The role of optimization geometry in single neuron learning

Nicholas M. Boffi
Courant Institute of Mathematical Sciences, New York University

Stephen Tu
Google Brain Robotics

Jean-Jacques Slotine
Nonlinear Systems Laboratory, Massachusetts Institute of Technology

Abstract

Recent numerical experiments have demonstrated that the choice of optimization geometry used during training can impact generalization performance when learning expressive nonlinear model classes such as deep neural networks. These observations have important implications for modern deep learning, but remain poorly understood due to the difficulty of the associated nonconvex optimization. Towards an understanding of this phenomenon, we analyze a family of pseudogradient methods for learning generalized linear models under the square loss – a simplified problem containing both nonlinearity in the model parameters and nonconvexity of the optimization which admits a single neuron as a special case. We prove nonasymptotic bounds on the generalization error that sharply characterize how the interplay between the optimization geometry and the feature space geometry sets the out-of-sample performance of the learned model. Experimentally, selecting the optimization geometry as suggested by our theory leads to improved performance in generalized linear model estimation problems such as nonlinear and nonconvex variants of sparse vector recovery and low-rank matrix sensing.

1 Introduction

Optimization geometry, whereby the loss gradient is computed with respect to a non-Euclidean metric, is a common tool in modern machine learning. Notable examples of algorithms that use non-Euclidean metrics to improve convergence include mirror descent (Beck and Teboulle, 2003; Nemirovski and Yudin, 1983; Krichene et al., 2015), natural gradient descent (Amari, 1998; Gunasekar et al., 2021), and adaptive gradient methods such as AdaGrad (Duchi et al., 2011) and Adam (Kingma and Ba, 2015).

Recently, Azizan et al. (2021) showed empirically that varying the optimization geometry through choice of a mirror descent potential can improve the generalization performance of expressive model classes such as deep neural networks, but a theoretical characterization of when and why this will occur is currently absent. One path towards explaining these observations could be to consider the effect of mirror descent in linear regression. However, nonconvexity of the optimization problem and nonlinearity in the parameters are hallmarks of deep networks, and it is not clear that conclusions about linear models lacking these properties will carry over to deep learning.

The simplest model class with both nonlinearity and nonconvexity is the class of generalized linear models (GLMs). GLMs extend linear models by incorporating a single "layer" of nonlinearity in the parameters (McCullagh and Nelder, 1989), wherein the dependent variables are assumed to be given as a known nonlinear activation of a linear predictor of the data. From a modern perspective, a GLM represents a single neuron, and as such can be seen as the simplest model of a neural network that allows for rigorous analysis yet still contains both nonconvexity and nonlinearity. As a result, guarantees for GLMs may provide insight into the theoretical properties of more complex models, an observation that has been exploited by several recent works (Maillard et al., 2021; Barbier et al., 2019; Frei et al., 2020).

In this work, we revisit the GLM-tron of Kakade et al. (2011). Leveraging recent developments in continuous-time optimization and adaptive control theory (Boffi and Slotine, 2021), we extend the continuous-time limit of the GLM-tron iteration to a mirror descent-like flow that we call the Reflectron. The Reflectron is specified by the choice of a pseudogradient ξ and a

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

strongly convex potential function ψ; these two ingredients together define a search direction and a search geometry. As particular cases, the Reflectron recovers both the GLM-tron and gradient descent.

Outline of results We first prove non-asymptotic generalization error bounds for mirror descent applied to a GLM in the full-batch setting. Our results highlight how the choice of ψ can improve the *statistical performance* of the model if selected in a way that respects the underlying geometric structure of the feature space. In the realizable setting, we further characterize the learned parameters as minimizing the Bregman divergence under ψ between the initialization and the interpolating manifold.

We then discretize the continuous dynamics via the forward-Euler method to obtain an implementable algorithm, and we show that the corresponding iteration enjoys guarantees that match those of the continuous-time flow. We further consider a stochastic gradient-like setting for learning GLMs in the realizable and bounded noise settings, and prove $\mathcal{O}(1/t)$ and $\mathcal{O}\left(1/\sqrt{t}\right)$ bounds for the generalization error, respectively. We conclude with experiments highlighting the ability of our theoretical results to capture the importance of optimization geometry in practice. By choosing the mirror descent potential as suggested by our analysis in nonlinear and nonconvex sparse vector and low-rank matrix recovery problems – both of which amount to GLM estimation under the square loss – we demonstrate improved generalization performance of the learned model.

1.1 Related work and significance

Applications of the GLM-Tron The GLM-tron of Kakade et al. (2011) was the first known computationally and statistically efficient algorithm for learning both GLMs and Single Index Models (SIMs). A recent extension known as the BregmanTron (Nock and Menon, 2020) obtains improved guarantees for the SIM problem by applying Bregman divergences to directly learn the loss function; here, we instead focus on the GLM-tron as a primitive that allows us to analyze the role of optimization geometry in learning nonlinear models. Frei et al. (2020) use similar proof techniques to Kakade et al. (2011) to analyze gradient descent on the square loss for learning a single neuron. Our work extends their results to the mirror descent and pseudogradient settings, and characterizes the role of optimization geometry in generalization. Foster et al. (2020) utilize the GLM-tron for system identification in a particular nonlinear discrete-time dynamics model, and Goel and Klivans (2019) use a kernelized GLM-tron to provably learn two-hidden-layer neural networks. Similar update laws have independently been developed in the adaptive control literature (Tyukin et al., 2007), along with mirror descent and momentum variants (Boffi and Slotine, 2021).

Implicit bias and generalization Modern machine learning frequently takes place in a high-dimensional regime with many more parameters than examples. It is now well-known that deep networks will interpolate noisy data, yet exhibit low generalization error *despite interpolation* when trained on meaningful data (Zhang et al., 2017). Defying classical statistical wisdom, an explanation for this apparent paradox has been given in the *implicit bias* (Soudry et al., 2018) of optimization algorithms and the double descent curve (Belkin et al., 2019; Bartlett et al., 2020; Muthukumar et al., 2019; Hastie et al., 2019). The notion of implicit bias captures the proclivity of a method to converge to a particular kind of interpolating solution – such as minimum norm – when many options exist.

Implicit bias has been categorized for gradient-based algorithms on separable classification problems (Soudry et al., 2018; Nacson et al., 2018), regression problems (Gunasekar et al., 2018b), and multilayer models (Gunasekar et al., 2018a; Woodworth et al., 2020; Gunasekar et al., 2017). Approximate results are also available for the implicit bias of gradient-based algorithms when used to train deep networks (Azizan et al., 2019). Moreover, it was shown empirically (Azizan et al., 2021) that the choice of mirror descent potential affects the generalization error of deep networks, and a qualitative explanation was provided in terms of changing the specific form of implicit bias. Our focus on the GLM setting allows us to distinguish between the generalization performance of models trained with different potentials, which provides a quantitative and geometric explanation. While we focus on the square loss, a number of recent works have investigated the possibility of using the square loss for training deep networks for classification (Demirkaya et al., 2020; Hui and Belkin, 2021; Han et al., 2022).

2 Problem setting and background

Our problem setting follows the original work of Kakade et al. (2011). Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ denote an i.i.d. dataset sampled from a distribution $\mathcal{D}$ supported on $\mathcal{X} \times [0,1]$, $\mathcal{X} \subseteq \mathbb{R}^d$, where $\mathbb{E}[y_i|\mathbf{x}_i] = u(\langle \boldsymbol{\theta}, \mathbf{x}_i \rangle)$ for $\boldsymbol{\theta} \in \mathbb{R}^d$ a fixed, unknown vector of parameters. $u : \mathbb{R} \to [0,1]$ is assumed to be a known, nondecreasing, and L-Lipschitz activation function. Our goal is to approximate $\mathbb{E}[y_i|\mathbf{x}_i]$ as measured by the expected square loss. To this end, for a hypothesis $h : \mathbb{R}^d \to \mathbb{R}$, we define the generalization error $\mathrm{err}(h)$ and the excess

risk compared to the Bayes-optimal predictor $\varepsilon(h)$ as

$$\operatorname{err}(h) = \mathbb{E}_{\mathbf{x},y}\left[\left(h\left(\mathbf{x}\right) - y\right)^2\right], \tag{1}$$

$$\varepsilon(h) = \mathbb{E}_{\mathbf{x},y}\left[\left(h(\mathbf{x}) - u\left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle\right)\right)^2\right], \tag{2}$$

with $\widehat{\operatorname{err}}(h)$ and $\widehat{\varepsilon}(h)$ their empirical counterparts over the dataset. Towards minimizing $\operatorname{err}(h)$, we present a family of mirror descent-like algorithms for minimizing $\widehat{\varepsilon}(h)$ over parametric hypotheses of the form $h(\mathbf{x}) = u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}\right\rangle\right)$. Via standard statistical techniques (Bartlett and Mendelson, 2002), we transfer our guarantees on $\widehat{\varepsilon}(h)$ to $\varepsilon(h)$, which in turn implies a small $\operatorname{err}(h)$. The starting point of our analysis is the GLM-tron of Kakade et al. (2011), which is an iterative update law of the form

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t - \frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right\rangle\right) - y_i\right)\mathbf{x}_i, \tag{3}$$

with initialization $\widehat{\boldsymbol{\theta}}_1 = \mathbf{0}$. Equation (3) is a pseudogradient-based update law obtained from gradient descent on the square loss $\widehat{\operatorname{err}}(h)$ by replacing all appearances of u' by the fixed value 1. It admits a continuous-time limit for an infinitesimal step size,

$$\frac{d}{dt}\widehat{\boldsymbol{\theta}} = -\frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right\rangle\right) - y_i\right)\mathbf{x}_i, \tag{4}$$

where (3) is obtained from (4) via a forward-Euler discretization with a timestep $\Delta t = 1$.

Notation. Throughout this paper, we will use the notation $\frac{d}{dt}\mathbf{x} = \dot{\mathbf{x}}$ interchangeably for any time-dependent function $\mathbf{x}(t)$. Moreover, we will denote

$$\mathcal{R}_n\left(\mathcal{F}\right) = \mathbb{E}_{\mathbf{x}_i,\epsilon_i}\left[\sup_{h\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i h(\mathbf{x}_i)\right]$$

the Rademacher complexity of a function class $\mathcal{F}$ on n samples, and the shorthand $\zeta(h) = \max\{\varepsilon(h), \widehat{\varepsilon}(h)\}$ in our generalization error bounds.

3 Continuous-time theory

In this section, we analyze a continuous-time flow that we will discretize to obtain implementable algorithms in Section 4. Our continuous-time analysis sketches the essence of the techniques required to obtain discrete-time guarantees, and provides intuition for our main results while avoiding discretization-specific details. The class of algorithms we consider is captured by the dynamics

$$\frac{d}{dt}\nabla\psi\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right)\mathbf{x}_i. \tag{5}$$

for $\psi : \mathcal{M} \to \mathbb{R}$, $\mathcal{M} \subseteq \mathbb{R}^d$, and $\xi : \mathcal{M}\times\mathcal{X} \to \mathbb{R}$ with $\xi \geq 0$. To obtain guarantees on the algorithms represented by (5), we require two primary assumptions.

Assumption 3.1. $\psi : \mathcal{M} \to \mathbb{R}$ is σ-strongly convex with respect to a norm $\|\cdot\|$. Moreover, $\min_{\mathbf{w}\in\mathcal{M}}\psi(\mathbf{w}) = 0$.

Note that any ψ with finite minimum can be shifted to satisfy the final requirement of Assumption 3.1, as our algorithms only depend on gradients and Bregman divergences of ψ.

Assumption 3.2. The activation function $u : \mathbb{R} \to [0,1]$ is known, nondecreasing, and L-Lipschitz.

The parameters of the hypothesis h_t at time t are computed by applying the inverse gradient of ψ, which is guaranteed to exist by strong convexity. The mirror descent generalization of the GLM-tron is obtained from (5) by setting $\xi\left(\mathbf{w}, \mathbf{x}\right) = 1$, while mirror descent itself is obtained by setting $\xi\left(\mathbf{w}, \mathbf{x}\right) = u'\left(\langle \mathbf{w}, \mathbf{x}\rangle\right)$. In order to outline the intuition behind our results, we focus exclusively on the case when $\xi\left(\mathbf{w}, \mathbf{x}\right) = 1$ and defer the analysis with arbitrary ξ to discrete-time.

3.1 Statistical guarantees

The following theorem gives a statistical guarantee for the Reflectron in continuous-time. It shows that for any choice of potential function ψ, the Reflectron eventually finds a nearly Bayes-optimal predictor.

