\sup_{x\in\mathcal{X}} k(x,x) \le L$, for any $h \in \mathcal{H}$ we have $\sup_{x\in\mathcal{X}}|h(x)| \le L\|h\|_{\mathcal{H}}$. Since $g_{\alpha_{t^\star}} \in \mathcal{F}_{R'+R}$, it follows that $\sup_{x\in\mathcal{X}}|g_{\alpha_{t^\star}}| \le L(R'+R)$. Applying Theorem 1 to the early-stopped mirror descent estimator $g_{\alpha_{t^\star}}$ (with $b = \max\{M, L(R'+R)\}$) completes the proof of this theorem. □

4.2 *In-Sample Linear Prediction Under ℓ_1 Constraints*

Let $Z \in \mathbb{R}^{n\times d}$ be a fixed-design matrix such that the ℓ_2 norms of columns of $Z/\sqrt{n}$ are bounded by some constant κ. Assume a well-specified model, i.e. the existence of a vector α' such that the observations $y \in \mathbb{R}^n$ follow the distribution $y = Z\alpha' + \xi$, where ξ is a vector with i.i.d. zero-mean σ^2-subGaussian components. We aim to find a vector $\alpha \in \mathbb{R}^d$ that achieves a small in-sample prediction error defined as $\frac{1}{n}\left\|Z\alpha - Z\alpha'\right\|_2^2$.

A candidate implicit regularization based algorithm, known to be minimax optimal for sparse recovery under restricted isometry assumption [59,69], is defined as follows. Let $\alpha_t \in \mathbb{R}^d$ denote the iterate obtained at time t, let $\odot$ denote the Hadamard product, and let 1 denote a vector with all entries equal to one. Consider the parametrization $\alpha_t = u_t \odot u_t - v_t \odot v_t$ where $u_t, v_t \in \mathbb{R}^d$. Instead of running gradient descent directly on α_t, the algorithm considered in the works [59,69] is defined by running gradient descent updates on the concatenated parameter vector (u, v), yielding the following updates

(where $\gamma \in \mathbb{R}$):

$$u_0 = v_0 = \sqrt{\gamma/2} \cdot 1, \quad \alpha_t = u_t \odot u_t - v_t \odot v_t,$$
$$u_{t+1} = u_t \odot (1 - 2\eta \nabla R_n(\alpha_t)), \quad v_{t+1} = v_t \odot (1 + 2\eta \nabla R_n(\alpha_t)).$$

We remark that the above updates were also studied in [65], albeit with a focus on how the initialization scale affects the gradient descent solution obtained at convergence. Noting that $1 + x \approx e^x$ for small x, we can approximate the above updates (with the step-size η rescaled by a constant factor) by the unconstrained EG± algorithm [31] whose updates are given by

$$\alpha_0^+ = \alpha_0^- = (\gamma/2)1, \quad \alpha_t = \alpha_t^+ - \alpha_t^-,$$
$$\alpha_{t+1}^+ = \alpha_t^+ \odot \exp(-\eta \nabla R_n(\alpha_t)), \quad \alpha_{t+1}^- = \alpha_{t+1}^- \odot \exp(\eta \nabla R_n(\alpha_t)).$$

It was shown in [24] that the above updates correspond to running unconstrained mirror descent initialized at 0 with the mirror map given by

$$\psi(\alpha) = \phi_\gamma(\alpha) = \sum_{i=1}^{d} \left(\alpha_i \operatorname{arcsinh}\left(\alpha_i/\gamma\right) - \sqrt{\alpha_i^2 + \gamma^2} \right).$$

See [4,5] for extended discussions on the above update rules. In the rest of the section, we denote ψ by ϕ_γ to make the dependence on γ explicit. We consider running mirror descent with the hyperbolic entropy mirror map ϕ_γ with any $0 < \gamma \le (\|\alpha'\|_1 \wedge 1)/(3e^2 d)$ and with any step-size η that satisfies $0 \le \eta \le \frac{1}{24\kappa^2 \|\alpha'\|_1 \log(3\gamma^{-1})} \wedge \frac{\|\alpha'\|_1}{2\sigma^2}$. The theorem below yields minimax-optimal rates [50] for the in-sample prediction error up to the multiplicative factor $\log(3\gamma^{-1})$.

THEOREM 5. Consider the set-up described above. There exists a data-dependent stopping time $t^\star \le \sqrt{n}/(\eta \cdot 3\kappa\sigma\sqrt{\log d})$ such that with probability at least $1 - 2e^{-nc} - \frac{1}{8d^3}$, where c is an absolute constant, we have

$$\frac{1}{n} \left\| Z\alpha_{t^\star} - Z\alpha' \right\|_2^2 \le 36 \cdot \frac{\kappa \left\|\alpha'\right\|_1 \sigma \sqrt{\log d}}{\sqrt{n}} \cdot \log(3\gamma^{-1}).$$

Before proving the above theorem, we state two lemmas, which relate the Bregman divergence induced by the mirror map ϕ_γ to the geometry induced by the ℓ_1 norm. We prove both lemmas at the end of this section.

