Basic Inequalities for Optimization Algorithms with Applications to Statistical Risk Analysis

Seunghoon Paik[†], Kangjie Zhou[‡], Ryan J. Tibshirani[†]

[†]University of California, Berkeley, [‡]Columbia University

Draft – Oct 2025

Abstract

Implicit regularization, the inductive bias of a learning algorithm to prefer simpler solutions over more complex ones, is a topic of huge interest in statistic modeling and modern machine learning. A prominent approach for its analysis is to relate it to the corresponding explicit regularization scheme, yet a unifying framework to formalize this connection has been absent. We introduce such a framework, using the so-called *basic inequality*, a key tool that connects the dynamics of an optimization algorithm to its explicit regularization counterpart. While related inequalities are fundamental in optimization theory, we isolate and highlight a specific form as a simple and versatile tool, which we believe has been underappreciated. Specifically, for a given iterative algorithm, a basic inequality provides an upper bound on the objective value at its last iterate, $f(\theta_T)$, with respect to an arbitrary reference point z. This bound is characterized by two factors: a geometry-aware distance between the initialization, the last iterate θ_T, and the reference point z; and accumulated step sizes, representing total elapsed time of the algorithm. We demonstrate the utility of this framework in both optimization and statistical perspective, in the application of training dynamics, prediction risk of statistical models, and randomized model selection. We supplement our theoretical results with experiments on generalized linear models with gradient descent and exponential gradient descent.

1 Introduction

This paper introduces *basic inequalities* for iterative optimization algorithms, a framework that connects the implicit and explicit regularization. Given an optimization problem, $\min_\theta f(\theta)$, our focus is on

$$\theta_T \quad \text{and} \quad \hat{\theta}_\lambda,$$

where θ_T is the last iterate of an algorithm, and $\hat{\theta}_\lambda = \arg\min_\theta f(\theta) + \lambda g(\theta)$ is an explicitly regularized solution with a penalty $g(\cdot)$ and regularization parameter $\lambda \geq 0$.

The concept of *implicit regularization* (Neyshabur et al., 2014), where the optimization algorithm itself implicitly induces the bias of the model, has been a persistent theme in statistics and optimization literature. This is a specific instance of broader *inductive bias* of models (Baxter, 2000), which describes model capacity and generalization. A well-known example of implicit regularization is *early stopping* of iterative algorithms, which appeared in the neural network community (Prechelt, 2002) and later more formally analyzed in various statistical context (Zhang and Yu, 2005; Yao et al., 2007; Raskutti et al., 2014). Implicit regularization is distinct from *explicit regularization*, a classical approach of adding a penalty term to the loss function, whose statistical properties have been studied more in the literature.

One major research direction has been to understand implicit regularization by connecting it to explicit regularization (Lemaire, 1996; Gunasekar et al., 2018). While connections have been usually established with algorithm- or loss-specific context (Suggala et al., 2018; Ji and Telgarsky, 2019; Ali et al., 2019), a general and unifying framework is absent. This paper proposes that basic inequalities can serve as this framework.

For a given algorithm, a basic inequality bounds the objective value at the last iterate, $f(\theta_T)$ relative to any reference point z. This bound is governed by two factors: a geometry-aware distance between the

initialization, last iterates θ_T, and z; and accumulated step sizes, which represents total elapsed time of the algorithm. For example, gradient descent with initialization at the origin in $\mathbb{R}^d$ and a constant step size η has the following basic inequality:

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

The utility of this inequality is its flexibility of strategically choosing z. Relationships akin to our basic inequalities are foundational in the convergence analysis of iterative algorithms (Nesterov, 2003; Nemirovski et al., 2009; Reddi et al., 2019) and are implicit in the analyses of several works on implicit regularization (Ji and Telgarsky, 2019; Ji et al., 2020; Wu et al., 2024, 2025). Our contribution is to highlight these inequalities and demonstrate their broad utility for both optimization and statistical analysis of implicit regularization.

Summary of Contributions. Our contributions are as follows.

- We introduce *basic inequalities* for the last iterate of iterative algorithms (Section 2 and 7), including gradient descnet and mirror descent. While related inequalities are fundamental in optimization theory, we present and highlight specific forms that provide a simple yet general framework for connecting implicit and explicit regularization. The form we present deserves greater attention, as they can be used to more universal applications.
- We demonstrate the utility of basic inequalities through several applications.
  - Training dynamics (Section 3). We bound the combined loss and penalty term of the iterates by explicit regularization estimators, and analyze the path of the iterates in spirit of Lemaire (1996).
  - Generalized linear models (Section 4 and 5). We derive high-probability prediction risk bounds for early-stopped gradient descent and exponentiated gradient descent, where the rates are matched with their explicit regularization counterparts, ridge and KL-divergence regularization, respectively.
  - Randomized model selection (Section 6). We derive excess risk bounds for random model selection problem with either using exponentiated gradient descent or KL-divergence regularization.
- We conduct simulation experiments across linear, logistic, and Poission regressions in both under-parametrized and overparamterized regimes (Section 8). The result back up our theory, showing a strong empirical similarity between implicit and explicit regularizations in their training dynamics, prediction risk curves, and solution paths.

Related Work. Connecting implicit regularization to corresponding well-understood explicit regularization has been studied in the literature for long time, includes characterizing the limit points and solution paths of algorithms. For overparametrized linear regression, gradient descent is known to converge to the min-L_2 norm solution (Lemaire, 1996). It is generalized to mirror descent, which converges to min-Bregman-divergence solution (Gunasekar et al., 2018; Azizan and Hassibi, 2019). In some cases, more direct equivalence of the entire solution path can be established, for instances, Least Angle Regression (LARS) algorithm generates the lasso path (Efron et al., 2004). Regarding an L_2-penalty, more recent work has focused on quantifying the L_2-distance between implicitly and explicitly regularized estimators for strongly convex loss functions (Suggala et al., 2018), while for linear regression, more tighter comparison is possible (Ali et al., 2019).

For classification problems, implicit regularization often becomes a max-margin solution. For instance, AdaBoost with an infinitesimal step size converges to L_1-max-margin binary classifier (Zhang and Yu, 2005). In logistic regression with linearly separable data, gradient descent converges in direction to the max-margin solution (Soudry et al., 2018; Ji and Telgarsky, 2019), with analogous results for mirror descent (Sun et al., 2023). For a broader loss class, when the risk does not achieve its infimum, the gradient descent path and a corresponding explicit regularization path converge to the same direction (Ji et al., 2020). Recent work on logistic regression has investigated the phase transitions of gradient descent with large step sizes (Wu et al., 2024) and the existence of beneficial early stopping time with respect to excess risk (Wu et al., 2025).

The analysis of explicit regularization is itself a mature field. For ridge regression, the out-of-sample and in-sample prediction risk bounds has been precisely characterized due to its closed form solution (Ali et al., 2019), with sharp asymptotic results available from Marchenko-Pastur theorem in random matrix theory

(Dobriban and Wager, 2018; Hastie et al., 2022). For L_2-regularized logistic regression, the self-concordance property of the loss has been used to prediction risk bounds (Bach, 2010). For lasso (Tibshirani, 1996), which uses an L_1-penalty, a key observation for non-asymptotic analysis is the *basic inequality* (Bühlmann and Van De Geer, 2011), derived from the estimator's zero-order optimality condition. For a squared loss $f(\theta) = (1/2n)\|Y - X\theta\|_2^2$ and the lasso estimator $\hat{\theta}_\lambda$,

$$\frac{1}{2n}\big\|X(\hat{\theta}_\lambda - \theta)\big\|_2^2 \le \frac{1}{n}\big\langle Y - X\theta, X(\hat{\theta}_\lambda - \beta)\big\rangle + \lambda\big(\|\theta\|_1 - \|\hat{\theta}_\lambda\|_1\big).$$

Regarding the in-sample risk of $\hat{\theta}_\lambda$, this inequality leads to the slow rate of $O(\sqrt{(\log d)/n})$, and a fast rate of $O((\log d)/n)$ under additional assumptions on X (van de Geer and Bühlmann, 2009; Bühlmann and Van De Geer, 2011). We present the above inequality to emphasize its structural similarity to the basic inequalities for iterative algorithms that we introduce.