Theorem 3.1. *Suppose that $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are drawn i.i.d. from a distribution $\mathcal{D}$ supported on $\mathcal{X}\times[0,1]$ where $\mathbb{E}\left[y|\mathbf{x}\right] = u\left(\langle\boldsymbol{\theta},\mathbf{x}\rangle\right)$, u satisfies Assumption 3.2, and $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown vector of parameters. Let ψ satisfy Assumption 3.1. Assume that $\left\|\frac{1}{n}\sum_{i=1}^n \left(u\left(\langle\boldsymbol{\theta},\mathbf{x}_i\rangle\right) - y_i\right)\mathbf{x}_i\right\|_* \leq \eta$ where $\|\cdot\|_*$ denotes the dual norm to $\|\cdot\|$. Then for any $\delta \in (0,1)$, there exists some time $t < \sqrt{\frac{\psi(\boldsymbol{\theta})\sigma}{2\eta^2}}$ such that the hypothesis $h_t = u\left(\left\langle\widehat{\boldsymbol{\theta}}(t), \mathbf{x}\right\rangle\right)$ satisfies*

$$\zeta\left(h_t\right) \leq \sqrt{\frac{8L^2\eta^2\psi(\boldsymbol{\theta})}{\sigma}} + 4\mathcal{R}_n\left(\mathcal{F}\right) + \sqrt{\frac{8\log(1/\delta)}{n}},$$

with probability at least $1-\delta$, where $\widehat{\boldsymbol{\theta}}(0) = \arg\min_{\mathbf{w}\in\mathcal{M}}\psi(\mathbf{w})$, and $\mathcal{F} = \{\mathbf{x}\mapsto\langle\mathbf{w},\mathbf{x}\rangle : \mathbf{w}\in\mathcal{M},\ d_\psi\left(\boldsymbol{\theta}\parallel\mathbf{w}\right) \leq \psi(\boldsymbol{\theta})\}$.

Proof. Consider the rate of change of the Bregman divergence between the parameters for the Bayes-optimal predictor $\boldsymbol{\theta}$ and the parameter estimates $\widehat{\boldsymbol{\theta}}(t)$,

$$\frac{d}{dt} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}\right) = \left\langle \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \nabla^2 \psi\left(\widehat{\boldsymbol{\theta}}\right) \dot{\widehat{\boldsymbol{\theta}}} \right\rangle .$$

Observe that $\frac{d}{dt} \nabla \psi\left(\widehat{\boldsymbol{\theta}}\right) = \nabla^2 \psi\left(\widehat{\boldsymbol{\theta}}\right) \dot{\widehat{\boldsymbol{\theta}}}$, so that

$$\begin{aligned} \frac{d}{dt} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}\right) &= \frac{1}{n} \sum_{i=1}^n \left(y_i - u\left(\langle \mathbf{x}_i, \boldsymbol{\theta} \rangle\right)\right) \left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\rangle \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left(u\left(\langle \mathbf{x}_i, \boldsymbol{\theta} \rangle\right) - u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}} \right\rangle\right)\right) \left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\rangle . \end{aligned}$$

Using that u is L-Lipschitz and nondecreasing, we may upper bound the second term by $-\frac{1}{L}\widehat{\varepsilon}(h_t)$,

$$\begin{aligned} &\frac{d}{dt} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}\right) \\ &\leq \frac{1}{n} \sum_{i=1}^n \left(y_i - u\left(\langle \mathbf{x}_i, \boldsymbol{\theta} \rangle\right)\right) \left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\rangle - \frac{1}{L}\widehat{\varepsilon}(h_t) . \quad (6) \end{aligned}$$

By assumption, $\left\| \frac{1}{n} \sum_{i=1}^n \left(y_i - u(\langle \mathbf{x}_i, \boldsymbol{\theta} \rangle)\right) \mathbf{x}_i \right\|_* \leq \eta$. Now, observe that by strong convexity of ψ and by the initialization,

$$\left\| \widehat{\boldsymbol{\theta}}(0) - \boldsymbol{\theta} \right\| \leq \sqrt{\frac{2 d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right)}{\sigma}} \leq \sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}} .$$

By induction, assume that $d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}(t)\right) \leq \psi(\boldsymbol{\theta})$ at time t. Then we have the bound

$$\frac{d}{dt} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}\right) \leq -\frac{1}{L}\widehat{\varepsilon}(h_t) + \eta \sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}} ,$$

so that either $\frac{d}{dt} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}\right) < -\eta\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$ or $\widehat{\varepsilon}(h_t) \leq 2L\eta\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$. In the latter case, we have obtained the desired bound on $\widehat{\varepsilon}(h_t)$. Otherwise, t cannot exceed

$$t_f = \frac{d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right)}{\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}\eta} = \sqrt{\frac{\psi(\boldsymbol{\theta})\sigma}{2\eta^2}}$$

to satisfy $\widehat{\varepsilon}(h_t) \leq 2L\eta\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$. Hence there is some h_t with $t < t_f$ such that $\widehat{\varepsilon}(h_t) \leq 2L\eta\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$. To transfer this bound on $\widehat{\varepsilon}$ to ε, we need to bound $|\widehat{\varepsilon}(h_t) - \varepsilon(h_t)|$. Application of a standard uniform convergence result (cf. Theorem B.3) to the square loss[1] implies

$$|\widehat{\varepsilon}(h_t) - \varepsilon(h_t)| \leq 4\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{8\log(1/\delta)}{n}}$$

with probability at least $1 - \delta$. □

[1] Note that while the square loss is neither bounded nor Lipschitz in general, it is both over the domain $[0, 1]$ with bound $b = 1$ and Lipschitz constant $L' = 1$.

Because $\varepsilon(h_t) = \mathrm{err}(h_t)$ up to a constant, we can find a good predictor by using a hold-out set to estimate $\mathrm{err}(h_t)$ throughout learning.

The statement of Theorem 3.1 uses a specific initialization strategy to write the generalization error bound in terms of $\psi(\boldsymbol{\theta})$; with an arbitrary initialization, $\psi(\boldsymbol{\theta})$ can be replaced by $d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right)$, and our definition of $\mathcal{F}$ can be modified accordingly. As the bound depends on $\psi(\boldsymbol{\theta})$, C, and η, the potential ψ may be chosen in correspondence with available knowledge on the problem structure to optimize the guarantee on the generalization error. In Corollaries 4.1-4.3, we provide explicit illustrations of this fact. In the experiments in Section 6, we show how this can be used for improved estimation over the GLM-tron in problems such as sparse vector and low-rank matrix recovery.

Our proof of Theorem 3.1 is similar to the corresponding proof for the GLM-tron (Kakade et al., 2011), but has two primary modifications. First, we consider the Bregman divergence under ψ between the Bayes-optimal parameters and the current parameter estimates, rather than the squared Euclidean distance. Our use of Bregman divergence critically relies on the Bayes-optimal parameters appearing in the first argument. Second, rather than analyzing the *iteration* on $\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|_2^2$ as in the discrete-time case, we analyze the *dynamics* of the Bregman divergence. Taking $\psi = \frac{1}{2}\|\cdot\|_2^2$ recovers the guarantee of the GLM-tron up to forward Euler discretization-specific details.

3.2 Implicit regularization

We now study how the choice of ψ impacts the model learned by (5). To do so, we require a realizability assumption on the dataset.

Assumption 3.3. There exists a fixed $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $y_i = u\left(\langle \boldsymbol{\theta}, \mathbf{x}_i \rangle\right)$ for all $i = 1, \ldots, n$.

In many cases, even the noisy dataset of Section 3.1 may satisfy Assumption 3.3 for a $\bar{\boldsymbol{\theta}} \neq \boldsymbol{\theta}$. We now begin by proving convergence of the training error.

Lemma 3.1. *Suppose that $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are drawn i.i.d. from a distribution $\mathcal{D}$ supported on $\mathcal{X} \times [0, 1]$. Let the dataset satisfy Assumption 3.3, let u satisfy Assumption 3.2, and let ψ satisfy Assumption 3.1. Suppose $\|\mathbf{x}_i\|_* \leq C$. Then $\widehat{\varepsilon}(h_t) \to 0$ where $h_t(\mathbf{x}) = u\left(\left\langle \widehat{\boldsymbol{\theta}}(t), \mathbf{x} \right\rangle\right)$ and $\widehat{\boldsymbol{\theta}}(0) = \arg\min_{\mathbf{w} \in \mathcal{M}} \psi(\mathbf{w})$. Furthermore, $\min_{t' \in [0,t]} \{\widehat{\varepsilon}(h_{t'})\} \leq \mathcal{O}(1/t)$.*

Proof. Under the assumptions, (6) implies

$$\frac{d}{dt} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}\right) \leq -\frac{1}{L}\widehat{\varepsilon}(h_t) \leq 0 .$$

Integrating both sides of the above gives the bound

$$\int_0^t \widehat{\varepsilon}(h_{t'})dt' \leq L d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right).$$

Explicit computation shows that $\frac{d}{dt}\widehat{\varepsilon}(h_t)$ is bounded, so that $\widehat{\varepsilon}(h_t)$ is uniformly continuous in t. By Barbalat's Lemma (cf. Lemma B.1), this implies that $\widehat{\varepsilon} \to 0$ as $t \to \infty$. Now, note that

$$\begin{aligned}\inf_{t' \in [0,t]} \{\widehat{\varepsilon}(h_{t'})\}\, t &= \int_0^t \inf_{t' \in [0,t]} \{\widehat{\varepsilon}(h_{t'})\}\, dt'' \\ &\leq \int_0^t \widehat{\varepsilon}(h_{t'})dt' \leq L d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right),\end{aligned}$$

so that $\inf_{t' \in [0,t]} \{\widehat{\varepsilon}(h_{t'})\} \leq \frac{L d_\psi(\boldsymbol{\theta} \,\|\, \widehat{\boldsymbol{\theta}}(0))}{t}$. □

Lemma 3.1 shows that (5) will converge to an interpolating solution for a realizable dataset, and that the best hypothesis up to time t does so at an $\mathcal{O}(1/t)$ rate; the proof is given in the appendix.

In general, there may be many possible vectors $\widehat{\boldsymbol{\theta}}$ consistent with the data. The following theorem provides insight into the parameters learned by (5). Our result is analogous to the characterization of the implicit bias of mirror descent due to Gunasekar et al. (2018a), and uses a continuous-time proof technique inspired by the discrete-time technique in Azizan et al. (2019). A similar continuous-time proof first appeared in Boffi and Slotine (2021) in the context of adaptive control.

Theorem 3.2. *Consider the setting of Lemma 3.1. Let $\mathcal{A} = \{\bar{\boldsymbol{\theta}} \in \mathcal{M} : u\left(\langle \bar{\boldsymbol{\theta}}, \mathbf{x}_i\rangle\right) = y_i, i = 1, \dots, n\}$ be the set of parameters that interpolate the data, and assume that $\widehat{\boldsymbol{\theta}}(t) \to \widehat{\boldsymbol{\theta}}_\infty \in \mathcal{A}$. Further assume that $u(\cdot)$ is invertible. Then $\widehat{\boldsymbol{\theta}}_\infty = \arg\min_{\bar{\boldsymbol{\theta}} \in \mathcal{A}} d_\psi\left(\bar{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right)$. In particular, if $\widehat{\boldsymbol{\theta}}(0) = \arg\min_{\mathbf{w} \in \mathcal{M}} \psi(\mathbf{w})$, then $\widehat{\boldsymbol{\theta}}_\infty = \arg\min_{\bar{\boldsymbol{\theta}} \in \mathcal{A}} \psi(\bar{\boldsymbol{\theta}})$.*

Proof. Let $\bar{\boldsymbol{\theta}} \in \mathcal{A}$ be arbitrary. Define the error on example i as $\tilde{y}_i\left(\widehat{\boldsymbol{\theta}}(t)\right) = \left(u\left(\left\langle \widehat{\boldsymbol{\theta}}(t), \mathbf{x}_i\right\rangle\right) - y_i\right)$. Then,

$$\begin{aligned}&\frac{d}{dt} d_\psi\left(\bar{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}(t)\right) \\ &= -\frac{1}{n}\sum_{i=1}^n \tilde{y}_i\left(\widehat{\boldsymbol{\theta}}(t)\right)\left\langle \widehat{\boldsymbol{\theta}}(t) - \bar{\boldsymbol{\theta}}, \mathbf{x}_i\right\rangle, \\ &= -\frac{1}{n}\sum_{i=1}^n \tilde{y}_i\left(\widehat{\boldsymbol{\theta}}(t)\right)\left(\left\langle \widehat{\boldsymbol{\theta}}(t), \mathbf{x}_i\right\rangle - u^{-1}(y_i)\right).\end{aligned}$$

Above, we used that $\bar{\boldsymbol{\theta}} \in \mathcal{A}$ and that $u(\cdot)$ is invertible, so that $u\left(\langle \bar{\boldsymbol{\theta}}, \mathbf{x}_i\rangle\right) = y_i$ implies that $\langle \bar{\boldsymbol{\theta}}, \mathbf{x}_i\rangle = u^{-1}(y_i)$. Integrating both sides of the above, we find that

$$\begin{aligned}&d_\psi\left(\bar{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}_\infty\right) - d_\psi\left(\bar{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right) = \\ &-\frac{1}{n}\sum_{i=1}^n \int_0^\infty \tilde{y}_i\left(\widehat{\boldsymbol{\theta}}(t)\right)\left(\left\langle \widehat{\boldsymbol{\theta}}(t), \mathbf{x}_i\right\rangle - u^{-1}(y_i)\right) dt.\end{aligned}$$

The above relation is true for any $\bar{\boldsymbol{\theta}} \in \mathcal{A}$. Furthermore, the integral on the right-hand side is independent of $\bar{\boldsymbol{\theta}}$. Hence the arg min of the two Bregman divergences must be equal, which shows that $\widehat{\boldsymbol{\theta}}_\infty = \arg\min_{\bar{\boldsymbol{\theta}} \in \mathcal{A}} d_\psi\left(\bar{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right)$. □

Theorem 3.2 elucidates the implicit bias of pseudogradient algorithms captured by (5). Out of all possible interpolating parameters, (5) finds those that minimize the Bregman divergence between the set of interpolating parameters and the initialization.