LEMMA 5. For any $0 < \gamma < (\left\|\alpha'\right\|_1 \wedge 1)/(3e^2 d)$ we have

$$\left\|\alpha'\right\|_1 \le D_{\phi_\gamma}(\alpha', 0) \le \left\|\alpha'\right\|_1 \log(3\gamma^{-1}).$$

Denote by $\mathcal{B}_R = \{w \in \mathbb{R}^d \mid \|w\|_1 \le R\}$ an ℓ_1 ball of radius R. The following lemma will be applied to show that before stopping, the mirror descent iterates $(\alpha_t)_{t \ge 0}$ stay inside an ℓ_1 ball with radius at most $6 \left\|\alpha'\right\|_1 \log(3\gamma^{-1})$.

LEMMA 6. For any $\alpha' \in \mathbb{R}^d$ and any $0 < \gamma < (\|\alpha'\|_1 \wedge 1)/(3e^2 d)$ we have

$$\left\{\alpha \in \mathbb{R}^d : D_{\phi_\gamma}(\alpha', \alpha) \leq 2D_{\phi_\gamma}(\alpha', 0)\right\} \subseteq \mathcal{B}_{6\|\alpha'\|_1 \log(3\gamma^{-1})}.$$

We are now ready to prove Theorem 5. We remark that since the slow rate $n^{-1/2}$ is minimax optimal [50] in the setting considered in Theorem 5, the localization effect provided by offset complexities is not needed in this example. However, we can apply Theorem 3 together with the *basic inequality* proof technique, as demonstrated in the proof below.

Proof. Proof of Theorem 5. First note that since the ℓ_2 norms of the columns of $Z/\sqrt{n}$ are bounded by κ, the empirical loss function R_n is $2\kappa^2$-smooth with respect to the ℓ_1 norm. Let

$$R^\star = 6\|\alpha'\|_1 \log(3\gamma^{-1}).$$

As shown in ([24], Lemma 4), ϕ_γ is also $\rho = (2R^\star)^{-1}$-strongly convex with respect to the ℓ_1 norm on $\mathcal{B}_{R^\star}$. Thus, we set the smoothness parameter $\beta = 2\kappa^2$ and the strong convexity parameter $\rho = (2R^\star)^{-1}$.

Condition on the event $A_1 = \{R_n(\alpha') \leq 2\sigma^2\}$. Since the noise random variables are σ^2-subGaussian, by sub-Exponential concentration we have $\mathbb{P}(A_1) \geq 1 - 2e^{-nc}$, where c is an absolute constant independent of any problem parameters. (cf. [61], Section 5.2.4)). By Theorem 3, Lemma 5 and $R_n(\alpha') \leq 2\sigma^2$, it is hence enough to set

$$\eta \leq \frac{1}{4\kappa^2 R^\star} \wedge \frac{\|\alpha'\|_1}{2\sigma^2} \leq \frac{\rho}{2\beta} \wedge \frac{D_\psi(\alpha', 0)}{\mathcal{L}(\alpha')}$$

so that there exists a stopping time

$$t^\star \leq \frac{2D_{\phi_\gamma}(\alpha', 0)}{\eta\varepsilon} \leq \frac{R^\star}{3\eta\varepsilon}$$

such that for all $t \leq t^\star$ it holds that

$$\alpha_t \in \{\alpha \in \mathbb{R}^d : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', 0) + \eta R_n(\alpha')\} \subseteq \mathcal{B}_{R^\star} \quad \text{(cf. Lemma 6)}$$

and also such that the following inequality holds:

$$R_n(\alpha_{t^\star}) - R_n(\alpha') + \frac{1}{n}\|Z\alpha_{t^\star} - Z\alpha'\|_2^2 \leq \varepsilon.$$

Rearranging the above inequality, as is typically done via the basic inequality proof technique (see, for example, ([62], Theorem 7.20)) we obtain

$$\frac{1}{n}\|Z\alpha_{t^\star} - Z\alpha'\|_2^2 \leq \left\langle \frac{1}{n} Z^\mathsf{T}\xi, \alpha_{t^\star} - \alpha' \right\rangle + \varepsilon \leq \frac{1}{n}\left\|Z^\mathsf{T}\xi\right\|_\infty \|\alpha_{t^\star} - \alpha'\|_1 + \varepsilon.$$

Define the event $A_2 = \{\frac{1}{n}\|Z^\mathsf{T}\xi\|_\infty \le 4\kappa\sigma\sqrt{\log d}/\sqrt{n}\}$. Since the ℓ_2 norms of the columns of $Z/\sqrt{n}$ are bounded by κ and since the noise vector ξ consists of independent σ^2 sub-Gaussian random variables, $\mathbb{P}(A_2) \ge 1 - \frac{1}{8d^3}$ by standard sub-Gaussian concentration.