The use of Kullback-Leibler (KL) divergence penalty as an explicit regularization in statistical learning is not as popular as ℓ^p regularization. Nevertheless, it serves as a powerful tool for certain tasks, especially those involving probability distributions over a collection of candidate models, and is a cornerstone of the *PAC-Bayes* framework (Alquier, 2024). Its two prominent applications are *model aggregation* (Wolpert, 1992; Breiman, 1996) and *randomized model selection* (Leung and Barron, 2006; Zhang, 2006), where we seek an optimal weights to predictors from a given base models. Model aggregation uses a convex combination of the base predictors as the final predictor, according to the learned probability vector. On the other hand, in randomized model selection, a single model is randomly drawn according to the probability vector for prediction. In both tasks, KL penalty regularizes the probability vector by penalizing its deviation from a prior distribution, which is typically chosen to be uniform.

2 Basic inequalities for iterative optimization algorithm

This section introduces a set of *basic inequalities*. While similar relationships are used in the convergence rate analysis in the optimization literature as mentioned in the introduction, their broad utility as a standalone framework has been largely overlooked. We believe these inequalities deserve greater attention as simple yet fundamental framework for implicit regularization analysis, as demonstrated in the sections that follow.

Definitions and notations. We introduce definition and notation being used in the paper. For $u, v \in \mathbb{R}^d$, their inner product is $\langle u, v\rangle := u^\top v$. For a set $S \subseteq \mathbb{R}^d$, its interior and boundary are denoted by $\mathrm{int}(S)$ and ∂S. The size of a set S is denoted as $|S|$. A d-dimensional ball of radius $r > 0$ centered at $p \in \mathbb{R}^d$ is defined as $\mathsf{B}_d(r; p) := \{\theta \in \mathbb{R}^d : \|\theta - p\|_2 \le r\}$, and $\mathsf{B}_d(r) := \mathsf{B}_d(r; 0)$. We denote $\mathbb{N}_0 = \{0\} \cup \mathbb{N} = \{0, 1, 2, \ldots\}$, and $[n] := \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$.

A function $f : \Omega \to \mathbb{R}$ is convex if $f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y)$ for any $x, y \in \Omega$ and $\alpha \in [0, 1]$. It is strictly convex if the inequality holds strictly for $x \ne y$ and $\alpha \in (0, 1)$. The subgradient of a convex function f at x is denoted by $\partial f(x)$. A function f is essentially strictly convex if it is strictly convex on all convex subsets of $\{x : \partial f(x) \ne \emptyset\}$. A function $f : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ is essentially smooth if it satisfies three conditions: (i) $\mathrm{int}(\Omega) \ne \emptyset$; (ii) f is differentiable on $\mathrm{int}(\Omega)$; and (iii) $\lim_{i\to\infty} \|\nabla f(x_i)\|_2 = \infty$ for any sequence $\{x_i\}_{i=1}^\infty \subset \Omega$ converging to a point $x \in \partial\Omega$. A function f is of *Legendre type* if it is both essentially smooth and essentially strictly convex (Rockafellar, 1997). A differentiable function f is called α-strongly convex with respect to a norm $\|\cdot\|$ with some $\alpha > 0$, if $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|x - y\|^2$. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, it is L-smooth with respect to a norm $\|\cdot\|$ if the gradient $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz with respect to $\|\cdot\|$ and its dual norm $\|\cdot\|_*$ with some $L > 0$, i.e., $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|$ for any $x, y \in \mathbb{R}^d$. When the norm is not specified, the Euclidean norm $\|\cdot\|_2$ is assumed.

A random variable Z is sub-Gaussian with parameter σ^2, if $\mathbb{E}[\exp(\alpha(Z - \mathbb{E}[Z]))] \le \exp(\alpha^2\sigma^2/2)$ for all $\alpha \in \mathbb{R}$, denoted $Z \sim \mathrm{sG}(\sigma^2)$. For a matrix $X \in \mathbb{R}^{n\times d}$, we define the empirical covariance matrix as $\widehat{\Sigma}_X := \frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$, often denoted $\widehat{\Sigma}$ when X is clear from context.

2.1 Basic inequality for early-stopped gradient descent

Gradient descent (Euler, 1792; Cauchy et al., 1847) is one of the most widely used algorithms for both convex and non-convex optimization problems. Given a differentiable loss function $f : \mathbb{R}^d \to \mathbb{R}$, the gradient descent

algorithm with an initialization $\theta_0 \in \mathbb{R}^d$ and step sizes $(\eta_t)_{t=0}^{\infty}$ generates iterates according to

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

Our first basic inequality, given below, describes a remarkably simple yet powerful statement about the objective value of the last iterate $\theta_T \in \mathbb{R}^d$ of gradient descent to any reference point $z \in \mathbb{R}^d$. This inequality will be a key for understanding how early stopping in gradient descent can act as a form or regularization, a theme we will explore in later sections.

Assumption As1 (Gradient descent setting)**.** *The function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, differentiable, and L-smooth for some $L > 0$.*

Theorem 1 (Basic inequality for gradient descent)**.** *Under Assumption As1, consider gradient descent with iterates (1) and step sizes $\eta_t \in (0, 1/L]$. Then, for any reference point $z \in \mathbb{R}^d$ and any stopping time $T \in \mathbb{N}$, it holds that*

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

In particular, for a constant step size $\eta_t = \eta$, this simplifies to

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

Proof of Theorem 1. The proof proceeds in three steps.
Step 1: Bounding the proximity difference at t and $t+1$. We measure proximity via the Euclidean distance. For any $z \in \mathbb{R}^d$,

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

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

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

The L-smoothness of f and $\eta_t \in (0, 1/L]$ guarantee, via the descent lemma (Lemma A1), that

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

This ensures $f(\theta_T) \leq f(\theta_t) - \frac{1}{2}\eta_t \|\nabla f(\theta_t)\|_2^2$ for any $t > 0$. Using this to lower bound $f(\theta_t) - f(z)$ by $f(\theta_T) - f(z)$,

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

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 concludes the proof. □

2.2 Basic inequality for early-stopped mirror descent

Mirror descent (Nemirovskij and Yudin, 1983; Beck and Teboulle, 2003) extends gradient descent to non-Euclidean geometries, using a Bregman divergence to measure proximity. This generalization is crucial for problems with specific domain and geometric constraints, for instances, the probability simplex and the distance between two probability measures. We develop a basic inequality for mirror descent that shares a structural resemblance with the one from gradient descent, as can be anticipated from their relationship.

Assumption As2 (Mirror descent setting)**.** *Let $\mathcal{K}$ and Ω be closed convex sets in $\mathbb{R}^d$ such that $\mathcal{K} \subseteq \Omega$, and their interiors are not empty. A function $f : \Omega \to \mathbb{R}$ is convex on $\mathcal{K}$, and it is differentiable on* $\operatorname{int}(\Omega)$*. A function $\phi : \Omega \to \mathbb{R}^d$ is of Legendre type, and it is continuous on Ω.*

The *Bregman divergence* induced by ϕ, for $u, v \in \Omega$, is

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

This is only well-defined if and only if $v \in \operatorname{int}(\Omega)$ since ϕ is of Legendre type (Lemma A2). Given initialization $\theta_0 \in \mathcal{K} \cap \operatorname{int}(\Omega)$ and the step sizes $(\eta_t)_{t=0}^\infty$, mirror descent generates iterates as

$$\theta_{t+1} := \arg\min_{\theta \in \mathcal{K}} \left\{ \eta_t \langle \nabla f(\theta_t), \theta\rangle + D_\phi(\theta, \theta_t) \right\}. \tag{2}$$

This update guarantees that $\theta_t \in \mathcal{K} \cap \operatorname{int}(\Omega)$ for any $t \geq 0$ (Lemma A4), keeping $D_\phi(\cdot, \theta_t)$ well-defined.