4 Discrete-time algorithms

Equation (5) can be discretized via Forward-Euler to form an implementation with a step size $\lambda > 0$,

$$\begin{aligned}&\nabla\psi\left(\widehat{\boldsymbol{\phi}}_{t+1}\right) - \nabla\psi\left(\widehat{\boldsymbol{\theta}}_t\right) \\ &= -\frac{\lambda}{n}\sum_{i=1}^n \left(u\left(\left\langle \widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right)\mathbf{x}_i, \quad (7)\end{aligned}$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \Pi_{\mathcal{C}}^\psi\left(\widehat{\boldsymbol{\phi}}_{t+1}\right). \quad (8)$$

In (8), $\mathcal{C}$ denotes a convex constraint set and $\Pi_{\mathcal{C}}^\psi(\mathbf{z}) = \arg\min_{\mathbf{x} \in \mathcal{C} \cap \mathcal{M}} d_\psi(\mathbf{x} \,\|\, \mathbf{z})$ denotes the Bregman projection. To analyze the iteration (7) & (8), we need two assumptions on ξ.

Assumption 4.1 (Adapted from Frei et al. (2020))**.** For any $a > 0$ and $b > 0$, there exists a $\gamma > 0$ such that $\inf_{\|\mathbf{w}\| \leq a, \|\mathbf{x}\|_* \leq b} \xi(\mathbf{w}, \mathbf{x}) \geq \gamma > 0$.

For mirror descent, Assumption 4.1 reduces to a requirement that the derivative of the activation remains nonzero over any compact set.

Assumption 4.2. There exists a constant $B > 0$ such that $\xi(\mathbf{w}, \mathbf{x}) \leq B$ for all $\mathbf{w} \in \mathcal{M}$, $\mathbf{x} \in \mathcal{X}$.

For mirror descent, we take $B = L$, while for the mirror descent generalization of GLM-tron, we take $B = 1$. We may now state our statistical guarantees.

Theorem 4.1. *Suppose that $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are drawn i.i.d. from a distribution $\mathcal{D}$ supported on $\mathcal{X} \times [0,1]$ where $\mathbb{E}[y|\mathbf{x}] = u(\langle \boldsymbol{\theta}, \mathbf{x}\rangle)$, u satisfies Assumptions 3.2 & 4.1, and $\boldsymbol{\theta} \in \mathcal{C}$ is an unknown vector of parameters. Let ψ satisfy Assumption 3.2, and let ξ satisfy Assumptions 4.1 & 4.2. Assume that $\|\mathbf{x}_i\|_* \leq C$, $\|\boldsymbol{\theta}\| \leq W$, and*

$\left\|\frac{1}{n}\sum_{i=1}^n \left(u\left(\langle\boldsymbol{\theta},\mathbf{x}_i\rangle\right)-y_i\right)\xi\left(\langle\boldsymbol{\theta},\mathbf{x}_i\rangle\right)\mathbf{x}_i\right\|_* \le \eta$. *Let* γ *correspond to* $a=C$ *and* $b=W+\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$ *in Assumption 4.1. Then for* $\lambda \le \frac{\sigma}{2C^2BL}$ *there exists some iteration* $t < \frac{1}{\lambda}\sqrt{\frac{\sigma\psi(\boldsymbol{\theta})}{2\eta^2}}$ *such that* $h_t = u\left(\left\langle\widehat{\boldsymbol{\theta}}_t,\mathbf{x}\right\rangle\right)$ *satisfies with probability at least* $1-\delta$

$$\zeta(h_t) \le \sqrt{\frac{32L^2\eta^2\psi(\boldsymbol{\theta})}{\gamma^2\sigma}}\left(\frac{2C^2LB+1}{2C^2LB}\right) + 4\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{8\log(1/\delta)}{n}},$$

where $\widehat{\boldsymbol{\theta}}_1 = \arg\min_{\mathbf{w}\in\mathcal{C}\cap\mathcal{M}}\psi(\mathbf{w})$, *and where* $\mathcal{F} = \{\mathbf{x}\mapsto\langle\mathbf{w},\mathbf{x}\rangle : \mathbf{w}\in\mathcal{M},\, d_\psi(\boldsymbol{\theta}\,\|\,\mathbf{w})\le\psi(\boldsymbol{\theta})\}$.

Theorem 4.1 shows that, for a suitable choice of step size, the discrete-time iteration (7) & (8) preserves the guarantees of the continuous-time flow (5). The proof (and all subsequent proofs) are given in the appendix. We now state several consequences of Theorem 4.1 in standard settings that highlight the impact of the potential on generalization in nonconvex learning.

Corollary 4.1 (p/q dual norm pairs with $p\in[2,\infty)$)**.** *Let* $\|\cdot\|_* = \|\cdot\|_p$ *with* $p\in[2,\infty)$. *Then* $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_q^2$ *is* $(q-1)$*-strongly convex with respect to* $\|\cdot\|_q$ *where* $1/q+1/p=1$. *The generalization error is bounded as*

$$\zeta(h_t) \le \frac{4LWC}{q-1}\left(\frac{\sqrt{2\log(4/\delta)(q-1)}+1}{\sqrt{n}}\right)\frac{2C^2LB+1}{2C^2LB} + \frac{4CW}{\sqrt{n(q-1)}}\left(1+\frac{1}{\sqrt{q-1}}\right)+\sqrt{\frac{8\log(2/\delta)}{n}}.$$

Corollary 4.2 ($\infty/1$ dual norm pairs, global setup)**.** *Let* $\|\cdot\| = \|\cdot\|_1$ *and* $\|\cdot\|_* = \|\cdot\|_\infty$. *Then* $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_q^2$ *with* $q=\frac{\log(d)}{\log(d)-1}$ *is* $\frac{1}{3\log(d)}$*-strongly convex with respect to* $\|\cdot\|_1$. *Then, the generalization error can be bounded*

$$\zeta(h_t) \le \frac{4CW(1+\sqrt{3\log d})^2}{n^{1/2}} + \sqrt{\frac{8\log(2/\delta)}{n}} + \frac{12LCW\sqrt{3\log(d)}(2C^2LB+1)}{C^2LB}\sqrt{\frac{\log(4d/\delta)}{n}}.$$

Corollary 4.3 ($\infty/1$ dual norm pairs, simplex setup)**.** *Let* $\|\cdot\| = \|\cdot\|_1$ *and* $\|\cdot\|_* = \|\cdot\|_\infty$. *Take* $\psi(\mathbf{w}) = d_{KL}(\mathbf{w}\,\|\,\mathbf{u})$ *where* $\mathbf{u}$ *is the discrete uniform distribution in* d *dimensions and where* d_{KL} *denotes the KL divergence. Then* $\psi(\mathbf{w})$ *is* 1*-strongly convex with respect to* $\|\cdot\|_1$ *over the probability simplex and* $\psi(\mathbf{w})\le\log(d)$ *for any* $\mathbf{w}$. *Then,*

$$\zeta(h_t) \le 4C\sqrt{\frac{2\log d}{n}} + \sqrt{\frac{8\log(2/\delta)}{n}} + \frac{3LC\sqrt{32\log d}(2C^2LB+1)}{C^2LB}\sqrt{\frac{\log(4d/\delta)}{n}}.$$

In the above results, the dimensionality dependence of the generalization error has been reduced by judicious choice of ψ. In particular, Corollaries 4.2 and 4.3 are merely logarithmic in dimension, while a bound for the GLM-tron would be polynomial in dimension.

Similar to Theorem 4.1, we now show that Lemma 3.1 and Theorem 3.2 are preserved when discretizing (5). We first state a convergence guarantee.

Lemma 4.1. *Suppose that* $\{\mathbf{x}_i,y_i\}_{i=1}^n$ *are drawn i.i.d. from a distribution* $\mathcal{D}$ *supported on* $\mathcal{X}\times[0,1]$. *Let the dataset satisfy Assumption 3.3 let* u *satisfy Assumption 3.2, and let* ψ *satisfy Assumption 3.1. Suppose* $\|\mathbf{x}_i\|_*\le C$. *Then for* $\lambda\le\frac{2\sigma}{C^2BL}$, $\widehat{\varepsilon}(h_t)\to 0$ *where* $h_t(\mathbf{x}) = u\left(\left\langle\widehat{\boldsymbol{\theta}}(t),\mathbf{x}\right\rangle\right)$ *is the hypothesis with parameters output by (7) & (8) at time* t *with* $\widehat{\boldsymbol{\theta}}_1 = \arg\min_{\mathbf{w}\in\mathcal{C}\cap\mathcal{M}}\psi(\mathbf{w})$. *Furthermore,* $\min_{t'\in[0,t]}\{\widehat{\varepsilon}(h_{t'})\}\le\mathcal{O}(1/t)$.

We conclude by showing that the implicit bias properties of (7) & (8) match those of (5).

Theorem 4.2. *Consider the setting of Lemma 4.1, and assume that* $u(\cdot)$ *is invertible. Let* $\mathcal{A} = \{\bar{\boldsymbol{\theta}}\in\mathcal{C}\cap\mathcal{M} : u\left(\left\langle\bar{\boldsymbol{\theta}},\mathbf{x}_i\right\rangle\right) = y_i, i=1,\dots,n\}$ *be the set of parameters that interpolate the data, and assume that* $\widehat{\boldsymbol{\theta}}_t\to\widehat{\boldsymbol{\theta}}_\infty\in\mathcal{A}$. *Then* $\widehat{\boldsymbol{\theta}}_\infty = \arg\min_{\bar{\boldsymbol{\theta}}\in\mathcal{A}} d_\psi\left(\bar{\boldsymbol{\theta}}\,\middle\|\,\widehat{\boldsymbol{\theta}}_1\right)$. *In particular, if* $\widehat{\boldsymbol{\theta}}_1 = \arg\min_{\mathbf{w}\in\mathcal{C}\cap\mathcal{M}}\psi(\mathbf{w})$, *then* $\widehat{\boldsymbol{\theta}}_\infty = \arg\min_{\bar{\boldsymbol{\theta}}\in\mathcal{A}}\psi(\bar{\boldsymbol{\theta}})$.

Taken together, the results in this section show that the continuous guarantees are preserved by discretization, though the analysis requires care of higher-order terms that vanish in the continuous limit.

5 Stochastic optimization

In this section, we provide guarantees for the iteration

$$\nabla\psi\left(\widehat{\boldsymbol{\phi}}_{t+1}\right) - \nabla\psi\left(\widehat{\boldsymbol{\theta}}_t\right) = -\lambda\left(u\left(\left\langle\widehat{\boldsymbol{\theta}}_t,\mathbf{x}_t\right\rangle\right)-y_t\right)\xi\left(\widehat{\boldsymbol{\theta}}_t,\mathbf{x}_t\right)\mathbf{x}_t, \tag{9}$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \Pi_{\mathcal{C}}^{\psi}\left(\widehat{\boldsymbol{\phi}}_{t+1}\right), \tag{10}$$

which is similar to stochastic gradient descent. We first consider the bounded noise setting, where we conclude a $\mathcal{O}(1/\sqrt{t})$ convergence rate of the generalization error.

Theorem 5.1. *Suppose that* $\{\mathbf{x}_t,y_t\}_{t=1}^\infty$ *are drawn i.i.d. from a distribution* $\mathcal{D}$ *supported on* $\mathcal{X}\times[0,1]$ *where* $\mathbb{E}[y|\mathbf{x}] = u(\langle\boldsymbol{\theta},\mathbf{x}\rangle)$, $\boldsymbol{\theta}\in\mathcal{C}$ *is an unknown vector of parameters, and* u *satisfies Assumption 3.2. Assume that* $\mathcal{C}$ *is compact, and let* $R=\mathsf{Diam}(\mathcal{C})$ *as measured in the norm* $\|\cdot\|$. *Suppose that* ψ *satisfies Assumption 3.1, and that* $\|\mathbf{x}_t\|_*\le C$ *for all* t. *Fix a*

horizon T, and choose $\lambda < \min\left\{\frac{2\sigma}{C^2LB}, \frac{1}{\sqrt{T}}\right\}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\min_{t<T} \varepsilon(h_t) \leq \mathcal{O}\left(\frac{L}{\sqrt{T}\gamma}\left(\psi(\boldsymbol{\theta}) + \sqrt{CBR\log(6/\delta)} + \frac{C^2B^2}{\sigma}\right)\right)$$

where h_t is the hypothesis output by (9) & (10) at iteration t with $\widehat{\boldsymbol{\theta}}_1 = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \psi(\boldsymbol{\theta})$, and γ corresponds to $a = C$ and $b = R + \|\boldsymbol{\theta}\|$ in Assumption 4.1.

We now consider the realizable setting, where we obtain fast $\mathcal{O}(1/t)$ rates.

