Basic Inequalities for First-Order Optimization Algorithms with Applications to Statistical Risk Analysis

Seunghoon Paik[†] Kangjie Zhou[‡] Matus Telgarsky[§] Ryan J. Tibshirani[†]
[†]University of California, Berkeley [‡]Columbia University [§]New York University

Abstract

In this work, we introduce *basic inequalities* for first-order iterative optimization algorithms, forming a simple yet versatile framework which connects implicit and explicit regularization. Building on related comparison inequalities for optimization iterates that already exist in the literature, we extend and unify these arguments to produce a general framework, which can be used as a tool for statistical analysis. In more detail, let f denote the objective function to be optimized. Given a first-order iterative algorithm initialized at θ_0, with current iterate θ_T, the basic inequality upper bounds $f(\theta_T) - f(z)$ for any reference point z in terms of the accumulated step sizes, and the distances between θ_0, θ_T, and z. These distances are measured in a geometry inherent to the optimization algorithm, which then translates into a notion of regularization being applied across the path of iterates.

In addition to refining existing results on gradient descent, we provide new results for mirror descent and other first-order methods. We then show how to use these basic inequalities to derive elementary yet useful bounds on the prediction risk of early-stopped gradient descent and exponentiated gradient descent iterates in generalized linear models. We also supplement these findings with numerical experiments.

1 Introduction

This paper introduces *basic inequalities* for first-order optimization algorithms, which connects implicit and explicit notions of regularization. We consider an optimization problem

$$\min_{\theta \in \mathbb{R}^d} \; f(\theta)$$

for an objective function f, consider a first-order optimization method which produces iterate θ_T at iteration T, and study the relationship between θ_T and $\hat{\theta}_\lambda$, where the latter solves

$$\min_{\theta \in \mathbb{R}^d} \; f(\theta) + \lambda g(\theta)$$

for a regularizer g, and a tuning parameter $\lambda > 0$. By construction, the solution $\hat{\theta}_\lambda$ is subject to traditional regularization, explicitly shaped by g. In comparison, the iterate θ_T is subject to implicit regularization, as we consider iterative algorithms which enforce inductive bias along the optimization trajectory, ultimately leading to specific structured solutions in overparametrized settings. The study of iterate behavior at a finite iteration T is often called *early stopping*.

The basic inequality derived in this paper upper bounds the gap $f(\theta_T) - f(z)$ for any reference point z in terms the distances between θ_0, θ_T, and z, and the accumulated step sizes, which represent the total elapsed "time" in the optimization process. As an example, for gradient descent initialized at the origin, $\theta_0 = 0 \in \mathbb{R}^d$, and a constant step size $\eta > 0$, we have the following basic inequality:

$$f(\theta_T) - f(z) \leq \frac{1}{2\eta T} \Big(\|z\|_2^2 - \|\theta_T - z\|_2^2 \Big), \tag{1}$$

for any $z \in \mathbb{R}^d$ and $T \geq 1$. Meanwhile, for an explicitly regularized solution, we have the following inequality:

$$f(\hat{\theta}_\lambda) - f(z) \leq \lambda \Big(g(z) - g(\hat{\theta}_\lambda) \Big), \tag{2}$$

for any $z \in \mathbb{R}^d$ and $\lambda > 0$, obtained by rearranging the statement that $f(\theta) + \lambda g(\theta)$ is minimized at $\hat{\theta}_\lambda$.

The inequality in (2) is often used as a starting point in the statistical analysis of regularized estimators. For example, consider a regression loss $f(\theta) = \frac{1}{2n}\|Y - X\theta\|_2^2$ for a response $Y \in \mathbb{R}^n$ and features $X \in \mathbb{R}^{n \times d}$, and ℓ_1 penalty $g(\theta) = \|\theta\|_1$; we can instantiate (2) at the regression coefficients $z = \theta^*$ in a population linear model for $Y|X$, and rearrange further to yield

$$\frac{1}{2n}\|X\hat{\theta}_\lambda - X\theta^*\|_2^2 \leq \frac{1}{n}\langle X^\mathsf{T}\epsilon, \hat{\theta}_\lambda - \theta^*\rangle + \lambda\Big(\|\theta^*\|_1 - \|\hat{\theta}_\lambda\|_1\Big),$$

where $\epsilon = Y - X\theta^* \in \mathbb{R}^n$ is a noise vector, and $\langle a, b\rangle = a^\mathsf{T}b$. The above is referred to as the basic inequality in the high-dimensional statistics literature (Buhlmann and van de Geer, 2011), and it is a central tool for analyzing the risk of $\hat{\theta}_\lambda$, as an estimate of θ^*. Given the close resemblance between (1) and (2), one might imagine that the former can thus also be used as a tool for analyzing the risk of the early-stopped estimate θ_T. This will be a large part of the focus in our paper.

1.1 Summary of contributions

The inequality (1) as well as numerous variations on this idea can be found in the optimization literature on convergence analyses of iterative algorithms (Nesterov, 2003; Nemirovski et al., 2009; Reddi et al., 2018), and the machine learning literature on implicit regularization and generalization (Ji and Telgarsky, 2019; Ji et al., 2020; Wu et al., 2024, 2025). In this paper, we isolate and highlight a certain technique for constructing such inequalities, and show that this provides a framework for the statistical analysis of early-stopped estimates. A summary of our specific contributions is as follows.

- *A single analytical framework.* We introduce *basic inequalities* for the iterates θ_T produced by popular first-order algorithms, including gradient descent and mirror descent (Section 2). These serve as well-rounded tools for understanding implicit regularization, as detailed in below.
- *Training envelopes and dynamics.* Starting from the basic inequality, we derive a training envelope that provides lower and upper bounds on the combined loss and penalty of the iterates relative to explicitly regularized solutions (Section 2). We also show how the basic inequality enables us to characterize the asymptotic behavior of solutions (as $T \to \infty$).
- *Statistical risk bounds.* We use the basic inequality to derive high-probability bounds on the prediction risk of early-stopped iterates, and demonstrate that they match the rates obtained through traditional analysis of their explicit-regularized counterparts. Specifically, given a sample size n, dimension d, and regularity bound b, our framework yields the following results (where $\tilde{O}(\cdot)$ ignores log factors):
  - For gradient descent on generalized linear model (GLM) losses, we obtain an excess risk bound of $\tilde{O}(b\sqrt{d/n})$, matching the rate obtained by explicit ridge regularization (Section 3).
  - For exponentiated gradient descent on GLM losses, we instead obtain a bound of $\tilde{O}(\sqrt{(b \log d)/n})$, matching the rate obtained by explicit KL regularization (Section 4).
  - For exponentiated gradient descent on linear losses in a randomized aggregation problem setting, we obtain an excess risk bound of $\tilde{O}(\sqrt{(b \log d)/n})$, and we note that this also applies to explicit KL-regularized solutions because here early stopping and explicit regularization actually result in identical paths (Section 5).
- *Experiments.* Complementing our theory, we run experiments with gradient descent and exponentiated gradient descent on linear, logistic, and Poisson regression tasks. Across both underparameterized and overparameterized regimes, the results show strong empirical alignment between implicit and explicit regularization in terms of training dynamics, prediction risk, and solution paths (Section 6).

Throughout this paper, we make use of standard definitions and notation, and we provide an overview in the appendix for completeness. Furthermore, with the exception some key proofs for the basic inequalities, all proofs will be deferred to the appendix.

1.2 Related work

In what follows, as we present our results, we will then highlight and discuss the relevant literature in detail. Here we give a broader overview of work related to our paper. At a high level, the term implicit regularization refers to the phenomenon in which an optimization algorithm enforces stronger properties (typically, enforces greater regularity) than what is suggested purely by the perspective of loss minimization. A major interest in the literature on implicit regularization has been the study of solutions an asymptotic sense, i.e., as the given algorithm approaches a limit point. This idea can be traced back to the literature on boosting and the study of max-margin classifiers (Schapire et al., 1998; Rosset et al., 2004; Telgarsky, 2013).

In the deep learning age, the observation that neural network models can and will often generalize despite significant overparametrization has reinvigorated interest in implicit regularization (Neyshabur et al., 2014). In the vein of earlier boosting analyses, much progress has been made on understanding the limiting behavior of gradient descent (Soudry et al., 2018; Ji and Telgarsky, 2019; Ji et al., 2020), and mirror descent (Gunasekar et al., 2018; Sun et al., 2023).

Distinct from the asymptotic behavior of an algorithm, early stopping considers the behavior of iterates at finite but variable number of iterations T. Originally viewed as a heuristic to prevent overfitting in neural networks (Prechelt, 2002), a formal understanding began to develop, with boosting again being the primary motivating framework early on (Buhlmann and Yu, 2003; Zhang and Yu, 2005). There is by now a significant body of work connecting and comparing early-stopped gradient descent to explicit ℓ_2 (or ridge) regularization (Yao et al., 2007; Raskutti et al., 2014; Wei et al., 2017; Suggala et al., 2018; Ali et al., 2019, 2020; Zou et al., 2021; Sonthalia et al., 2024; Wu et al., 2024, 2025).

Similarly, a central focus in the current paper is to connect gradient descent to ridge regularization, and more generally, mirror descent to Bregman divergence regularization (Section 2). Our results are less refined than what can be found in the literature, which analyzes the precise behavior of descent iterates for particular problems—particular loss functions, often paired with assumptions on the data at hand. That said, our take is more general: we provide a general recipe for connecting optimization iterates and explicit regularization through what we *basic inequalities*. This is based purely on optimization-theoretic arguments, and therefore only relies on high-level properties of the loss (e.g., Lipschitz smoothness). The inequalities are deterministic, and while general and hence somewhat coarse, they are tight enough to enable meaningful corollaries about the statistical risk of early-stopped estimates in various settings (Sections 3–5). Finally, the framework for basic inequalities extends beyond gradient and mirror descent to other first-order algorithms (Section 7), and we believe will be a fruitful platform for future development.

2 Basic inequalities

This section introduces basic inequalities for gradient descent and mirror descent, and shows how they can be used to understand training dynamics. While the basic inequality for mirror descent strictly generalizes that for gradient descent, we cover gradient descent first. This is done for the sake of exposition, and also because the ℓ_2 geometry underlying gradient descent leads to some special corollaries.

2.1 Gradient descent

Given a differentiable loss function $f : \mathbb{R}^d \to \mathbb{R}$, gradient descent with an initial point $\theta_0 \in \mathbb{R}^d$, and step sizes $\{\eta_t\}_{t=0}^{\infty}$, generates iterates for $t = 0, 1, 2, \ldots$ via

$$\theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t). \tag{3}$$

Theorem 1 presents our first result, a basic inequality for gradient descent, which bounds the objective value of the iterate $\theta_T \in \mathbb{R}^d$ with respect to any reference point $z \in \mathbb{R}^d$.

Theorem 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and L-smooth, for $L > 0$. Consider gradient descent (3) with step sizes $\eta_t \in (0, 1/L]$, $t = 0, 1, 2, \ldots$. For any reference point $z \in \mathbb{R}^d$, and any iteration $T \geq 1$,*

$$f(\theta_T) - f(z) \leq \frac{1}{2\sum_{t=0}^{T-1} \eta_t} \Big(\|\theta_0 - z\|_2^2 - \|\theta_T - z\|_2^2 \Big).$$

In particular, for a constant step size $\eta_t = \eta$, $t = 0, 1, 2, \ldots$, this simplifies to

$$f(\theta_T) - f(z) \leq \frac{1}{2\eta T}\Big(\|\theta_0 - z\|_2^2 - \|\theta_T - z\|_2^2\Big). \tag{4}$$

Proof. The proof can be broken down into three steps.

Step 1: *Tracking the proximity difference across adjacent iterations.* Measuring proximity via the Euclidean distance, note that

$$\|\theta_t - z\|_2^2 - \|\theta_{t+1} - z\|_2^2 = \|\theta_t - z\|_2^2 - \|\theta_t - \eta_t \nabla f(\theta_t) - z\|_2^2 = 2\eta_t \langle \nabla f(\theta_t), \theta_t - z\rangle - \eta_t^2 \|\nabla f(\theta_t)\|_2^2.$$

Step 2: *Bounding the objective difference $f(\theta_t) - f(z)$.* By convexity, we have $f(\theta_t) - f(z) \leq \langle \nabla f(\theta_t), \theta_t - z\rangle$. Substituting this into the result from Step 1,

$$2\eta_t\Big(f(\theta_t) - \frac{\eta_t}{2}\|\nabla f(\theta_t)\|_2^2 - f(z)\Big) \leq \|\theta_t - z\|_2^2 - \|\theta_{t+1} - z\|_2^2.$$

By L-smoothness of f and $\eta_t \in (0, 1/L]$, the descent lemma (Lemma B1) guarantees

$$f(\theta_{t+1}) \leq f(\theta_t) - \eta_t\Big(1 - \frac{L}{2}\eta_t\Big)\|\nabla f(\theta_t)\|_2^2 \leq f(\theta_t) - \frac{\eta_t}{2}\|\nabla f(\theta_t)\|_2^2.$$

This ensures $f(\theta_T) \leq f(\theta_t) - (\eta_t/2)\|\nabla f(\theta_t)\|_2^2$ for $t < T$. Using this in the second-to-last display,

$$2\eta_t(f(\theta_T) - f(z)) \leq \|\theta_t - z\|_2^2 - \|\theta_{t+1} - z\|_2^2.$$

Step 3: *Aggregating bounds over iterations.* Summing the result of Step 2 over $t < T$ gives a telescoping sum:

$$2\sum_{t=0}^{T-1} \eta_t(f(\theta_T) - f(z)) \leq \|\theta_0 - z\|_2^2 - \|\theta_T - z\|_2^2,$$

which completes the proof after rearrangement. □

Some readers will see (4) as a familiar result, as it is just a few steps away from a "textbook" result for gradient descent. To derive the standard $1/T$ convergence rate on the suboptimality gap of gradient descent under Lipschitz smoothness, we can set z to be a minimizer of f, and then drop the negative term involving $\|\theta_T - z\|_2^2$ on the right-hand side. This gives the familiar bound $f(\theta_T) - f^\star \leq \frac{b}{2\eta T}$, where $f^\star$ is the optimal objective value, and b bounds the distance between θ_0 and the solution set.

Inspired by use of similar techniques in Ji and Telgarsky (2019); Ji et al. (2020); Wu et al. (2024, 2025), in (4) we recognize the importance of keeping $z \in \mathbb{R}^d$ as a free variable, and maintaining both initial $\|\theta_0 - z\|_2^2$ and current $\|\theta_T - z\|_2^2$ distances in the bound. Though these are seemingly simple choices, the specific form of (4) will lead to numerous insights on the behavior of gradient descent iterates, as we will see in what follows. As our first example, we show how the basic inequality (4) leads to a ridge-regularized training envelope for gradient descent iterates.

Corollary 1. *Under the assumptions of Theorem 1, consider gradient descent (3) with a constant step size $\eta_t = \eta \in (0, 1/L]$, $t = 0, 1, 2, \ldots$. Then for any $T \geq 1$ and $\lambda_T = 1/(\eta T)$,*

$$\min_{z \in \mathbb{R}^d}\Big(f(z) + \frac{\lambda_T}{4}\|\theta_0 - z\|_2^2\Big) \leq f(\theta_T) + \frac{\lambda_T}{4}\|\theta_0 - \theta_T\|_2^2 \leq \min_{z \in \mathbb{R}^d}\Big(f(z) + \lambda_T\|\theta_0 - z\|_2^2\Big). \tag{5}$$

Proof. Observe that

$$\begin{aligned}
\|\theta_0 - z\|_2^2 - \|\theta_T - z\|_2^2 &= 2\langle \theta_0 - z, \theta_0 - \theta_T\rangle - \|\theta_0 - \theta_T\|_2^2 \\
&\leq 2\|\theta_0 - z\|_2\|\theta_0 - \theta_T\|_2 - \|\theta_0 - \theta_T\|_2^2 \\
&\leq 2\|\theta_0 - z\|_2^2 - \frac{1}{2}\|\theta_T - \theta_0\|_2^2,
\end{aligned}$$

where the last line uses Young's inequality, $2ab \leq ca^2 + b^2/c$, with $c = 1/2$. Plugging this into (4), and then minimizing over $z \in \mathbb{R}^d$, gives the upper bound in (5). The lower bound is immediate. □

The result in (5) translates the "elapsed time" ηT in gradient descent to an effective ridge regularization parameter λ_T. For simplicity, set $\theta_0 = 0$, and let $\hat{\theta}_\lambda \in \mathbb{R}^d$ be the (unique) minimizer of the ridge-penalized criterion $f(z) + \lambda\|z\|_2^2$ over $z \in \mathbb{R}^d$. Rephrased, the result in (5) says that for all $T \geq 1$,

$$f(\hat{\theta}_{\lambda_T/4}) + \frac{\lambda_T}{4}\|\hat{\theta}_{\lambda_T/4}\|_2^2 \leq f(\theta_T) + \frac{\lambda_T}{4}\|\theta_T\|_2^2 \leq f(\hat{\theta}_{\lambda_T}) + \lambda_T\|\hat{\theta}_{\lambda_T}\|_2^2.$$

In other words, while (by definition) the iterate θ_T must be suboptimal with respect to the ridge-penalized criterion at a penalty factor $\lambda_T/4$, its achieved criterion value is no more than the optimal ridge-penalized criterion at an inflated penalty factor λ_T. These lower and upper bounds together provide an envelope for the ridge-penalized criteria achieved by gradient descent across the full path, providing interesting evidence of the implicit ℓ_2 regularity present in the gradient descent iterates. Notably, this is true for any convex and Lipschitz smooth loss f.

The next result demonstrates how the basic inequality is powerful enough to recover various standard facts about the training dynamics of gradient descent. Our approach and presentation here is inspired by the work of Lemaire (1996) on gradient flow.

Corollary 2. *Under the assumptions of Theorem 1, consider gradient descent (3) with step sizes $\eta_t \in (0, 1/L]$, $t = 0, 1, 2, \ldots$. The following holds, where $\Pi_S(u) = \arg\min_{s \in S} \|u - s\|_2$ denotes the projection of a point u onto a set S, and $\mathrm{dist}_S(u) = \min_{s \in S} \|u - s\|_2$ denotes the distance from u to S.*

1. *(Objective convergence.) Let $f^\star = \inf_{\theta \in \mathbb{R}^d} f(\theta) \in [-\infty, \infty)$. If $\sum_{t=0}^\infty \eta_t = \infty$, then $\lim_{t \to \infty} f(\theta_t) = f^\star$.*
2. *(Nonincreasing distance to solution set.) Define the solution set $S = \{\theta \in \mathbb{R}^d : f(\theta) = f^\star\}$. Notice that S is closed and convex. If $S \neq \emptyset$, then*

$$\{\|\theta_t - s\|_2\}_{t=0}^\infty \text{ is nonincreasing for any } s \in S, \quad \text{and thus,} \quad \{\mathrm{dist}_S(\theta_t)\}_{t=0}^\infty \text{ is nonincreasing.}$$

3. *(Iterate convergence.) If $S \neq \emptyset$ and $\sum_{t=0}^\infty \eta_t = \infty$, then $\lim_{t \to \infty} \theta_t = \theta_\infty \in S$. Furthermore,*

$$\|\theta_\infty - \Pi_S(\theta_0)\|_2 \leq \mathrm{dist}_S(\theta_0), \quad \text{and thus,} \quad \|\theta_\infty - \theta_0\|_2 \leq 2 \cdot \mathrm{dist}_S(\theta_0).$$

4. *(Minimum norm solution.) If $S \neq \emptyset$, $\sum_{t=0}^\infty \eta_t = \infty$, and S is affine, then $\theta_\infty = \Pi_S(\theta_0)$.*

Proof. We prove each part separately.