Assumption As3 (Additional mirror descent setting)**.** *The function ϕ is α-strongly convex for $\alpha > 0$ with respect to a norm $\|\cdot\|$ on $\mathcal{K}$. The function f is L-smooth with respect to $\|\cdot\|$ for $L > 0$ on $\mathcal{K} \cap \operatorname{int}(\Omega)$.*

Theorem 2 (Basic inequality for mirror descent)**.** *Under Assumptions As2 and As3, consider mirror descent with iterates (2) and step sizes $\eta_t \in (0, \alpha/L]$. Then, for any reference point $z \in \mathcal{K}$ and stopping time $T \in \mathbb{N}$, it holds that*

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

In particular, for a constant step size $\eta_t = \eta$, this simplifies to

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

Proof of Theorem 2. The proof parallels that of the gradient descent case, but leverages properties of the Bregman divergence.
Step 1: Bounding the proximity difference at t and $t+1$. We measure proximity via the Bregman divergence. The well-known "three-point identity" for Bregman divergence (see Lemma A5) states that

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

Step 2: Bounding the criterion difference $f(\theta_t) - f(z)$. Convexity of f on $\mathcal{K}$ implies that

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

Multiplying both sides by η_t and using Step 1,

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

α-strong convexity of ϕ implies that $D_\phi(\theta_{t+1}, \theta_t) \geq (\alpha/2)\|\theta_{t+1} - \theta_t\|^2$ (Lemma A3), and L-smoothness of f yields $f(\theta_{t+1}) \leq f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle + (L/2)\|\theta_{t+1} - \theta_t\|^2$. Thus, we can upper bound $\eta_t(f(\theta_{t+1}) - f(z))$ as the following:

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

where $\eta_t \leq \alpha/L$ is used in the last inequality. The descent lemma for mirror descent (Lemma A7) shows $f(\theta_t)$ is non-increasing, we have

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

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

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

which concludes the proof. □

Theorem 2 indeed implies Theorem 1, its gradient descent counterpart, as a special case. Specifically, by selecting $\phi(v) = \frac{1}{2}\|v\|_2^2$, for which the Bregman divergence $D_\phi(u, v) = \frac{1}{2}\|u - v\|_2^2$ and the strong convexity parameter $\alpha = 1$, the mirror descent iterates and basic inequality precisely reduce to those of gradient descent.

Despite this direct relationship, presenting the two theorems separately, as we have done, highlights distinct operational mechanics. The derivation for gradient descent directly leverages the natural pairing of its updates with the inner product and Euclidean norm: $\langle \nabla f(\theta_t), \theta_{t+1} - \theta_t \rangle = \eta_t \|\nabla f(\theta_t)\|_2^2$. In contrast, the analysis of mirror descent with an arbitrary norm requires more nuanced arguments to navigate the interplay between the chosen geometry and the update rule of mirror descent. Comparing gradient and mirror decent pathways provides a richer picture of the two algorithms.

The basic inequalities for gradient descent and mirror descent in Theorem 1 and 2 are cornerstones of this paper, which lead us to the connection between implicit and explicit regularization in the remaining sections.

3 Application: Training dynamics

The basic inequalities can provide insights into the algorithm's training dynamics. In this section, we will observe the evolution of the training loss and the distance between iterates and the solution set, and the limit points of the iterates. Corollary 1 and 2, whose part (b)-(e) are motivated by Corollary 2.2 in Lemaire (1996), reveal the connection to explicit regularization and clarify convergence towards specific solution under certain conditions. As before, gradient descent results are the special cases of those in mirror descent, but comparing their proofs highlights the underlying geometries of the algorithms. The proofs are provided in-text, while part (d)-(e) have more details in Appendix B.

Corollary 1 (Gradient descent)**.** *Under Assumption As1, consider the gradient descent updates in (1) with step sizes $\eta_t \in (0, 1/L]$.*

(a) (Training loss bound with explicit ridge regularization.) For any and $T \in \mathbb{N}$,

$$f(\theta_T) + \frac{1/4}{\sum_{t=0}^{T-1} \eta_t}\|\theta_0 - \theta_T\|_2^2 \;\le\; \min_{z \in \mathbb{R}^d}\left[f(z) + \frac{1}{\sum_{t=0}^{T-1} \eta_t}\|\theta_0 - z\|_2^2\right].$$

(b) (Asymptotic training loss.) Define $\inf f := \inf_{\theta \in \mathbb{R}^d} f(\theta)$, which may be negatively infinite. If $\sum_{t=0}^{\infty} \eta_t = \infty$, then

$$\lim_{t \to \infty} f(\theta_t) = \inf f.$$

(c) (Non-increasing distance to solution set.) Define the solution set $S := \{\theta^* \in \mathbb{R}^d : f(\theta^*) = \inf_{\theta \in \mathbb{R}^d} f(\theta)\}$, which is closed and convex, but possibly empty. Let $\mathrm{Dist}_S(u) := \min_{s \in S} \|u - s\|_2$ denotes the distance from u to S. If $S \neq \emptyset$, then

$$\forall s \in S,\ \{\|\theta_t - s\|_2\}_{t=1}^{\infty} \text{ is non-increasing}, \quad \text{and thus}, \quad \{\mathrm{Dist}_S(\theta_t)\}_{t=1}^{\infty} \text{ is non-increasing}.$$

(d) (Limit of updates.) If $S \neq \emptyset$ and $\sum_{t=0}^{\infty} \eta_t = \infty$, then

$$\lim_{t \to \infty} \theta_t = \theta_\infty \in S.$$

Moreover, defining $\mathrm{Proj}_S(u) := \arg\min_{s \in S} \|u - s\|_2$ as the projection of u onto S, we have

$$\|\theta_\infty - \mathrm{Proj}_S(\theta_0)\|_2 \le \mathrm{Dist}_S(\theta_0) \quad \text{and thus} \quad \|\theta_\infty - \theta_0\|_2 \le 2\mathrm{Dist}_S(\theta_0).$$

(e) (Minimum-norm solution.) If S is a non-empty affine subspace and $\sum_{t=0}^{\infty} \eta_t = \infty$, then

$$\theta_\infty = \mathrm{Proj}_S(\theta_0).$$

Proof of Corollary 1. (a). By Young's inequality, $2ab \le ca^2 + b^2/c$ for any $c > 0$, we get

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

Using this to upper bound the basic inequality from Theorem 1, we obtain that

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

for any $z \in \mathbb{R}^d$, completing the proof.

(b). Theorem 1 implies that for any $z \in \mathbb{R}^d$, $f(\theta_T) \le f(z) + \|\theta_0 - z\|_2^2/(2\sum_{t=0}^{T-1}\eta_t)$. Given that $\sum_{t=0}^{\infty}\eta_t = \infty$, taking the limit superior yields $\limsup_{T\to\infty} f(\theta_T) \le f(z)$. Since this holds for any z, we get $\limsup_{T\to\infty} f(\theta_T) \le \inf f$. Combined with the trivial inequality $\inf f \le f(\theta_T)$, we conclude that $\lim_{T\to\infty} f(\theta_T) = \inf f$.