By the union bound, the events A_1 and A_2 happen simultaneously with probability at least $1-2e^{-nc_6} - \frac{1}{8d^3}$. Setting $\varepsilon = R^\star\kappa\sigma\sqrt{\log d}/\sqrt{n}$ concludes our proof. □

Proof of Lemma 5. The upper-bound is shown in ([24], Section 3). We proceed as follows to prove the lower-bound:

$$\begin{aligned}
D_{\phi_\gamma}(\alpha',\alpha) &= \sum_{i=1}^d \left[\alpha_i'\left(\operatorname{arcsinh}\left(\frac{\alpha_i'}{\gamma}\right) - \operatorname{arcsinh}\left(\frac{\alpha_i}{\gamma}\right)\right) - \sqrt{(\alpha_i')^2+\gamma^2} + \sqrt{\alpha_i^2+\gamma^2}\right] \\
&\ge \sum_{i=1}^d \alpha_i'\left(\operatorname{arcsinh}\left(\frac{\alpha_i'}{\gamma}\right) - \operatorname{arcsinh}\left(\frac{\alpha_i}{\gamma}\right)\right) - 2\|\alpha'\|_1 + \|\alpha\|_1 \\
&\ge \sum_{i=1}^d \alpha_i'\operatorname{arcsinh}\left(\frac{\alpha_i'}{\gamma}\right) - \|\alpha'\|_1 \operatorname{arcsinh}\left(\frac{\|\alpha\|_1}{\gamma}\right) - 2\|\alpha'\|_1 + \|\alpha\|_1 \\
&= \sum_{i=1}^d |\alpha_i'| \log\frac{|\alpha_i'| + \sqrt{(\alpha_i')^2+\gamma^2}}{\gamma} - \|\alpha'\|_1 \operatorname{arcsinh}\left(\frac{\|\alpha\|_1}{\gamma}\right) - 2\|\alpha'\|_1 + \|\alpha\|_1 \\
&\ge \|\alpha'\|_1 \log\frac{\|\alpha_1'\|}{d\gamma} - \|\alpha'\|_1 \operatorname{arcsinh}\left(\frac{\|\alpha\|_1}{\gamma}\right) - 2\|\alpha'\|_1 + \|\alpha\|_1 \\
&= \|\alpha'\|_1 \log\frac{\|\alpha_1'\|}{e^2 d\gamma} - \|\alpha'\|_1 \operatorname{arcsinh}\left(\frac{\|\alpha\|_1}{\gamma}\right) + \|\alpha\|_1, \qquad (4.4)
\end{aligned}$$

where the penultimate line follows via an application of the log sum inequality. The result follows by plugging in $\|\alpha\|_1 = 0$ and using $\gamma \le \|\alpha'\|_1/(e^3 d)$. □

Proof of Lemma 6. Note that for any $x > \gamma > 0$, we have $\operatorname{arcsinh}(x/\gamma) \le \log(3x/\gamma)$. Hence, continuing from Equation (4.4) we have

$$\begin{aligned}
&\|\alpha\|_1 > \|\alpha'\|_1 \\
\implies &\|\alpha\|_1 \le D_{\phi_\gamma}(\alpha',\alpha) + \|\alpha'\|_1\left(\log\left(3e^2 d\right) + \log\frac{\|\alpha\|_1}{\|\alpha'\|_1}\right) \\
&\qquad \le D_{\phi_\gamma}(\alpha',\alpha) + \|\alpha'\|_1\left(\log\frac{1}{\gamma} + \frac{1}{2}\frac{\|\alpha\|_1}{\|\alpha_1'\|}\right) \\
\implies &\|\alpha\|_1 \le 2D_{\phi_\gamma}(\alpha',\alpha) + 2\|\alpha'\|_1 \log\frac{1}{\gamma}.
\end{aligned}$$

The result follows by applying the upper-bound proved in Lemma 5, namely, $D_{\phi_\gamma}(\alpha',\alpha) \le 2D_{\phi_\gamma}(\alpha',0) \le 2\|\alpha'\|_1\log(3\gamma^{-1})$. □

4.3 *Statistical Guarantees Along the Optimization Path*

In this section, we show that through the lens of offset Rademacher complexity, the iterates of the mirror descent algorithm satisfy similar excess risk guarantees to a family of explicitly constrained empirical risk minimization estimators. As discussed in the introduction, such results are of interest from the computational point of view: computation of a regularization path corresponds to solving a new optimization problem for each regularization parameter. In contrast, the computation of the mirror descent optimization path is relatively cheap in comparison, amounting to the cost of one gradient-based update to obtain a new candidate estimator.