Part 1. By the first result in Theorem 1, we know that $f(\theta_T) \leq f(z) + \|\theta_0 - z\|_2^2/(2\sum_{t=0}^{T-1} \eta_t)$ for any $z \in \mathbb{R}^d$. Since $\sum_{t=0}^\infty \eta_t = \infty$, we have $\limsup_{T \to \infty} f(\theta_T) \leq f(z)$, and taking an infimum over z gives the result.

Part 2. By the first result in Theorem 1, for any $s \in S$,

$$\|\theta_T - s\|_2^2 \leq \|\theta_0 - s\|_2^2 + 2\bigg(\sum_{t=0}^{T-1} \eta_t\bigg)(f(s) - f(\theta_T)) \leq \|\theta_0 - s\|_2^2,$$

where the last inequality is due to $f(s) \leq f(\theta_T)$. Now, θ_T can also be seen as the result of running gradient descent from an initial point θ_t for $T - t$ iterations, and hence by the same argument $\|\theta_T - s\|_2 \leq \|\theta_t - s\|_2$, which proves the first claim. Taking an infimum over $s \in S$ proves the second claim.

Part 3. Since $\{\|\theta_t - s\|_2\}_{t=0}^\infty$ is nonincreasing and bounded below by zero, the Bolzano-Weierstrass theorem implies that there exists a subsequence which converges, i.e., $\lim_{k \to \infty} \theta_{t_k} = \theta_\infty$. By part 1 and continuity of f, we know that $\theta_\infty \in S$. Then once again as $\{\|\theta_t - s\|_2\}_{t=0}^\infty$ is nonincreasing, we know that $\lim_{t \to \infty} \theta_t = \theta_\infty$, the first claim. The second claim is a follows from $\{\|\theta_t - s\|_2\}_{t=0}^\infty$ being nonincreasing for $s = \Pi_S(\theta_0)$.

Part 4. Let $p = \Pi_S(\theta_0)$ and $v = p - \theta_\infty$. Assume $v \neq 0$. For any $c \geq 0$, define $\beta_c = p + c \cdot \mathrm{dist}_S(\theta_0) \cdot v/\|v\|_2$. As S is affine, we know that $\beta_c \in S$. Then by parts 2 and 3, we know that $\|\theta_\infty - \beta_c\|_2 \leq \|\theta_0 - \beta_c\|_2$. Since the three points $\beta_c, \theta_\infty, p$ are collinear by construction,

$$\|\theta_\infty - \beta_c\|_2 = \|\theta_\infty - p\|_2 + \|p - \beta_c\|_2 = \|v\|_2 + c \cdot \mathrm{dist}_S(\theta_0).$$

Meanwhile, by the Pythagorean theorem,

$$\|\theta_0 - \beta_c\|_2^2 = \|\theta_0 - p\|_2^2 + \|p - \beta_c\|_2^2 = \mathrm{dist}_S(\theta_0)^2 + (c \cdot \mathrm{dist}_S(\theta_0))^2 = (1 + c^2) \cdot \mathrm{dist}_S(\theta_0)^2.$$

Therefore, we finally have

$$\|v\|_2 + c \cdot \mathrm{dist}_S(\theta_0) = \|\theta_\infty - \beta_c\|_2 \leq \|\theta_0 - \beta_c\|_2^2 = \sqrt{1 + c^2} \cdot \mathrm{dist}_S(\theta_0).$$

Thus $\|v\|_2 \leq (\sqrt{1 + c^2} - c) \cdot \mathrm{dist}_S(\theta_0)$. Taking $c \to \infty$ implies $\|v\|_2 = 0$, which means $p = \theta_\infty$. □

We note that part 4 of the above collorary reproduces a well-known "folklore" result in overparameterized linear regression: gradient descent initialized at the origin converges to the minimum ℓ_2 norm solution. This also covers any loss f that depends on θ only through a linear transformation $X\theta$, such as generalized linear model losses (as introduced formally in Section 3).

2.2 Mirror descent

Mirror descent is a generalization of gradient descent where Euclidean distance is replaced by the *Bregman divergence* induced by a convex, differentiable function $\phi : \Omega \to \mathbb{R}$. For points $u, v \in \mathrm{int}(\Omega)$ (the interior of the domain Ω, or relative interior in the case that Ω is not full-dimensional), this is defined by

$$D_\phi(u, v) = \phi(u) - \phi(v) - \langle \nabla\phi(v), u - v \rangle,$$

For a constraint $\mathcal{C} \subseteq \Omega$, an initial point $\theta_0 \in \mathcal{C} \cap \mathrm{int}(\Omega)$, and step sizes $\{\eta_t\}_{t=0}^\infty$, the mirror descent update is

$$\theta_{t+1} = \operatorname*{arg\,min}_{\theta \in \mathcal{C}} \Big(\eta_t \langle \nabla f(\theta_t), \theta \rangle + D_\phi(\theta, \theta_t) \Big). \tag{6}$$

This has an equivalent two-stage formulation, where we first perform a gradient step in a transformed space given by the "mirror map" $\nabla\phi$, and then take a Bregman projection onto $\mathcal{C}$:

$$\nabla\phi(\tilde{\theta}_{t+1}) = \nabla\phi(\theta_t) - \eta_t \nabla f(\theta_t), \qquad \theta_{t+1} = \operatorname*{arg\,min}_{\theta \in \mathcal{C}} D_\phi(\theta, \tilde{\theta}_{t+1}).$$

A popular example of mirror descent is the exponentiated gradient descent algorithm, which is designed for optimization over the probability simplex, $\Delta_d = \{a \in \mathbb{R}^d : a_i \geq 0, \sum_{i=1}^d a_i = 1\}$. By choosing the negative entropy function $\phi(u) = \sum_{i=1}^d u_i \log u_i$, which is 1-strongly convex with respect to $\|\cdot\|_1$, on $\Omega = \mathcal{C} = \Delta_d$, the induced Bregman divergence becomes the Kullback-Leibler (KL) divergence, $D_{\mathrm{KL}}(a, b) = \sum_{i=1}^d a_i \log(a_i/b_i)$. The corresponding mirror descent update, where $\odot$ denotes coordinatewise multiplication, is

$$\tilde{\theta}_{t+1} = \theta_t \odot \exp(-\eta_t \nabla f(\theta_t)), \qquad \theta_{t+1} = \tilde{\theta}_{t+1} / \|\tilde{\theta}_{t+1}\|_1. \tag{7}$$

Mirror descent can also recover projected gradient descent with constraint set $\mathcal{C}$, which of course reduces to ordinary gradient descent when $\mathcal{C} = \mathbb{R}^d$. Taking $\phi(u) = \frac{1}{2}\|u\|_2^2$, the induced Bregman divergence is simply $D_\phi(u, v) = \frac{1}{2}\|u - v\|_2^2$, and the mirror descent update is

$$\theta_{t+1} = \Pi_C(\theta_t - \eta_t \nabla f(\theta_t)), \tag{8}$$

where recall Π_C denotes Euclidean projection onto $\mathcal{C}$.

Below we introduce a basic inequality for mirror descent iterates in (6). Just as mirror descent generalizes gradient descent, Theorem 2 generalizes Theorem 1.

Theorem 2. *Let $\Omega, \mathcal{C}$ be closed, convex sets in $\mathbb{R}^d$ with nonempty interior, and $\mathcal{C} \subseteq \Omega$. Assume $f : \Omega \to \mathbb{R}$ is convex on $\mathcal{C}$, and L-smooth with respect a norm $\|\cdot\|$ on $\mathcal{C} \cap \mathrm{int}(\Omega)$, for $L > 0$. Assume $\phi : \Omega \to \mathbb{R}^d$ is of Legendre type, and α-strongly convex with respect to $\|\cdot\|$ on Ω, for $\alpha > 0$. Consider mirror descent (6) with step sizes $\eta_t \in (0, \alpha/L]$, $t = 0, 1, 2, \ldots$. For any reference point $z \in \mathcal{C}$, and any $T \geq 1$,*

$$f(\theta_T) - f(z) \leq \frac{1}{\sum_{t=0}^{T-1} \eta_t} \Big(D_\phi(z, \theta_0) - D_\phi(z, \theta_T) \Big).$$

In particular, for a constant step size $\eta_t = \eta$, $t = 0, 1, 2, \ldots$, this simplifies to

$$f(\theta_T) - f(z) \leq \frac{1}{\eta T} \Big(D_\phi(z, \theta_0) - D_\phi(z, \theta_T) \Big). \tag{9}$$

Proof. The proof parallels that from the gradient descent case.

Step 1: *Tracking the proximity difference across adjacent iterations.* Measuring proximity via the Bregman divergence, the three-point identity for the Bregman divergence states that

$$D_\phi(z, \theta_{t+1}) + D_\phi(\theta_{t+1}, \theta_t) - D_\phi(z, \theta_t) = \langle \nabla\phi(\theta_t) - \nabla\phi(\theta_{t+1}), z - \theta_{t+1} \rangle.$$

Note $\langle \eta_t \nabla f(\theta_t) + \nabla\phi(\theta_{t+1}) - \nabla\phi(\theta_t), z - \theta_{t+1} \rangle \geq 0$, by the first-order optimality condition for θ_{t+1} in (6). Rearranging terms, this implies $\eta_t \langle \nabla f(\theta_t), \theta_{t+1} - z \rangle \leq \langle \nabla\phi(\theta_{t+1}) - \nabla\phi(\theta_t), z - \theta_{t+1} \rangle$, and combining with the above identity,

$$\eta_t \langle \nabla f(\theta_t), \theta_{t+1} - z \rangle \leq D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}) - D_\phi(\theta_{t+1}, \theta_t).$$

Step 2: *Bounding the objective difference* $f(\theta_t) - f(z)$. By convexity,

$$f(\theta_t) - f(z) \leq \langle \nabla f(\theta_t), \theta_t - z \rangle = \langle \nabla f(\theta_t), \theta_t - \theta_{t+1} \rangle + \langle \nabla f(\theta_t), \theta_{t+1} - z \rangle.$$

Combining this with the result of Step 1, we have

$$\eta_t (f(\theta_t) - f(z)) \leq \eta_t \langle \nabla f(\theta_t), \theta_t - \theta_{t+1} \rangle + D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}) - D_\phi(\theta_{t+1}, \theta_t).$$

Meanwhile, the L-smoothness of f gives $f(\theta_{t+1}) \leq f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t \rangle + (L/2)\|\theta_{t+1} - \theta_t\|^2$, and the α-strong convexity of ϕ implies $D_\phi(\theta_{t+1}, \theta_t) \geq (\alpha/2)\|\theta_{t+1} - \theta_t\|^2$. Substituting these into the above display,

$$\begin{aligned} \eta_t (f(\theta_{t+1}) - f(z)) &\leq D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}) - \Big(\frac{\alpha}{2} - \frac{L}{2}\eta_t\Big)\|\theta_{t+1} - \theta_t\|^2 \\ &\leq D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}), \end{aligned}$$

where we use $\eta_t \leq \alpha/L$ in the last inequality. The descent lemma for mirror descent (Lemma B2) shows that $f(\theta_t)$ is nonincreasing in t, thus we have

$$\eta_t (f(\theta_T) - f(z)) \leq D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}).$$

Step 3: *Aggregating bounds over iterations.* Summing the result of Step 2 over $t < T$ gives a telescoping sum:

$$\sum_{t=0}^{T-1} \eta_t (f(\theta_T) - f(z)) \leq D_\phi(z, \theta_0) - D_\phi(z, \theta_T),$$

which completes the proof after rearrangement. □

Next we present the mirror descent analog of Corollary 1, on a regularized training envelope.

Corollary 3. *Under the assumptions of Theorem 2, consider mirror descent (6) with a constant step size $\eta_t = \eta \in (0, \alpha/L]$, $t = 0, 1, 2, \ldots$. Further, assume that there exists $G > 0$ such that $D_\phi(z, \theta_0) \leq \frac{G}{2}\|\theta_0 - z\|^2$ for any $z \in \mathcal{C}$ (note that this is implied by ϕ being G-smooth with respect to $\|\cdot\|$ on $\mathcal{C}$). Then for any $T \geq 1$ and $\lambda_T = 1/(\eta T)$,*

$$\min_{z \in \mathcal{C}} \Big(f(z) + \frac{\alpha\lambda_T}{4}\|\theta_0 - z\|^2 \Big) \leq f(\theta_T) + \frac{\alpha\lambda_T}{4}\|\theta_0 - \theta_T\|^2 \leq \min_{z \in \mathcal{C}} \Big(f(z) + \frac{(G+\alpha)\lambda_T}{2}\|\theta_0 - z\|^2 \Big). \tag{10}$$

Proof. Observe that

$$\begin{aligned} D_\phi(z, \theta_0) - D_\theta(z, \theta_T) &\leq \frac{G}{2}\|\theta_0 - z\|^2 - \frac{\alpha}{2}\|\theta_T - z\|^2 \\ &\leq \frac{G}{2}\|\theta_0 - z\|^2 - \frac{\alpha}{2}\Big(\frac{1}{2}\|\theta_T - \theta_0\|^2 - \|\theta_0 - z\|^2\Big) \\ &= \frac{G+\alpha}{2}\|\theta_0 - z\|_2^2 - \frac{\alpha}{4}\|\theta_T - \theta_0\|_2^2, \end{aligned}$$

where the first line uses the given assumption on ϕ combined with α-strong convexity, and the second uses $\frac{1}{2}\|\theta_T - \theta_0\|^2 \leq \|\theta_T - z\|^2 + \|\theta_0 - z\|^2$, based on the triangle inequality and Young's inequality $2ab \leq a^2 + b^2$. We can plug this into (9), and minimize over $z \in \mathcal{C}$, to derive the upper bound in (10). As before, the lower bound is immediate. □

We remark that (10) recovers the previous gradient descent result (10) when we take $\phi = \frac{1}{2}\|\cdot\|_2^2$, which implies $G = \alpha = 1$. In this case, we have a tight envelope, with a 1:4 ratio of the penalty factors used in the comparison between the regularized criteria achieved by gradient descent (penalty factor $\lambda_T/4$) and an optimal solution (penalty factor λ_T). In general, the envelope in (10) admits a ratio of $1{:}(2G/\alpha + 2)$, which can be loose when G is large relative to α. In particular, for exponentiated gradient descent (7) the function ϕ is the negative entropy, in which case we have smoothness and strong convexity parameters $G = d$ and $\alpha = 1$ with respect to the ℓ_1 norm. This leads to a ratio of $1{:}(2d + 2)$, and therefore the upper bound in (10) quickly becomes informative in high dimensions. Our empirical results in Section 6 suggest that this is overly conservative in practice. Furthermore, using a reverse Pinsker-type inequality for the KL divergence (Sason, 2015), we can derive an alternative upper bound that (in certain regimes) has a better dimension dependence. This is presented below, and proved in the appendix.

Corollary 4. *Under the assumptions of Theorem 2, consider now exponentiated gradient descent (7) with a constant step size $\eta_t = \eta \in (0, 1/L]$, $t = 0, 1, 2, \ldots$ and initial point $\theta_0 = \pi$, where $\pi = (1/d, \ldots, 1/d) \in \Delta_d$ is the uniform distribution. Then for any $T \geq 1$ and $\lambda_T = 1/(\eta T)$,*

$$\min_{z \in \Delta_d} \Big(f(z) + \frac{\lambda_T}{4}\|\pi - z\|_1^2 \Big) \leq f(\theta_T) + \frac{\lambda_T}{4}\|\pi - \theta_T\|_1^2 \\ \leq \min_{z \in \Delta_d} \Big(f(z) + \lambda_T \cdot \min\Big\{ \frac{d+1}{2}\|\pi - z\|_1^2,\ \frac{1}{2}\|\pi - z\|_1^2 + \frac{\log d}{2(1-1/d)}\|\pi - z\|_1 \Big\} \Big). \quad (11)$$

Under mild conditions, we can reformulate the bound in this corollary using a pure ℓ_1 penalty. Provided we run enough iterations of exponentiated gradient descent so that θ_T is a bounded away from the uniform distribution π, i.e., $\|\theta_T - \pi\|_1 \geq c$ for any constant $c > 0$, we can use this lower bound in (11) and the simple upper bound $\|x - y\|_1 \leq 2$ for any $x, y \in \Delta_d$, to obtain

$$\min_{z \in \Delta_d} \Big(f(z) + \frac{c\lambda_T}{4}\|\pi - z\|_1 \Big) \leq f(\theta_T) + \frac{c\lambda_T}{4}\|\pi - \theta_T\|_1 \leq \min_{z \in \Delta_d} \Big(f(z) + \lambda_T \Big[\frac{\log d}{2(1-1/d)} + 1 \Big] \|\pi - z\|_1 \Big\} \Big). \quad (12)$$

The envelope in (12) clearly has a much better dimension dependence, with a ratio of $1{:}O(\log d)$ between the penalty factors for exponentiated gradient descent and an optimal (ℓ_1-regularized) solution.

Lastly we present the mirror descent analog of Corollary 2, on training dynamics. The proof is overall similar to that of Corollary 2, and deferred to the appendix.

Corollary 5. *Under the assumptions of Theorem 2, consider mirror descent (6) with step sizes $\eta_t \in (0, \alpha/L]$, $t = 0, 1, 2, \ldots$. The following holds, where $\Pi_S^\phi(u) = \arg\min_{s \in S} D_\phi(s, u)$ denotes the Bregman projection of a point u onto a set S, and $\mathrm{dist}_S^\phi(u) = \min_{s \in S} D_\phi(s, u)$ denotes the Bregman distance from u to S.*

1. *(Objective convergence.) Let $f^\star = \inf_{\theta \in \mathbb{R}^d} f(\theta) \in [-\infty, \infty)$. If $\sum_{t=0}^\infty \eta_t = \infty$, then $\lim_{t \to \infty} f(\theta_t) = f^\star$.*
2. *(Nonincreasing distance to solution set.) Define the solution set $S = \{\theta \in \mathcal{C} : f(\theta) = f^\star\}$. Notice that S is closed and convex. If $S \neq \emptyset$, then*

$$\{D_\phi(s, \theta_t)\}_{t=0}^\infty \text{ is nonincreasing for any } s \in S, \quad \text{and thus,} \quad \{\mathrm{dist}_S^\phi(\theta_t)\}_{t=0}^\infty \text{ is nonincreasing.}$$

3. *(Iterate convergence.) Suppose $S \neq \emptyset$ and $\sum_{t=0}^\infty \eta_t = \infty$. If either $S \cap \mathrm{int}(\Omega) \neq \emptyset$, or D_ϕ is continuous in its second argument relative to $\mathrm{int}(\Omega)$[1], then $\lim_{t \to \infty} \theta_t = \theta_\infty \in S$.*
4. *(Minimum Bregman divergence solution.) If $S \neq \emptyset$, $\sum_{t=0}^\infty \eta_t = \infty$, and S is affine with $S \subseteq \mathrm{int}(\Omega)$, then $\theta_\infty = \Pi_S^\phi(\theta_0)$.*

We note that that part 4 of the above collorary reproduces a result of Gunasekar et al. (2018); this covers losses f which depend on θ only through a linear transformation $X\theta$, such as generalized linear model losses, which we study in detail next.

[1]To be precise, this means that for every sequence $\{y_k\}_{k=1}^\infty \subseteq \mathrm{int}(\Omega)$ such that $y_k \to y$, we have $D_\phi(y, y_k) \to 0$. Note this is not generally true for ϕ of Legendre type, see, e.g., Remark 3.4 and Example 7.32 in Bauschke and Borwein (1997).