Theorem 5.2. *Suppose that $\{\mathbf{x}_t, y_t\}_{t=1}^{\infty}$ are drawn i.i.d. from a distribution $\mathcal{D}$ supported on $\mathcal{X}$. Let Assumption 3.3 be satisfied with $\boldsymbol{\theta} \in \mathcal{C}$ an unknown vector of parameters, let u satisfy Assumption 3.2, let ψ satisfy Assumption 3.1, and assume that $\|\mathbf{x}_t\|_* \leq C$ for all t. Fix $\lambda < \frac{2\sigma}{LC^2B}$. Then for any $\delta \in (0,1)$, for all $T \geq 1$, with probability at least $1-\delta$*

$$\min_{t<T} \varepsilon(h_t) \leq \mathcal{O}\left(\frac{L^2C^2B\psi(\boldsymbol{\theta})\log(1/\delta)}{\sigma T\gamma}\right),$$

where h_t is the hypothesis output by (9) & (10) at iteration t and where γ corresponds to $a = C$ and $b = \|\boldsymbol{\theta}\| + \sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$ in Assumption 4.1.

6 Experiments

We now illustrate our theoretical results in two concrete problem settings. We first study a scalar-valued output problem where the Bayes-optimal parameter vector is sparse. We then consider a vector-valued system identification problem where the Bayes-optimal parameter matrix is low-rank.

We compare three variants of the Reflectron with three different choices of potential. The first is the GLM-tron, which is equivalent to the use of the Euclidean potential $\psi_2(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$. The second is the p-norm algorithm (Gentile, 2003), which uses the potential $\psi_p(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_p^2$ for $p \in [1,\infty]$. The third variant is the hypentropy algorithm (Ghai et al., 2020), which generalizes the setup considered in Corollary 4.3 beyond the probability simplex and uses the potential $\psi_\beta(\mathbf{x}) = \sum_{i=1}^d (\mathbf{x}_i \operatorname{arcsinh}(\mathbf{x}_i/\beta) - \sqrt{\mathbf{x}_i^2 + \beta^2})$ for $\beta \in (0,\infty)$. A complete description of the experimental setup is given in the appendix.

6.1 Sparse vector GLMs

In this setting, the learner receives measurements $y_i = \sigma(\langle\boldsymbol{\theta}, \mathbf{x}_i\rangle) + w_i$ with $\mathbf{x}_i \sim \mathrm{Unif}([-1,1]^d)$, $w_i \sim \mathrm{Unif}([-\sigma_w, \sigma_w]^d)$, and where $\sigma(\cdot)$ is the sigmoid activation. $\boldsymbol{\theta}$ is taken to be an s-sparse vector with $s \ll d$, and we compare the GLM-tron with *explicit* ℓ_1 projection to the Reflectron with matching *explicit* ℓ_1 projection and the *implicit* regularization due to the either the hypentropy or p-norm potentials.

The learner has knowledge that $\boldsymbol{\theta}$ is sparse, as well as access to the upper bounds $\|\boldsymbol{\theta}\|_p \leq W_p$. In the experiments, we set $W_p = 2\|\boldsymbol{\theta}\|_p$. Let $\mathbb{B}_p(r)$ denote the closed ℓ_p-ball in $\mathbb{R}^d$ of radius r centered at the origin. For the GLM-tron, we set $\mathcal{C} = \mathbb{B}_1(W_1)$ and the projection onto $\mathcal{C}$ is Euclidean. For the p-norm algorithm, we set $\mathcal{C} = \mathbb{B}_p(W_p)$ and apply a Bregman projection onto $\mathcal{C}$. For hypentropy, we set $\mathcal{C} = \mathbb{B}_1(W_1)$ and again use the Bregman projection onto $\mathcal{C}$. We compare each algorithm in two experimental regimes.

In the first regime, the ambient dimension $d = 10000$ and sparsity $s = 100$ are fixed, and we study the performance of each algorithm as a function of the number of data points n. For each pair of (n, alg), we run the full-batch pseudogradient algorithm for 5000 iterations over a grid of hyperparameters, and we tune the step size λ and the p value for the p-norm algorithm (resp. β for hypentropy). As suggested by Theorem 4.1, we use a holdout set of size $n_{\text{hold}} = 500$ to select the parameters with lowest validation error over 5000 iterations. Each algorithm is run for 5 trials and the configuration that achieves the lowest median test error (over the 5 trials) is shown in the figures. The size of the test set is $n_{\text{test}} = 1000$, and the error bars correspond to the min/median/max over the 5 trials.

Figure 1a shows the training error and holdout error of the best configuration for each algorithm with $n = 1000$. Each algorithm overfits, and the holdout set is necessary to find the predictor with lowest generalization error. Figure 1b shows the resulting test error of each algorithm. For each value of n, both the p-norm and hypentropy algorithms have lower test error when compared to the GLM-tron, in line with the generalization error predictions of Theorem 4.1.

In the second regime, the number of data points is fixed at $n = 1000$ while the ambient dimension d is varied for fixed $s/d = 0.01$. Figure 1c shows the test error for each algorithm, which increases with the ambient dimension d in all cases. As in Figure 1b, for fixed d, both the p-norm and hypentropy algorithms have lower test error than the GLM-tron. Taken together, Figures 1(a)-(c) validate the claims of Theorem 4.1.

To verify the predictions of Theorem 4.2, we remove the explicit projection onto $\mathcal{C}$ for both the GLM-tron and hypentropy and visualize the structure of the learned parameter vector in Figure 1d ($d = 1000$, $s = 10$, and $n = 1000$). Figure 1d shows that hypentropy recovers a much sparser solution than the GLM-tron despite the lack of an explicit projection

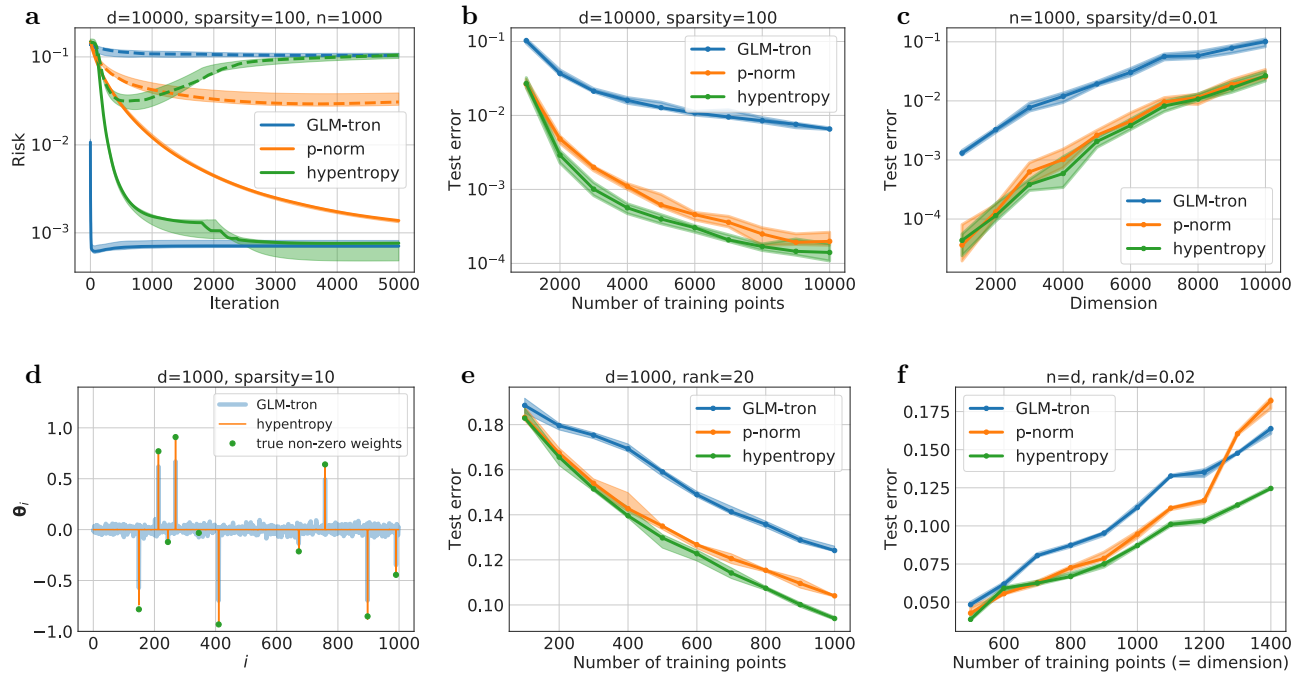

Figure 1: (a) Risk curves. Solid indicates training error and dashed indicates holdout error. (b) Test error with d and s fixed as n varies. (c) Test error with n and s/d fixed as d varies. (d) Weights learned without explicit projection. (e) Test error with d and r fixed as n varies. (f) Test error with fixed r/d as $n = d$ varies.

onto the ℓ_1-ball. In particular, 971 coordinates have absolute value greater than 0.001 for the parameters found by the GLM-tron, while there are only 56 for hypentropy. Moreover, the qualitative structure of the parameter vector found by hypentropy is much closer to that of the true parameters, and quantitatively $\left\|\widehat{\boldsymbol{\theta}}_{\mathrm{glm}} - \boldsymbol{\theta}\right\|_1 = 24.609$ while $\left\|\widehat{\boldsymbol{\theta}}_{\mathrm{hyp}} - \boldsymbol{\theta}\right\|_1 = 0.421$. The parameters found by the p-norm algorithm have similar structure to those found by hypentropy and are omitted for visual clarity.

6.2 Low rank system identification

We now consider a nonlinear system identification problem similar to Foster et al. (2020), where the system dynamics are given by a vector-valued GLM and the parameters may be identified using a spectral variant of the Reflectron. In this setting, the learner observes n trajectories $\{\mathbf{x}_t^i\}_{t=0,i=1}^{T,n}$ from the discrete-time dynamical system $\mathbf{x}_{t+1}^i = \rho \mathbf{x}_t^i + \sigma(\boldsymbol{\Theta}\mathbf{x}_t^i) + w_t^i$. The system is initialized from $\mathbf{x}_0^i \sim \mathrm{Unif}([-1,1]^d)$, the process noise is given by $w_t^i \sim \mathrm{Unif}([-\sigma_w, \sigma_w]^d)$, $\sigma(\cdot)$ is the element-wise sigmoid activation, and $\boldsymbol{\Theta}$ is a $d \times d$ matrix with $r = \mathrm{rank}(\boldsymbol{\Theta}) \ll d$. This model is motivated by applications in computational neuroscience, where the system state can be interpreted as a vector of firing rates in a recurrent neural network, and the learned parameters represent the synaptic weights (Rutishauser et al., 2015). In the experiments, we fix $\rho = 0.9$, $T = 5$, and $\sigma_w = 0.1$.

The generalization error for an estimate $\widehat{\boldsymbol{\Theta}}$ is $\varepsilon(\widehat{\boldsymbol{\Theta}}) = \frac{1}{2T}\sum_{t=0}^{T-1} \mathbb{E}_{\mathbf{x}_0}\left[\left\|\mathbf{x}_{t+1} - \rho\mathbf{x}_t - \sigma(\widehat{\boldsymbol{\Theta}}\mathbf{x}_t)\right\|^2\right]$, which measures the ability of the learned connectivity to correctly predict a new random trajectory in a mean square sense. We search for $\widehat{\boldsymbol{\Theta}}$ by minimizing the empirical loss $\widehat{\varepsilon}(\widehat{\boldsymbol{\Theta}}) = \frac{1}{2nT}\sum_{i=1,t=0}^{n,T-1}\left\|\mathbf{x}_{t+1}^i - \rho\mathbf{x}_t^i - \sigma(\widehat{\boldsymbol{\Theta}}\mathbf{x}_t^i)\right\|^2$.

Similar to Section 6.1, the learner has knowledge that $\boldsymbol{\Theta}$ is low-rank, and we compare how the implicit bias of each method impacts its generalization performance. Let $\lambda(\mathbf{M})$ denote the vector of singular values of a matrix $\mathbf{M}$. For both the GLM-tron and hypentropy algorithms, we project onto $\mathcal{C} = \left\{\widehat{\boldsymbol{\Theta}} \in \mathbb{R}^{d\times d} : \left\|\lambda(\widehat{\boldsymbol{\Theta}})\right\|_1 \le 2\left\|\lambda(\boldsymbol{\Theta})\right\|_1\right\}$. For the p-norm algorithm, we project onto $\mathcal{C} = \left\{\widehat{\boldsymbol{\Theta}} \in \mathbb{R}^{d\times d} : \left\|\lambda(\widehat{\boldsymbol{\Theta}})\right\|_p \le 2\left\|\lambda(\boldsymbol{\Theta})\right\|_p\right\}$. Hyperparameters are tuned just as in the sparse vector setting.

In Figure 1e, the ambient dimension and rank are fixed to be $d = 1000$ and $r = 20$, and we study the impact of the number of trajectories n on the generalization

error. Both the p-norm and hypentropy algorithms achieve lower test error than the GLM-tron algorithm for fixed n. In Figure 1f, the ambient dimension d is varied with fixed $r/d = 0.02$, and the number of trajectories is held equal to the dimension $n = d$. A heuristic explanation of this scaling is provided in Appendix A.3. As the dimension increases, the gap between the test error for the GLM-tron and hypentropy increases. This trend also holds for the p-norm algorithm for $n < 1200$, which begins to become brittle to the choice of hyperparameter for large n. As in the previous section, these results validate the predictions of Theorem 4.1, now with vector-valued outputs.