We consider the following set-up. As in the rest of this paper, Assumptions 1 and 2 hold. In addition, in order for the general offset Rademacher complexity excess risk upper-bound (cf. Theorem 1) to be applicable, we assume that the data-generating distribution P is supported on $\mathcal{X} \times [-b, b]$ for some $b > 0$, and for any $y \in [-b, b]$ the loss function $\ell(\cdot, y)$ is C_b-Lipschitz.

Fix an arbitrary $\alpha_0 \in \mathbb{R}^m$ and for $R > 0$ let

$$\mathcal{F}_R = \mathcal{F}(\alpha_0, R) = \{\alpha \in \mathbb{R}^m : D_\psi(\alpha, \alpha_0) \leq R\}.$$

We define any optimal parameter in the space $\mathcal{F}_R$ by $\alpha_R^\star$:

$$\alpha_R^\star \in \operatorname{argmin}_{\alpha \in \mathcal{F}_R} R(\alpha),$$

where in the case of multiple minimizers, the ties may be broken arbitrarily.

REMARK 1. In the conference version of this paper, the result ([60], Theorem 5) is incorrect as stated: its proof is only correct for symmetric Bregman divergences. The correct formulation of this result is stated below in Proposition 6.

Henceforth, we restrict our analysis to mirror maps ψ for which the corresponding Bregman divergences D_ψ are symmetric: for any α, α' it holds that $D_\psi(\alpha, \alpha') = D_\psi(\alpha', \alpha)$. Consider the family of constrained empirical risk minimization estimators

$$\widehat{\alpha}_R^{(\mathrm{erm})} \in \operatorname{argmin}_{\alpha \in \mathcal{F}_R} R_n(\alpha).$$

For example, when $\psi(\alpha) = \|\alpha\|_2^2$, the estimators $(\widehat{\alpha}_R^{(\mathrm{erm})})_{R \geq 0}$ correspond to the ridge regression regularization path. For any $R > 0$, the convexity of the class $\mathcal{F}_R$ implies that the triplet $(P, \mathcal{F}_R, \widehat{\alpha}_R^{(\mathrm{erm})})$ satisfies the offset condition (cf. Definition 2) with parameters $\varepsilon = 0$ and $\gamma/2$, where recall that γ is equal to the strong-convexity parameter of ℓ stated in Assumption 1. In particular, by Theorem 1, for any $R > 0$ and for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ it holds that

$$\mathcal{E}(g_{\widehat{\alpha}_R^{(\mathrm{erm})}}, \mathcal{F}_R) \leq c_1 C_b' \mathbb{E}_{D_n}\Big[\mathfrak{R}_{D_n^{\mathrm{x}}}\Big(\operatorname{star}(\mathcal{F}_R - g_{\alpha_R^\star}), (C_b')^{-1}\frac{\gamma}{2}\Big)\Big] + c_2 \frac{2\gamma^{-1}(C_b')^2 \log(1/\delta)}{n}, \tag{4.5}$$

where $c_1, c_2 > 0$ are universal constants appearing in the statement of Theorem 1 and $C_b' = C_b + \gamma b/2$.

We now show that continuous-time mirror descent iterates $(\alpha_t)_{t \geq 0}$ satisfy nearly identical bound to (4.5): the bound stated below is the same modulo enlargement of the class $\mathcal{F}_R$ by $\mathcal{F}_{4R}$ and an extra additive term $\varepsilon > 0$, which can be chosen to be arbitrarily small at the expense of deteriorating upper-bounds on the stopping time.

PROPOSITION 6. Let $(\alpha_t)_{t\geq 0}$ be the continuous-time mirror descent optimization path (cf. Theorem 2). Suppose that for any α, α' we have $D_\psi(\alpha, \alpha') = D_\psi(\alpha', \alpha)$ (i.e. $D_\psi(\cdot,\cdot)$ is a squared Mahalanobis distance) and fix any $\varepsilon > 0$. Then, there exists a data-dependent stopping time $t_R^\star \leq 2R/\varepsilon$ such that for any $\delta \in (0,1)$ with probability at least $1-\delta$ it holds that

$$\mathcal{E}(g_{\alpha_{t_R^\star}}, \mathcal{F}_R) \leq c_1 C_b' \mathbb{E}_{D_n}\Big[\mathfrak{R}_{D_n^x}\Big(\mathrm{star}(\mathcal{F}_{4R} - g_{\alpha_R^\star}), (C_b')^{-1}\frac{\gamma}{2}\Big)\Big] + c_2 \frac{2\gamma^{-1}(C_b')^2 \log(1/\delta)}{n} + \varepsilon.$$