3 Generalized linear models and ridge regularity

We now shift to a statistical perspective to analyze the prediction risk of the iterates. This section focuses on generalized linear models (GLMs) and gives a comparative analysis of two regularization schemes: implicit via gradient descent and explicit via ridge regularization.

We adopt a standard GLM setup with an identity sufficient statistic, where $Y \in \mathbb{R}^n$ denotes the response vector and $X \in \mathbb{R}^{n \times d}$ the feature matrix. The negative log-likelihood function is thus (after rescaling by $\frac{1}{n}$):

$$f(\theta) = \frac{1}{n}\Big(- Y^\top X\theta + \mathsf{A}(X\theta)\Big). \tag{13}$$

for a map $\mathsf{A} : \mathbb{R}^n \to \mathbb{R}$ which acts coordinatewise, as in $\mathsf{A}(v) = \sum_{i=1}^n A(v_i)$. Here $A : \mathbb{R} \to \mathbb{R}$ is the cumulant generating function of the univariate exponential family associated with the given GLM. Key examples are $A(u) = u^2/2$ for the Gaussian distribution (linear regression), $A(u) = \log(1+e^u)$ for the Bernoulli distribution (logistic regression), and $A(u) = e^u$ for the Poisson distribution (Poisson regression). An important property is that, for any exponential family, the cumulant generating function A is convex.

In what follows, we consider a fixed-X analysis of prediction risk in GLMs: we generally treat $X \in \mathbb{R}^{n \times d}$ as fixed (nonrandom). We denote by P_i the distribution of each response Y_i, and write

$$\mu_i = \mathbb{E}[Y_i], \quad \text{and} \quad \epsilon_i = Y_i - \mu_i.$$

We allow for model misspecification in the sense that we do not require the mean μ_i to follow the canonical link of the given GLM. That is, when analyzing linear regression we do not require the mean μ_i to be linear in x_i (the i^{th} row of X), when analyzing logistic regression we do not require the log odds $\log(\mu_i/(1-\mu_i))$ to be linear in x_i, and so on. We define the *prediction risk* (or simply risk) of an estimator $\hat{\theta} = \hat{\theta}(X, Y)$ as the expected negative log-likelihood evaluated at a test copy $Y^* \in \mathbb{R}^n$ of the training response vector, where each $Y_i^* \sim P_i$ (independent of Y):

$$\mathrm{Risk}(\hat{\theta}) = \frac{1}{n}\mathbb{E}\Big[- (Y^*)^\top X\hat{\theta} + \mathsf{A}(X\hat{\theta}) \,\Big|\, Y\Big]. \tag{14}$$

An important observation is that $\mathrm{Risk}(\hat{\theta}) = \frac{1}{n}(-\mu^\top X\hat{\theta} + \mathsf{A}(X\hat{\theta}))$, and we can therefore add and subtract ϵ to μ and expand to obtain:

$$\mathrm{Risk}(\hat{\theta}) = f(\hat{\theta}) + \frac{1}{n}\epsilon^\top X\hat{\theta}.$$

This expresses the prediction risk as the sum of the training error and a stochastic term linear in the noise. This structure will allow us to apply the basic inequality to derive high-probability risk bounds.

3.1 Risk analysis: ridge-regularized solution

The ridge-regularized GLM estimator is defined by augmenting the GLM loss (13) with a squared ℓ_2 penalty, denoted

$$\hat{\theta}_\lambda = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \Big(f(\theta) + \lambda\|\theta\|_2^2 \Big). \tag{15}$$

for a regularization parameter $\lambda > 0$. (The solution exists and is unique as the objective is strongly convex.) The purpose of this subsection is to obtain relatively straightforward but nonetheless meaningful risk bounds for ridge-regularized GLM estimates, in order to compare what is achievable for gradient descent estimates from the basic inequality, in the next subsection.

As we will see, the risk bounds for explicitly- and implicitly-regularized estimates will be comparable, and further, the strategies we use to derive them will be similar. As a jumping off point for the ridge analysis, by definition of $\hat{\theta}_\lambda$ in (15), we know that for any $\theta \in \mathbb{R}^d$,

$$f(\hat{\theta}_\lambda) + \lambda\|\hat{\theta}_\lambda\|_2^2 \le f(\theta) + \lambda\|\theta\|_2^2.$$

This can be rewritten as

$$f(\hat{\theta}_\lambda) - f(\theta) \le \lambda\Big(\|\theta\|_2^2 - \|\hat{\theta}_\lambda\|_2^2 \Big).$$

Adding $\frac{1}{n}\epsilon^\top X(\hat{\theta}_\lambda - \theta)$ to each side, and recalling the decomposition of risk in (14), we get

$$\text{Risk}(\hat{\theta}_\lambda) - \text{Risk}(\theta) \leq \frac{1}{n}\epsilon^\top X(\hat{\theta}_\lambda - \theta) + \lambda\Big(\|\theta\|_2^2 - \|\hat{\theta}_\lambda\|_2^2\Big). \tag{16}$$

The next proposition develops this further to provide more salient deterministic and high-probability bounds on the excess risk. Here and henceforth, a random variable Z is said to be sub-Gaussian with parameter σ^2, written $Z \sim \text{sG}(\sigma^2)$, if $\mathbb{E}[\exp(t(Z - \mathbb{E}[Z]))] \leq \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$.

Proposition 1. *For any $\lambda > 0$ and any reference point $\theta \in \mathbb{R}^d$, the prediction risk of $\hat{\theta}_\lambda$ in (15) satisfies*

$$\text{Risk}(\hat{\theta}_\lambda) - \text{Risk}(\theta) \leq \frac{1}{2\lambda}\Big\|\frac{X^\top\epsilon}{n}\Big\|_2^2 + 2\lambda\|\theta\|_2^2.$$

Furthermore, assume that each $\epsilon_i \sim \text{sG}(\sigma_i^2)$. Let $\sigma^2 = \max_{i=1,\ldots,n}\sigma_i^2$ and $\widehat{\Sigma} = \frac{1}{n}X^\top X$. Then, for any $\delta > 0$ and $b > 0$, choosing

$$\lambda = \frac{\sigma}{2b\sqrt{n}}\sqrt{\text{tr}(\widehat{\Sigma}) + 2\sqrt{\delta}\|\widehat{\Sigma}\|_F + 2\delta\|\widehat{\Sigma}\|_{\text{op}}}, \tag{17}$$

the excess risk of $\hat{\theta}_\lambda$ with respect the class of parameters with ℓ_2 norm at most b satisfies

$$\text{Risk}(\hat{\theta}_\lambda) - \min_{\theta:\|\theta\|_2\leq b}\text{Risk}(\theta) \leq \frac{2b\sigma}{\sqrt{n}}\sqrt{\text{tr}(\widehat{\Sigma}) + 2\sqrt{\delta}\|\widehat{\Sigma}\|_F + 2\delta\|\widehat{\Sigma}\|_{\text{op}}}, \tag{18}$$

with probability at least $1 - e^{-\delta}$.

The bound (18) is governed by the spectral properties of the empirical covariance $\widehat{\Sigma}$. Under the scaling

$$\text{tr}(\widehat{\Sigma}) = O(d), \quad \|\widehat{\Sigma}\|_F = O(\sqrt{d}), \quad \text{and} \quad \|\widehat{\Sigma}\|_{\text{op}} = O(1), \tag{19}$$

which holds, e.g., with high probability when X has independent sub-Gaussian entries (Vershynin, 2018), we can choose $\delta = \log n$ to yield an excess risk bound of $\tilde{O}(b\sigma\sqrt{d/n})$ with probability at least $1 - 1/n$.

Next we give our main result for the ridge-penalized GLM estimator. This is accomplshed by tailoring the general excess risk bound in Proposition 1 to specific GLMs by identifying the sub-Gaussian parameter for the noise ϵ_i. Notably, Poisson regression requires a truncation argument due to the exponential nature of the noise tail, resulting in a slightly weaker guarantee than that for linear and logistic regression.

Theorem 3. *Consider the ridge-regularized GLM estimator $\hat{\theta}_\lambda$ in (15) with $\lambda > 0$. Consider the following cases, which determine the response distribution for each $i = 1, \ldots, n$:*

1. *Gaussian (linear regression): $P_i = \mathcal{N}(\mu_i, \sigma_i^2)$.*
2. *Bernoulli (logistic regression): $P_i = \text{Bern}(\mu_i)$.*
3. *Poisson (Poisson regression): $P_i = \text{Pois}(\mu_i)$.*

In case 1, we define $\sigma = \max_{i=1,\ldots,n}\sigma_i$; in case 2, $\sigma = \frac{1}{2}$; and in case 3, $\sigma = (6\|\mu\|_\infty + 2)\log n + 3\sqrt{\|\mu\|_\infty}$. Then, for any $\delta > 0$ and $b > 0$, choosing λ as in (17) yields the excess risk bound in (18), which holds with probability at least $1 - e^{-\delta}$ in cases 1 and 2, and at least $1 - e^{-\delta} - 1/n$ in case 3 provided $\delta \geq 1$ and $n \geq 3$.

As before, under the spectral assumptions on $\widehat{\Sigma}$ in (19), the result in Theorem 3 provides an excess risk bound of $\tilde{O}(b\sigma\sqrt{d/n})$ with high probability. An important feature of the theorem is its robustness to model misspecification. To reiterate, the mean μ_i under the response distribution P_i need not follow the canonical link of the given GLM. The analysis solely relies on sub-Gaussian tail bounds for ϵ_i (and in the Poisson case, an additional truncation argument).

As explained above, the intent of Theorem 3 is not to provide the sharpest possible results on excess risk in regularized GLMs. It is instead to provide a reasonable benchmark to which we can compare risk bounds for early-stopped gradient descent iterates in the next subsection. Still, it is worth a brief discussion on how the results in Theorem 3 compares to what we should expect given the existing literature.

Comparison with existing literature. There is of course a huge amount of literature on analyzing risk for regularized GLMs, and we only provide a brief overview of some of the most relevant points of comparison. First, taking a broader perspective, an excess prediction risk rate of $\sqrt{d/n}$ is standard in statistical learning theory. Such rates—and generalizations thereof—are to be expected, with the feature dimension d replaced by a suitable notion of complexity such as VC dimension (Wainwright, 2019).

From the point of view of classical asymptotic theory for maximum likelihood, one would expect a rate of d/n for estimating GLM parameters in well-specified settings, though this would be in an asymptotic regime with d fixed and $n \to \infty$ (Wasserman, 2004). Moving to finite-sample prediction risk analyses for GLMs, in some cases, a rate of d/n is achievable. For example, in well-specified linear regression without regularization, a standard calculation yields an expected excess prediction risk of $\sigma^2 d/(2n)$. For logistic regression, the story is more nuanced. Using sophisticated analyses that leverage the self-corcondant property of the logistic loss, rates "close" to d/n have been established in a variety of different settings, including misspecified ones (Bach, 2010; Ostrovskii and Bach, 2021). For Poisson regression, we are not aware of a finite-sample analysis which produces a rate of d/n for the excess risk.

3.2 Risk analysis: early-stopped gradient descent

Moving from explicit to implicit regularization, we now study gradient descent on the unpenalized GLM loss in (13), initialized at $\theta_0 = 0$. For simplicity, we focus on a constant step size $\eta_t = \eta$, $t = 0, 1, 2, \ldots$, but the results can be generalized to arbitrary step sizes. We also consider projected gradient descent on the closed Euclidean ball centered at the origin $\mathsf{B}_d(r) = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ of a given radius $b > 0$.

Toward deriving risk bounds, recall the basic inequality for gradient descent in Theorem 1: by (4), for the initial point $\theta_0 = 0$ and any $\theta \in \mathbb{R}^d$,

$$f(\theta_T) + \frac{\lambda_T}{2}\|\theta_T - \theta\|_2^2 \leq f(\theta) + \frac{\lambda_T}{2}\|\theta\|_2^2.$$

As projected gradient descent is a special case of mirror descent, by Theorem 2 and (9), the above display is also true for this algorithm with $\mathcal{C} = \mathsf{B}_d(b)$ and any $\theta \in \mathsf{B}_d(b)$. In either case, this can be rewritten as

$$f(\theta_T) - f(\theta) \leq \frac{\lambda_T}{2}\Big(\|\theta\|_2^2 - \|\theta_T - \theta\|_2^2\Big).$$

Following the same calculations as in given for (16), we now add $\frac{1}{n}\epsilon^\top X(\hat{\theta}_\lambda - \theta)$ to each side. Recalling the decomposition of risk in (14), this gives

$$\mathrm{Risk}(\theta_T) - \mathrm{Risk}(\theta) \leq \frac{1}{n}\epsilon^\top X\big(\theta_T - \theta\big) + \frac{\lambda_T}{2}\Big(\|\theta\|_2^2 - \|\theta_T - \theta\|_2^2\Big). \tag{20}$$

The next proposition uses this to derive excess risk bounds which closely resemble those in Proposition 1.

Proposition 2. *Assume that the GLM loss f in (13) is L-smooth, for $L > 0$. Consider gradient descent (3) with $\eta_t = \eta \in (0, 1/L]$, $t = 0, 1, 2, \ldots$, initialized at $\theta_0 = 0$. Then for any reference point $\theta \in \mathbb{R}^d$, and for any $T \geq 1$ and $\lambda_T = 1/(\eta T)$,*

$$\mathrm{Risk}(\theta_T) \leq \mathrm{Risk}(\theta) + \frac{1}{2\lambda_T}\Big\|\frac{X^\top \epsilon}{n}\Big\|_2^2 + \frac{\lambda_T}{2}\|\theta\|_2^2.$$

If f is only L-smooth on $\mathsf{B}_d(b)$ for some $b > 0$, then the same result holds for projected gradient descent (8) with $\mathcal{C} = \mathsf{B}_d(b)$, and any reference point $\theta \in \mathsf{B}_d(b)$.

Furthermore, assume that each $\epsilon_i \sim \mathrm{sG}(\sigma_i^2)$. Define $\sigma^2 = \max_{i=1,\ldots,n} \sigma_i^2$ and $\widehat{\Sigma} = \frac{1}{n}X^\top X$. For any $\delta > 0$ and $b > 0$, define the target regularization parameter

$$\lambda_T^* = \frac{\sigma}{2b\sqrt{n}}\sqrt{\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta}\|\widehat{\Sigma}\|_F + 2\delta\|\widehat{\Sigma}\|_{\mathrm{op}}}, \tag{21}$$

and define $T^* = 1/(\eta\lambda_T^*)$. If T^* is an integer, then setting $T = T^*$ the excess risk of θ_T with respect the class of parameters with ℓ_2 norm at most b satisfies

$$\mathrm{Risk}(\theta_T) - \min_{\theta:\|\theta\|_2 \leq b} \mathrm{Risk}(\theta) \leq \frac{b\sigma}{\sqrt{n}}\sqrt{\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta}\|\widehat{\Sigma}\|_F + 2\delta\|\widehat{\Sigma}\|_{\mathrm{op}}}, \tag{22}$$

with probability at least $1 - e^{-\delta}$. *If* T^* *is not an integer, then setting* $T = \lceil T^* \rceil$, *the above bound holds with an additional (lower-order) term* $e_T = (\eta\sigma^2/(2n)) \cdot (\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta}\|\widehat{\Sigma}\|_F + 2\delta\|\widehat{\Sigma}\|_{\mathrm{op}})$ *on the right-hand side, which is due to discretization.*

Under the standard spectral assumptions on $\widehat{\Sigma}$ in (19), we can take $\delta = \log n$, and then the bound in (22) becomes $\tilde{O}(b\sigma\sqrt{d/n})$, as before. Moreover, in this case the additional error term (for non-integral T^*) scales as $e_T = \tilde{O}(\sigma^2 d/n)$, confirming that it is indeed lower-order in the regime $d/n \to 0$.

Building on this proposition, our next result derives excess risk bounds for gradient descent for common GLMs. This mirrors Theorem 3 for ridge-regularized GLMs.

Theorem 4. *Let the response distribution* P_i *and* σ *be defined as in Theorem 3, in case 1: Gaussian, case 2: Bernoulli, and case 3: Poisson. Note in cases 1 and 2, the GLM loss is* L*-smooth on* $\mathbb{R}^d$ *with* $L = \|\widehat{\Sigma}\|_{\mathrm{op}}$ *and* $L = \frac{1}{4}\|\widehat{\Sigma}\|_{\mathrm{op}}$, *respectively; in case 3, it is* L*-smooth on* $\mathsf{B}_d(b)$ *with* $L = \exp(b \cdot \max_{i=1,\ldots,n} \|x_i\|_2) \cdot \|\widehat{\Sigma}\|_{\mathrm{op}}$, *for any* $b > 0$. *Consider gradient descent* (3) *in cases 1 and 2, and projected gradient descent* (8) *on* $\mathcal{C} = \mathsf{B}_d(b)$ *in case 3. In each case, we set* $\eta_t = \eta \in (0, 1/L]$, $t = 0, 1, 2, \ldots$ *and initialize at* $\theta_0 = 0$. *Fixing any* $\delta > 0$ *and* $b > 0$, *let* λ_T^* *be as in* (21) *and* $T^* = 1/(\eta\lambda_T^*)$. *The following holds.*

- *If* T^* *is an integer, then for* $T = T^*$ *the excess risk bound in* (22) *holds with probability at least* $1 - e^{-\delta}$ *in cases 1 and 2, and at least* $1 - e^{-\delta} - 1/n$ *in case 3 provided that* $\delta \geq 1$ *and* $n \geq 3$.
- *If* T^* *is not an integer, then for* $T = \lceil T^* \rceil$ *the excess risk bound holds with the additional (lower-order) additive term* e_T *as defined in Proposition 2.*

Just like Theorem 3, Theorem 4 translates into a high probability excess risk bound of $\tilde{O}(b\sigma\sqrt{d/n})$ under the standard spectral assumptions in (19). The main point we intend to convey in the above theorem is that the basic inequality for gradient descent leads to relatively simple but useful risk bounds for GLMs in general (misspecified) settings, comparable to those achievable via explicit ℓ_2 regularization. Below we briefly discuss related literature for broader context.

Comparison with existing literature. For linear regression or quadratic loss more generally there has been a number of analyses for early-stopped gradient descent/flow that yield characterizations more precise than the results in the above theorem, such as Yao et al. (2007); Raskutti et al. (2014); Wei et al. (2017); Ali et al. (2019). In particular, the latter authors establish that gradient flow stopped at time T obtains a risk at most 1.69 times that of ridge regression for $\lambda = 1/T$. For logistic regression, time-asymptotic analyses are more common, which show, e.g., that for linearly separable data gradient descent converges to the maximum ℓ_2-margin direction (Soudry et al., 2018; Ji and Telgarsky, 2019). In concurrent work, Wu et al. (2025) gave refined risk bounds for early-stopped gradient descent in logistic regression. This provides sharper control of the excess risk in certain well-specified settings, at the rate d/n. We are not of work analyzing the gradient descent path and its excess risk for Poisson regression.

4 Model aggregation with KL regularity

In this section, we examine mirror descent and its explicit regularization counterpart, Bregman-divergence-regularization. As our primary application, we focus on exponentiated gradient descent and Kullback-Leibler (KL) divergence regularization. Results for general Bregman divergences are given in the appendix.

The notation in this section inherits from the last, with $X \in \mathbb{R}^{n\times d}$ denoting a fixed (nonrandom) feature matrix and $Y \in \mathbb{R}^n$ a response vector. We again consider the GLM negative log-likelihood function f in (13), but now we restrict our attention to parameters $\theta \in \Delta_d$, where recall $\Delta_d = \{a \in \mathbb{R}^d : a_i \geq 0, \sum_{i=1}^d a_i = 1\}$ is the probability simplex. As motivation, we may think of this setup as representing *model aggregation*, with each X_j (the j^{th} column of X) denoting a predictor of the underlying mean μ, and their convex combination $X\theta = \sum_{j=1}^d \theta_j X_j$, $\theta \in \Delta_d$ representing an aggregate predictor we seek to learn by optimizing the loss f. This approach is often called *linear stacking* in the literature (Wolpert, 1992; Breiman, 1996).