(c). We can write $S = f^{-1}(\{\inf f\})$. Assume S is non-empty, which implies $\inf f$ is finite. Since f is convex and continuous, S is a closed convex set. Consequently, the distance function $\mathrm{Dist}_S(u)$ is well-defined. By Theorem 1, we know that

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

for any $s \in S$ and $T \in \mathbb{N}$. Since $s \in S$, we have $f(s) = \inf f \le f(\theta_T)$, making the summation term non-positive. Thus, $\|\theta_T - s\|_2^2 \le \|\theta_0 - s\|_2^2$. This argument applies more generally: starting the gradient descent process at iterate θ_τ and running for $\omega - \tau$ steps (where $\omega \ge \tau$) yields $\|\theta_\omega - s\|_2^2 \le \|\theta_\tau - s\|_2^2$. As this holds for any $s \in S$, taking the infimum over s on both sides gives $\mathrm{Dist}_S(\theta_\omega) \le \mathrm{Dist}_S(\theta_\tau)$. The sequence $\{\mathrm{Dist}_S(\theta_t)\}_{t=0}^{\infty}$ is therefore non-increasing.

(d). (Proof sketch.) Choose any $s \in S$. Due to the decreasing nature of $\|\theta_t - s\|_2^2$ from part (c), there is a limit point θ_∞ as of a subsequence $\{\theta_{t_i}\}_{i=0}^{\infty}$. Then $\theta_\infty \in S$ due to part (b). Thus, $\|\theta_t - \theta_\infty\|_2^2$ is also decreasing, which concludes $\theta_t \to \theta_\infty$ as $t \to \infty$.

(e). (Proof sketch.) Let $P := \mathrm{Proj}_S(\theta_0)$ and $v = P - \theta_\infty$. For any $c \ge 0$, define $\beta_c := P + c\cdot\mathrm{Dist}_S(\theta_0)\cdot(v/\|v\|_2) \in S$. Since $\beta_c \in S$, due to part (c) and (d), we must have $\|\theta_\infty - \beta_c\|_2 \le \|\theta_0 - \beta_c\|_2$. Based on the collinear structure of three points θ_∞, P, and β_c, observe that

$$\|v\|_2 + c\cdot\mathrm{Dist}_S(\theta_0) \le \sqrt{1+c^2}\cdot\mathrm{Dist}_S(\theta_0).$$

As $c \to \infty$, the term $\sqrt{1+c^2} - c \to 0$. Therefore, we have $\|v\|_2 \le 0$, implying $v = 0$, and thus, $P = \theta_\infty$. □

Before we present the results for mirror descent, let's look into Corollary 1 more deeply. Part (a) shows the resemblance of the structure of explicit regularization, yet their parallel is not exact. An explicitly regularized estimator, $\hat{\theta}$, is a minimizer of a composite objective $f(\theta) + \lambda\|\theta_0 - \theta\|_2^2$ for $\lambda > 0$. By its definition,

$$f(\hat{\theta}) + \lambda\|\theta_0 - \hat{\theta}\|_2^2 \;\le\; \min_{z\in\mathbb{R}^d}\left[f(z) + \lambda\|\theta_0 - z\|_2^2\right].$$

However, the bound from part (a) has a slightly different form, since the effective coefficients for the penalty term on the left- and right-hand sides have a fixed 1:4 ratio. Despite these distinctions, part (a) compellingly demonstrates an algorithm-inherent regularization effect.

Part (b)-(d) describe behaviors of $f(\theta_T)$ and $\|\theta_0 - \theta_T\|_2$ separately, while part (a) jointly treated them. Part (b) establishes the consistency of the algorithm. Part (c) addresses the stability of the iterates with respect to the solution set S. Part (d) provides a bound of the limit points of the iterates, which is not arbitrarily far from the projection of the initial point to S.

Part (e) characterizes the limit point of the iterates for an affine solution set. This finding resonates and generalizes a widely-known phenomenon in the overparametrized linear regression: gradient descent

initialized at $0 \in \mathbb{R}^d$ converges to the min-norm solution. Part (e) demonstrates that such specific convergence behavior is not unique to linear regression but a consequence of a broader principle captured by the corollary. In particular, the implication of part (e) covers the generalized linear models (GLMs). While its formal definition is deferred to Section 4, where our focus shifts from the optimization to the statistical property of the estimator, the application of part (e) to GLMs is presented in Corollary 4.

As hinted earlier several times, mirror descent algorithm exhibits analogous properties as gradient descent, through the same lens of the basic inequality. The proofs are in-text, while part (d)-(e) have more details in Appendix B. Note that Gunasekar et al. (2018) discussed a similar result as part (e), with a specific form of $f(\theta) = \sum_{i\in[n]} f(\langle x_i, \theta\rangle, y_i)$, which enables the gradient descent update always lies on the row space of x's.

Corollary 2 (Mirror descent)**.** *Under Assumption As2 and As3, consider a mirror descent update in (2) with step sizes $\eta_t \in (0, \alpha/L]$.*

(a) (Training loss bound with explicit norm-regularization.) If (i) ϕ is G-smooth with respect to $\|\cdot\|$ in $\mathcal{K}$, or (ii) $D_\phi(z, \theta_0) \le \frac{G}{2}\|\theta_0 - z\|^2$ for any $z \in \mathcal{K}$, then for any given $T \in \mathbb{N}$,

$$f(\theta_T) + \frac{\alpha/4}{\sum_{t=0}^{T-1}\eta_t}\|\theta_0 - \theta_T\|^2 \;\le\; \min_{z\in\mathcal{K}}\left[f(z) + \frac{(G+\alpha)/2}{\sum_{t=0}^{T-1}\eta_t}\|\theta_0 - z\|^2\right].$$

(b) (Asymptotic training loss.) Define $\inf f := \inf_{\theta\in\mathcal{K}} f(\theta)$, which may be negatively infinite. If $\sum_{t=0}^{\infty}\eta_t = \infty$, then

$$\lim_{t\to\infty} f(\theta_t) = \inf f.$$

(c) (Non-increasing Bregman distance to solution set.) Define the solution set $S := \{\theta^* \in \mathcal{K} : f(\theta^*) = \inf_{\theta\in\mathcal{K}} f(\theta)\} \subseteq \mathcal{K}$, which is closed and convex, but possibly empty. Then $\mathrm{BregDist}_S(u) := \min_{s\in S} D_\phi(s, u)$, which denotes the Bregman distance from u to S, is well-defined for any $u \in \mathcal{K} \cap \mathrm{int}(\Omega)$. If $S \neq \emptyset$, then

$$\forall s \in S,\ \{D_\phi(s, \theta_t)\}_{t=1}^{\infty} \text{ is non-increasing}, \quad \text{and thus}, \quad \{\mathrm{BregDist}_S(\theta_t)\}_{t=1}^{\infty} \text{ is non-increasing}.$$

(d) (Limit of updates.) Suppose $S \neq \emptyset$ and $\sum_{t=0}^{\infty}\eta_t = \infty$. Further assume either one of the following:

(i) $S \cap \mathrm{int}(\Omega) \neq \emptyset$;

(ii) for any $y \in \Omega$ and for any sequence $\{y_n\}_{n=1}^{\infty} \subset \mathrm{int}(\Omega)$ converging to y, $D_\phi(y, y_n) \to 0$.[1]

Then

$$(i)\ \lim_{t\to\infty}\theta_t = \theta_\infty \in S \cap \mathrm{int}(\Omega); \qquad (ii)\ \lim_{t\to\infty}\theta_t = \theta_\infty \in S.$$