Proof. We apply Theorem 2 with the reference point $\alpha' = \alpha_R^\star$. By Theorem 2, there exists a data-dependent stopping time $t_R^\star \leq 2D_\psi(\alpha_R^\star, \alpha_0)/\varepsilon \leq 2R/\varepsilon$ such that $D_\psi(\alpha^\star, \alpha_{t_R^\star}) \leq D_\psi(\alpha^\star, \alpha_0) \leq R$ and the estimator $g_{\alpha_{t_R^*}}$ satisfies the offset condition with parameters $\varepsilon, \gamma/2$. By the assumption that D_ψ is symmetric, we have $D_\psi(\alpha, \alpha') = \|\alpha - \alpha'\|_A^2 = (\alpha-\alpha')^\top A(\alpha-\alpha')$ for some positive semi-definite matrix A (see [46], Lemma 2) for a proof of this claim). It follows that

$$D_\psi(\alpha_{t_R^\star}, \alpha_0) = \|\alpha_{t_R^\star} - \alpha_0\|_A^2 \leq 2\|\alpha_{t_R^\star} - \alpha_R^\star\|_A^2 + 2\|\alpha_R^\star - \alpha_0\|_A^2 \leq 4R.$$

In particular, in addition to satisfying the offset condition, the estimator $g_{\alpha_{t_R^\star}}$ is contained in the set $\mathcal{F}_{4R}$. The result follows by Theorem 1. □

By replacing the application of Theorem 2 with Theorem 3, a corresponding result may be obtained for discrete-time mirror descent iterates.

The above proposition complements some recently obtained results that connect optimization and regularization paths, as discussed below. First, ([58], Theorem 3) establishes that the optimization paths of *continuous-time* mirror descent algorithms and the regularization paths of *corresponding* regularized problems, when suitably aligned via some mapping between the number of mirror descent iterations and the regularization parameter of the penalized problem, are point-wise close. This allows the authors of the paper mentioned above to port existing results on explicitly regularized estimators to early-stopped mirror descent algorithms. However, their proof crucially depends on two assumptions. The first one requires strong-convexity and smoothness of the mirror map ψ with respect to the Euclidean norm; note that mirror maps of the form $\psi(\alpha) = \frac{1}{2}\alpha^\top A\alpha$ are symmetric, yet Proposition 6 does not depend on the conditioning of A. Their second assumption requires the empirical risk function $R_n(\alpha)$ to be strongly convex with respect to the Euclidean norm. Again, such an assumption is not present in our work, as explained in Section 1.2.

Concerning connections between regularization and optimization paths, another related paper is [2]. Therein, the authors study the optimization path of *continuous-time* gradient descent on a least-squares objective and show that the solution at time t has risk at most 1.69 times the risk of the *ridge* solution with regularization parameter $\lambda = 1/t$. Their results are based on the analytic tractability of gradient descent flow on least squares objective for a well-specified Gaussian model. Because we work under different assumptions (the loss is not quadratic, and the model is misspecified), our result obtained above is necessarily based on different tools.

5. Future Directions

Our work provides a simple and transparent framework for simultaneously analysing statistical and computational properties of iterates traced by a family of mirror descent algorithms applied to the

i.i.d. batch statistical learning setting. Among the research directions that would yield additional computational savings are extensions of our results to stochastic and accelerated frameworks, where connections between early stopping and localized complexity measures are yet to be established, even in the restricted setting of Euclidean gradient descent updates.

Beyond the computational savings, our main results reveal a curious property of mirror descent. For an unknown parameter of interest denoted by α', the statistical complexity of an appropriately stopped mirror descent iterate is given by the offset complexity of the class $\{g_\alpha - g_{\alpha'} : D_\psi(\alpha', \alpha) \leq D_\psi(\alpha', \alpha_0)\}$. Thus, g_α is *implicitly constrained* to lie in a *possibly non-convex* Bregman ball *centred at* the unknown α' with *unknown radius* $D_\psi(\alpha', \alpha_0)$. Therefore, in general, solutions traced by mirror descent iterates cannot be practically expressed as solutions of *explicitly constrained* optimization problems. Consequently, early-stopped mirror descent can potentially solve problems that cannot be tractably solved by the means of explicit regularization. This observation necessitates further investigation.

Data Availability Statement

No new data were generated or analysed in support of this research.

Acknowledgments

Tomas Vaškevičius was supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). Varun Kanade and Patrick Rebeschini were supported in part by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