The flow of this section is analogous to the last: we first derive simple but meaningful excess risk bounds for the KL-regularized estimator, which serve as a benchmark for the early-stopped mirror descent analysis to follow. In each case, it is not our intention to derive the sharpest results possible, but instead to demonstrate the power of the basic inequality.

4.1 Risk analysis: KL-regularized solution

The KL-regularized GLM estimator is defined by augmenting the GLM loss (13) with a KL penalty, denoted

$$\hat{\theta}_\lambda = \operatorname*{arg\,min}_{\theta \in \Delta_d} \Big(f(\theta) + \lambda D_{\mathrm{KL}}(\theta, u) \Big), \tag{23}$$

where recall $D_{\mathrm{KL}}(a, b) = \sum_{i=1}^d a_i \log(a_i/b_i)$, and $u \in \Delta_d$ is an arbitrary anchor point. (Note that the solution exists and is unique because the objective is strictly convex.)

Using similar arguments to those leading up to (16) and Proposition 1 in the ridge regularization setting, the proposition below establishes determinstic and high-probability excess risk bounds.

Proposition 3. *For any $\lambda > 0$, anchor point $u \in \Delta_d$, and reference point $\theta \in \Delta_d$, the prediction risk of $\hat{\theta}_\lambda$ in (23) satisfies*

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \mathrm{Risk}(\theta) \leq \frac{1}{\lambda} \Big\| \frac{X^\top \epsilon}{n} \Big\|_\infty^2 + 2\lambda D_{\mathrm{KL}}(\theta, u).$$

Furthermore, assume that each $\epsilon_i \sim \mathrm{sG}(\sigma_i^2)$, and each $\|X_j\|_2 \leq \sqrt{n}$. Let $\sigma^2 = \max_{i=1,\ldots,n} \sigma_i^2$. Then, for any $\delta > 0$ and $b > 0$, choosing

$$\lambda = \sigma \sqrt{\frac{\log(2d) + \delta}{bn}}, \tag{24}$$

the excess risk of $\hat{\theta}_\lambda$ with respect the class of parameters with $D_{\mathrm{KL}}(\theta, u) \leq b$ satisfies

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \min_{\theta : D_{\mathrm{KL}}(\theta,u) \leq b} \mathrm{Risk}(\theta) \leq 4\sigma \sqrt{\frac{b(\log(2d) + \delta)}{n}}, \tag{25}$$

with probability at least $1 - e^{-\delta}$.

Next we specialize this to common GLMs, in a structure similar to Theorem 3 for ridge regularization.

Theorem 5. *Consider the KL-regularized GLM estimator $\hat{\theta}_\lambda$ in (23) with $\lambda > 0$. Consider cases 1, 2, and 3 as described in Theorem 3, and assume that each $\|X_j\|_2 \leq \sqrt{n}$. Then, for any $\delta > 0$ and $b > 0$, choosing λ as in (24) yields the excess risk bound in (25), which holds with probability at least $1 - e^{-\delta}$ in cases 1 and 2, and at least $1 - e^{-\delta} - 1/n$ in case 3 provided $\delta \geq 1$ and $n \geq 3$.*

Choosing $\delta = \log n$, the bound in (25) becomes $\tilde{O}(\sigma\sqrt{b(\log d)/n})$, which holds with probability at least $1 - 1/n$ in the Gaussian and Bernoulli cases, and at least $1 - 2/n$ in the Poisson case. We defer discussion of related work until after presenting the analogous implicit regularization result, in the next subsection.

4.2 Risk analysis: early-stopped mirror descent

We now study early-stopped exponentiated gradient descent as the implicit regularization counterpart to KL regularization. The proposition below provides excess risk bounds closely resembling those in Proposition 3.

Proposition 4. *Assume that the GLM loss f in (13) is L-smooth with respect to $\|\cdot\|_1$, for $L > 0$. Consider exponentiated gradient descent (7) with $\eta_t = \eta \in (0, 1/L]$, $t = 0, 1, 2, \ldots$, initialized at $\theta_0 = u \in \Delta_d$. Then for any reference point $\theta \in \Delta_d$, and for any $T \geq 1$ and $\lambda_T = 1/(\eta T)$,*

$$\mathrm{Risk}(\theta_T) - \mathrm{Risk}(\theta) \leq \frac{1}{2\lambda_T} \Big\| \frac{X^\top \epsilon}{n} \Big\|_\infty^2 + \lambda_T D_{\mathrm{KL}}(\theta, u).$$

Furthermore, assume that each $\epsilon_i \sim \mathrm{sG}(\sigma_i^2)$, and each $\|X_j\|_2 \leq \sqrt{n}$. Let $\sigma^2 = \max_{i=1,\ldots,n} \sigma_i^2$. For any $\delta > 0$ and $b > 0$, define the target regularization parameter

$$\lambda_T^* = \sigma \sqrt{\frac{\log(2d) + \delta}{bn}}. \tag{26}$$

and define $T^* = 1/(\eta\lambda_T^*)$. If T^* is an integer, then setting $T = T^*$ the excess risk of θ_T with respect the class of parameters with $D_{\mathrm{KL}}(\theta, u) \le b$ satisfies

$$\mathrm{Risk}(\theta_T) - \min_{\theta: D_{\mathrm{KL}}(\theta,u)\le b} \mathrm{Risk}(\theta) \le 2\sigma\sqrt{\frac{b(\log(2d)+\delta)}{n}}. \tag{27}$$

with probability at least $1 - e^{-\delta}$. If T^* is not an integer, then setting $T = \lceil T^* \rceil$, the above bound holds with an additional (lower-order) term $e_T = \eta\sigma^2(\log(2d)+\delta)/n$ on the right-hand side, which is due to discretization.

Finally, we specialize this to common GLMs, providing results analogous to those in Theorem 5.

Theorem 6. *Consider exponentiated gradient descent (7) on the GLM loss in (13), under cases 1, 2, and 3 as described in Theorem 3, and assume that each $\|X_j\|_2 \le \sqrt{n}$. These losses are L-smooth on Δ_d with respect to $\|\cdot\|_1$, where the smoothness parameter is $L = \frac{1}{n}\sum_{i=1}^n \|x_i\|_\infty^2$ in case 1 (Gaussian), $L = \frac{1}{4n}\sum_{i=1}^n \|x_i\|_\infty^2$ in case 2 (Bernoulli), and $L = \frac{1}{n}\sum_{i=1}^n e^{\|x_i\|_\infty}\|x_i\|_\infty^2$ in case 3 (Poisson). Set $\eta_t = \eta \in (0, 1/L]$, $t = 0, 1, 2, \ldots$ and $\theta_0 = u \in \Delta_d$. Fixing any $\delta > 0$ and $b > 0$, let λ_T^* as in (26) and $T^* = 1/(\eta\lambda_T^*)$. The following holds.*

- *If T^* is an integer, then for $T = T^*$ the excess risk bound in (27) holds with probability at least $1 - e^{-\delta}$ in cases 1 and 2, and at least $1 - e^{-\delta} - 1/n$ in case 3 provided that $\delta \ge 1$ and $n \ge 3$.*
- *If T^* is not an integer, then for $T = \lceil T^* \rceil$ the excess risk bound holds with the additional (lower-order) additive term e_T as defined in Proposition 4.*

As before, the bound in (27) gives a high probability excess risk bound of $\tilde{O}(\sigma\sqrt{b(\log d)/n})$. Below we briefly discuss the broader literature.

Comparison with existing literature. The problem of model aggregation has been studied extensively in the statistics literature, e.g., Tsybakov (2003); Juditsky et al. (2005); Lecué (2007); Juditsky et al. (2008); Dalalyan and Tsybakov (2009); Lecué and Mendelson (2013); Lecué and Rigollet (2014). A rate of $\sqrt{(\log d)/n}$ for excess risk when the comparator class is all convex combinations of base predictors ($X\theta$ for $\theta \in \Delta_d$, in our notation) is somewhat standard. The influential paper Juditsky et al. (2005) shows that this rate is achieved by an estimator called mirror averaging. This peforms one epoch of *stochastic* exponentiated gradient descent followed by an online-to-batch conversion (in the language of online learning). It is interesting to emphasize the differences to our results above—we show that this rate is achievable using *batch* mirror descent, provided that we use early stopping.

It is also worth noting that faster rates are possible under stronger assumptions. Juditsky et al. (2008) show that mirror averaging can also produce an excess risk bound of order $(\log d)/n$ when the excess risk is defined with respect to the single best predictor (rather than the best convex combination of predictors) in the given finite class. This can also be achieved by a method called Q-aggregation (Lecué and Rigollet, 2014), which applies KL-type regularization to a hybrid loss composed of a linear stacking objective (as studied in this section) and a randomized prediction objective (as studied next).

5 Randomized prediction with KL regularity

We now study *randomized prediction*, an alternative meta-learning approach to model aggregation as studied previously. While model aggregation outputs a convex combination of base predictors, randomized prediction samples a single model according to a learned distribution over the base predictors.

Using similar notation to the last section, let $X_j \in \mathbb{R}^n$, $j = 1, \ldots, d$ be base predictors and let $Y \in \mathbb{R}^n$ be a response vector. Given a parameter $\theta \in \Delta_d$ on the probability simplex, we define the randomized predictor $\beta = \beta(X)$ by setting $\beta(X) = X_j$ with probability θ_j. As shorthand, we will write the corresponding predictor as $\beta \sim \theta$. We define the training risk as

$$\widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^n \ell(\beta(X)_i, Y_i)$$

where ℓ is an arbitrary loss function.

A randomized predictor can also make test predictions, and hence we can also define a suitable notion of test risk. Let $X_j^* \in \mathbb{R}^n$, $j = 1, \ldots, d$ and $Y^* \in \mathbb{R}^n$ denote test instances of the base predictors and response variable, respectively. Then the randomized predictor $\beta \sim \theta$ sets $\beta(X^*) = X_j^*$ with probability θ_j. We define the test risk as

$$R(\beta) = \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n \ell(\beta(X^*)_i, Y_i^*)\bigg],$$

where the expectation is with respect to any joint distribution P on (X^*, Y^*). This setting is quite general. For ℓ taking the GLM form, and the joint distribution P placing a point mass at $X^* = X$, this recovers the setup in the previous two sections, with $R(\beta) = \mathrm{Risk}(\theta_j)$ and $\widehat{R}(\beta) = f(\theta_j)$ for a particular (random) index j. However, the current setup is even broader, allowing X^* to be random, and ℓ to be arbitrary.

To construct the sampling distribution, we will consider exponential weighting based on the empirical risk evaluated at each base model:

$$\hat{\theta}_\lambda(\mathrm{d}\beta) \propto \exp(-\widehat{R}(\beta)/\lambda) \cdot u(\mathrm{d}\beta), \tag{28}$$

where u is some prior measure and $\lambda > 0$ is a tuning parameter. With no prior information, one may take the uniform $u = \pi$ on Δ_d. The above estimator is called the Gibbs posterior in the Bayesian statistics literature. It can be equivalently defined by solving a KL-regularized optimization problem (Alquier, 2024):

$$\hat{\theta}_\lambda = \operatorname*{arg\,min}_{\theta \in \Delta_d} \Big(\mathbb{E}_{\beta\sim\theta}[\widehat{R}(\beta)] + \lambda D_{\mathrm{KL}}(\theta, u)\Big), \tag{29}$$

where the expectation is only with respect to the randomness in drawing $\beta \sim \theta$. This problem is also referred to as information risk minimization in the some parts of the literature (Leung and Barron, 2006; Zhang, 2006; Xu and Raginsky, 2017).

The flow of the previous two sections would suggest that we would here produce a risk analysis of the explicitly-regularized estimator (29) and then compare to what is possible to derive from the basic inequality for its implicitly-regularized counterpart, exponentiated gradient descent on the loss $f(\theta) = \mathbb{E}_{\beta\sim\theta}[\widehat{R}(\beta)]$. In this section, however, we depart from this strategy for one important reason: these two estimators actually coincide exactly.

Proposition 5. *Consider the Gibbs posterior $\hat{\theta}_\lambda$ in (29) with an arbitrary prior $u \in \Delta_d$, and exponentiated gradient descent (7) with linear loss $f(\theta) = \mathbb{E}_{\beta\sim\theta}[\widehat{R}(\beta)]$, constant step sizes $\eta_t = \eta > 0$, $t = 0, 1, 2, \ldots$, and initialization $\theta_0 = u$. Then for $\lambda = 1/(\eta T)$, these two estimators coincide: $\hat{\theta}_\lambda = \theta_T$.*

This means that to derive excess risk bounds, we can apply basic inequality arguments either to $\hat{\theta}_\lambda$ or to θ_T. In what follows, we pursue the latter, because it results in sharper constants, i.e., the excess risk bounds are sharper by a factor of 2 (which is unsurprising, since this was also the case in the last two sections). We restrict our attention to the case in which the optimal tuning corresponds to an integral time T, for brevity (and analogous discretization results would follow as before).

Proposition 6. *For any $\lambda > 0$, prior $u \in \Delta_d$, and reference point $\theta \in \Delta_d$, the risk of $\hat{\theta}_\lambda$ in (29) satisfies*

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}[R(\beta)] - \mathbb{E}_{\beta\sim\theta}[R(\beta)] \leq \frac{1}{2\lambda}\|\widehat{R} - R\|_\infty^2 + \lambda D_{\mathrm{KL}}(\theta, u),$$

where $\|\widehat{R} - R\|_\infty = \sup_\beta |\widehat{R}(\beta) - R(\beta)|$. The same bound applies to exponentiated gradient descent (7) with the loss $f(\theta) = \mathbb{E}_{\beta\sim\theta}[\widehat{R}(\beta)]$ (and constant step size η and $\theta_0 = u$), provided $T = 1/(\eta\lambda)$ is an integer.

Furthermore, if the loss ℓ is bounded by $C > 0$ and the training samples are i.i.d., then for any $\delta > 0$ and $b > 0$, it holds that

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}[R(\beta)] - \min_{\theta: D_{\mathrm{KL}}(\theta,u)\leq b} \mathbb{E}_{\beta\sim\theta}[R(\beta)] \leq \frac{C^2(\log(2d) + \delta)}{4n\lambda} + \lambda b.$$

with probability at least $1 - e^{-\delta}$. In particular, choosing

$$\lambda = \frac{C}{2}\sqrt{\frac{\log(2d) + \delta}{bn}},$$

it holds that

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}[R(\beta)] - \min_{\theta: D_{\mathrm{KL}}(\theta,u)\le b} \mathbb{E}_{\beta\sim\theta}[R(\beta)] \le C\sqrt{\frac{b\log(2d)+\delta}{n}}.$$

with probability at least $1-e^{-\delta}$. *The same result again applies to exponentiated gradient descent provided provided* $T = 1/(\eta\lambda)$ *is an integer.*

Lastly, we briefly discuss related literature.

Comparison with existing literature. The most directly related analysis given by Alquier (2024); under the same conditions as our result, this author establishes that the Gibbs posterior estimator satisfies

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}[R(\beta)] - \min_{\theta: D_{\mathrm{KL}}(\theta,u)\le b} \mathbb{E}_{\beta\sim\theta}[R(\beta)] \le \frac{C^2}{4n\lambda} + 2\lambda(b+\log 2+\delta),$$

with probability at least $1-e^{-\delta}$. Compared to the first result in Proposition 6, we can see that this is sharper by a factor of $\log d$ in the first term on the right-hand side (and so in particular, it applies in situations where the number of base models is infinite). This is due to a more sophisticated analysis of the empirical process term in Alquier (2024); they use the Donsker-Varadhan formula, whereas we simply use Young's inequality and Hoeffding's inequality. Still, it is interesting to see that the basic inequality for mirror descent produces an excess risk rate of $\sqrt{(\log d)/n}$, which is a meaningful result for randomization prediction and only slightly worse than (for finite d) than the rate $1/\sqrt{n}$ achievable via more advanced techniques.

6 Experiments

We present empirical results to supplement our theoretical findings on the relationship between explicit and implicit regularization, investigating both training dynamics and prediction risk. Python code to reproduce our experiments is available at https://github.com/100shpaik/.

Experimental setup. We evaluate gradient descent (GD) and exponentiated gradient descent (EGD), as in (3) and (7), respectively. In all experiments that follow, we initialize GD is at the origin, and EGD at the uniform distribution. We apply GD to three GLM losses: linear, logistic, and Poisson (as in Section 3), and compare its iterates to explicit ridge regularization (15). We run EGD on these same the losses, in a model aggregation context (as in Section 4), and compare its iterates to explicit KL regularization (23). In general, we use variable step sizes η_t, $t = 0, 1, 2, \ldots$, and define the total elapsed time as $\tau = \sum_{t=0}^{T-1} \eta_t$, where we align $1/\tau$ with the regularization parameter λ. Further implementation details are provided in the appendix.

Data distributions. We study both underparameterized ($n > d$) and overparameterized ($n < d$) regimes, and generate features $X \in \mathbb{R}^{n\times d}$ and responses $Y \in \mathbb{R}^n$ as follows. We sample the entries of X independently from $\mathcal{N}(0,1)$, and then sample Y given X from a well-specified model for each GLM, with parameter $\theta \in \mathbb{R}^d$. In the regression tasks (where we compare GD and ridge regularization), we sample the entries of θ uniformly on $[-1,1]$. In model aggregation tasks (where we compare EGD and KL regularization), we sample uniformly on $[0,1]$, and then renormalize so that $\theta \in \Delta_d$.

We introduce an additional parameter $\gamma > 0$ to control the signal-to-noise ratio. To be precise, for each $i = 1, \ldots, n$, we independently sample $Y_i|x_i$ in the linear, logisitic, and Poisson regression cases as follows:

- linear: $Y_i|x_i \sim \mathcal{N}(x_i^\top\theta, \gamma^2)$;
- logistic: $Y_i|x_i \sim \mathrm{Bernoulli}(p_i)$ with $p_i = 1/(1+\exp(-\gamma x_i^\top\theta))$;
- Poisson: $Y_i|x_i \sim \mathrm{Pois}(\mu_i)$ with $\mu_i = \gamma x_i^\top\theta$.

Table 1 summarizes the values of n, d, and γ used in our experiments. (The specific values of γ are chosen to produce "U-shaped" prediction risk curves, avoiding regimes where the risk is uninterestingly monotonic, which typically occurs when γ is too small or large.)

Table 1: Summary of n, d, and γ values used in our experiments.

GLM	**Regression (GD vs. ridge)**		**Model aggregation (EGD vs. KL)**	
	Underparametrized $(n, d) = (200, 20)$	Overparametrized $(n, d) = (100, 200)$	Underparametrized $(n, d) = (200, 20)$	Overparametrized $(n, d) = (30, 60)$
Linear	$\gamma = 5.0$	$\gamma = 5.0$	$\gamma = 1.0$	$\gamma = 0.1$
Logistic	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 1.5$	$\gamma = 10.0$
Poisson	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 1.2$	$\gamma = 3.5$

6.1 Training dynamics

Figure 1 plots the penalized training loss:

$$P(\theta, \lambda) = f(\theta) + \lambda g(\theta),$$

across implicitly- and explicitly-regularized estimates. In panel (a), we consider the regression tasks, where $g(\theta) = \|\theta\|_2^2$, and in panel (b), we consider model aggregation, where $g(\theta) = \|\theta - \pi\|_1^2$. Each row corresponds to a different GLM loss f: linear, logistic, and Poisson.