(e) (Minimum-Bregman-divergence solution.) If S is a non-empty affine subspace, $S \subset \mathcal{K} \cap \mathrm{int}(\Omega)$, and $\sum_{t=0}^{\infty}\eta_t = \infty$, then

$$\theta_\infty = \mathrm{BregProj}_S(\theta_0).$$

Proof. (a). Note that each of two assumptions for ϕ in the theorem statement gives that $D_\phi(z, \theta_0) = \phi(z) - \phi(\theta_0) - \langle\nabla\phi(\theta_0), z - \theta_0\rangle \le \frac{G}{2}\|z - \theta_0\|^2$. The α-strong convexity of ϕ (via Lemma A3) implies $D_\phi(z, \theta_T) \ge \frac{\alpha}{2}\|z - \theta_T\|^2$. By the triangle inequality, easily observe that $\|\theta_0 - \theta_T\|^2 \le 2\|\theta_0 - z\|^2 + 2\|z - \theta_T\|^2$. Rearranging this gives us $\|z - \theta_T\|^2 \ge \frac{1}{2}\|\theta_0 - \theta_T\|^2 - \|z - \theta_0\|^2$. Substituting this into the lower bound for $D_\phi(z, \theta_T)$, we have $D_\phi(z, \theta_T) \ge \frac{\alpha}{2}(\frac{1}{2}\|\theta_0 - \theta_T\|^2 - \|z - \theta_0\|^2) = \frac{\alpha}{4}\|\theta_0 - \theta_T\|^2 - \frac{\alpha}{2}\|z - \theta_0\|^2$. Finally, combining the upper bound for $D_\phi(z, \theta_0)$ and the lower bound for $D_\phi(z, \theta_T)$:

$$D_\phi(z, \theta_0) - D_\phi(z, \theta_T) \le \left(\frac{G}{2}\|z - \theta_0\|^2\right) - \left(\frac{\alpha}{4}\|\theta_0 - \theta_T\|^2 - \frac{\alpha}{2}\|z - \theta_0\|^2\right) = \frac{G+\alpha}{2}\|z - \theta_0\|^2 - \frac{\alpha}{4}\|\theta_0 - \theta_T\|^2.$$

Then the basic inequality in Theorem 2 completes the proof.

[1] This is not generally true for any Legendre type ϕ, see e.g., Remark 3.4 and Example 7.32 in Bauschke et al. (1997).

(b). From Theorem 2, for any $z \in \mathcal{K}$, we know that $f(\theta_T) \le f(z) + D_\phi(z, \theta_0)/\sum_{t=0}^{T-1} \eta_t$. Since $\sum_{t=0}^{\infty} \eta_t = \infty$, taking the limit superior gives $\limsup_{T\to\infty} f(\theta_T) \le f(z)$. This yields $\limsup_{T\to\infty} f(\theta_T) \le \inf f$. Since $\inf f \le f(\theta_T)$ trivially holds, we conclude $\lim_{T\to\infty} f(\theta_T) = \inf f$.

(c). We can write $S = \mathcal{K} \cap f^{-1}(\{\inf f\})$. Assume S is non-empty. As both $\mathcal{K}$ and $f^{-1}(\{\inf f\})$ are closed and convex, S is also closed and convex.

First we show that $\mathrm{BregDist}_S(u)$ is well-defined for any $u \in \mathcal{K} \cap \mathrm{int}(\Omega)$. Choose any $x \in S \subseteq \mathcal{K}$ and define a set $\tilde{S} := S \cap \{y \in \mathbb{R}^d : \|y - u\| \le \sqrt{(2/\alpha) D_\phi(x, u)}\}$, which is bounded in $\|\cdot\|$ and closed. Since all norms on finite-dimensional real vector space are equivalent, $\tilde{S}$ is also bounded with respect to $\|\cdot\|_2$, hence compact. Note that $D_\phi(\cdot, u)$ is continuous on $\mathrm{int}(\Omega)$ since ϕ is continuous on $\mathrm{int}(\Omega)$. Therefore, $D_\phi(\cdot, u)$ attains its minimum on the compact set $\tilde{S}$. Moreover, as $D_\phi(s, u) \ge \frac{\alpha}{2}\|s - u\|^2$ for $s \in \mathcal{K}$ by Lemma A3, we know that $x \in \tilde{S}$ and the minimizer of $D_\phi(\cdot, u)$ over $\tilde{S}$ is the minimizer over S. Thus, $\mathrm{BregDist}_S(u)$ is well-defined.

Now we prove that $D_\phi(\cdot, \theta_t)$ is non-increasing. The basic inequality in Theorem 2 says that for any $s \in S$:

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

By the definition of S, clearly $f(s) = \inf f \le f(\theta_T)$, so the left-hand side is non-negative. This implies $D_\phi(s, \theta_T) \le D_\phi(s, \theta_0)$. Applying this argument iteratively from time τ to $\omega \ge \tau$, we find $D_\phi(s, \theta_\omega) \le D_\phi(s, \theta_\tau)$. Taking the infimum over $s \in S$ yields that the sequence $\{\mathrm{BregDist}_S(\theta_t)\}_{t=0}^{\infty}$ is non-increasing.

(d). (Proof sketch.) Fix $s \in S$. From part (c), $\{\theta_t\}_{t=0}^{\infty}$ has a convergent subsequence $\{\theta_{t_i}\}_{i=1}^{\infty}$ with limit $\theta_\infty := \lim_{i\to\infty} \theta_{t_i} \in \mathcal{K}$. By part (b), $\theta_\infty \in S$. Each of two assumptions given in the theorem statement implies $D_\phi(\theta_\infty, \theta_{t_i}) \to 0$ as $i \to \infty$. Using this we can prove a contradiction if the entire sequence $\{\theta_t\}_{t=0}^{\infty}$ does not converge to θ_∞.

(e). (Proof sketch.) Define $P := \mathrm{BregProj}_S(\theta_0) \in S$. Let $v := P - \theta_\infty \neq 0$, then $P + cv \in S$ for any $c \in \mathbb{R}$ since S is affine. Since S is affine, the generalized Pythagorean theorem for Bregman projection holds with equality: $D_\phi(P + cv, \theta_0) = D_\phi(P + cv, P) + D_\phi(P, \theta_0)$. Using two other inequalities regarding D_ϕ, we can prove that $\langle \nabla\phi(P) - \nabla\phi(\theta_\infty), cv\rangle \le D_\phi(P, \theta_0) - D_\phi(P, \theta_\infty)$ for any $c \in \mathbb{R}$. Since $D_\phi(P, \theta_\infty) \le D_\phi(P, \theta_0)$ by part (c) and (d), we conclude $\nabla\phi(P) = \nabla\phi(\theta_\infty)$, which implies $\theta_\infty = P$. □

3.1 Notable example: exponentiated gradient descent algorithm

A prominent instance of mirror descent beyond Euclidean geometry is the exponentiated gradient descent algorithm (Helmbold et al., 1995; Kivinen and Warmuth, 1997). Exponentiated gradient descent is particularly suited for optimization problems constrained to the probability simplex $\Delta_d := \{a \in \mathbb{R}^d : a_i \ge 0, \sum_{i=1}^d a_i = 1\}$, serving a general role in various areas such as portfolio selection (Helmbold et al., 1998; De Rooij et al., 2014), solving max-margin or log-linear problem (Bartlett et al., 2004; Collins et al., 2008), and aggregation of models or estimators (Juditsky et al., 2005, 2008).

Exponentiated gradient descent shows that mirror descent efficiently and naturally updates under a specific constraint set and geometry. To view this algorithm as mirror descent, one chooses $\mathcal{K} = \Omega = \Delta_d$ and the negative entropy function $\phi(a) = \sum_{i=1}^d a_i \log a_i$. Note that ϕ is of Legendre type on Δ_d, and 1-strongly convex on Δ_d with respect to $\|\cdot\|_1$, due to Pinsker's inequality. The Bregman divergence induced by ϕ has a special form and name: Kullback–Leibler (KL) divergence, $D_{\mathrm{KL}}(a, b) = \sum_{i=1}^d a_i \log(a_i/b_i)$. Then, mirror descent update (2) yields the exponentiated gradient descent update: given $\theta_t \in \mathrm{int}(\Delta_d)$,

$$\tilde{\theta}_{t+1} = \theta_t \odot \exp(-\eta_t \nabla f(\theta_t)), \quad \theta_{t+1} = \frac{1}{\|\tilde{\theta}_{t+1}\|_1} \tilde{\theta}_{t+1}. \tag{3}$$

The first half of the Corollary 3 is derived by part (a) of Corollary 2. The other half is due to specific structure of KL divergence. The proof is provided in Appendix B.