1. Abernethy, J. D., Hazan, E. & Rakhlin, A. (2008) Competing in the dark: an efficient algorithm for bandit linear optimization. *COLT*. https://dblp.uni-trier.de/rec/conf/colt/AbernethyHR08.html?view=bibtex.
2. Ali, A., Kolter, J. Z. & Tibshirani, R. J. (2019) A continuous-time view of early stopping for least squares regression. In Chaudhuri, K. & Sugiyama, M. (eds) *International Conference on Artificial Intelligence and Statistics*. PMLR, **89**, pp 1370–1378.
3. Ali, A., Dobriban, E. & Tibshirani, R. J. (2020) The implicit regularization of stochastic gradient flow for least squares. *Proceedings of Machine Learning Research*. PMLR, **119**, 233–244.
4. Amid, E. & Warmuth, M. K. (2020a) Interpolating between gradient descent and exponentiated gradient using reparameterized gradient descent. arXiv preprint arXiv:2002.10487.
5. Amid Ehsan & Warmuth M. K.. (2020b) Winnowing with gradient descent. In Abernethy J. D. & Agarwal S (eds) *Conference on Learning Theory*, pages 163–182. PMLR.
6. Arora, S., Cohen, N., Hu, W. & Luo, Y. (2019) Implicit regularization in deep matrix factorization. *Adv. Neural Inf. Process. Syst.*, **32**, 7411–7422.
7. Audibert, J.-Y. (2008) Progressive mixture rules are deviation suboptimal. *Adv. Neural Inf. Process. Syst.*, **20**, 41–48.
8. Azizan, N. & Hassibi, B. (2019) Stochastic gradient/mirror descent: minimax optimality and implicit regularization. In *International Conference on Learning Representations*. https://dblp.uni-trier.de/rec/conf/iclr/RuhiH19.html?view=bibtex.
9. Bansal, N. & Gupta, A. (2019) Potential-function proofs for gradient methods. *Theory Comput.*, **15**, 1–32.
10. Bartlett, P. L. & Mendelson, S. (2002) Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, **3**, 463–482.
11. Bartlett, P. L. & Mendelson, S. (2006) Empirical minimization. *Probab. Theory Relat. Fields*, **135**, 311–334.
12. Bartlett, P. L. & Traskin, M. (2007) Adaboost is consistent. *J. Mach. Learn. Res.*, **8**, 2347–2368.
13. Bartlett, P. L., Bousquet, O. & Mendelson, S. (2005) Local Rademacher complexities. *Ann. Stat.*, **33**, 1497–1537.

14. Bauer, F., Pereverzev, S. & Rosasco, L. (2007) On regularization algorithms in learning theory. *J. Complexity*, **23**, 52–72.
15. Beck, A. & Teboulle, M. (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, **31**, 167–175.
16. Bickel, P. J., Ritov, Y.'a., Zakai, A. & Yu, B. (2006) Some theory for generalized boosting algorithms. *J. Mach. Learn. Res.*, **7**, 705–732.
17. Blanchard, G. & Krämer, N. (2016) Convergence rates of kernel conjugate gradient for random design regression. *Anal. Appl.*, **14**, 763–794.
18. Bubeck, S. (2015) Convex optimization: algorithms and complexity. *Foundations and trends®. Mach. Learn.*, **8**, 231–357.
19. Bubeck, S., Cohen, M. B., Lee, Y. T., Lee, J. R. & Mądry, A. (2018) K-server via multiscale entropic regularization. In Diakonikolas, I., Kempe, D. & Henzinger, M. (eds) *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*. ACM, 3–16.
20. Bubeck, S., Cohen, M. B., Lee, J. R. & Lee, Y. T. (2019) Metrical task systems on trees via mirror descent and unfair gluing. In Diakonikolas, I., Chan, T. M (ed) *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp 89–97.
21. Bühlmann, P. & Bin, Y. (2003) Boosting with the l2 loss: regression and classification. *J. Am. Stat. Assoc.*, **98**, 324–339.
22. Chen, Y., Jin, C. & Yu, B. (2018) Stability and convergence trade-off of iterative optimization algorithms. arXiv preprint arXiv:1804.01619.
23. Engl, H. W., Hanke, M. & Neubauer, A. (1996) *Regularization of inverse problems*, vol. **375**. Springer Dordrecht.
24. Ghai, U., Hazan, E. & Singer, Y. (2019) Exponentiated gradient meets gradient descent. arXiv preprint arXiv:1902.01903.
25. Gidel, G., Bach, F. & Lacoste-Julien, S. (2019) Implicit regularization of discrete gradient dynamics in deep linear neural networks. arXiv preprint arXiv:1904.13262.
26. Gunasekar, S., Woodworth, B. E, Bhojanapalli, S., Neyshabur, B. & Srebro, N. (2017) Implicit regularization in matrix factorization. In. *Advances in Neural Information Processing Systems*, **30**, 6151–6159.
27. Gunasekar, S., Lee, J., Soudry, D. & Srebro, N. (20189) Characterizing implicit bias in terms of optimization geometry. In Dy, J. G. & Krause, A. (eds) *Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research*. Stockholmsmässan, Stockholm Sweden: PMLR, pp 1832–1841.
28. Hazan, E. (2016) Introduction to online convex optimization. *Found. Trends Optim.*, **2**, 157–325.
29. Jiang, W. (2004) Process consistency for adaboost. *Ann. Stat.*, **32**, 13–29.
30. Kanade, V., Rebeschini, P. & Vaškevičius, T. (2022) Exponential tail local rademacher complexity risk bounds without the bernstein condition. arXiv:2202.11461.
31. Kivinen, J. & Warmuth, M. K. (1997) Exponentiated gradient versus gradient descent for linear predictors. *Inform. omput.*, **132**, 1–63.
32. Koltchinskii, V. (2006) Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Stat.*, **34**, 2593–2656.
33. Landweber, L. (1951) An iteration formula for Fredholm integral equations of the first kind. *Am. J. Math.*, **73**, 615–624.
34. Lecué, G., Mitchell, C., et al. (2012) Oracle inequalities for cross-validation type procedures. *Electron. J. Stat.*, **6**, 1803–1837.
35. Li, Y., Ma, T. & Zhang, H. (2018) Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *Conf. Learn. Theory*, 2–47.
36. Liang, T., Rakhlin, A. & Sridharan, K. (2015) Learning with square loss: localization through offset Rademacher complexity. *Conf. Learn. Theory*, 1260–1285.
37. Lin, J. & Cevher, V. (2018) Optimal distributed learning with multi-pass stochastic gradient methods. In Dy, J. G. & Krause, A. (eds) *Proceedings of the 35th International Conference on Machine Learning, number CONF*. PML.