Examining first the results in panel (a), we compare three curves: one for GD iterates, $P(\theta_T, \frac{1}{4\tau})$ as the total elapsed time τ varies (plotted in ridge), and two for ridge estimates, $P(\hat{\theta}_\lambda, \lambda)$ as λ varies, using $\lambda = \frac{1}{4\tau}$ (in green) or $\lambda = \frac{1}{\tau}$ (in blue). According to Corollary 1, we know that the red curve must lie in between the green and blue, and we see this is indeed true in the figure. Notably, the red line closely tracks the green line for $\lambda = \frac{1}{4\tau}$. This reveals a stronger correspondence between θ_T and $\hat{\theta}_\lambda$ with $\lambda = \frac{1}{4\tau}$ than anticipated by the exisiting theory, which will be revisited in the prediction risk analysis shortly.

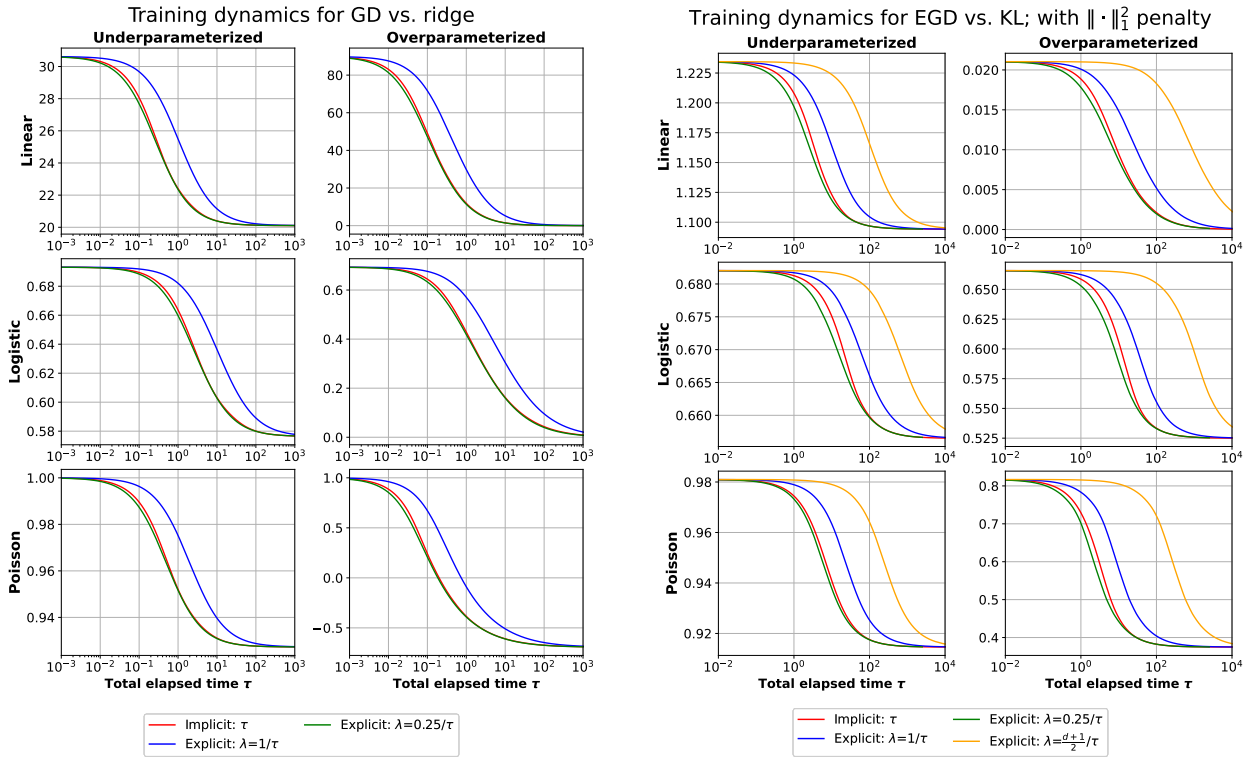
The results in panel (b) are analogous, except with squared ℓ_1 penalties used in $P(\theta, \lambda)$, as suggested by Corollary 3. Here we have an additional curve corresponding to explicit (KL) regularization, where $\lambda = \frac{d+1}{2\tau}$ (as in the upper bound in the corollary, recalling $G = d$ and $\alpha = 1$ in the present setting). We can see that this is overly conservative in practice, and the correspondence is tightest between θ_T and $\hat{\theta}_\lambda$ with $\lambda = \frac{1}{4\tau}$. Moreover, results where $P(\theta, \lambda)$ is defined in terms of a KL penalty (not covered by any existing theory) are similar, and given in the appendix.

6.2 Prediction risk

Figure 2 plots prediction risk curves (where this is defined in a fixed-X sense, as explained in Section 3) in a format analogous to that used in Figure 1. Recall that unlike our results on training dynamics, the risk theory in our paper does not draw direct comparisons between the curves for implicitly- and explicitly regularized estimates; instead, Theorems 3–6 derive excess risk bounds, which end up being very similar for GD and ridge regularization, and EGD and KL regularization. Nevertheless, we can see in panel (a) that the risk curves for GD and ridge estimates are qualitatively quite similar, with the red and green curves—which correspond to θ_T and $\hat{\theta}_\lambda$ with $\lambda = \frac{1}{4\tau}$—being overall the closest across the settings. This is also broadly true in panel (b), for the EGD and KL-regularized estimates, but now the red and blue curves—which correspond to θ_T and $\hat{\theta}_\lambda$ with $\lambda = \frac{1}{\tau}$—being closer. Finally, the minimum risk obtained by implicit and explicit regularization is quite close in all cases, with neither one dominating the other.

6.3 Solution path

Figure 3 displays solution paths for implicitly- and explicitly-regularized estimates. In a given plot, one curve represents one coordinate path, either θ_{Ti} for varying τ or $\hat{\theta}_{\lambda i}$ for varying λ. Although none of our theory in this paper compares solution paths directly, we generally see a striking resemblance between GD paths and ridge-regularized paths, and between EGD paths and KL-regularized paths. This provides further compelling evidence of the deep connections between implicit and explicit regularization. Meanwhile, it is interesting to note that the paths for EGD and KL regularization converge to a sparse solution in the limit, as $\tau \to \infty$ or $\lambda \to 0$. Studying connections between these regularization mechanisms and lasso (ℓ_1) regularization may be a topic of future work.



(a) GD vs. ridge, plotted with $\|\theta\|_2^2$ penalty.

(b) EGD vs. KL, plotted with $\|\theta - \pi\|_1^2$ penalty.

Figure 1: Training envelopes for (a) regression and (b) model aggregation tasks.

7 Other basic inequalities

We also discuss basic inequalities for two additional first-order iterative algorithms, proximal gradient descent and NoLips (essentially, a version of mirror descent which uses different assumptions).

7.1 Proximal gradient descent

Proximal gradient descent applies to composite objectives $f = f_0 + f_1$ with f_0, f_1 convex and f_0 differentiable. Given an initial point $\theta_0 \in \mathbb{R}^d$, and step sizes $\{\eta_t\}_{t=0}^{\infty}$, this method generates iterates for $t = 0, 1, 2, \ldots$ via

$$\theta_{t+1} = \mathrm{prox}_{\eta_t f_1}\Big(\theta_t - \eta_t \nabla f_0(\theta_t)\Big), \tag{30}$$

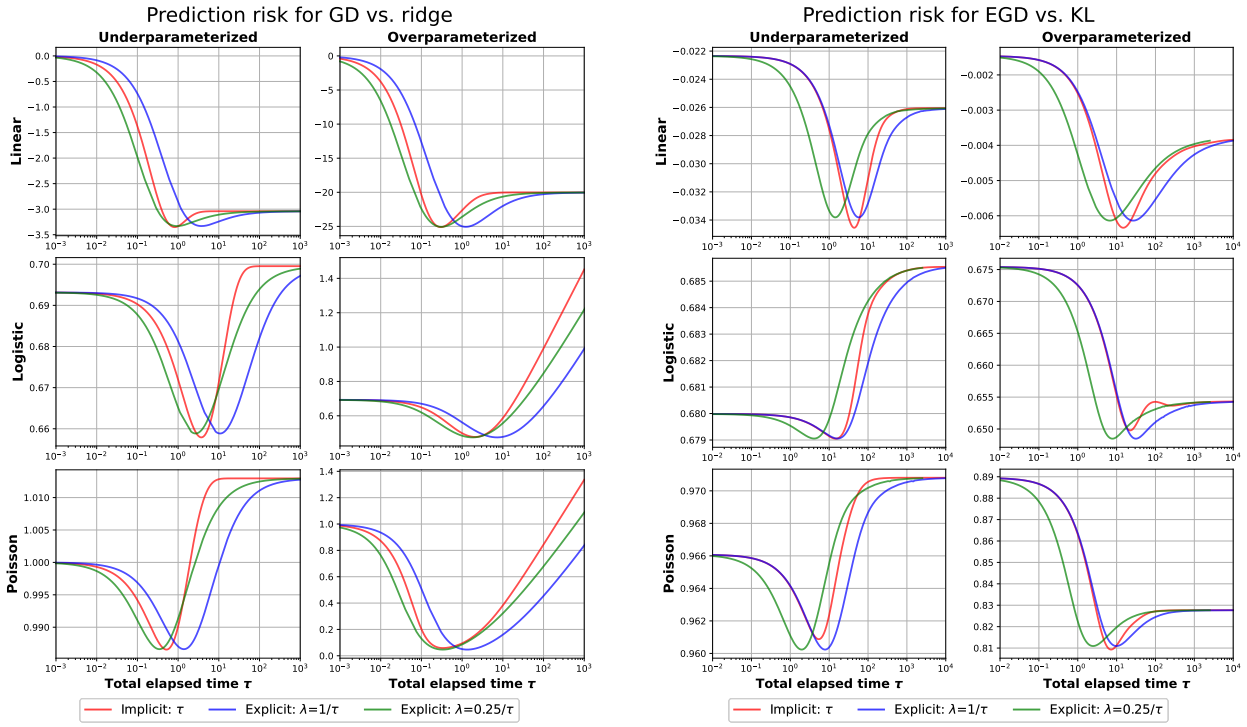
where the proximal operator defined as

$$\mathrm{prox}_{\eta f_1}(\theta) = \operatorname*{arg\,min}_{z \in \mathbb{R}^d} \left(\frac{1}{2}\|\theta - z\|_2^2 + \eta f_1(z)\right).$$

A basic requirement for using this algorithm is that this proximal operator can be computed in closed-form or efficiently approximated. The following theorem derives a basic inequality for proximal gradient descent, which generalizes Theorem 1 (which corresponds to $f_1 = 0$).

Theorem 7. *Let $f_0 : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and L-smooth for $L > 0$, and $f_1 : \mathbb{R}^d \to \mathbb{R}$ be convex. Consider proximal gradient descent (30) with step sizes $\eta_t \in (0, 1/L]$, $t = 0, 1, 2, \ldots$. For any reference point $z \in \mathbb{R}^d$, and any iteration $T \geq 1$, it holds for $f = f_0 + f_1$ that*

$$f(\theta_T) - f(z) \leq \frac{1}{2\sum_{t=0}^{T-1} \eta_t}\Big(\|\theta_0 - z\|_2^2 - \|\theta_T - z\|_2^2\Big),$$



(a) GD vs. ridge

(b) EGD vs. KL

Figure 2: Prediction risk curves for (a) regression and (b) model aggregation tasks.

By the same arguments used to derive training envelopes in Corollary 1, the result in Theorem 7 implies

$$\min_{z\in\mathbb{R}^d}\left(f_0(z)+f_1(z)+\frac{\lambda_T}{4}\|\theta_0-z\|_2^2\right)\le f_0(\theta_T)+f_1(\theta_T)+\frac{\lambda_T}{4}\|\theta_0-\theta_T\|_2^2 \le \min_{z\in\mathbb{R}^d}\left(f(z)+f_1(z)+\lambda_T\|\theta_0-z\|_2^2\right).$$

When $f_1(\theta)=\lambda\|\theta\|_1$ is a lasso penalty, the corresponding proximal operator is known as soft-thresholding, and proximal gradient descent is called the iterative soft-thresholding algorithm (ISTA). In this case, further specifying, e.g., $f_0(\theta)=\frac{1}{2n}\|Y-X\theta\|_2^2$ and initializing $\theta_0=0$, the above display becomes

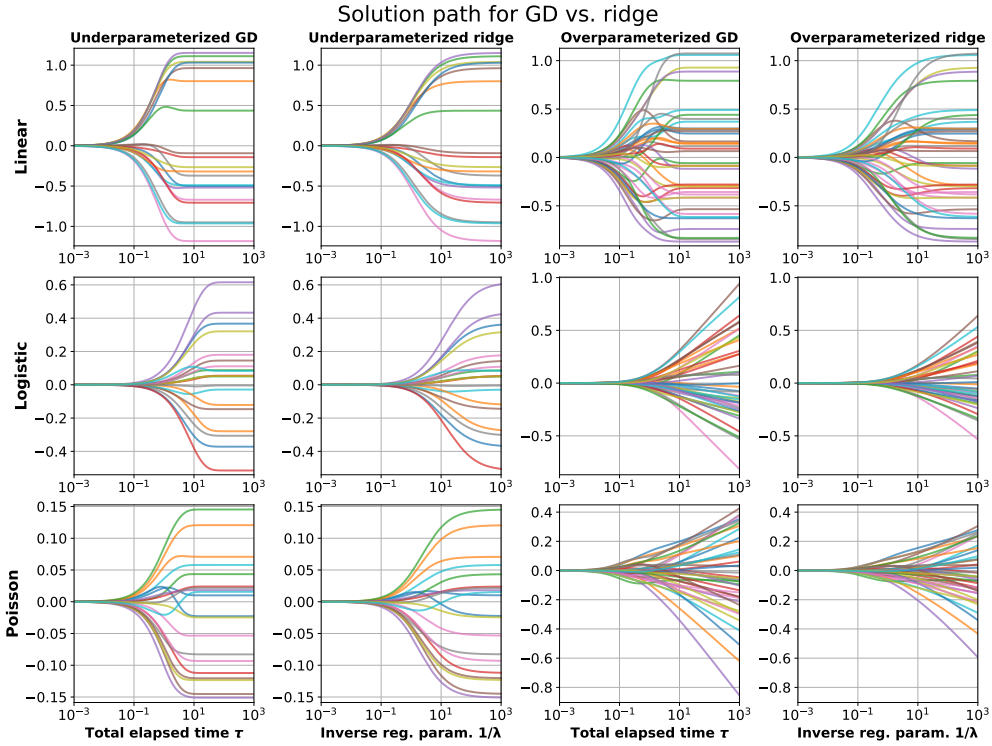
$$\min_{z\in\mathbb{R}^d}\left(\frac{1}{2n}\|Y-Xz\|_2^2+\lambda\|z\|_1+\frac{\lambda_T}{4}\|z\|_2^2\right)\le \frac{1}{2n}\|Y-X\theta_T\|_2^2+\lambda\|\theta_T\|_1+\frac{\lambda_T}{4}\|\theta_T\|_2^2 \le \min_{z\in\mathbb{R}^d}\left(\frac{1}{2n}\|Y-Xz\|_2^2+\lambda\|z\|_1+\lambda_T\|z\|_2^2\right).$$

The interpretation is that early-stopped ISTA provides regularization akin to explicit elastic net (composite lasso and ridge penalties) regularization.

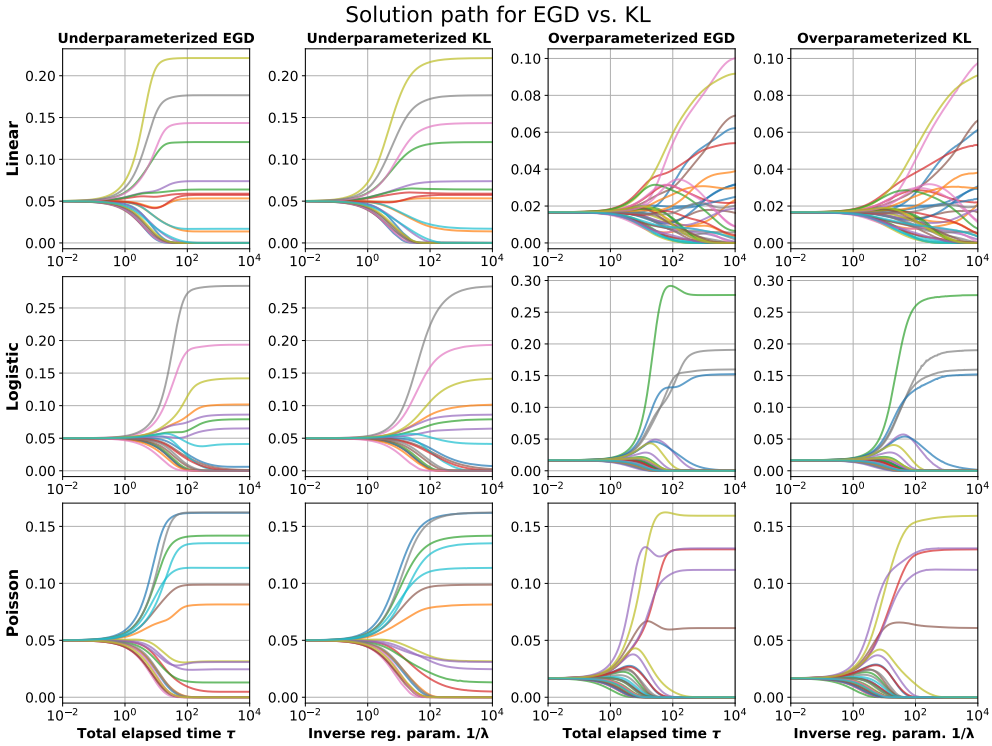
7.2 NoLips algorithm

Bauschke et al. (2017) proposed an iterative algorithm called NoLips, which can be viewed as an instance of mirror descent but operates under nonstandard assumptions; notably, it does not require Lipschitz smoothness of the objective f or strong convexity of the potential function ϕ. Below we derive a basic inequality for this algorithm.

Theorem 8. *Let $\Omega, \mathcal{C}$ be closed, convex sets in $\mathbb{R}^d$ with nonempty interior, and $\mathcal{C}\subseteq\Omega$. Assume $f:\Omega\to\mathbb{R}$ is convex on $\mathcal{C}$, and differentiable on on* $\mathrm{int}(\Omega)$. *Assume $\phi:\Omega\to\mathbb{R}^d$ is of Legendre type. Moreover, assume that*



(a) GD vs. ridge



(b) EGD vs. KL

Figure 3: Solution paths for (a) regression and (b) model aggregation. We plot all $d = 20$ coordinate paths for underparameterized regimes and the first 40 coordinates for overparameterized regimes, to avoid overcrowding the visualization.