Corollary 3 (Training loss bound for the last iterate of EGD)**.** *Under Assumption As2 and As3, consider EGD with iterates in (3) with an initialization of* $\theta_0 = (1/d, \ldots, 1/d)^\top \in \mathbb{R}^d$ *and step sizes* $\eta_t \in (0, 1/L]$*. Then for any* $T \in \mathbb{N}$ *and* $z \in \Delta_d$*:*

$$f(\theta_T) + \frac{1/4}{\sum_{t=0}^{T-1} \eta_t} \|\theta_0 - \theta_T\|_1^2 \;\le\; f(z) + \frac{1}{\sum_{t=0}^{T-1} \eta_t} \min\left[\frac{d+1}{2}\|\theta_0 - z\|_1^2,\; \frac{1}{2}\|\theta_0 - z\|_1^2 + \frac{\log d}{2}\|\theta_0 - z\|_1\right).$$

3.2 Notable example: generalized linear model (GLM)

As briefly explained after Corollary 1, we can directly apply Corollary 1 and 2, particularly their part (e) regarding the limit point of iterates, to the generalized linear models (GLM). The GLM, which will be formally defined in Section 4, is a broad class of model that includes linear regression as a special case. Corollary 4 now details the application on GLM, specifically considering optimization within a general affine set $\mathcal{K} \subseteq \mathbb{R}^d$.

Corollary 4 (Limit of gradient descent and mirror descent on GLM)**.** *Consider the GLM loss function $\ell(\theta)$ specified in Definition 1, with a relaxed condition: the optimization domain is not necessarily $\mathbb{R}^d$, but an affine subset $\mathcal{K} \subseteq \mathbb{R}^d$. Write the solution set $S := \{s \in \mathcal{K} : \ell(s) = \inf_{\theta \in \mathcal{K}} \ell(\theta)\}$. Suppose that Assumption As2 and As3 holds with $\Omega = \mathcal{K}$ and $f = \ell$, with appropriate ϕ and α. Further assume $S \neq \emptyset$ and $S \subset \operatorname{int}(\mathcal{K})$. Consider the mirror descent update generated by (2) initialized at $\theta_0 \in \operatorname{int}(\mathcal{K})$ with step sizes $\eta_t \in (0, \alpha/L]$. Then, $\lim_{t\to\infty} \theta_t \to \operatorname{BregProj}_S(\theta_0)$.*

Proof. Lemma B1 proves that $S = \mathcal{K} \cap (\{s\} + \{v \in \mathbb{R}^d : Xv = 0\})$: a rough intuition for this observation is that the GLM loss function depends on θ only through $X\theta$. Since S is an intersection of two affine sets, it is also an affine set. Therefore, the part (e) of Corollary 2 (and its gradient descent counterpart in Corollary 1) directly applies, concluding the proof. □

4 Application: GLMs with gradient descent

Having observed the application of basic inequalities for analyzing training dynamics in Section 3, we now shift gears to their statistical perspective: the prediction risk of the estimators. This section will focuses on generalized linear models (GLMs) and compared two regularization methods: an explicit regularization via ridge penalties; and an implicit regularization by early-stopped gradient descent. Meanwhile, an analogous analysis for mirror descent will be presented in the subsequent Section 5.

GLMs refer to a broader model class related to the exponential family, whose formal definition can be found in Appendix C. The loss function for our analysis is defined below, as a special case of GLMs with an identity sufficient statistic. This is general enough to include linear, logistic, and Poisson regression, which are related to Gaussian, Bernoulli, and Poisson distribution, respectively.

Definition 1 (GLM loss and estimator; special case of (19))**.** *Let $(X, Y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$. The GLM loss function and estimator, with an identity sufficient statistics, are defined as*

$$\hat{\theta}_0 := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \ell(\theta) \quad \text{where} \quad \ell(\theta) := \frac{1}{n}\Big(- Y^\top X\theta + A(X\theta)\Big). \tag{4}$$

Note that $A : \mathbb{R}^n \to \mathbb{R}$ acts component-wisely as $A(v) = \sum_{i=1}^n A(v_i)$, where $A : \mathbb{R} \to \mathbb{R}$ is the cumulant function for the corresponding univariate exponential family. Note that, for Gaussian, Bernoulli, and Poisson distribution, $A(\xi) = \xi^2/2$, $A(\xi) = \log(1 + e^\xi)$, and $A(\xi) = e^\xi$, respectively.

Now we introduce the prediction risk, under the fixed-design setting. Where training data is (X, Y), we evaluate the prediction risk of an estimator $\theta = \theta(X, Y)$ on a fresh response vector W, an independent copy of Y given fixed X. The formal data generating process and the prediction risk definition are as follows.

Assumption As4 (Data Generating Process)**.** *The features $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times d}$ is fixed. The responses $Y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ consist of mutually independent samples $y_i \sim P_i$. The distribution P_i may depend on x_i. Importantly, P_i is not required to be a member of the exponential family that defines the GLM estimator (allowing for model misspecification). Let $\mu_i := \mathbb{E}[y_i]$ under P_i be the true conditional mean of y_i. Let $\mu := (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$ be the vector of true means, and $\epsilon := Y - \mu$ be the zero-mean noise vector.*

Definition 2 (Prediction risk)**.** *Under Assumption As4, let $\theta = \theta(X, Y)$ be an estimator derived from training data (X, Y). Let $W = (w_1, \ldots, w_n)^\top$ be a vector of fresh, independent test responses, where $w_i \sim P_i$ is identically distributed to y_i, and W is independent of Y. The* prediction risk *of an estimator θ is the expected GLM loss on test data, conditional on training data (X, Y) is defined as*

$$\operatorname{Risk}(\theta) := \frac{1}{n}\mathbb{E}_W\Big[- W^\top X\theta + A(X\theta) \,\Big|\, X, Y\Big] = \frac{1}{n}\Big(- \mu^\top X\theta + A(X\theta)\Big). \tag{5}$$

4.1 Risk analysis: Ridge-penalized GLM estimator

We begin our risk analysis with the widely-used explicit regularization, the ridge penalty, in order to later compare with gradient descent. The ridge-penalized GLM estimator is defined by augmenting the GLM loss (4) with an additional $\|\cdot\|_2^2$-penalty on the coefficients. For a user-chosen regularization parameter $\lambda \geq 0$, the ridge estimator $\hat{\theta}_\lambda$ is the minimizer of

$$\ell_\lambda(\theta) := \ell(\theta) + \lambda\|\theta\|_2^2, \quad \text{i.e.,} \quad \hat{\theta}_\lambda := \arg\min_{\theta\in\mathbb{R}^d} \ell_\lambda(\theta). \tag{6}$$

Note that $\ell_0(\theta) = \ell(\theta)$, making $\hat{\theta}_0$ defined in (4) be consistent to the definition in (6) with $\lambda = 0$. The cumulant function A used in a GLM is convex, which implies $\ell(\theta)$ is also convex. Consequently, $\ell_\lambda(\theta)$ is 2λ-strongly convex for $\lambda > 0$, whose standard proof using properties of A can be found in Appendix D.

Our first result, Proposition 1, provides a general bound on the prediction risk of the ridge-penalized GLM estimator $\hat{\theta}_\lambda$ against the one for an arbitrary parameter θ.

Proposition 1 (Risk bound for ridge-penalized GLM estimator)**.** *For any $\lambda > 0$ and any reference parameter $\theta \in \mathbb{R}^d$, the prediction risk of $\hat{\theta}_\lambda$ is bounded by:*

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

The above proposition decomposes the excess risk $\mathrm{Risk}(\hat{\theta}_\lambda) - \mathrm{Risk}(\theta)$ into two parts, a 'variance' term $\frac{1}{n}\|X^\top\epsilon/n\|_2^2$ and a 'regularization' term $\|\theta\|_2^2$, which can be balanced by choosing λ. To make this bound more statistically sound, we assume sub-Gaussian noise and obtain the high-probability bound.