38. Lin, J., Camoriano, R. & Rosasco, L. (2016a) Generalization properties and implicit regularization for multiple passes sgm. *Int. Conf. Mach. Learn.*, 2340–2348.
39. Lin, J., Rosasco, L. & Zhou, D.-X. (2016b) Iterative regularization for learning with convex loss functions. *J. Mach. Learn. Res.*, **17**, 2718–2755.
40. Lugosi, G. & Mendelson, S. (2020) Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc*, **22**, 925–965.
41. Matet, S., Rosasco, L., Villa, S. & Vu, B. L. (2017) Don't relax: early stopping for convex regularization. arXiv preprint arXiv:1707.05422.
42. Mendelson, S. (2014) Learning without concentration. *Conf. Learn. Theory*, **62**, 1–25.
43. Mendelson, S. (2020) Extending the scope of the small-ball method. *Studia Math.*, **256**, 147–167.
44. Nemirovsky, A. & Yudin, D. (1983) *Problem complexity and method efficiency in optimization*. New York: Wiley.
45. Neu, G. & Rosasco, L. (2018) Iterate averaging as regularization for stochastic gradient descent. arXiv preprint arXiv:1802.08009.
46. Nielsen, F., Boissonnat, J.-D. & Nock, R. (2007) Bregman voronoi diagrams: properties, algorithms and applications. arXiv preprint arXiv:0709.2196.
47. Pagliana, N. & Rosasco, L. (2019) Implicit regularization of accelerated methods in hilbert spaces. *Adv. Neural Inf. Process. Syst.*, **32**, 14454–14464.
48. Prechelt, L. (1998) Early stopping-but when? In Orr, G. B. & Müller, K.-R. (eds) *Neural Networks: Tricks of the trade*. Springer, pp. 55–69.
49. Rakhlin, A. & Sridharan, K. (2014) Online non-parametric regression. In Balcan, M.-F., Feldman, V. & Szepesvári, C. (eds) *Conference on Learning Theory*, JMLR, 1232–1264.
50. Raskutti, G., Wainwright, M. J. & Bin, Y. (2011) Minimax rates of estimation for high-dimensional linear regression over l_q balls. *IEEE Trans. Inf. Theory*, **57**, 6976–6994.
51. Raskutti, G., Wainwright, M. J. & Bin, Y. (2014) Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The. J. Mach. Learn. Res.*, **15**, 335–366.
52. Richards, D. & Rebeschini, P. (2019) Optimal statistical rates for decentralised non-parametric regression with linear speed-up. *Adv. Neural Inf. Process. Syst.*, **32**, 1214–1225.
53. Richards, D. & Rebeschini, P. (2020) Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *J. Mach. Learn. Res.*, **21**, 1–44.
54. Robbins, H. & Monro, S. (1951) A stochastic approximation method. *Ann. Math. Stat.*, **22**, 400–407.
55. Rosasco, L. & Villa, S. (2015) Learning with incremental iterative regularization. *Adv. Neural Inf. Process. Syst.*, **28**, 1630–1638.
56. Scholkopf, B. & Smola, A. J. (2001) *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, Mass.: MIT press.
57. Shamir, O. (2015) The sample complexity of learning linear predictors with the squared loss. *J. Mach. Learn. Res.*, **16**, 3475–3486.
58. Suggala, A., Prasad, A. & Ravikumar, P. K. (2018) Connecting optimization and regularization paths. *Adv. Neural Inf. Process. Syst.*, **31**, 10608–10619.
59. Vaškevičius, T., Kanade, V. & Rebeschini, P. (2019) Implicit regularization for optimal sparse recovery. *Adv. Neural Inf. Process. Syst.*, **32**, 2968–2979.
60. Vaškevičius, T., Kanade, V. & Rebeschini, P. (2020) The statistical complexity of early-stopped mirror descent. *Adv. Neural Inf. Process. Syst.*, **33**, 253–264.
61. Vershynin, R. (2010) Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027.
62. Wainwright, M. J. (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, vol. **48**. Cambridge, England: Cambridge University Press.
63. Wegkamp, M. (2003) Model selection in nonparametric regression. *Ann. Stat.*, **31**, 252–273.
64. Wei, Y., Yang, F. & Wainwright, M. J. (2019) Early stopping for kernel boosting algorithms: a general analysis with localized complexities. *IEEE Trans. Inf. Theory*, **65**, 6685–6703.
65. Woodworth, B., Gunasekar, S., Lee, J., Soudry, D. & Srebro, N. (2019) Kernel and deep regimes in overparametrized models. arXiv preprint arXiv:1906.05827.