$L\phi - f$ is convex on $\mathcal{C} \cap \text{int}(\Omega)$, for some $L > 0$. Consider mirror descent (6)—which in this case, is known as NoLips—with step sizes $\eta_t \in (0, 1/L]$, $t = 0, 1, 2, \ldots$. For any reference point $z \in \mathcal{C}$, and any $T \geq 1$,

$$f(\theta_T) - f(z) \leq \frac{1}{\sum_{t=0}^{T-1} \eta_t} \Big(D_\phi(z, \theta_0) - D_\phi(z, \theta_T) \Big).$$

An interesting example where the assumptions of Theorem 8 hold but those of Theorem 2 (the standard mirror descent assumptions) do not is the Poisson linear inverse (PLI) problem. Unlike a Poisson GLM which uses the canonical link $\log(\mu) = X\theta$, in PLI we use a linear link $\mu = X\theta$. This is typically done in situations where the elements of X are nonnegative and we constrain those of θ to be nonnegative, ensuring that μ itself has nonnegative entries. The Poisson negative log-likelihood under this link function is (after scaling by $\frac{1}{n}$):

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \Big(- Y_i \log(x_i^\top \theta) + x_i^\top \theta \Big).$$

This is not Lipschitz smooth for over the nonnegative orthant $\mathbb{R}^d_+$, because the derivative of $\log(u)$ blows up as $u \to 0^+$. This means that the traditional mirror descent analysis cannot be applied. However, as shown in Bauschke et al. (2017), if we choose Burg's entropy as the potential function $\phi(\theta) = -\sum_{i=1}^d \log(\theta_i)$, then one can show that $L\phi - f$ is convex on $\mathbb{R}^d_{++}$ for any $L \geq \|Y\|_1$. Thus Theorem 8 applies, and the interpretation loosely speaking is that early-stopping the NoLips algorithm for the PLI problem provides regularization akin to the Bregman divergence of Burg's entropy: this is $D_\phi(\theta, \theta_0) = \sum_{i=1}^d \frac{\theta_i}{\theta_{0i}} - \log(\frac{\theta_i}{\theta_{0i}})$ (up to constants), and is known as the Itakura-Saito divergence.

8 Discussion

In this paper, we examined basic inequalities for first-order optimization algorithms, which provide a unified framework for analyzing implicit regularization through both computational and statistical perspectives. We demonstrated the broad applicability of this framework by showing it can be used to infer properties about training dynamics and also to derive excess risk bounds, for different algorithms in various settings.

We close by highlighting a few directions for future work. The first is to derive refined basic inequalities under stronger assumptions, such as strong convexity. We believe this may enable us to derive faster excess risk rates for early-stopped gradient or mirror descent (i.e., of order d/n for GLM regression, or $\log d/n$ for model aggregation). A second direction of interest is to extend the basic inequality framework to stochastic gradient methods, ideally permitting some degree of nonconvexity in the objective. A third direction would be to establish basic inequalities for steepest descent algorithms with respect to non-Euclidean norms, such as the ℓ_1 norm, in which case steepest descent becomes greedy coordinate descent, also called forward stagewise regression. Given the rich tradition of work connecting stagewise algorithms to ℓ_1 regularization in statistics (Efron et al., 2004), it may be interesting to revisit this problem and see if a perspective based on something like the basic inequality can shed new light.

Acknowledgements

SP and RJT were supported by Office of Naval Research (ONR) grant number N00014-20-1-2787. KZ was supported by the Founder's Postdoctoral Fellowship in Statistics at Columbia University.

References

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Alnur Ali, Edgar Dobriban, and Ryan J. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, 2020.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:383–414, 2010.

Heinz H. Bauschke and Jonathan M. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

Peter Buhlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

Peter Buhlmann and Bin Yu. Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Conference on Learning Theory*, 2009.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, 2018.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, 2019.

Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, 2020.

Anatoli B. Juditsky, Alexander V. Nazin, Alexandre B. Tsybakov, and Nicolas Vayatis. Recursive aggregation of estimators via the mirror descent algorithm with averaging. *Methods of Signal Processing*, 31: 368–384, 2005.

Anatoli B. Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *Annals of Statistics*, 36(5):2183–2206, 2008.

Guillaume Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13 (4):1000–1022, 2007.

Guillaume Lecué and Shahar Mendelson. On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli*, 19(3):646–675, 2013.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *Annals of Statistics*, 42(51): 211–224, 2014.

Bastien Lemaire. An asymptotical variational principle associated with the steepest descent method for a convex function. *Journal of Convex Analysis*, 3(1):63–70, 1996.

Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.

Kevin Lin, James Sharpnack, Alessandro Rinaldo, and Ryan J. Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Neural Information Processing Systems*, 2017.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2003.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Dmitrii M. Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15:326–391, 2021.

David Pollard. A few good inequalities. *Lecture Notes: Statistics 600: Advanced Probability*, 2017. URL http://www.stat.yale.edu/~pollard/Courses/600.spring2017/Handouts/Basic.pdf.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*. 2002.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

Igal Sason. On reverse Pinsker inequalities. *arXiv preprint arXiv:1503.07118*, 2015.

Robert E. Schapire, Yoav Freund, Peter L. Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26(5):1651–1686, 1998.

Aaron Sidford. Smooth convex generalizations. *Lectures Notes, MSE 213: Introduction to Optimization Theory*, 2020. URL https://web.stanford.edu/~sidford/courses/20fa_opt_theory/sidford_mse213_2020fa_chap_5_extensions.pdf.

Rishi Sonthalia, Jackie Lok, and Elizaveta Rebrova. On regularization via early stopping for least squares. *arXiv preprint arXiv:2406.04425*, 2024.

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

Arun S. Suggala, Adarsh Prasad, and Pradeep Ravikumar. Connecting optimization and regularization paths. In *Neural Information Processing Systems*, 2018.

Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24(393):1–58, 2023.

Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, 2013.

Ryan J. Tibshirani. Empirical process theory for nonparametric analysis. *Lecture Notes, Stat 241B: Advanced Topics in Statistical Learning*, 2023. URL https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/emp_process.pdf.

Alexandre B. Tsybakov. Optimal rates of aggregation. In *Conference on Learning Theory*, 2003.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge, 2018.

Martin J. Wainwright. *High-dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, 2019.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In *Neural Information Processing Systems*, 2017.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *Conference on Learning Theory*, 2024.

Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Benefits of early stopping in gradient descent for overparameterized logistic regression. *arXiv preprint arXiv:2502.13283*, 2025.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Neural Information Processing Systems*, 2017.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P. Foster, and Sham M. Kakade. The benefits of implicit regularization from SGD in least squares problems. In *Neural Information Processing Systems*, 2021.

A Definitions and notation

We overview the standard definitions and notation used in the main paper. For a set $S \subseteq \mathbb{R}^d$, we denote its interior and boundary by $\text{int}(S)$ and ∂S, respectively. When S is finite, we denote its cardinality by $|S|$.

For a convex set Ω, a function is said to be convex if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$ for any $x, y \in \Omega$ and $\alpha \in [0,1]$, and strictly convex if the inequality is strict for $x \neq y$ and $\alpha \in (0,1)$. We denote the subdifferential (set of subgradients) of f at x by $\partial f(x)$. A function f is said to be essentially strictly convex if it is strictly convex on every convex subset of $\{x \in \Omega : \partial f(x) \neq \emptyset\}$. A function $f : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ is said to be essentially smooth provided it satisfies three conditions: (i) $\text{int}(\Omega) \neq \emptyset$; (ii) f is differentiable on $\text{int}(\Omega)$; and (iii) $\lim_{k\to\infty} \|\nabla f(x_k)\|_2 = \infty$ for any sequence $\{x_k\}_{k=1}^\infty \subset \Omega$ converging to a point in $\partial\Omega$. Finally, a function f is said to be of Legendre type if it is both essentially smooth and essentially strictly convex.

A differentiable function f is called α-strongly convex with respect to a norm $\|\cdot\|$, for a given $\alpha > 0$, if $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|x-y\|^2$. A differentiable function f is called L-smooth with respect to a norm $\|\cdot\|$, for a given $L > 0$, if the gradient map ∇f is L-Lipschitz with respect to $\|\cdot\|$, which recall means that $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x-y\|$ for all x, y, where $\|\cdot\|_*$ denotes the dual norm.

B Proofs for Section 2

B.1 Descent lemmas

For completeness, we state and prove the standard descent lemma for gradient descent.

Lemma B1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and L-smooth for $L > 0$. For the gradient descent update in (3) with step size $\eta_t \in (0, 2/L]$,*

$$f(\theta_{t+1}) \leq f(\theta_t) - \eta_t\Big(1 - \frac{L}{2}\eta_t\Big)\|\nabla f(\theta_t)\|_2^2 \leq f(\theta_t).$$

Proof. By L-smoothness,

$$f(\theta_{t+1}) \leq f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|_2^2.$$

Substituting $\theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t)$ into the above and simplifying gives the desired result. □

Next we state and prove the corresponding descent lemma for mirror descent.

Lemma B2. *Let $\Omega, \mathcal{C}$ be closed, convex sets in $\mathbb{R}^d$ with nonempty interior, and $\mathcal{C} \subseteq \Omega$. Assume $f : \Omega \to \mathbb{R}$ is convex on $\mathcal{C}$, and L-smooth with respect a norm $\|\cdot\|$ on $\mathcal{C} \cap \text{int}(\Omega)$, for $L > 0$. Assume $\phi : \Omega \to \mathbb{R}^d$ is of Legendre type, and α-strongly convex with respect to $\|\cdot\|$ on Ω, for $\alpha > 0$. For the mirror descent update in (6) with step size $\eta_t \in (0, 2\alpha/L]$,*

$$f(\theta_{t+1}) \leq f(\theta_t) + \Big(\frac{L}{2} - \frac{\alpha}{\eta_t}\Big)\|\theta_t - \theta_{t+1}\|^2 \leq f(\theta_t).$$

Proof. By L-smoothness,

$$f(\theta_{t+1}) \leq f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2.$$

Meanwhile, by the three-point lemma and the first-order condition for convexity (as derived in part 1 of the proof of Theorem 2),

$$\eta_t\langle \nabla f(\theta_t), \theta_{t+1} - z\rangle \leq D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}) - D_\phi(\theta_{t+1}, \theta_t).$$

and plugging in $z = \theta_t$ gives

$$\eta_t\langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle \leq -D_\phi(\theta_t, \theta_{t+1}) - D_\phi(\theta_{t+1}, \theta_t).$$

Combining this with the L-smoothness inequality, we obtain

$$f(\theta_{t+1}) \le f(\theta_t) - \frac{1}{\eta_t} D_\phi(\theta_t, \theta_{t+1}) - \frac{1}{\eta_t} D_\phi(\theta_{t+1}, \theta_t) + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2.$$

By α-strong convexity, the terms $D_\phi(\theta_t, \theta_{t+1})$ and $D_\phi(\theta_{t+1}, \theta_t)$ are each lower bounded by $\frac{\alpha}{2}\|\theta_{t+1} - \theta_t\|^2$. Plugging this in and simplifying gives the desired result. □

B.2 Proof of Corollary 4

We will establish the two upper bounds given in (11) separately. For the first upper bound, which involves the penalty $\frac{d+1}{2}\|\pi - z\|_1^2$, this is a consequence of (10) after checking the appropriate strong convexity and smoothness properties of the negative entropy ϕ. By Pinsker's inequality, we know that $\alpha = 1$ is the strong convexity parameter of ϕ with respect to $\|\cdot\|_1$. For smoothness, we can calculate, letting $s = z - \pi$,

$$\begin{aligned} DKL(z, \pi) &= \sum_{i=1}^d z_i \log(dz_i) \\ &= \sum_{i=1}^d (s_i + 1/d) \log(1 + ds_i) \\ &\le \sum_{i=1}^d (s_i + 1/d) ds_i \\ &= d \sum_{i=1}^d s_i^2, \end{aligned}$$

where the inequality uses $\log(1 + x) \le x$ for $x \ge -1$. Meanwhile, denoting $C = \sum_{i=1}^d s_i \cdot \mathbb{1}(s_i \ge 0)$, note that $\sum_{i=1}^d s_i \cdot \mathbb{1}(s_i < 0) = -C$ and $\|s\|_1 = 2C$ since $\sum_{i=1}^d s_i = 0$; therefore

$$\begin{aligned} \sum_{i=1}^d s_i^2 &= \sum_{i=1}^d s_i^2 \cdot \mathbb{1}(s_i \ge 0) + \sum_{i=1}^d s_i^2 \cdot \mathbb{1}(s_i < 0) \\ &\le \left(\sum_{i=1}^d |s_i| \cdot \mathbb{1}(s_i \ge 0) \right)^2 + \left(\sum_{i=1}^d |s_i| \cdot \mathbb{1}(s_i < 0) \right)^2 \\ &= 2C^2 \\ &= \|s\|_1^2/2. \end{aligned}$$

Combining this with the second-to-last display therefore gives

$$D_{\mathrm{KL}}(z, \pi) \le \frac{d}{2}\|s\|_1^2,$$

which proves the desired smoothness result with $G = d$.

For the second upper bound, which involves the penalty $\frac{1}{2}\|\pi - z\|_1^2 + \frac{\log d}{2}\|\pi - z\|_1$, we now use a reverse Pinsker-type inequality due to Theorem 1 in Sason (2015):

$$D_{\mathrm{KL}}(z, \pi) \le \frac{\log d}{2(1 - 1/d)}\|z - \pi\|_1.$$

By strong convexity, the triangle inequality, and Young's inequality,

$$D_{\mathrm{KL}}(z, \theta_T) \ge \frac{1}{4}\|\pi - \theta_T\|_1^2 - \frac{1}{2}\|z - \pi\|_1^2.$$

Combining the previous two bounds gives

$$D_{\mathrm{KL}}(z, \pi) - D_{\mathrm{KL}}(z, \theta_T) \le \frac{\log d}{2(1 - 1/d)}\|z - \pi\|_1 - \frac{1}{4}\|\pi - \theta_T\|_1^2 + \frac{1}{2}\|z - \pi\|_1^2.$$

Applying (9) and rearranging proves the desired result.

B.3 Proof of Corollary 5

Parts 1 and 2 follow from arguments analogous to those for Corollary 2. We prove parts 3 and 4 below.

Part 3. By part 2 $\{D_\phi(s,\theta_t)\}_{t=0}^\infty$ is nonincreasing, and by strong convexity $\|s-\theta_t\|^2 \le \frac{2}{\alpha} D_\phi(s,\theta_t)$ for each t. Thus by the Bolzano-Weierstrass theorem there exists a subsequence which converges, i.e., $\lim_{k\to\infty}\theta_{t_k} = \theta_\infty$. By part 1 and continuity of f, we know that $\theta_\infty \in S$.

To see that the whole sequence $\{\theta_t\}_{t=0}^\infty$ converges, suppose not. Then there exists $\delta > 0$ and a subsequence $\{\theta_{t_k}\}_{k=0}^\infty$ such that $\|\theta_{t_k} - \theta_\infty\| \ge \delta$ for all k. With the same argument as used above, we know there exists a subsubsequence $\{\theta_{t_{k_m}}\}_{m=0}^\infty$ which converges to another limit point $\tilde\theta_\infty \in S$ such that $\|\tilde\theta_\infty - \theta_\infty\| \ge \delta$. Now let case (i) denote the assumption that $S \cap \mathrm{int}(\Omega) \neq \emptyset$, and case (ii) denote the assumption that D_ϕ is continuous in its second argument relative to $\mathrm{int}(\Omega)$. Consider the following.

- In case (i), picking any $s \in S \cap \mathrm{int}(\Omega)$, since $\{D_\phi(s,\theta_{t_k})\}_{k=0}^\infty$ is nonincreasing and hence bounded, we know by Theorem 3.8(ii) in Bauschke and Borwein (1997) that $\theta_\infty \in S$ and $D_\phi(\theta_\infty,\theta_{t_k}) \to 0$.
- In case (ii), by definition of relative continuity, and since each $\theta_{t_k} \in \mathrm{int}(\Omega)$ (due to the nature of the mirror descent updates and the Legendre property of ϕ), we have $D_\phi(\theta_\infty,\theta_{t_k}) \to 0$.

In either case, $D_\phi(\theta_\infty,\theta_{t_k}) \to 0$ as $k\to\infty$. By the nonincreasing property of $\{D_\phi(s,\theta_t)\}_{t=0}^\infty$, we have that $D_\phi(\theta_\infty,\theta_t) \to 0$ as $t\to\infty$. By a similar argument applied to the subsubsequence $\{\theta_{t_{k_m}}\}_{m=0}^\infty$, we hence also $D_\phi(\tilde\theta_\infty,\theta_t) \to 0$ as $t\to\infty$. But this leads to the desired contradiction, as the same sequence cannot converge to distinct points $\theta_\infty, \tilde\theta_\infty$. Formally,

$$\max\Big\{D_\phi(\theta_\infty,\theta_t), D_\phi(\tilde\theta_\infty,\theta_t)\Big\} \ge \max\Big\{\frac{\alpha}{2}\|\theta_\infty-\theta_t\|^2, \frac{\alpha}{2}\|\tilde\theta_\infty-\theta_t\|^2\Big\} \ge \frac{\alpha}{2}\Big(\frac{\delta}{2}\Big)^2.$$

Hence, the entire sequence $\{\theta_t\}_{t=0}^\infty$ converges, and $\lim_{t\to\infty}\theta_t = \theta_\infty \in S$.

Part 4. Let $p = \Pi_S^\phi(\theta_0)$ and $v = p - \theta_\infty$. Assume $v \neq 0$. For any $c \in \mathbb{R}$, note that $p + cv \in S$ since S affine. We can obtain three relations regarding p, θ_∞, and $p+cv$. First, the three-point identity for D_ϕ gives

$$D_\phi(p+cv,\theta_\infty) - D_\phi(p+cv,p) - D_\phi(p,\theta_\infty) = \langle \nabla\phi(p) - \nabla\phi(\theta_\infty), p+cv-p\rangle.$$

Second, by part 2, we know $D_\phi(p+cv,\theta_\infty) \le D_\phi(p+cv,\theta_0)$. Third, as S is affine, the generalized Pythagorean theorem for Bregman projection holds with equality:

$$D_\phi(p+cv,\theta_0) = D_\phi(p+cv,p) + D_\phi(p,\theta_0).$$

Combining these three relations, we have

$$\begin{aligned}\langle \nabla\phi(p) - \nabla\phi(\theta_\infty), cv\rangle &= D_\phi(p+cv,\theta_\infty) - D_\phi(p+cv,p) - D_\phi(p,\theta_\infty)\\ &\le D_\phi(p+cv,\theta_0) - D_\phi(p+cv,p) - D_\phi(p,\theta_\infty)\\ &= D_\phi(p,\theta_0) - D_\phi(p,\theta_\infty).\end{aligned}$$

As this holds for all $c \in \mathbb{R}$, we conclude $\nabla\phi(p) = \nabla\phi(\theta_\infty)$. By α-strong convexity,

$$0 = \langle \nabla\phi(p) - \nabla\phi(\theta_\infty), p-\theta_\infty\rangle \ge \alpha\|p-\theta_\infty\|^2,$$

so $v = 0$ and $p = \theta_\infty$. This completes the proof.

C Proofs for Section 3

C.1 Proof of Proposition 1

From (16), we can apply Young's inequality $2ab \le ca^2 + b^2/c$, with $c = \lambda$, to the first term on the right-hand side, yielding

$$\frac{1}{n}\epsilon^\top X(\hat\theta_\lambda - \theta) + 2\lambda\Big(\|\theta\|_2^2 - \|\hat\theta_\lambda\|_2^2\Big) \le \frac{1}{2\lambda}\Big\|\frac{X^\top\epsilon}{n}\Big\|_2^2 + \frac{\lambda}{2}\|\hat\theta_\lambda - \theta\|_2^2 + \lambda\Big(\|\theta\|_2^2 - \|\hat\theta_\lambda\|_2^2\Big)$$

$$\leq \frac{1}{2\lambda}\Big\|\frac{X^\top \epsilon}{n}\Big\|_2^2 + \lambda\Big(\|\hat{\theta}_\lambda\|_2^2 + \|\theta\|_2^2\Big) + \lambda\Big(\|\theta\|_2^2 - \|\hat{\theta}_\lambda\|_2^2\Big)$$
$$= \frac{1}{2\lambda}\Big\|\frac{X^\top \epsilon}{n}\Big\|_2^2 + 2\lambda\|\theta\|_2^2.$$

This proves the first claim in the proposition.