Proposition 2 (Oracle risk bound for ridge-penalized GLM estimator)**.** *Assume each ϵ_i in Assumption As4 is sub-Gaussian with parameter σ_i^2. Let $\sigma^2 := \max_i \sigma_i^2$. Recall that $\widehat{\Sigma} = \frac{1}{n}X^\top X$. Then, for any $\delta > 0$ and $b > 0$, choosing*

$$\lambda = \frac{\sigma}{2b\sqrt{n}}\sqrt{\mathrm{tr}(\widehat{\Sigma}) + 2\left\|\widehat{\Sigma}\right\|_F\sqrt{\delta} + 2\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}\delta} \tag{7}$$

yields the following bound with probability at least $1 - e^{-\delta}$:

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \inf_{\theta:\ \|\theta\|_2\leq b}\mathrm{Risk}(\theta) \leq \frac{2b\sigma}{\sqrt{n}}\sqrt{\mathrm{tr}(\widehat{\Sigma}) + 2\left\|\widehat{\Sigma}\right\|_F\sqrt{\delta} + 2\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}\delta}. \tag{8}$$

The above bound depends on the spectrum of the empirical covariance matrix $\widehat{\Sigma}$. Regarding the spectral properties of $\widehat{\Sigma}$, consider a following common scenario which holds, for instance, for random designs of features where entries of X are independent sub-Gaussian (see Vershynin (2018) for concentration results):

$$\mathrm{tr}(\widehat{\Sigma}) = O(d), \quad \left\|\widehat{\Sigma}\right\|_F = O(\sqrt{d}), \quad \text{and} \quad \left\|\widehat{\Sigma}\right\|_{\mathrm{op}} = O(1), \quad \text{with respect to } d. \tag{9}$$

Then, letting $\delta = \log n$ in (8) yields an excess risk bound of $\tilde{O}(b\sigma\sqrt{d/n})$ with probability at least $1 - 1/n$.

We now state our main result for the ridge-penalized GLM estimator. Theorem 4 tailors an oracle inequality from Proposition 2 to specific GLMs by identifying their respective sub-Gaussian noise parameter. The proof, presented in Appendix D, adapts the one for Proposition 2 but requires particular observation for the Poisson regression, which leads to a slightly less probability guarantee than the Gaussian and Bernoulli distributions.

Theorem 3 (Specific cases of ridge-penalized GLM estimator)**.** *Under Assumption As4, consider the ridge-penalized GLM with the parameter λ, i.e., with the loss function of ℓ_λ in (6), where ℓ uses (i) Gaussian distribution (linear regression), (ii) Bernoulli distribution (logistic regression), or (iii) Poisson distribution (Poisson regression). Further assume, respectively, one of the following for P_i in Assumption As4:*

(i) Gaussian distribution where $P_i = \mathcal{N}(\mu_i, \sigma_i)$, for all $i \in [n]$;

(ii) Bernoulli distribution where $P_i = \mathrm{Bernoulli}(\mu_i)$, for all $i \in [n]$;

(iii) Poission distribution where $P_i = \mathrm{Pois}(\mu_i)$, for all $i \in [n]$, and $n \geq 3$.

Define σ_{Dist} *as the following, respectively:*

(i) For Gaussian P_i, $\sigma_{\text{Dist}} = \max_{1\le i\le n} \sigma_i$;

(ii) For Bernoulli P_i, $\sigma_{\text{Dist}} = 1/2$;

(iii) For Poisson P_i, $\sigma_{\text{Dist}} = (2\|\mu\|_\infty + 2/3)\log n + \|\mu\|_\infty/2$, *where* $\|\mu\|_\infty := \max_{1\le i\le n} \mu_i$.

Finally, for any $\delta > 0$ *and* $b > 0$, *if we choose*

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

then with the probability at least $1 - e^{-\delta}$ *for (i)-(ii), or* $1 - 1/n - e^{-\delta}$ *for (iii),*

$$\text{Risk}(\hat{\theta}_\lambda) - \inf_{\theta:\ \|\theta\|_2\le b} \text{Risk}(\theta) \le \frac{2b\sigma_{\text{Dist}}}{\sqrt{n}}\sqrt{\operatorname{tr}(\widehat{\Sigma}) + 2\|\widehat{\Sigma}\|_F\sqrt{\delta} + 2\|\widehat{\Sigma}\|_{\text{op}}\delta}.$$

Theorem 3 establishes a unified prediction risk bound applicable for widely used GLMs. The derived high-probability upper bound of $\tilde{O}(b\sigma\sqrt{d/n})$, contingent upon the spectral properties of $\widehat{\Sigma}$ in (9), provides a general benchmark for estimator performance.

A notable strength of this theorem lies in its robustness to model misspecification. For example, instead of Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$, any sub-Gaussian distributions with parameter σ_i^2 is valid for case (i). Also, instead of Bernoulli(μ_i), any bounded $y_i \in [0, 1]$ is still valid for case (ii). This aspect underscores broad applicability of the theorem, and more fundamentally, the basic inequality.

Comparison with existing literature. While this unified bounds offers a useful general perspective, its comparison with the literature reveals nuances for each GLM.

Considering ordinary linear regression (i.e., $\lambda = 0$; no penalty), we know the closed form solution is $\hat{\theta} = (X^\top X)^{-1}X^\top Y$, when it is well-defined. Starting from this closed form solution, we can derive tighter analysis: in a well-specified case of $\mathbb{E}[Y] = X\theta_0$, we can prove that $\text{Risk}(\hat{\theta}) - \text{Risk}(\theta_0) = \frac{1}{2n}\|H\epsilon\|_2^2$ where $H = X(X^\top X)^{-1}X^\top$ is the projection matrix. Moreover, when $\epsilon_i \sim \text{sG}(\sigma^2)$, we can prove that $\text{Risk}(\hat{\theta}) - \text{Risk}(\theta_0)$ has a high-probability upper bound of $O(\sigma^2 d/n)$. Compared to our bound of $\tilde{O}(b\sigma\sqrt{d/n})$, this upper bound does not depend on b when $b > \|\theta_0\|_2$ but instead σ has a squared term, and more importantly, the magnitude of d/n factor is not a square root. Also, this does not cover general $b < \|\theta_0\|_2$. Related calculations are provided in Appendix D.

Regarding the ridge regression (i.e., $\lambda > 0$), the literature has considered a couple notion of prediction risks, which also yields $O(\sqrt{d/n})$ bound. Ali et al. (2019) defined in-sample and out-of-sample prediction risk for ridge regression, similar but not identical to our definition of the risk. For brevity, here we discuss their out-of-sample risk bound only. They derived the exact upper bound instead of a high-probability bound, whose form is simplified to $\frac{\lambda^2}{(1+\lambda)^2}\|\beta_0\|_2^2 + \frac{\sigma^2 d}{n(1+\lambda)^2}$ when $\widehat{\Sigma} = \Sigma = I$. The bound achieves the minimum value of $O(\|\beta_0\|\sigma\sqrt{d/n})$ with the best choice of λ. Meanwhile, $\lambda = 0$ covers the ordinary linear regression case, with the bound $O(\sigma^2 d/n)$. Yet this observation does not apply to general $b < \|\beta_0\|_2$.

Moreover, regarding out-of-sample prediction risk for ordinary linear regression and ridge regression, their asymptotic value, in almost surely limit, has been studied leveraging the closed form solutions and Marchenko-Pastur theorem in random matrix theory (Hastie et al. (2022); Dobriban and Wager (2018); nicely summarized in Tibshirani (2023)).