7 Conclusions and future directions

In this work, we studied the effect of optimization geometry on the statistical performance of generalized linear models trained with the square loss. We obtained strong non-asymptotic guarantees that identify how the interplay between optimization geometry and feature space geometry can reduce dimensionality dependence of both the training and generalization errors. We demonstrated the validity of our theoretical results on sparse vector and low-rank matrix recovery problems, where it was shown that pairing the optimization geometry with the feature space geometry as suggested by our analysis consistently led to improved out-of-sample performance.

Single neurons and GLMs highlight important aspects of more complex deep models, so that our work provides insight into the observations by Azizan et al. (2021) that the choice of mirror descent potential affects the generalization performance of deep networks. Moreover, our results provide a quantitative characterization of this effect.

There are a number of natural directions for future work. A first goal is to classify the typical feature space geometry for neural networks on standard datasets. Given such a classification, a well-tailored potential function could be developed to improve generalization performance. A second question is whether there are pseudogradient methods suitable for multilayer architectures, and if they could lead to improved performance or a simpler analysis relative to gradient descent.

References

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276.

Azizan, N., Lale, S., and Hassibi, B. (2019). Stochastic mirror descent on overparameterized nonlinear models: Convergence, implicit regularization, and generalization. *arXiv:1906.03830*.

Azizan, N., Lale, S., and Hassibi, B. (2021). Stochastic mirror descent on overparameterized nonlinear models. *IEEE Transactions on Neural Networks and Learning Systems*.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*.

Boffi, N. M. and Slotine, J.-J. E. (2021). Implicit regularization and momentum algorithms in nonlinearly parameterized adaptive control and prediction. *Neural Computation*, 33(3):590–673.

Demirkaya, A., Chen, J., and Oymak, S. (2020). Exploring the role of loss functions in multiclass classification. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandar, T. (2008). Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning*.

Foster, D. J., Rakhlin, A., and Sarkar, T. (2020). Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*.

Frei, S., Cao, Y., and Gu, Q. (2020). Agnostic learning of a single neuron with gradient descent. In *Neural Information Processing Systems*.

Gentile, C. (2003). The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299.

Ghai, U., Hazan, E., and Singer, Y. (2020). Exponentiated gradient meets gradient descent. In *31st International Conference on Algorithmic Learning Theory*.

Goel, S. and Klivans, A. (2019). Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory*.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. (2018a). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. (2018b). Implicit bias of gradient descent on linear convolutional networks. In *Neural Information Processing Systems*.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. In *Neural Information Processing Systems*.

Gunasekar, S., Woodworth, B., and Srebro, N. (2021). Mirrorless mirror descent: A natural derivation of mirror descent. In *International Conference on Artificial Intelligence and Statistics*.

Han, X. Y., Papyan, V., and Donoho, D. L. (2022). Neural collapse under mse loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv:1903.08560*.

Hui, L. and Belkin, M. (2021). Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*.

Ji, Z. and Telgarsky, M. (2020). Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*.

Kakade, S. M., Kalai, A. T., Kanade, V., and Shamir, O. (2011). Efficient learning of generalized linear and single index models with isotonic regression. In *Neural Information Processing Systems*.

Kakade, S. M., Sridharan, K., and Tewari, A. (2009). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Neural Information Processing Systems*.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Krichene, W., Bayen, A., and Bartlett, P. L. (2015). Accelerated mirror descent in continuous and discrete time. In *Neural Information Processing Systems*.

Maillard, A., Arous, G. B., and Biroli, G. (2021). Landscape complexity for the empirical risk of generalized linear models. *arxiv:1912.02143*.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. CRC Press.

Muthukumar, V., Vodrahalli, K., and Sahai, A. (2019). Harmless interpolation of noisy data in regression. In *2019 IEEE International Symposium on Information Theory (ISIT)*.

Nacson, M. S., Lee, J. D., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. (2018). Convergence of gradient descent on separable data. In *International Conference on Artificial Intelligence and Statistics*.

Nemirovski, A. and Yudin, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley.

Nock, R. and Menon, A. K. (2020). Supervised learning: No loss no cry. In *International Conference on Machine Learning*.

Rutishauser, U., Slotine, J.-J. E., and Douglas, R. (2015). Computation in dynamically bounded asymmetric systems. *PLOS Computational Biology*, 11(1):1–22.

Slotine, J.-J. E. and Li, W. (1991). *Applied Nonlinear Control*. Prentice Hall.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(1):2822–2878.

Tyukin, I. Y., Prokhorov, D. V., and van Leeuwen, C. (2007). Adaptation and parameter estimation in systems with unstable target dynamics and nonlinear parametrization. *IEEE Transactions on Automatic Control*, 52(9):1543–1559.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. (2020). Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*.

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

A Details on experimental setup

A.1 Projections

Euclidean projection onto ℓ_1-ball. For the GLM-tron algorithm, we use the following projection step after every iteration:

$$\operatorname*{arg\,min}_{\mathbf{x}:\|\mathbf{x}\|_1 \le R} \|\mathbf{x} - \mathbf{y}\| \ .$$

The algorithm used to compute this is described in Figure 1 of Duchi et al. (2008).

ℓ_p projection onto ℓ_p-ball. For the p-norm algorithm, we use the following Bregman projection:

$$\operatorname*{arg\,min}_{\mathbf{x}:\|\mathbf{x}\|_p \le R} d_{\psi_p}(\mathbf{x}, \mathbf{y}) \ .$$

The solution is $\mathbf{y}$ for $\|\mathbf{y}\|_p \le R$ and $\frac{\mathbf{y}}{\|\mathbf{y}\|_p} R$ otherwise. Note that we did not implement the Bregman projection

$$\operatorname*{arg\,min}_{\mathbf{x}:\|\mathbf{x}\|_1 \le R} d_{\psi_p}(\mathbf{x}, \mathbf{y}) \ .$$

since we are not aware of an efficient (nearly linear time in dimension) algorithm for doing so.

Hypentropy Bregman projection onto ℓ_1-ball. For the hypentropy algorithm, we use the following Bregman projection:

$$\operatorname*{arg\,min}_{\mathbf{x}:\|\mathbf{x}\|_1 \le R} d_{\psi_\beta}(\mathbf{x}, \mathbf{y}) \ .$$

To implement this projection, we use the following bisection search algorithm communicated to us by Udaya Ghai, which was also used in Ghai et al. (2020). Define the shrinkage function $s_\theta^\beta : \mathbb{R}^d \to \mathbb{R}^d$ as:

$$s_\theta^\beta(\mathbf{x}) = \operatorname{sign}(\mathbf{x}) \max\left\{ \frac{\theta(\sqrt{\mathbf{x}^2 + \beta^2} + |\mathbf{x}|)}{2} - \frac{\sqrt{\mathbf{x}^2 + \beta^2} - |\mathbf{x}|}{2\theta}, 0 \right\} ,$$

where the operations above are all elementwise. One can show that there must exist a $\theta \in (0, 1]$ such that:

$$s_\theta^\beta(\mathbf{y}) = \operatorname*{arg\,min}_{\mathbf{x}:\|\mathbf{x}\|_1 \le R} d_{\psi_\beta}(\mathbf{x}, \mathbf{y}) \ .$$

From the above considerations, we can use bisection to search for a $\theta \in (0, 1]$ such that

$$\left\| s_\theta^\beta(\mathbf{y}) \right\|_1 = R.$$

A.2 Hyperparameter values

In this section, we list the hyperparameters that were gridded over for each figure.

Parameters for Figure 1a

Parameter	Values
λ	$\{1.0, 0.1, 0.01, 0.001\}$
β	$\{1.0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$

Parameters for Figure 1b

Parameter	Values
λ	$\{1.0, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$
β	$\{1.0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$
p	$\{1.1, 1.2, 1.3, 1.4, 1.5\}$

Parameters for Figure 1c Same parameters as Figure 1b.

Parameters for Figure 1e

Parameter	Values
λ	$\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$
β	$\{1.0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$
p	$\{1.1, 1.2, 1.3, 1.4, 1.5\}$

Parameters for Figure 1f Same parameters as Figure 1e.

A.3 Heuristic argument for keeping $n = d$ in Figure 1f

Recall that the empirical loss is

$$\widehat{\varepsilon}(\widehat{\mathbf{\Theta}}) = \frac{1}{2nT}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\left\|\mathbf{x}_{t+1}^i - \rho\mathbf{x}_t^i - \sigma(\widehat{\mathbf{\Theta}}\mathbf{x}_t^i)\right\|^2 ,$$

while the pseudogradient $g(\widehat{\mathbf{\Theta}})$ is

$$g(\widehat{\mathbf{\Theta}}) = \frac{1}{nT}\sum_{t=0}^{T-1}\sum_{i=1}^{n}(\sigma(\widehat{\mathbf{\Theta}}\mathbf{x}_t^i) - \mathbf{x}_{t+1}^i + \rho\mathbf{x}_t^i)(\mathbf{x}_t^i)^\mathsf{T} .$$

A key term in the statistical bound for the Reflectron is the dual norm of the pseudogradient $g(\widehat{\mathbf{\Theta}})$. For the GLM-tron, this is the Frobenius norm $\left\|g(\widehat{\mathbf{\Theta}})\right\|_F$, while for hypentropy this is the operator norm $\left\|g(\widehat{\mathbf{\Theta}})\right\|$. The p-norm case is similiar to hypentropy for the purpose of this discussion, and we omit the details.

Estimating these norms in general is non-trivial due to both the nonlinearity of the activation function and the time-dependence of the trajectory. Instead, we consider a simpler problem based on random matrices to heuristically understand relevant scalings with n and d. In particular, we set $T = 1$ and consider the $d \times d$ matrix $\mathbf{H}$ defined as:

$$\mathbf{H} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T} , \quad \mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}) .$$

Above, each of the $\mathbf{x}_i$'s are independent. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix with i-th row given by $\mathbf{x}_i$; then we have $\mathbf{H} = \frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$. We first estimate a bound on $\mathbb{E}\left\|\mathbf{H}\right\|_F$ via Jensen's inequality

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{H}\right\|_F \le \sqrt{\mathbb{E}\left\|\mathbf{H}\right\|_F^2} &= \sqrt{\frac{1}{n^2}\operatorname{Tr}\left(\mathbb{E}\sum_{i,j}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{x}_j\mathbf{x}_j^\mathsf{T}\right)} = \sqrt{\frac{1}{n^2}\sum_{i,j}\mathbb{E}\left\langle\mathbf{x}_i, \mathbf{x}_j\right\rangle^2} \\
&= \sqrt{\frac{1}{n^2}\left(n\mathbb{E}\left\|\mathbf{x}_1\right\|^4 + n(n-1)\mathbb{E}\left\langle\mathbf{x}_1, \mathbf{x}_2\right\rangle^2\right)} \\
&= \sqrt{\frac{1}{n^2}\left(n(d^2+2d) + n(n-1)d\right)} \\
&= \sqrt{\frac{d^2}{n} + \left(1 + \frac{1}{n}\right)d} \\
&\asymp \sqrt{d} + \frac{d}{\sqrt{n}} .
\end{aligned}$$

On the other hand, $\|\mathbf{X}\| \lesssim \sqrt{n} + \sqrt{d}$ w.h.p. Therefore,

$$\|\mathbf{H}\| = \frac{1}{n}\|\mathbf{X}\|^2 \lesssim \frac{1}{n}(\sqrt{n} + \sqrt{d})^2 \asymp 1 + \frac{d}{n} .$$

Now consider setting $n \asymp d$. Then as $n \to \infty$, we have that $\|\mathbf{H}\| \lesssim 1$ while $\|\mathbf{H}\|_F$ tends to ∞.

B Preliminary results

In this section, we present some results required for our proofs.

The following theorem gives a bound on the Rademacher complexity of a linear predictor, where the weights in the linear function class admit a bound in terms of a strongly convex potential function.

Theorem B.1 (Kakade et al. (2009)). *Let S be a closed convex set and let $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_* \leq C\}$. Let $\psi : S \to \mathbb{R}$ be σ-strongly convex with respect to $\|\cdot\|$ such that $\inf_{\mathbf{w}\in S} \psi(\mathbf{w}) = 0$. Define $\mathcal{W} = \{\mathbf{w} \in S : \psi(\mathbf{w}) \leq W^2\}$, and let $\mathcal{F}_{\mathcal{W}} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \mathbf{w} \in \mathcal{W}\}$ Then,*

$$\mathcal{R}_n(\mathcal{F}_{\mathcal{W}}) \leq CW\sqrt{\frac{2}{\sigma n}}$$

The following theorem is useful for bounding the Rademacher complexity of the generalized linear models considered in this work, as well as for bounding the generalization error in terms of the Rademacher complexity of a function class.

Theorem B.2 (Bartlett and Mendelson (2002)). *Let $\phi : \mathbb{R} \to \mathbb{R}$ be L_ϕ-Lipschitz, and assume that $\phi(0) = 0$. Let $\mathcal{F}$ be a class of functions. Then $\mathcal{R}_n(\phi \circ \mathcal{F}) \leq 2L_\phi \mathcal{R}_n(\mathcal{F})$.*

The following theorem allows for a bound on the generalization error if bounds on the empirical risk and the Rademacher complexity of the function class are known.