To prove the second claim, since each $\epsilon_i \sim \mathrm{sG}(\sigma^2)$, Theorem 1 of Hsu et al. (2012) implies

$$\mathbb{P}\bigg(\Big\|\frac{X^\top \epsilon}{n}\Big\|_2^2 > \frac{\sigma^2}{n}\Big[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta}\|\widehat{\Sigma}\|_F + 2\delta\|\widehat{\Sigma}\|_{\mathrm{op}}\Big]\bigg) \leq e^{-\delta}, \tag{31}$$

Applying this to the first result of the proposition gives

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \min_{\theta:\|\theta\|_2 \leq b} \mathrm{Risk}(\theta) \leq \frac{R}{2\lambda} + 2\lambda b^2,$$

with probability at least $1 - e^{-\delta}$, where

$$R = \frac{\sigma^2}{n}\Big[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta}\|\widehat{\Sigma}\|_F + 2\delta\|\widehat{\Sigma}\|_{\mathrm{op}}\Big]. \tag{32}$$

Choosing $\lambda = \sqrt{R}/(2b)$, we get that

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \min_{\theta:\|\theta\|_2 \leq b} \mathrm{Risk}(\theta) \leq 2b\sqrt{R},$$

with probability at least $1 - e^{-\delta}$, which completes the proof.

C.2 Proof of Theorem 3

In the Gaussian case $\epsilon_i \sim \mathrm{sG}(\sigma_i^2)$, and in the Bernoulli case $\epsilon_i \sim \mathrm{sG}(1/4)$. The result in either of these cases is then a direct application of Proposition 1. In the Poisson case, the noise $\epsilon_i = y_i - \mu_i$ is not sub-Gaussian so we will use a truncation argument, similar to that in Appendix A.4 of Lin et al. (2017). Define an event

$$\mathcal{E} = \{\epsilon_i < M,\ i = 1, \ldots, n\}, \quad \text{where} \quad M = 4\big(\|\mu\|_\infty + 1/3\big)\log n.$$

Note that $M > 1$ for $n \geq 3$. Further,

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{i=1}^n \mathbb{P}(y_i - \mu_i \geq M).$$

Now using a Poisson concentration result from Pollard (2017), for $X \sim \mathrm{Pois}(\mu)$, and for all $x > 0$,

$$\mathbb{P}(X - \mu \geq x) \leq \exp\bigg(-\frac{x^2}{2\mu}\psi\Big(\frac{x}{\mu}\Big)\bigg), \quad \text{where} \quad \psi(x) = \frac{(1+x)\log(1+x) - x}{x^2/2}.$$

Moreover, when $x \geq 1$,

$$\frac{x^2}{2\mu}\psi\Big(\frac{x}{\mu}\Big) \geq \frac{1/2}{\mu + 1/3}x.$$

Therefore, we have the following for each i,

$$\mathbb{P}(\epsilon_i \geq M) \leq \exp\bigg(-\frac{1/2}{\mu_i + 1/3}M\bigg) \leq \exp\bigg(-\frac{1/2}{\|\mu\|_\infty + 1/3}M\bigg) = 1/n^2,$$

which means that $\mathbb{P}(\mathcal{E}^c) \leq \sum_{i=1}^n 1/n^2 = 1/n$. Now define an event

$$\mathcal{S}_\delta = \bigg\{\Big\|\frac{X^\top \epsilon}{n}\Big\|_2^2 > \frac{\sigma^2}{n}\Big[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta}\|\widehat{\Sigma}\|_F + 2\delta\|\widehat{\Sigma}\|_{\mathrm{op}}\Big]\bigg\},$$

where σ^2 is yet to be specified. Note $\mathbb{P}(\mathcal{S}_\delta) = \mathbb{P}(\mathcal{S}_\delta \cap \mathcal{E}^c) + \mathbb{P}(\mathcal{S}_\delta \cap \mathcal{E}) \leq \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{S}_\delta \cap \mathcal{E}) \leq 1/n + \mathbb{P}(\mathcal{S}_\delta \cap \mathcal{E})$. It remains to bound the latter probability.

Define $\tilde{\epsilon}_i = \epsilon_i \mathbb{1}(\epsilon_i < M)$, and $b_i = \mathbb{E}[\tilde{\epsilon}_i]$. As $\mathbb{E}[\epsilon_i] = 0$, we have $b_i = -\mathbb{E}[\epsilon \mathbb{1}(\epsilon_i \geq M)]$. By Cauchy-Schwarz, $|b_i|^2 \leq \mu_i \mathbb{P}(\epsilon_i \geq M) \leq \mu_i / n^2$. Finally, let $\bar{\epsilon}_i = \tilde{\epsilon}_i - b_i$, and $\bar{M} = M + \sqrt{\|\mu\|_\infty}/n$. Then $\bar{\epsilon}_i$ has mean zero and is bounded in magnitude by $\bar{M}$, and so it is sub-Gaussian with parameter $\bar{\sigma}^2 = \bar{M}^2$.

This means that we can apply (31) to bound the tail probability of $\|X^\top \bar{\epsilon}/n\|_2^2$, concretely

$$\mathbb{P}\left(\left\| \frac{X^\top \bar{\epsilon}}{n} \right\|_2^2 > \frac{\bar{\sigma}^2}{n} \left[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta} \|\widehat{\Sigma}\|_F + 2\delta \|\widehat{\Sigma}\|_{\mathrm{op}} \right] \right) \leq e^{-\delta}.$$

Now observe

$$\begin{aligned}
\left\| \frac{X^\top \tilde{\epsilon}}{n} \right\|_2^2 &\leq 2 \left\| \frac{X^\top \bar{\epsilon}}{n} \right\|_2^2 + 2 \left\| \frac{X^\top b}{n} \right\|_2^2 \\
&\leq 2 \left\| \frac{X^\top \bar{\epsilon}}{n} \right\|_2^2 + \frac{2}{n} \|b\|_2^2 \|\widehat{\Sigma}\|_{\mathrm{op}} \\
&\leq 2 \left\| \frac{X^\top \bar{\epsilon}}{n} \right\|_2^2 + \frac{2}{n^2} \|\mu\|_\infty \|\widehat{\Sigma}\|_{\mathrm{op}}.
\end{aligned}$$

This means that

$$\mathbb{P}\left(\left\| \frac{X^\top \tilde{\epsilon}}{n} \right\|_2^2 > \frac{2\bar{\sigma}^2}{n} \left[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta} \|\widehat{\Sigma}\|_F + 2\delta \|\widehat{\Sigma}\|_{\mathrm{op}} \right] + \frac{2\|\mu\|_\infty}{n^2} \|\widehat{\Sigma}\|_{\mathrm{op}} \right) \leq e^{-\delta}.$$

Provided $\delta \geq 1$,

$$\begin{aligned}
\frac{2\bar{\sigma}^2}{n} \left[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta} \|\widehat{\Sigma}\|_F + 2\delta \|\widehat{\Sigma}\|_{\mathrm{op}} \right] + \frac{2\|\mu\|_\infty}{n^2} \|\widehat{\Sigma}\|_{\mathrm{op}} &\leq \frac{2\bar{\sigma}^2 + \|\mu\|_\infty/n}{n} \left[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta} \|\widehat{\Sigma}\|_F + 2\delta \|\widehat{\Sigma}\|_{\mathrm{op}} \right] \\
&\leq \frac{\sigma^2}{n} \left[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta} \|\widehat{\Sigma}\|_F + 2\delta \|\widehat{\Sigma}\|_{\mathrm{op}} \right],
\end{aligned}$$

where the last line follows by calculating and defining $\sqrt{2\bar{\sigma}^2 + \|\mu\|_\infty/n} \leq \sqrt{2}(M + \sqrt{\|\mu\|_\infty}/n) + \sqrt{\|\mu\|_\infty/n}$ $\leq 6(\|\mu\|_\infty + 1/3) \log n + 3\sqrt{\|\mu\|_\infty} := \sigma$. Putting these pieces together, we have shown

$$\mathbb{P}\left(\left\| \frac{X^\top \tilde{\epsilon}}{n} \right\|_2^2 > \frac{\sigma^2}{n} \left[\mathrm{tr}(\widehat{\Sigma}) + 2\sqrt{\delta} \|\widehat{\Sigma}\|_F + 2\delta \|\widehat{\Sigma}\|_{\mathrm{op}} \right] \right) \leq e^{-\delta}.$$

Note that on $\mathcal{E}$, each $\tilde{\epsilon}_i = \epsilon_i$, and therefore $\mathbb{P}(\mathcal{S}_\delta \cap \mathcal{E}) \leq e^{-\delta}$ as well. This proves that $\mathbb{P}(\mathcal{S}_\delta) \leq 1/n + e^{-\delta}$, as desired. The remainder of the proof follows by choosing λ as before, and bounding the subsequent excess risk.

C.3 Proof of Proposition 2

From (20), we can apply Young's inequality $2ab \leq ca^2 + b^2/c$, with $c = \lambda$, to the first term on the right-hand side. This proves the first claim in the proposition.

To prove the second claim, combining (31) with the first result of the proposition gives

$$\mathrm{Risk}(\theta_T) - \min_{\theta : \|\theta\|_2 \leq b} \mathrm{Risk}(\theta) \leq \frac{1}{2\lambda_T} R + \frac{\lambda_T}{2} b^2,$$

with probability at least $1 - e^{-\delta}$, where R is defined in (32). At $\lambda_T^* = \sqrt{R}/b$, this becomes

$$\mathrm{Risk}(\theta_T) - \min_{\theta : \|\theta\|_2 \leq b} \mathrm{Risk}(\theta) \leq b\sqrt{R},$$

with probability at least $1 - e^{-\delta}$. In the case $\lambda_T^* = \sqrt{R}/b$ is obtainable by $\lambda_T = 1/(\eta T)$ for an integer T, this establishes the result in (22) for gradient descent. Otherwise, we set $T = \lceil \frac{1}{\eta \lambda_T^*} \rceil$; then $1/(\eta \lambda_T) \leq 1/(\eta \lambda_T^*) + 1$, i.e., $1/\lambda_T \leq 1/\lambda_T^* + \eta$, so by Lemma C1 (presented below),

$$\left(\frac{1}{2\lambda_T} R + \frac{\lambda_T}{2} b^2 \right) - \left(\frac{1}{2\lambda_T^*} R + \frac{\lambda^*}{2} b^2 \right) \leq \frac{\eta R}{2}.$$

Thus in this case, we have an additional discretization error term of at most $e_T = \eta R/2$. This completes the proof of the proposition.

Lemma C1. *Consider a function $g(x) = \frac{a}{x} + bx$ on $(0, \infty)$ with $a, b > 0$. This is minimized at $x^* = \sqrt{a/b}$. If $y > 0$ satisfies $1/y = 1/x^* + c$ with $c \geq 0$, then*

$$g(y) - g(x^*) = ac^2 y \leq ac.$$

Proof. Observe

$$\begin{aligned} g(y) - g(x^*) &= a\Big(\frac{1}{y} - \frac{1}{x^*}\Big) + b(y - x^*) \\ &= ac - b\Big(\frac{1}{y} - \frac{1}{x^*}\Big)yx^* \\ &= c(a - byx^*) \\ &= c\Big(a - \frac{ay}{x^*}\Big) \\ &= ac^2 y \\ &\leq ac, \end{aligned}$$

where the fourth line uses $bx^* = a/x^*$, and the last uses $cy = (1/y - 1/x^*)y \leq 1$. □

C.4 Proof of Theorem 4

The theorem follows from Proposition 2 in the same way that Theorem 3 follows from Proposition 1. It only remains to identify (local) Lipschitz parameters L for the individual GLMs. Note that in general

$$\nabla^2 f(\theta) = \frac{1}{n} X^\top \nabla^2 \mathsf{A}(X\theta) X = \frac{1}{n} \sum_{i=1}^n x_i A''(x_i^\top \theta) x_i^\top.$$

- For case 1: Gaussian, we have $A(u) = u^2/2$ and $A''(u) = 1$. Hence $\nabla^2 f(\theta) = \widehat{\Sigma}$ and $L = \|\widehat{\Sigma}\|_{\mathrm{op}}$ is the global Lipschitz parameter.
- For case 2: Bernoulli, we have $A(u) = \log(1 + e^u)$ and $A''(u) = e^u/(1 + e^u)^2 \leq \frac{1}{4}$. Hence $\nabla^2 f(\theta) \preceq \frac{1}{4}\widehat{\Sigma}$ and $L = \frac{1}{4}\|\widehat{\Sigma}\|_{\mathrm{op}}$ is the global Lipschitz parameter.
- For case 3: Poisson, we have $A(u) = e^u$ and $A''(u) = e^u$. Note that $x_i^\top \theta \leq \|x_i\|_2 \|\theta\|_2 = bm$, for $\|\theta\|_2 \leq b$ and $m = \max_{i=1,\ldots,n} \|x_i\|_2$. Hence $\nabla^2 f(\theta) \preceq e^{bm}\widehat{\Sigma}$ and $L = e^{bm}\|\widehat{\Sigma}\|_{\mathrm{op}}$ is the local Lipschitz parameter on $\mathsf{B}_d(b)$.

D Proofs for Section 4

D.1 Proof of Proposition 3

The first claim is a special case of a more general result for Bregman-divergence-regularized GLMs, given below in Proposition 7. In particular, this claim follows by setting ϕ to be the negative entropy function on $\Omega = \mathcal{C} = \Delta_d$. This is 1-strongly convex with respect to $\|\cdot\|_1$ by Pinsker's inequality.

To prove the second claim, note that each $X_j^\top \epsilon \sim \mathrm{sG}(\sigma^2 \|X_j\|_2^2)$. Thus, by a standard maximal inequality for sub-Gaussian random variables (Tibshirani, 2023),

$$\mathbb{P}\Big(\|X^\top \epsilon\|_\infty > \sigma\sqrt{2n(\log(2d) + \delta)}\Big) \leq e^{-\delta}. \tag{33}$$

Applying this to the first result of the proposition gives

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \min_{\theta : D_{\mathrm{KL}}(\theta, u) \leq b} \mathrm{Risk}(\theta) \leq \frac{R^2}{\lambda n^2} + 2\lambda b,$$

with probability at least $1 - e^{-\delta}$, where

$$R = \sigma\sqrt{2n(\log(2d) + \delta)}. \tag{34}$$

Choosing $\lambda = R/(\sqrt{2b}n)$, we get that

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \min_{\theta:\|\theta\|_2 \le b} \mathrm{Risk}(\theta) \le \frac{2\sqrt{2b}R}{n},$$

with probability at least $1 - e^{-\delta}$, which completes the proof.

Proposition 7. *Assume that $\phi : \Omega \to \mathbb{R}$ is α-strongly convex on $\mathcal{C}$ with respect to a norm $\|\cdot\|$, for $\alpha > 0$, and let $\|\cdot\|_*$ be the corresponding dual norm. Consider the regularized GLM estimator defined by augmenting the GLM loss (13) with a Bregman divergence penalty, denoted*

$$\hat{\theta}_\lambda = \operatorname*{arg\,min}_{\theta \in \mathcal{C}} \Big(f(\theta) + \lambda D_\phi(\theta, u) \Big),$$

for $\lambda > 0$, and an anchor point $u \in \mathcal{C}$. Then for any reference point $\theta \in \mathcal{C}$, the prediction risk of $\hat{\theta}_\lambda$ satisfies

$$\mathrm{Risk}(\hat{\theta}_\lambda) \le \mathrm{Risk}(\theta) + \frac{1}{\alpha\lambda} \Big\| \frac{X^\top \epsilon}{n} \Big\|_*^2 + 2\lambda D_\phi(\theta, u).$$

Proof. Following analogous steps to those leading up to (16),

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \mathrm{Risk}(\theta) \le \frac{1}{n} \epsilon^\top X(\hat{\theta}_\lambda - \theta) + \lambda \Big(D_\phi(\theta, u) - D_\phi(\hat{\theta}_\lambda, u) \Big).$$

Using Hölder's inequality (with respect to the pair $\|\cdot\|$ and $\|\cdot\|_*$) and Young's inequality $2ab \le ca^2 + b^2/c$, with $c = \alpha\lambda$, we have

$$\begin{aligned}
\frac{1}{n} \epsilon^\top X(\hat{\theta}_\lambda - \theta) &\le \Big\| \frac{X^\top \epsilon}{n} \Big\|_* \|\hat{\theta}_\lambda - \theta\| \\
&\le \Big\| \frac{X^\top \epsilon}{n} \Big\|_* \|\hat{\theta}_\lambda - u\| + \Big\| \frac{X^\top \epsilon}{n} \Big\|_* \|\theta - u\| \\
&\le \frac{1}{2\alpha\lambda} \Big\| \frac{X^\top \epsilon}{n} \Big\|_*^2 + \frac{\alpha\lambda}{2} \|\hat{\theta}_\lambda - u\|^2 + \Big\| \frac{X^\top \epsilon}{n} \Big\|_* \|\theta - u\|.
\end{aligned}$$

By strong convexity $D_\phi(u, v) \ge (\alpha/2)\|u - v\|^2$, we obtain

$$\begin{aligned}
\mathrm{Risk}(\hat{\theta}_\lambda) - \mathrm{Risk}(\theta) &\le \frac{1}{2\alpha\lambda} \Big\| \frac{X^\top \epsilon}{n} \Big\|_*^2 + \frac{\alpha\lambda}{2} \|\hat{\theta}_\lambda - u\|^2 + \Big\| \frac{X^\top \epsilon}{n} \Big\|_* \|\theta - u\| + \lambda \Big(D_\phi(\theta, u) - D_\phi(\hat{\theta}_\lambda, u) \Big) \\
&\le \frac{1}{2\alpha\lambda} \Big\| \frac{X^\top \epsilon}{n} \Big\|_*^2 + \Big\| \frac{X^\top \epsilon}{n} \Big\|_* \|\theta - u\| + \lambda D_\phi(\theta, u) \\
&\le \frac{1}{\alpha\lambda} \Big\| \frac{X^\top \epsilon}{n} \Big\|_*^2 + \frac{\alpha\lambda}{2} \|\theta - u\|^2 + \lambda D_\phi(\theta, u) \\
&= \frac{1}{\alpha\lambda} \Big\| \frac{X^\top \epsilon}{n} \Big\|_*^2 + 2\lambda D_\phi(\theta, u),
\end{aligned}$$

where the second-to-last step uses Young's inequality again, and the last step uses strong convexity. □

D.2 Proof of Theorem 5

This follows precisely as in the proof of Theorem 3: for the Gaussian and Bernoulli cases we simply identify σ in the sub-Gaussian parameterization, and for the Poisson case, we use a truncation argument.

D.3 Proof of Proposition 4

The first claim is again a special case of a more general result for mirror descent on a GLM loss, given below in Proposition 8, applied to the case where ϕ is the negative entropy function on $\Omega = \mathcal{C} = \Delta_d$. The second

claim follows from arguments just as in the proof of Proposition 3, using the sub-Gaussian maximal inequality (33). This gives

$$\mathrm{Risk}(\theta_T) - \min_{\theta: D_{\mathrm{KL}}(\theta,u)\le b} \mathrm{Risk}(\theta) \le \frac{R^2}{2\lambda n^2} + \lambda b,$$

with probability at least $1 - e^{-\delta}$, where R is defined in (34). At $\lambda_T^* = R/(\sqrt{2b}n)$, this gives

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \min_{\theta: \|\theta\|_2 \le b} \mathrm{Risk}(\theta) \le \frac{\sqrt{2b}R}{n},$$

which proves the result in (27) for the case of integral T^*. For non-integral T^*, we then apply an analogous discretization argument based on Lemma C1, which gives

$$\Big(\frac{R^2}{2\lambda_T n^2} + \lambda_T b\Big) - \Big(\frac{R^2}{2\lambda_T^* n^2} + \lambda_T^* b\Big) \le \frac{R^2}{2n^2}\eta.$$

This verifies the additional discretization error term of at most $e_T = \eta R^2/(2n^2)$, and completes the proof.