66. Yang, J., Sun, S. & Roy, D. M. (2019) Fast-rate pac-bayes generalization bounds via shifted rademacher processes. *Adv. Neural Inf. Process. Syst.*, **32**, 10803–10813.
67. Yao, Y., Rosasco, L. & Caponnetto, A. (2007) On early stopping in gradient descent learning. *Constr. Approx.*, **26**, 289–315.
68. Zhang, T. & Bin, Y. (2005) Boosting with early stopping: convergence and consistency. *Ann. Stat.*, **33**, 1538–1579.
69. Zhao, P., Yang, Y. & He, Q.-C. (2022) High-dimensional linear regression via implicit regularization. *Biometrika*, Oxford University Press, **109**, 1033–1046.
70. Zhivotovskiy, N. & Hanneke, S. (2018) Localization of vc classes: beyond local rademacher complexities. *Theor. Comput. Sci.*, **742**, 27–49.

A. Table of Notation

Table A1 *Table of notation*

Symbol	Description
n	The number of data points.
P	The data-generating distribution supported on $\mathcal{X} \times \mathcal{Y}$.
(x_i, y_i)	The i^{th} data point sampled independently from the distribution P.
D_n	A collection of n data points $(x_i, y_i)_{i=1}^n$ sampled i.i.d. from P.
D_n^x	A collection of n input points $(x_i)_{i=1}^n$, where $D_n = (x_i, y_i)_{i=1}^n$.
ℓ	The loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$.
$R(g)$	The population risk of a function g defined as $\mathbb{E}_{(X,Y)\sim P}[\,\ell(g(X), Y)]$.
$R_n(g)$	The empirical risk of a function g defined as $\frac{1}{n}\sum_{i=1}^n \ell(g(x_i), y_i)$.
$\widehat{g}$	An estimator, which maps datasets D_n to some set of functions $\mathcal{G}$.
$\mathcal{G}$	The set of possible values of functions, which some estimator $\widehat{g}$ can select.
$\mathcal{F}$	Some generic class of functions.
$\mathrm{star}(\mathcal{F})$	A star hull around 0 of $\mathcal{F}$, defined by $\{\lambda f : f \in \mathcal{F}, \lambda \in [0, 1]\}$.
$\mathcal{E}(\widehat{g}, \mathcal{F})$	The excess risk $R(\widehat{g}) - \inf_{g\in\mathcal{F}} R(g)$ of an estimator $\widehat{g}$ with respect to $\mathcal{F}$.
$g_{\mathcal{F}}$	A function $g \in \mathcal{F}$ such that $R(g) = \inf_{g\in\mathcal{F}} R(g)$.
$\|g - f\|_P^2$	Population ℓ_2 distance between g and f defined as $\mathbb{E}[(g(X) - f(X))^2]$
$\|g - f\|_n^2$	Empirical ℓ_2 distance between g and f defined as $\frac{1}{n}\sum_{i=1}^n (g(x_i) - f(x_i))^2$.
$\mathfrak{R}_{D_n^x}(\mathcal{G}, c)$	The offset Rademacher complexity of $\mathcal{G}$ (cf. Equation (1.2)).
m	The dimensionality of the parameter space.
g_α	A function parameterized by $\alpha \in \mathbb{R}^m$.
$Z \in \mathbb{R}^{n\times m}$	A matrix such that conditionally on D_n, $g_\alpha(x_i) = (Z\alpha)_i$ for any $\alpha \in \mathbb{R}^m$.
$R_n(\alpha)$	A shorthand notation for $R_n(g_\alpha)$.
ψ	A mirror map.
D_ψ	Bregman divergence induced by the mirror map ψ.
α_0	The initialization point of the mirror descent iterates.
α_t	The mirror descent iterate at time t.
α'	An arbitrarily chosen reference point.
δ_t	A shorthand notation for $R_n(\alpha_t) - R_n(\alpha')$.
r_t	A shorthand notation for $\frac{\gamma}{2}\|g_{\alpha_t} - g_{\alpha'}\|_n^2$.