Theorem B.3 (Bartlett and Mendelson (2002)). *Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be an i.i.d. sample from a distribution P over $\mathcal{X} \times \mathcal{Y}$ and let $\mathcal{L} : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}$ be an L-Lipschitz and b-bounded loss function in its first argument. Let $\mathcal{F} = \{f \mid f : \mathcal{X} \to \mathcal{Y}'\}$ be a class of functions. For any positive integer $n \geq 0$ and any scalar $\delta \geq 0$,*

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \mathcal{L}(f(\mathbf{x}_i), y_i) - \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\mathcal{L}(f(\mathbf{x}), y)\right] \right| \leq 4L\mathcal{R}_n(\mathcal{F}) + 2b\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}$$

with probability at least $1-\delta$ over the draws of the $\{\mathbf{x}_i, y_i\}$.

The following lemma is a technical result from functional analysis which has seen widespread application in adaptive control theory (Slotine and Li, 1991).

Lemma B.1 (Barbalat's Lemma). *Assume that $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$ is such that $\mathbf{x} \in \mathcal{L}_1$. If $\mathbf{x}(t)$ is uniformly continuous in t, then $\lim_{t\to\infty} \mathbf{x}(t) = 0$.*

Note that a sufficient condition for uniform continuity of $\mathbf{x}(t)$ is that $\dot{\mathbf{x}}(t) \in \mathcal{L}_\infty$.

The following two results will be used to obtain concentration inequalities in arbitrary p norms for empirical averages of random vectors.

Lemma B.2. *Let $\{X_i\}_{i=1}^n$ be random variables in a Banach space $\mathcal{X}$ equipped with a norm $\|\cdot\|$ such that $\|X_i\| \leq C$. Then for any $\delta > 0$, with probability at least $1-\delta$,*

$$\left| \left\| \frac{1}{n}\sum_{i=1}^n X_i \right\| - \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n X_i \right\| \right] \right| \leq \sqrt{\frac{2C^2}{n}\log(2/\delta)}$$

Proof. Observe that by the reverse triangle inequality, $f(X_1, X_2, \ldots, X_n) = \|\sum_{i=1}^n X_i\|$ satisfies the bounded differences inequality with uniform bound $2C$. □

Lemma B.3. *Let $\{\mathbf{X}_i\}_{i=1}^n$ be random vectors in Euclidean space $\mathbf{X}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ such that $\|\mathbf{X}_i\|_p \leq C$ and $\mathbb{E}[\mathbf{X}_i] = 0$ with $p \in [1, \infty]$. Then the following bound holds*

$$\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \right\|_p \right] \leq \begin{cases} \frac{d^{2/p-1}2^{1/2}C}{\sqrt{n}} & p \in [1,2) \\ \frac{C}{\sqrt{n(q-1)}} & p \in [2,\infty) \\ 4C\sqrt{\frac{\log(d)}{n}} & p = \infty \end{cases}$$

Proof. Let ϵ_i denote a Rademacher random variable. By a standard symmetrization argument,

$$\mathbb{E}_{\mathbf{X}_i}\left[\left\|\sum_{i=1}^n \left(\mathbf{X}_i - \mathbb{E}_{\mathbf{X}_i}\left[\mathbf{X}_i\right]\right)\right\|_p\right] \leq 2\mathbb{E}_{\mathbf{X}_i,\epsilon_i}\left[\left\|\sum_{i=1}^n \epsilon_i \mathbf{X}_i\right\|_p\right].$$

Let $\mathcal{F} = \{\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{w}\rangle : \|\mathbf{w}\|_q \leq 1\}$ with $\frac{1}{q} + \frac{1}{p} = 1$. Observe that by definition of the dual norm

$$\mathbb{E}_{\mathbf{X}_i,\epsilon_i}\left[\left\|\sum_{i=1}^n \epsilon_i \mathbf{X}_i\right\|_p\right] = n\mathcal{R}_n(\mathcal{F}).$$

By Theorem B.1, noting that $\|\cdot\|_q^2$ is $\frac{1}{2(q-1)}$-strongly convex with respect to $\|\cdot\|_q$ for $q \in (1,2]$, we then have that

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\right\|_p\right] \leq \frac{C}{\sqrt{n(q-1)}},$$

where $q \in (1,2]$ implies that $p \in [2,\infty)$.

Now consider the case $p = \infty$. Because each $\|\mathbf{X}_i\|_\infty \leq C$, each component of each $\mathbf{X}_i$ is sub-Gaussian. Hence,

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\right\|_\infty\right] \leq 4C\sqrt{\frac{\log(d)}{n}}.$$

Last, consider $p \in [1,2)$. Then we have the elementary bound via equivalence of norms

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\right\|_p\right] \leq d^{1/p-1/2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\right\|_2\right] \leq \frac{d^{2/p-1}2^{1/2}C}{\sqrt{n}}.$$

This completes the proof. □

To analyze our discrete-time iterations, we require three basic properties of the Bregman divergence.

Lemma B.4 (Bregman three-point identity)**.** *Let $\psi : \mathcal{M} \to \mathbb{R}^p$ denote a σ-strongly convex function with respect to some norm $\|\cdot\|$. Then for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{M}$,*

$$\langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{z}\rangle = d_\psi\left(\mathbf{x} \parallel \mathbf{y}\right) + d_\psi\left(\mathbf{z} \parallel \mathbf{x}\right) - d_\psi\left(\mathbf{z} \parallel \mathbf{y}\right). \tag{11}$$

Lemma B.5 (Generalized Pythagorean Theorem)**.** *Let $\psi : \mathcal{M} \to \mathbb{R}$ denote a σ-strongly convex function with respect to some norm $\|\cdot\|$. Let $\mathbf{x}_0 \in \mathcal{M}$ and let $\mathbf{x}^* = \Pi_{\mathcal{C}}^{\psi}(\mathbf{x}_0)$ be its projection onto a closed and convex set $\mathcal{C}$. Then for any $\mathbf{y} \in \mathcal{C}$,*

$$d_\psi\left(\mathbf{y} \parallel \mathbf{x}_0\right) \geq d_\psi\left(\mathbf{y} \parallel \mathbf{x}^*\right) + d_\psi\left(\mathbf{x}^* \parallel \mathbf{x}_0\right). \tag{12}$$

Lemma B.6 (Bregman duality)**.** *Let $\psi : \mathcal{M} \to \mathbb{R}$ denote a σ-strongly convex function with respect to some norm $\|\cdot\|$, and let ψ^* denote its Fenchel conjugate. Then ψ^* is $1/\sigma$-smooth with respect to $\|\cdot\|_*$, and moreover*

$$d_\psi\left(\mathbf{x} \parallel \mathbf{y}\right) = d_\psi\left(\nabla\psi^*(\mathbf{y}) \parallel \nabla\psi^*(\mathbf{x})\right) \tag{13}$$

To obtain fast rates in the realizable online learning setting, we require the following martingale Bernstein bound, which has been used in similar analyses prior to this work (Ji and Telgarsky, 2020; Frei et al., 2020).

Lemma B.7 (Beygelzimer et al. (2011))**.** *Let $\{Y_t\}_{t=1}^\infty$ be a martingale adapted to the filtration $\{\mathcal{F}_t\}_{t=1}^\infty$. Let $\{D_t\}_{t=1}^\infty$ be the corresponding martingale difference sequence. Define*

$$V_t = \sum_{k=1}^t \mathbb{E}\left[D_k^2 | \mathcal{F}_{k-1}\right],$$

and assume that $D_t \leq R$ almost surely. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$Y_t \leq R\log(1/\delta) + (e-2)V_t/R.$$

C Omitted proofs

C.1 Proof of Theorem 4.1

To make progress in the general setting, we require a definition of a modified error function for a parametric hypothesis $h(\mathbf{x}) = u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}\right\rangle\right)$. The empirical version over the dataset $\widehat{H}(h)$ is defined analogously.

$$H(h) = \mathbb{E}\left[\left(u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}\right\rangle\right) - u\left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle\right)\right)^2 \xi\left(\widehat{\boldsymbol{\theta}}, \mathbf{x}\right)\right]$$

Intuitively, under Assumption 4.1, we can relate H to ε. The following lemma makes this rigorous, and is adapted from Frei et al. (2020). The proof is a trivial modification of the proof given in their work.

Lemma C.1. *Let ξ satisfy Assumption 4.1. Let h denote a parametric hypothesis of the form $h(\mathbf{x}) = u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}\right\rangle\right)$. Then if $\left\|\widehat{\boldsymbol{\theta}}\right\| \leq a$ and $\|\mathbf{x}\|_* \leq b$, we have the bound $\widehat{\varepsilon}(h) \leq \widehat{H}(h)/\gamma$ where γ is a fixed constant defined in Assumption 4.1.*

We now begin the proof of Theorem 4.1.

Proof. By the Bregman three-point identity (11), with $\mathbf{z} = \boldsymbol{\theta}$, $\mathbf{x} = \widehat{\boldsymbol{\phi}}_{t+1}$, and $\mathbf{y} = \widehat{\boldsymbol{\theta}}_t$,

$$\begin{aligned}
d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\phi}}_{t+1}\right) &= d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - d_\psi\left(\widehat{\boldsymbol{\phi}}_{t+1} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + \left\langle \nabla\psi(\widehat{\boldsymbol{\phi}}_{t+1}) - \nabla\psi(\widehat{\boldsymbol{\theta}}_t), \widehat{\boldsymbol{\phi}}_{t+1} - \boldsymbol{\theta}\right\rangle, \\
&= d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - d_\psi\left(\widehat{\boldsymbol{\phi}}_{t+1} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - \left\langle \frac{\lambda}{n}\sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i, \widehat{\boldsymbol{\phi}}_{t+1} - \boldsymbol{\theta}\right\rangle, \\
&= d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - d_\psi\left(\widehat{\boldsymbol{\phi}}_{t+1} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - \left\langle \frac{\lambda}{n}\sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\right\rangle \\
&\quad - \left\langle \frac{\lambda}{n}\sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i, \widehat{\boldsymbol{\phi}}_{t+1} - \widehat{\boldsymbol{\theta}}_t\right\rangle.
\end{aligned}$$

After grouping the second and final terms in the above expression as in the proof of Theorem 4.1, the iteration becomes

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\phi}}_{t+1}\right) = d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + d_\psi\left(\widehat{\boldsymbol{\theta}}_t \,\middle\|\, \widehat{\boldsymbol{\phi}}_{t+1}\right) + \frac{\lambda}{n}\left\langle \sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle.$$

By the generalized Pythagorean Theorem (12),

$$\begin{aligned}
d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) &\leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + d_\psi\left(\widehat{\boldsymbol{\theta}}_t \,\middle\|\, \widehat{\boldsymbol{\phi}}_{t+1}\right) \\
&\quad + \frac{\lambda}{n}\left\langle \sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle. \qquad (14)
\end{aligned}$$

By duality (13), we may replace $d_\psi\left(\widehat{\boldsymbol{\theta}}_t \,\middle\|\, \widehat{\boldsymbol{\phi}}_{t+1}\right)$ by $d_\psi\left(\nabla\psi^*\left(\widehat{\boldsymbol{\phi}}_{t+1}\right) \,\middle\|\, \nabla\psi^*\left(\widehat{\boldsymbol{\theta}}_t\right)\right)$,

$$\begin{aligned}
d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) &\leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + d_{\psi^*}\left(\nabla\psi\left(\widehat{\boldsymbol{\phi}}_{t+1}\right) \,\middle\|\, \nabla\psi\left(\widehat{\boldsymbol{\theta}}_t\right)\right) \\
&\quad + \frac{\lambda}{n}\left\langle \sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle.
\end{aligned}$$

Because ψ is σ-strongly convex with respect to $\|\cdot\|$, ψ^* is $\frac{1}{\sigma}$-smooth with respect to $\|\cdot\|_*$. Thus,

$$\begin{aligned} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) &\leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + \frac{1}{2\sigma}\left\|\nabla\psi\left(\widehat{\boldsymbol{\phi}}_{t+1}\right) - \nabla\psi\left(\widehat{\boldsymbol{\theta}}_t\right)\right\|_*^2 \\ &\quad + \frac{\lambda}{n}\left\langle \sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle, \\ &= d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + \frac{\lambda^2}{2\sigma}\left\|\frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i\right\|_*^2 \\ &\quad + \frac{\lambda}{n}\left\langle \sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle. \end{aligned} \tag{15}$$

Above, we applied $\frac{1}{\sigma}$-smoothness and then used (7) to express the increment in $\nabla\psi$. The second term in (15) can be bounded as

$$\begin{aligned} &\left\|\frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i\right\|_*^2 \\ &\leq 2\left\|\frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_i, \boldsymbol{\theta}\rangle\right)\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i\right\|_*^2 + 2\left\|\frac{1}{n}\sum_{i=1}^n \left(u\left(\langle \mathbf{x}_i, \boldsymbol{\theta}\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i\right\|_*^2. \end{aligned}$$