Proposition 8. *Under the assumptions of Proposition 7, assume additionally also that the GLM loss f is L-smooth on $\mathcal{C}$. Consider mirror descent (6) with $\eta_t = \eta \in (0, 1/L]$, $t = 0, 1, 2, \ldots$, initialized at $\theta_0 = u \in \mathcal{C}$. Then for any reference point $\theta \in \mathcal{C}$, and for any $T \ge 1$ and $\lambda_T = 1/(\eta T)$,*

$$\mathrm{Risk}(\theta_T) - \mathrm{Risk}(\theta) \le \frac{1}{2\alpha\lambda_T}\Big\|\frac{X^\top \epsilon}{n}\Big\|_*^2 + \lambda_T D_\phi(\theta, u).$$

Proof. By the basic inequality (9) in Theorem 2,

$$f(\theta_T) - f(\theta) \le \lambda_T D_\phi(\theta, u) - \lambda_T D_\phi(\theta, \theta_T).$$

Following analogous steps to those leading up to (20), and then arguments as in the proof of Proposition 7,

$$\begin{aligned}
\mathrm{Risk}(\theta_T) - \mathrm{Risk}(\theta) &\le \frac{1}{n}\epsilon^\top X(\theta_T - \theta) + \lambda_T\Big(D_\phi(\theta, u) - D_\phi(\theta, \theta_T)\Big) \\
&\le \Big\|\frac{X^\top \epsilon}{n}\Big\|_* \|\theta_T - \theta\| + \lambda_T\Big(D_\phi(\theta, u) - D_\phi(\theta, \theta_T)\Big) \\
&\le \frac{1}{2\alpha\lambda_T}\Big\|\frac{X^\top \epsilon}{n}\Big\|_*^2 + \frac{\alpha\lambda_T}{2}\|\theta_T - \theta\|^2 + \lambda_T\Big(D_\phi(\theta, u) - D_\phi(\theta, \theta_T)\Big) \\
&\le \frac{1}{2\alpha\lambda_T}\Big\|\frac{X^\top \epsilon}{n}\Big\|_*^2 + \lambda_T D_\phi(\theta, u).
\end{aligned}$$

□

D.4 Proof of Theorem 6

After identifying the Lipschitz constant for each GLM loss, the arguments are the identical as those in the proof of Theorem 4. Recall the following equivalent characterization of L-smoothness with respect to $\|\cdot\|_1$ (Sidford, 2020):

$$|z^\top \nabla^2 f(\theta) z| \le L\|z\|_1^2,$$

for all $\theta, z \in \Delta_d$. Furthermore,

$$|z^\top \nabla^2 f(\theta) z| = \Big|\frac{1}{n}\sum_{i=1}^n A''(x_i^\top\theta)(x_i^\top z)^2\Big| \le \frac{\|z\|_1^2}{n}\sum_{i=1}^n A''(x_i^\top\theta)\|x_i\|_\infty^2,$$

by Hölder's inequality. Thus it suffices to take $L \ge \frac{1}{n}\sum_{i=1}^n A''(x_i^\top\theta)\|x_i\|_\infty^2$.

- For case 1: Gaussian, we have $A(u) = u^2/2$ and $A''(u) = 1$. Hence we may take $L = \frac{1}{n}\sum_{i=1}^n \|x_i\|_\infty^2$.
- For case 2: Bernoulli, we have $A(u) = \log(1 + e^u)$ and $A''(u) = e^u/(1 + e^u)^2 \le \frac{1}{4}$. Hence we may take $L = \frac{1}{4n}\sum_{i=1}^n \|x_i\|_\infty^2$.
- For case 3: Poisson, we have $A(u) = e^u$ and $A''(u) = e^u$. Again by Hölder's, we have $x_i^\top\theta \le \|x_i\|_\infty$ for $\theta \in \Delta_d$. Hence we may take $L = \frac{1}{n}\sum_{i=1}^n e^{\|x_i\|_\infty}\|x_i\|_\infty^2$.

E Proofs for Section 5

E.1 Proof of Proposition 5

Note $f(\theta) = \mathbb{E}_{\beta\sim\theta}[\widehat{R}(\beta)] = \sum_{j=1}^d \theta_j \widehat{R}(X_j)$. As this is linear in θ, its gradient is constant: $\nabla_j f(\theta) = \widehat{R}(X_j)$. The exponentiated gradient descent iterates in (7) can therefore be written as

$$\theta_{t+1}(\mathrm{d}\beta) \propto \exp(-\eta\widehat{R}(\beta)) \cdot \theta_t(\mathrm{d}\beta).$$

Starting at $\theta_0 = u$, we see that this matches (28) at $1/\lambda = \eta T$.

E.2 Proof of Proposition 6

By the basic inequality (9) in Theorem 2,

$$\mathbb{E}_{\beta\sim\theta_T}[\widehat{R}(\beta)] - \mathbb{E}_{\beta\sim\theta}[\widehat{R}(\beta)] \le \lambda_T\Big(D_{\mathrm{KL}}(\theta,u) - D_{\mathrm{KL}}(\theta,\theta_T)\Big).$$

Write $\nu = \theta - \theta_T$, and correspondingly

$$\mathbb{E}_{\beta\sim\theta_T}[\widehat{R}(\beta)] - \mathbb{E}_{\beta\sim\theta}[\widehat{R}(\beta)] = \int -\widehat{R}(\beta)\nu(\mathrm{d}\beta).$$

Using the same representation for test risk,

$$\begin{aligned}
\mathbb{E}_{\beta\sim\theta_T}[R(\beta)] - \mathbb{E}_{\beta\sim\theta}[R(\beta)] &= \int -R(\beta)\nu(\mathrm{d}\beta) \\
&= \int \Big[\widehat{R}(\beta) - R(\beta)\Big]\nu(\mathrm{d}\beta) + \int -\widehat{R}(\beta)\nu(\mathrm{d}\beta) \\
&\le \int \Big[\widehat{R}(\beta) - R(\beta)\Big]\nu(\mathrm{d}\beta) + \lambda_T\Big(D_{\mathrm{KL}}(\theta,u) - D_{\mathrm{KL}}(\theta,\theta_T)\Big) \\
&\le \|\widehat{R} - R\|_\infty \|\nu\|_1 + \lambda_T\Big(D_{\mathrm{KL}}(\theta,u) - D_{\mathrm{KL}}(\theta,\theta_T)\Big) \\
&\le \frac{1}{2\lambda_T}\|\widehat{R} - R\|_\infty^2 + \frac{\lambda_T}{2}\|\nu\|_1^2 + \lambda_T\Big(D_{\mathrm{KL}}(\theta,u) - D_{\mathrm{KL}}(\theta,\theta_T)\Big) \\
&\le \frac{1}{2\lambda_T}\|\widehat{R} - R\|_\infty^2 + \lambda_T D_{\mathrm{KL}}(\theta,u).
\end{aligned}$$

The third line uses the basic inequality, the fourth Hölder's inequality, the fifth Young's inequality, and the last Pinsker's inequality. This proves the first claim in the proposition.

To prove the second claim, by a union bound and Hoeffding's inequality,

$$\begin{aligned}
P\left(\|\widehat{R} - R\|_\infty > C\sqrt{\frac{\log(2d)+\delta}{2n}}\right) &\le d \cdot \sup_\beta \mathbb{P}\left(|\widehat{R}(\beta) - R(\beta)| > C\sqrt{\frac{\log(2d)+\delta}{2n}}\right) \\
&\le 2d\exp\left(-\frac{2C^2(\log(2d)+\delta)/2n}{C^2/n}\right) = e^{-\delta}.
\end{aligned}$$

Combining this with the first result in the proposition, it follows that with probability at least $1 - e^{-\delta}$,

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}[R(\beta)] - \min_{\theta: D_{\mathrm{KL}}(\theta,u)\le b} \mathbb{E}_{\beta\sim\theta}[R(\beta)] \le \frac{C^2(\log(2d)+\delta)}{4n\lambda} + \lambda b.$$

Finally, setting λ as defined in the proposition establishes the second claim, and completes the proof.

F Further details for Section 6

Optimization details: implicit regularization. For GD and EGD, we use learning rate schedules to cover small τ (i.e., early training) with high resolution and to reach large τ (i.e., later training) with fewer iterations. Tables 2 and 3 list the schedules for GD and EGD across three GLMs and two parameterization regimes. The schedule $\{(\eta^{(k)}, T^{(k)})\}_{k=1}^m$ means that the learning rate $\eta^{(1)}$ is used for $T^{(1)}$ iterations, then $\eta^{(2)}$ is used for the next $T^{(2)}$ iterations, and so on.

GLM	GD	
	Underparameterized $(n, d) = (200, 20)$	Overparameterized $(n, d) = (100, 200)$
Linear	$\{(10^{-4}, 10^4),\ (10^{-3}, 10^5),\ (10^{-2}, 10^5)\}$	same as Underparmeterized
Logistic	same as Linear	same as Underparmeterized
Poisson	same as Linear	$\{(10^{-4}, 10^5),\ (2 \times 10^{-4}, 2 \times 10^5),\ (5 \times 10^{-4}, 2 \times 10^6)\}$

Table 2: GD learning rate schedules.

GLM	EGD	
	Underparameterized $(n, d) = (200, 20)$	Overparameterized $(n, d) = (30, 60)$
Linear	$\{(10^{-4}, 10^5), (10^{-3}, 10^5), (10^{-2}, 10^5), (10^{-1}, 10^5)]\}$	same as Underparmeterized
Logistic	same as Linear	same as Underparmeterized
Poisson	same as Linear	same as Underparmeterized

Table 3: EGD learning rate schedules.

Optimization details: explicit regularization. In each GLM task and in each parameterization regime, we solve 500 regularized problems with λ values log-spaced on $[10^{-4}, 10^4]$. We use `scipy.optimize.minimize` from the `SciPy` library; in particular, we use the `L-BFGS-B` solver for ridge-regularized estimates and we use the `SLSQP` solver for KL-regularized estimates.

Training dynamics with KL penalty. Recall in Figure 1 panel (b) we investigated training dynamics for EGD vs. KL-regularized estimates, where the training envelope $P(\theta, \lambda) = f(\theta) + \lambda g(\theta)$ is defined using a penalty $g(\theta) = \|\theta - \pi\|_1^2$ (as per Corollary 3). Figure 4 examines training dynamics with the envelope defined using a KL penalty. The results are overall similar to Figure 1.

G Proofs for Section 7

G.1 Proof of Theorem 7

The proximal gradient descent update (30) can be written as

$$\theta_{t+1} = \theta_t - \eta_t G_{\eta_t}(\theta_t), \tag{35}$$

where $G_\eta(\theta) = \frac{1}{\eta}(\theta - \mathrm{prox}_{\eta f_1}(\theta - \eta \nabla f_0(\theta)))$ is called the generalized gradient. Using this notation, the proof is similar that for the gradient descent case in Theorem 1, with $G_{\eta_t}(\theta_t)$ in place of $\nabla f(\theta_t)$.

Step 1: *Tracking the proximity difference across adjacent iterations.* As before,

$$\|\theta_t - z\|_2^2 - \|\theta_{t+1} - z\|_2^2 = \|\theta_t - z\|_2^2 - \|\theta_t - \eta_t G_{\eta_t}(\theta_t) - z\|_2^2 = 2\eta_t \langle G_{\eta_t}(\theta_t), \theta_t - z\rangle - \eta_t^2 \|G_{\eta_t}(\theta_t)\|_2^2.$$

Step 2: *Bounding the objective difference* $f(\theta_t) - f(z)$. The generalized descent lemma for proximal gradient descent (Lemma G1 below) guarantees

$$2\eta_t (f(\theta_{t+1}) - f(z)) \le 2\eta_t \langle G_{\eta_t}(\theta_t), \theta_t - z\rangle - \eta_t^2 \|G_{\eta_t}(\theta_t)\|_2^2.$$

Moreover, the same lemma with $z = \theta_t$ gives $f(\theta_{t+1}) \le f(\theta_t) - (\eta_t/2)\|G_{\eta_t}(\theta_t)\|_2^2$, which in particular shows that f is noncreasing along the iterate sequence, and

$$2\eta_t (f(\theta_T) - f(z)) \le 2\eta_t \langle G_{\eta_t}(\theta_t), \theta_t - z\rangle - \eta_t^2 \|G_{\eta_t}(\theta_t)\|_2^2.$$

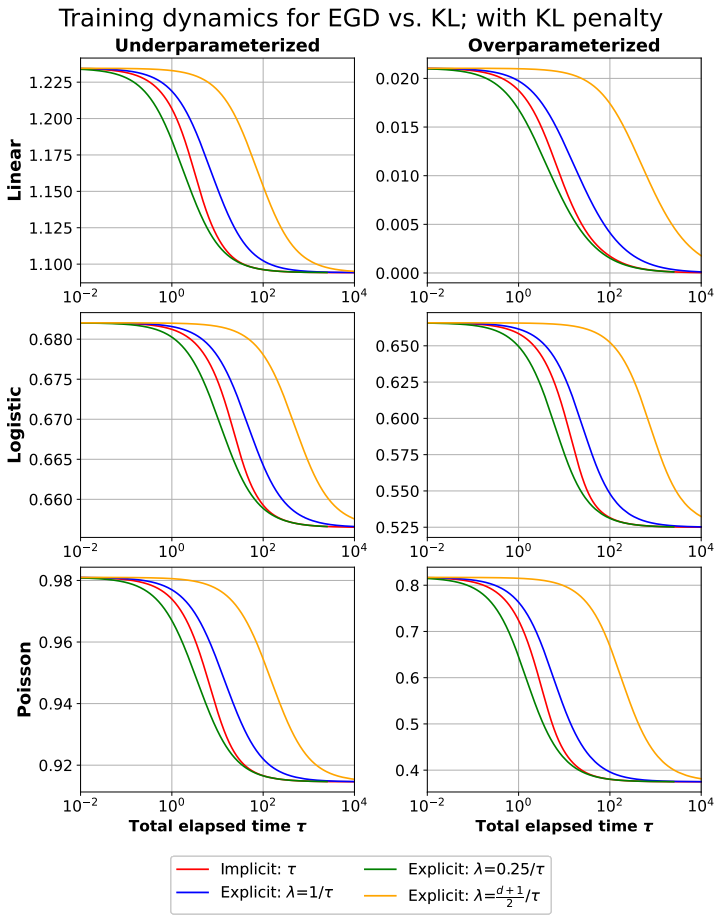


Figure 4: Training envelope for model aggregation, with penalty $g(\theta) = D_{\mathrm{KL}}(\theta, \pi)$.

Combined with the result of Step 1,

$$2\eta_t(f(\theta_T) - f(z)) \le \|\theta_t - z\|_2^2 - \|\theta_{t+1} - z\|_2^2.$$

The remainder of the proof (Step 3) follows exactly as before.

Lemma G1. *Let $f_0 : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and L-smooth with $L > 0$, and let $f_1 : \mathbb{R}^d \to \mathbb{R}$ be convex. For $f = f_0 + f_1$ and any $z \in \mathbb{R}^d$, the proximal gradient update in (35) with $\eta_t \in (0, 2/L]$ satisfies*

$$f(\theta_{t+1}) \le f(\theta_t) + \langle G_{\eta_t}(\theta_t), \theta_t - z\rangle - \eta_t\Big(1 - \frac{L}{2}\eta_t\Big)\|G_{\eta_t}(\theta_t)\|_2^2.$$

For $\eta_t \le 1/L$, this implies

$$f(\theta_{t+1}) \le f(\theta_t) + \langle G_{\eta_t}(\theta_t), \theta_t - z\rangle - \frac{\eta_t}{2}\|G_{\eta_t}(\theta_t)\|_2^2.$$

Proof. By L-smoothness,

$$\begin{aligned} f_0(\theta_{t+1}) &\le f_0(\theta_t) + \langle \nabla f_0(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|_2^2 \\ &\le f_0(\theta_t) - \langle \nabla f_0(\theta_t), \eta_t G_{\eta_t}(\theta_t)\rangle + \frac{L\eta_t^2}{2}\|G_{\eta_t}(\theta_t)\|_2^2 \\ &\le f_0(z) + \langle \nabla f_0(\theta_t), \theta_t - \eta_t G_{\eta_t}(\theta_t) - z\rangle + \frac{L\eta_t^2}{2}\|G_{\eta_t}(\theta_t)\|_2^2, \end{aligned}$$

where in the last line we used $f(\theta_t) \leq f_0(z) + \nabla f_0(\theta_t)^\mathsf{T}(\theta_t - z)$ by convexity. Moreover, by the subgradient optimality condition defining the proximal operator,

$$\theta_t - \eta_t \nabla f_0(\theta_t) - \theta_{t+1} \in \eta_t \partial f_1(\theta_{t+1}),$$

which may be equivalently written as

$$G_{\eta_t}(\theta_t) - \nabla f_0(\theta_t) \in \partial f_1(\theta_{t+1}).$$

This means that

$$f_1(\theta_{t+1}) \leq f_1(z) + \langle G_{\eta_t}(\theta_t) - \nabla f_0(\theta_t), \theta_t - \eta_t G_{\eta_t}(\theta_t) - z \rangle.$$

Combining this with the above inequality on f_0, and using $f = f_0 + f_1$, we obtain

$$f(\theta_{t+1}) \leq f(\theta_t) + \langle G_{\eta_t}(\theta_t), \theta_t - z \rangle - \eta_t \|G_{\eta_t}(\theta_t)\|_2^2 + \frac{L\eta_t^2}{2} \|G_{\eta_t}(\theta_t)\|_2^2,$$

which can be regrouped to give the desired result. □

G.2 Proof of Theorem 8

The proof is similar that for the mirror descent case in Theorem 2.

Step 1: *Tracking the proximity difference across adjacent iterations.* By the same argument as before,

$$\eta_t \langle \nabla f(\theta_t), \theta_{t+1} - z \rangle \leq D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}) - D_\phi(\theta_{t+1}, \theta_t).$$

Step 2: *Bounding the objective difference* $f(\theta_t) - f(z)$. By convexity of $L\phi - f$,

$$L\phi(\theta_{t+1}) - f(\theta_{t+1}) \geq L\phi(\theta_t) - f(\theta_t) + \langle L\nabla\phi(\theta_t) - \nabla f(\theta_t), \theta_{t+1} - \theta_t \rangle,$$

and rearranging gives

$$f(\theta_{t+1}) \leq f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t \rangle + L D_\phi(\theta_{t+1}, \theta_t).$$

Moreover, $f(\theta_t) \leq f(z) + \langle \nabla f(\theta_t), \theta_t - z \rangle$ by convexity, thus

$$f(\theta_{t+1}) \leq f(z) + \langle \nabla f(\theta_t), \theta_{t+1} - z \rangle + L D_\phi(\theta_{t+1}, \theta_t).$$

Combining this with the result of Step 1, we have

$$\begin{aligned} \eta_t(f(\theta_{t+1}) - f(z)) &\leq \eta_t D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}) - (1 - L\eta_t) D_\phi(\theta_{t+1}, \theta_t) \\ &\leq D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}), \end{aligned}$$

where we used $\eta_t \leq 1/L$ in the last inequality. Taking $z = \theta_t$, this shows us that f is noncreasing along the iterate sequence, and therefore

$$\eta_t(f(\theta_T) - f(z)) \leq D_\phi(z, \theta_t) - D_\phi(z, \theta_{t+1}).$$

The remainder of the proof (Step 3) follows exactly as before.