By Jensen's inequality, and using that $\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right) \leq B$,

$$\begin{aligned} \left\|\frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_i, \boldsymbol{\theta}\rangle\right)\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i\right\|_*^2 &\leq \frac{1}{n}\sum_{i=1}^n \left\|\left(u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right\rangle\right) - u\left(\langle \boldsymbol{\theta}, \mathbf{x}_i\rangle\right)\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i\right\|_*^2, \\ &= \frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right\rangle\right) - u\left(\langle \boldsymbol{\theta}, \mathbf{x}_i\rangle\right)\right)^2 \xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)^2 \|\mathbf{x}_i\|_*^2, \\ &\leq C^2 B\widehat{H}(h_t). \end{aligned}$$

By assumption, $\|\frac{1}{n}\sum_{i=1}^n \left(y_i - u(\langle \mathbf{x}_i, \boldsymbol{\theta}\rangle)\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i\|_* \leq \eta$. Combining this with the above, we find that

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\mathbf{x}_i\right\|_*^2 \leq 2\left(C^2 B\widehat{H}(h_t) + \eta^2\right).$$

By an induction argument identical to that used in the proof of Theorem 4.1, the iteration of the Bregman divergence between the Bayes-optimal parameters and the parameters of our hypothesis becomes

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) \leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - \lambda\widehat{H}(h_t)\left(\frac{1}{L} - \frac{\lambda BC^2}{\sigma}\right) + \eta\lambda\left(\frac{\lambda\eta}{\sigma} + \sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}\right).$$

Assume that $\eta \leq \sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$. For $\lambda \leq \frac{\sigma}{2BC^2L}$, we have

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) \leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - \frac{\lambda}{2L}\widehat{H}(h_t) + \eta\lambda\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}\left(\frac{2BC^2L+1}{2BC^2L}\right).$$

Thus, at each iteration we either have the decrease condition

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) - d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) \leq -\eta\lambda\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}\left(\frac{2BC^2L+1}{2BC^2L}\right)$$

or the error bound

$$\widehat{H}(h_t) < 4L\eta\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}\left(\frac{2BC^2L+1}{2BC^2L}\right).$$

In the former case there can be at most

$$t_f = \frac{d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}(0)\right)}{\eta\lambda\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}\left(\frac{2BC^2L+1}{2BC^2L}\right)} \leq \frac{1}{\lambda}\sqrt{\frac{\sigma\psi(\boldsymbol{\theta})}{2\eta^2}}$$

iterations before $\widehat{H}(h_t) \leq 4L\eta\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}\left(\frac{2BC^2L+1}{2BC^2L}\right)$. Furthermore, note that $\left\|\widehat{\boldsymbol{\theta}}(t)\right\| \leq \sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}} + \|\boldsymbol{\theta}\| \leq (1+W)\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$. Then by Lemma C.1, $\widehat{\varepsilon}(h_t) \leq \frac{4L\eta}{\gamma}\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}\left(\frac{2BC^2L+1}{2BC^2L}\right)$ where γ corresponds to $a = (1+W)\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$ and $b = C$ in Lemma C.1. The conclusion of the theorem now follows by application of Theorem B.3 to transfer the bound on $\widehat{\varepsilon}(h_t)$ to $\varepsilon(h_t)$. □

C.2 Proof of Corollary 4.1

Proof. Note that $\mathcal{F} \subseteq \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\|_q \leq W\left(1 + \frac{1}{\sqrt{q-1}}\right)\right\}$. Hence by Theorem B.1, $\mathcal{R}_n(\mathcal{F}) \leq \frac{CW}{\sqrt{n(q-1)}}\left(1 + \frac{1}{\sqrt{q-1}}\right)$. By Lemmas B.2 and B.3, $\eta = C\left(\sqrt{\frac{2\log(4/\delta)}{n}} + \frac{1}{\sqrt{n(q-1)}}\right)$. □

C.3 Proof of Corollary 4.2

Proof. Observe that we have the inclusion

$$\mathcal{F} \subseteq \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\|_1 \leq W\left(1 + \sqrt{3\log(d)}\right)\right\} \subseteq \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\|_q \leq W\left(1 + \sqrt{3\log(d)}\right)\right\}.$$

Hence $\mathcal{R}_n(\mathcal{F}) \leq \frac{CW(1+\sqrt{3\log d})^2}{n^{1/2}}$ by Theorem B.1. By Lemmas B.2 and B.3,

$$\eta = C\left(\sqrt{\frac{2\log(4/\delta)}{n}} + 4\sqrt{\frac{\log(d)}{n}}\right)$$

. □

C.4 Proof of Corollary 4.3

Proof. Note that $\mathcal{F} \subseteq \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \psi(\mathbf{w}) \leq \log(d)\}$ and $\mathcal{R}_n \leq C\sqrt{\frac{2\log d}{n}}$. By Lemmas B.2 and B.3, $\eta = C\left(\sqrt{\frac{2\log(4/\delta)}{n}} + 4\sqrt{\frac{\log(d)}{n}}\right)$. □

C.5 Proof of Lemma 4.1

Proof. From (15), we have a bound on the iteration for the Bregman divergence between the interpolating parameters and the current parameter estimates,

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) \leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + \frac{\lambda^2}{2\sigma}\left\|\frac{1}{n}\sum_{i=1}^n\left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\mathbf{x}_i\xi\left(\widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right)\right\|_2^2$$
$$+ \frac{\lambda}{n}\left\langle\sum_{i=1}^n\left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right)\mathbf{x}_i, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle.$$

Under the realizability assumption of the lemma, we may bound the second term above as

$$\frac{\lambda^2}{2\sigma}\left\|\frac{1}{n}\sum_{i=1}^n\left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right)\mathbf{x}_i\right\|_2^2 \leq \frac{\lambda^2C^2B}{2\sigma}\widehat{H}(h_t).$$

We may similarly bound the final term, exploiting monotonicity and Lipschitz continuity of $u(\cdot)$, as

$$\frac{\lambda}{n}\left\langle\sum_{i=1}^n\left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}, \mathbf{x}_i\right)\mathbf{x}_i, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle \leq -\frac{\lambda}{L}\widehat{H}(h_t).$$

Putting these together, we have the refined bound on the iteration

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) \leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + \lambda\left(\frac{\lambda C^2 B}{2\sigma} - \frac{1}{L}\right)\widehat{H}(h_t).$$

Let $0 < \alpha < 1$. For $\lambda \leq \frac{2\sigma(1-\alpha)}{C^2 BL}$,

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) \leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - \frac{\lambda\alpha}{L}\widehat{H}(h_t).$$

Note that this shows $d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) \leq d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right)$ for all t, so that $\left\|\widehat{\boldsymbol{\theta}}_t\right\| \leq \|\boldsymbol{\theta}\| + \sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$. Summing both sides of the above inequality from $t = 1$ to T reveals that

$$\sum_{t=1}^{T} \widehat{H}(h_t) \leq \frac{L}{\lambda\alpha}\left(d_\psi\left(\boldsymbol{\theta} \,\|\, \boldsymbol{\theta}_1\right) - d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{T+1}\right)\right) \leq \frac{L}{\lambda\alpha} d_\psi\left(\boldsymbol{\theta} \,\|\, \boldsymbol{\theta}_1\right).$$

Because T was arbitrary and the upper bound is independent of T, $\sum_{t=1}^{\infty} \widehat{H}(h_t)$ exists and hence $\widehat{H}(h_t) \to 0$ as $t \to \infty$. Furthermore,

$$\min_{t' \in [1,T]}\left\{\widehat{H}(h_{t'})\right\} T = \sum_{t=1}^{T} \min_{t' \in [1,T]}\left\{\widehat{H}(h_{t'})\right\} \leq \sum_{t=1}^{T} \widehat{H}(h_t) \leq \frac{L}{\lambda\alpha} d_\psi\left(\boldsymbol{\theta} \,\|\, \boldsymbol{\theta}_1\right),$$

so that $\min_{t' \in [1,T]}\left\{\widehat{H}(h_{t'})\right\} \leq \frac{L d_\psi(\boldsymbol{\theta} \,\|\, \boldsymbol{\theta}_1)}{\alpha\lambda T}$. By taking $\alpha \to 0$, we obtain the requirement $\lambda < \frac{2\sigma}{C^2 BL}$. To conclude the proof, apply Lemma C.1 with $a = \sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}} + \|\boldsymbol{\theta}\|$ and $b = C$. □

C.6 Proof of Theorem 4.2

The proof discretizes the proof of Theorem 3.2, and is similar to the proof of implicit regularization for mirror descent due to Azizan et al. (2019).

Proof. Let $\bar{\boldsymbol{\theta}} \in \mathcal{A}$ be arbitrary. From (14),

$$\begin{aligned} d_\psi\left(\bar{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) &\leq d_\psi\left(\bar{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + d_\psi\left(\widehat{\boldsymbol{\theta}}_t \,\middle\|\, \widehat{\boldsymbol{\phi}}_{t+1}\right) + \frac{\lambda}{n}\sum_{i=1}^{n}\left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\left\langle \mathbf{x}_i, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_t\right\rangle, \\ &= d_\psi\left(\bar{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + d_\psi\left(\widehat{\boldsymbol{\theta}}_t \,\middle\|\, \widehat{\boldsymbol{\phi}}_{t+1}\right) \\ &\quad + \frac{\lambda}{n}\sum_{i=1}^{n}\left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_i\right)\left(u^{-1}\left(y_i\right) - \left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right), \end{aligned}$$

where we have used that $\overline{\boldsymbol{\theta}} \in \mathcal{A}$ and applied invertibility of $u(\cdot)$. Summing both sides from $t = 1$ to ∞,

$$\begin{aligned} d_\psi\left(\overline{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}_\infty\right) &\leq d_\psi\left(\overline{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right) + \sum_{t=1}^{\infty} d_\psi\left(\widehat{\boldsymbol{\theta}}_t \,\middle\|\, \widehat{\boldsymbol{\phi}}_{t+1}\right) \\ &\quad + \frac{\lambda}{n}\sum_{i=1}^{n}\sum_{t=1}^{\infty}\left(u\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_i\right)\xi\left(\widehat{\boldsymbol{\theta}}_t, \xi_i\right)\left(u^{-1}\left(y_i\right) - \left\langle \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_t\right\rangle\right). \end{aligned}$$

The above relation is true for any $\overline{\boldsymbol{\theta}} \in \mathcal{A}$. Furthermore, the only dependence of the right-hand side on $\overline{\boldsymbol{\theta}}$ is through the first Bregman divergence. Hence the arg min of the two Bregman divergences involving $\overline{\boldsymbol{\theta}}$ must be equal, which shows that $\widehat{\boldsymbol{\theta}}_\infty = \arg\min_{\overline{\boldsymbol{\theta}} \in \mathcal{A}} d_\psi\left(\overline{\boldsymbol{\theta}} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right)$. Choosing $\widehat{\boldsymbol{\theta}}_1 = \arg\min_{\mathbf{w} \in \mathcal{C} \cap \mathcal{M}} \psi(\mathbf{w})$ completes the proof. □

C.7 Proof of Theorem 5.1

Proof. Let $\xi_t = \xi\left(\widehat{\boldsymbol{\theta}}_t, \mathbf{x}_t\right)$. From (15) adapted to the stochastic optimization setting, we have the bound

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) \le d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) + \frac{\lambda^2}{2\sigma}\left\|\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_t\right)\mathbf{x}_t\xi_t\right\|_*^2 + \lambda\left\langle\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_t\right)\mathbf{x}_t\xi_t, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle.$$

Note that we can write

$$\begin{aligned}\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_t\right)^2 &= \left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)^2 + \left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)^2 \\ &\quad + 2\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)\left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right).\end{aligned}$$

Using that u is nondecreasing and L-Lipschitz,

$$\begin{aligned}\left\langle\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - y_t\right)\mathbf{x}_t\xi_t, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle &\le -\frac{1}{L}\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)^2\xi_t \\ &\quad + \left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)\xi_t\left\langle \mathbf{x}_t, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle.\end{aligned}$$

Putting these together, we conclude the bound,

$$\begin{aligned}d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) &\le d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - \lambda\left(\frac{1}{L} - \frac{\lambda C^2 B}{2\sigma}\right)\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)^2\xi_t \\ &\quad + \lambda\xi_t\left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)\left(\left\langle \mathbf{x}_t, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle + \frac{\lambda C^2 B}{\sigma}\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)\right) \\ &\quad + \frac{\lambda^2 C^2 B^2}{2\sigma}\left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)^2.\end{aligned}$$

Summing both sides from $t = 1$ to T,

$$\begin{aligned}d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{T+1}\right) &\le d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right) - \lambda\left(\frac{1}{L} - \frac{\lambda C^2 B}{2\sigma}\right)\sum_{t=1}^T\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)^2\xi_t \\ &\quad + \lambda\sum_{t=1}^T\xi_t\left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)\left(\left\langle \mathbf{x}_t, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle + \frac{\lambda C^2 B}{\sigma}\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)\right) \\ &\quad + \frac{\lambda^2 C^2 B^2}{2\sigma}\sum_{t=1}^T\left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)^2.\end{aligned}$$

Define the filtration $\{\mathcal{F}_t = \sigma(\mathbf{x}_1, y_1, \mathbf{x}_2, y_2, \ldots, \mathbf{x}_t, y_t, \mathbf{x}_{t+1})\}_{t=1}^\infty$, and note that

$$\begin{aligned}D_t^{(1)} &= \xi_t\left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)\left\langle \mathbf{x}_t, \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\rangle, \\ D_t^{(2)} &= \xi_t\left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right),\end{aligned}$$

are martingale difference sequences adapted to $\{\mathcal{F}_t\}$. Furthermore, note that $|D_t^{(1)}| \le CBR$ and $|D_t^{(2)}| \le LCBR$ almost surely where $R = \mathsf{Diam}(C)$. Hence, by an Azuma-Hoeffding bound, with probability at least $1 - \delta/3$,

$$\begin{aligned}\sum_{t=1}^T D_t^{(1)} &\le \sqrt{CBRT\log(6/\delta)}, \\ \sum_{t=1}^T D_t^{(2)} &\le \sqrt{LCBRT\log(6/\delta)}.\end{aligned}$$

The variance term is trivially bounded almost surely,

$$\sum_{t=1}^T\left(u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right) - y_t\right)^2 \le T.$$

Putting these bounds together and rearranging, we conclude that with probability at least $1-2\delta/3$,

$$\lambda\left(\frac{1}{L}-\frac{\lambda C^2 B}{2\sigma}\right)\sum_{t=1}^{T}\left(u\left(\left\langle \mathbf{x}_t,\widehat{\boldsymbol{\theta}}_t\right\rangle\right)-u\left(\langle \mathbf{x}_t,\boldsymbol{\theta}\rangle\right)\right)^2\xi_t \le d_\psi\left(\boldsymbol{\theta}\,\middle\|\,\widehat{\boldsymbol{\theta}}_1\right)-d_\psi\left(\boldsymbol{\theta}\,\middle\|\,\widehat{\boldsymbol{\theta}}_{T+1}\right)$$
$$+\lambda\sqrt{CBRT\log(6/\delta)}+\frac{\lambda^2C^2}{\sigma}\sqrt{LCBRT\log(6/\delta)}+\frac{\lambda^2C^2B^2T}{2\sigma}.$$

Let $\beta\in(0,1)$ and take $\lambda=\min\left\{\frac{2\sigma(1-\beta)}{C^2BL},\frac{1}{\sqrt{T}}\right\}$. Define $\beta'=1-\frac{C^2LB}{2\sigma\sqrt{T}}$, and define $\bar{\beta}=\max\{\beta,\beta'\}$. Then $\frac{1}{L}-\frac{\lambda C^2B}{2\sigma}=\frac{\bar{\beta}}{L}>0$. Defining $h_t=\left(u\left(\left\langle \mathbf{x}_t,\widehat{\boldsymbol{\theta}}_t\right\rangle\right)-u\left(\langle \mathbf{x}_t,\boldsymbol{\theta}\rangle\right)\right)^2\xi_t$, we find

$$\sum_{t=1}^{T}h_t\le\frac{L}{\bar{\beta}}\max\left\{\sqrt{T},\frac{C^2BL}{2\sigma(1-\beta)}\right\}d_\psi\left(\boldsymbol{\theta}\,\middle\|\,\widehat{\boldsymbol{\theta}}_1\right)+\frac{L}{\bar{\beta}}\sqrt{CBRT\log(6/\delta)}$$
$$+\frac{C^2L}{\sigma\bar{\beta}}\sqrt{LCBR\log\left(6/\delta\right)}+\frac{C^2L\sqrt{T}B^2}{2\sigma\bar{\beta}}. \tag{16}$$

By Assumption 4.1, noting that $\|\mathbf{x}_t\|_*\le C$ and $\left\|\widehat{\boldsymbol{\theta}}_t\right\|\le R+\|\boldsymbol{\theta}\|$, there exists a fixed $\gamma>0$ such that $\sum_{t=1}^{T}\varepsilon_t\le\frac{1}{\gamma}\sum_{t=1}^{T}h_t$. We now want to transfer this bound to a bound on $\varepsilon(h_t)$ via Lemma B.7. Define $D_t^{(3)}=\varepsilon(h_t)-\varepsilon_t$, and note that this is a martingale difference sequence adapted to the filtration $\{\mathcal{F}_t=\sigma(\mathbf{x}_1,y_1,\mathbf{x}_2,y_2,\ldots,\mathbf{x}_t,y_t)\}$. $D_t^{(3)}$ satisfies the following inequalities almost surely,

$$D_t^{(3)}\le\frac{1}{2}L^2C^2R,$$
$$\mathbb{E}\left[\left(D_t^{(3)}\right)^2|\mathcal{F}_{t-1}\right]\le\frac{1}{2}L^2C^2R\varepsilon(h_t).$$

Thus, by Lemma B.7, with probability at least $1-\delta/3$,

$$\sum_{\tau=1}^{T}\varepsilon(h_\tau)\le\frac{L^2C^2R}{2(3-e)}\log(3/\delta)+\frac{1}{3-e}\sum_{\tau=1}^{T}\varepsilon_\tau.$$

Using (16), we then have with probability at least $1-\delta$,

$$\sum_{\tau=1}^{T}\varepsilon(h_\tau)\le\frac{L^2C^2R}{2(3-e)}\log(3/\delta)+\frac{L}{\bar{\beta}\gamma(3-e)}\max\left\{\sqrt{T},\frac{C^2BL}{2\sigma(1-\beta)}\right\}d_\psi\left(\boldsymbol{\theta}\,\middle\|\,\widehat{\boldsymbol{\theta}}_1\right)$$
$$+\frac{L}{\bar{\beta}\gamma(3-e)}\sqrt{CBRT\log(6/\delta)}+\frac{C^2L}{\sigma\bar{\beta}\gamma(3-e)}\sqrt{LCBR\log\left(6/\delta\right)}+\frac{C^2L\sqrt{T}B^2}{2\sigma\bar{\beta}\gamma(3-e)}.$$

Noting that $\min_{t<T}\varepsilon(h_\tau)\le\frac{1}{T}\sum_{\tau=1}^{T}\varepsilon(h_\tau)$ completes the proof. □

C.8 Proof of Theorem 5.2

Proof. Again from (15) adapted to the stochastic optimization setting, we have the bound

$$\begin{aligned}
d_\psi\left(\boldsymbol{\theta}\,\middle\|\,\widehat{\boldsymbol{\theta}}_{t+1}\right)&\le d_\psi\left(\boldsymbol{\theta}\,\middle\|\,\widehat{\boldsymbol{\theta}}_t\right)+\frac{\lambda^2}{2\sigma}\left\|\left(u\left(\left\langle \mathbf{x}_t,\widehat{\boldsymbol{\theta}}_t\right\rangle\right)-u\left(\langle \mathbf{x}_t,\boldsymbol{\theta}\rangle\right)\right)\xi_t\mathbf{x}_t\right\|_*^2\\
&\qquad+\lambda\left\langle\left(u\left(\left\langle \mathbf{x}_t,\widehat{\boldsymbol{\theta}}_t\right\rangle\right)-u\left(\langle \mathbf{x}_t,\boldsymbol{\theta}\rangle\right)\right)\xi_t\mathbf{x}_t,\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_t\right\rangle,\\
&\le d_\psi\left(\boldsymbol{\theta}\,\middle\|\,\widehat{\boldsymbol{\theta}}_t\right)+\frac{\lambda^2C^2B}{2\sigma}\left(u\left(\left\langle \mathbf{x}_t,\widehat{\boldsymbol{\theta}}_t\right\rangle\right)-u\left(\langle \mathbf{x}_t,\boldsymbol{\theta}\rangle\right)\right)^2\xi_t\\
&\qquad-\frac{\lambda}{L}\left(u\left(\left\langle \mathbf{x}_t,\widehat{\boldsymbol{\theta}}_t\right\rangle\right)-u\left(\langle \mathbf{x}_t,\boldsymbol{\theta}\rangle\right)\right)^2\xi_t,\\
&=d_\psi\left(\boldsymbol{\theta}\,\middle\|\,\widehat{\boldsymbol{\theta}}_t\right)-\frac{\lambda}{L}\left(1-\frac{\lambda LC^2B}{2\sigma}\right)\left(u\left(\left\langle \mathbf{x}_t,\widehat{\boldsymbol{\theta}}_t\right\rangle\right)-u\left(\langle \mathbf{x}_t,\boldsymbol{\theta}\rangle\right)\right)^2\xi_t.
\end{aligned}$$

Let $0 < \beta < 1$. Taking $\lambda = \frac{(1-\beta)2\sigma}{LC^2B}$,

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) \le d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) - \frac{2\sigma(1-\beta)\beta}{L^2C^2B}\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)^2 \xi_t,$$

so that $d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_{t+1}\right) \le d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) \le \ldots \le d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right)$. Let W be such that $\|\boldsymbol{\theta}\| = W\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$. Then $d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right) \le \psi(\boldsymbol{\theta})$ so that $\left\|\widehat{\boldsymbol{\theta}}_t\right\| \le (1+W)\sqrt{\frac{2\psi(\boldsymbol{\theta})}{\sigma}}$ by σ-strong convexity of ψ with respect to $\|\cdot\|$. Summing both sides from 1 to $T-1$ leads to the inequality

$$d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_T\right) \le d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right) - \frac{2\sigma(1-\beta)\beta}{L^2C^2B}\sum_{t=1}^{T-1}\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)^2 \xi_t.$$

Rearranging, using positivity of the Bregman divergence, and defining $h_t = \left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)^2 \xi_t$, we conclude that

$$\sum_{t=1}^{T-1} h_t \le \frac{L^2C^2B}{2\sigma(1-\beta)\beta} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right).$$

Applying Assumption 4.1 shows that there exists a $\gamma > 0$ such that

$$\sum_{t=1}^{T-1} \varepsilon_t \le \frac{L^2C^2B}{2\sigma(1-\beta)\beta\gamma} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right). \tag{17}$$

We would now like to transfer the bound (17) to a bound on $\varepsilon(h_t)$. Define $D_t = \varepsilon(h_t) - \varepsilon_t$, and note that $\{D_t\}_{t=1}^\infty$ is a martingale difference sequence adapted to the filtration $\{\mathcal{F}_t = \sigma(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)\}_{t=1}^\infty$. Note that, almost surely,

$$\begin{aligned} D_t \le \varepsilon(h_t) &= \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left(u\left(\left\langle \widehat{\boldsymbol{\theta}}_t, \mathbf{x}\right\rangle\right) - u\left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle\right)\right)^2\right] \\ &\le \frac{1}{2}L^2C^2\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|^2 \le \frac{L^2C^2}{\sigma} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_t\right) \le \frac{L^2C^2}{\sigma} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right) \end{aligned}$$

where we have applied σ-strong convexity of ψ with respect to $\|\cdot\|$ to upper bound $\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|^2$ by the corresponding Bregman divergence. Now, consider the following bound on the conditional variance

$$\begin{aligned} \mathbb{E}\left[D_t^2 | \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\varepsilon(h_t)^2 - 2\varepsilon(h_t)\varepsilon_t + \varepsilon_t^2 | \mathcal{F}_{t-1}\right], \\ &= \varepsilon(h_t)^2 - 2\varepsilon(h_t)^2 + \mathbb{E}\left[\varepsilon_t^2 | \mathcal{F}_{t-1}\right], \\ &\le \mathbb{E}\left[\varepsilon_t^2 | \mathcal{F}_{t-1}\right], \\ &= \frac{1}{4}\mathbb{E}\left[\left(u\left(\left\langle \mathbf{x}_t, \widehat{\boldsymbol{\theta}}_t\right\rangle\right) - u\left(\langle \mathbf{x}_t, \boldsymbol{\theta}\rangle\right)\right)^4 | \mathcal{F}_{t-1}\right], \\ &\le \frac{L^2C^2 d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right)}{\sigma}\varepsilon(h_t). \end{aligned}$$

Hence by Lemma B.7, with probability at least $1-\delta$,

$$\sum_{\tau=1}^{t}\left(\varepsilon(h_\tau) - \varepsilon_\tau\right) \le \frac{L^2C^2}{\sigma} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right)\log(1/\delta) + (e-2)\sum_{\tau=1}^{t}\varepsilon(h_\tau).$$

Rearranging terms,

$$(3-e)\sum_{\tau=1}^{t}\varepsilon(h_\tau) \le \frac{L^2C^2}{\sigma} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right)\log(1/\delta) + \sum_{\tau=1}^{t}\varepsilon_\tau.$$

Applying the bound from (17),

$$(3-e)\sum_{\tau=1}^{t}\varepsilon(h_\tau) \leq \frac{L^2C^2}{\sigma} d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right)\left(\log(1/\delta) + \frac{1}{2(1-\beta)\beta\gamma}\right).$$

We then conclude

$$\min_{t<T}\varepsilon(h_t) \leq \frac{1}{T}\sum_{\tau=1}^{T}\varepsilon(h_\tau) \leq \frac{L^2C^2 d_\psi\left(\boldsymbol{\theta} \,\middle\|\, \widehat{\boldsymbol{\theta}}_1\right)}{\sigma(3-e)T}\left(\log(1/\delta) + \frac{B}{2(1-\beta)\beta\gamma}\right),$$

which completes the proof. □