In the context of L_2-regularized logistic regression, the $\tilde{O}(b\sqrt{d/n})$ rate provides a better or comparable rate to known results. Bach (2010) leverages the self-concordance property of the logistic loss, to bound prediction risks for misspecified and well-specified models. The misspecified case offers high-probability bound of $O(b^2 d/\sqrt{n})$ if Gaussian x_i's are further assumed. The well-specified model result is not universal in the sense that it needs many conditions to be satisfied related to the data. Yet with the strongest heuristic assumption that all matrices appearing are isometric, the high-probability upper bound that we can deduce is $O(bd/\sqrt{n})$ with the best choice of $\lambda = O(d/\sqrt{n})$.

In summary, while Theorem 3 offers a valuable unified perspective on prediction risk for a range of GLMs, it may not achieve the optimal rate for every specific model instance when benchmarked against highly specialized analyses, but still provides interesting and useful new results.

4.2 Risk analysis: Early-stopped gradient descent GLM estimator

Changing the point of view from the explicit regularization to implicit regularization, now we focus on GLM estimators obtained by early stopping of gradient descent with the original, non-penalized GLM loss $\ell(\theta)$ in (4). We initialize with $\theta_0 = 0$. Let $\theta_T^{(\mathrm{gd})}$ denote the iterate after T steps of either, (i) standard gradient descent on $\mathbb{R}^d$, or (ii) projected gradient descent on $\mathsf{B}_d(b)$ with some $b > 0$. Mirroring the role of λ in ridge regression, we define an effective regularization parameter for early stopping with constant step size η:

$$\lambda_T = \frac{1}{\eta T}. \tag{10}$$

This definition of λ is motivated by our basic inequality for gradient descent in Theorem 1. If variable step sizes η_t are used, then $\lambda_T = 1/\sum_{t=0}^{T-1} \eta_t$.

Before we introduce our main result, we introduce another variant of gradient descent, *projected gradient descent*, over a closed convex set $\mathcal{K}$ with an initialization $\theta_0 \in \mathbb{R}$ and step size η_t, whose iterates follow

$$\tilde{\theta}_{t+1} = \theta_t - \eta_t \nabla f(\theta_t), \quad \theta_{t+1} = \operatorname*{arg\,min}_{\theta \in \mathcal{K}} \|\tilde{\theta}_{t+1} - \theta\|_2^2. \tag{11}$$

It is known that the projected gradient descent is a special case of mirror descent iterates in (2) with $\phi = \frac{1}{2}\|\cdot\|_2^2$, which is also a special case of an equivalent two-step update form of mirror descent.[2] Thus Theorem 2 holds for the projected gradient descent as well: $f(\theta_T) - f(z) \le \frac{1}{2\eta T}(\|z - \theta_0\|_2^2 - \|z - \theta_T\|_2^2)$, which has the same form as Theorem 1 but only for $z \in \mathcal{K}$.

Therefore, we will consider both gradient and projected gradient descent together in this section. The risk bound for early-stopped (projected) gradient descent in Proposition 3 closely resembles that for ridge regression, a testament to the connection of these two regularization methods.

Proposition 3 (Risk bound for early-stopped (projected) gradient descent GLM estimator)**.** *Assume the GLM loss $\ell(\theta)$ is L-smooth in either (i) $\mathbb{R}^d$ or (ii) $\mathsf{B}_d(b)$ for some $b > 0$. For each assumption, respectively, consider $\theta_T^{(\mathrm{gd})}$ obtained by T iterations of*

$$\textit{(i) gradient descent as (1)} \quad \textit{or} \quad \textit{(ii) projected gradient descent over } \mathsf{B}_d(b) \textit{ for } b > 0 \textit{ as (11)}, \tag{12}$$

with initalization $\theta_0 = 0$ and constant step size $\eta \in (0, 1/L]$. Then, for any stopping time $T \in \mathbb{N}$ and any reference point (i) $\theta \in \mathbb{R}^d$ or (ii) $\theta \in \mathsf{B}_d(b)$, respectively:

$$\mathrm{Risk}\big(\theta_T^{(\mathrm{gd})}\big) \le \mathrm{Risk}(\theta) + \frac{1}{2\lambda_T}\Big\|\frac{X^\top \epsilon}{n}\Big\|_2^2 + \frac{\lambda_T}{2}\|\theta\|_2^2. \tag{13}$$

Note that the similar results holds for arbitrary stepsizes $\eta_t \in (0, 1/L]$, as the proof in Appendix E explains. The resemblance of the above proposition with Proposition 1 is remarkable, which is originated from the connection between gradient descent and ridge penalty in the basic inequality in Theorem 1. As Proposition 1 led to 2 in the ridge-GLM case, we have an analogous oracle inequality for early-stopped gradient descent estimator for GLM.

Proposition 4 (Oracle risk bound with sub-Gaussian noise for early-stopped GLM estimator)**.** *Assume that ϵ_i in Assumption As4 is sub-Gaussian with parameter σ_i^2. Write $\sigma := \max(\sigma_1, \ldots, \sigma_n)$. Further assume that the loss function ℓ is either L-smooth in either (i) $\mathbb{R}^d$ or (ii) $\mathsf{B}_d(b)$ for some $b > 0$. Consider $\theta_T^{(\mathrm{gd})}$ obtained by T iterations of (12), with initalization $\theta_0 = 0$ and constant step size $\eta \in (0, 1/L]$. For any $\delta > 0$, define*

$$\lambda_{\mathrm{gd}}^* = \frac{\sigma}{b\sqrt{n}}\sqrt{\mathrm{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F\sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\mathrm{op}}\delta}.$$

Suppose the following stopped-time T is an integer:

$$T = \frac{1}{\eta\lambda_{\mathrm{gd}}^*}.$$

[2] Step 1: $\theta'_{t+1} = (\nabla\phi)^{-1}[\nabla\phi(\theta_t) - \eta_t \nabla f(\theta_t)]$. Step 2: $\theta_{t+1} = \arg\min_{\theta \in \mathcal{K}} D_\phi(x, \theta'_{t+1})$.

Then, the following holds with probability at least $1 - e^{-\delta}$*:*

$$\mathrm{Risk}(\theta_T^{(\mathrm{gd})}) - \inf_{\theta:\ \|\theta\|_2 \le b} \mathrm{Risk}(\theta) \le \frac{b\sigma}{\sqrt{n}} \sqrt{\mathrm{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F \sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\mathrm{op}} \delta}.$$

In general, for

$$T = \left\lceil \frac{1}{\eta \lambda^*_{\mathrm{gd}}} \right\rceil, \quad i.e. \quad T = \arg\min_{t \in \mathbb{N}} \{\lambda_t \le \lambda^*_{\mathrm{gd}}\},$$

the same bound holds as above with an additional discretization error term on the right-hand side of $(\sigma^2 \eta/(2n)) \cdot (\mathrm{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F \sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\mathrm{op}} \delta)$*.*

Using the above proposition, we have the main theorem for the early-stopped GLM estimator, similar to Theorem 3 for the ridge-penalized estimator. The proof can be found in Appendix E.3.

Theorem 4 (Specific cases of early-stopped GLM estimator)**.** *Under Assumption As4, consider the GLM with the loss function* ℓ *in* (4)*, using (i) Gaussian, (ii) Bernoulli, or (iii) Poisson distributions, as in Theorem 3. For each (i)-(iii), respectively, further assume that the distribution* P_i *and* σ_{Dist} *are those in Theorem 3. Then, for each distribution,* ℓ *is* L_{Dist}*-smooth in certain domains:*

(i) For Gaussian P_i*'s,* ℓ *is* $\|\widehat{\Sigma}\|_{\mathrm{op}}$*-smooth in* $\mathbb{R}^d$*;*

(ii) For Bernoulli P_i*'s,* ℓ *is* $\frac{1}{4}\|\widehat{\Sigma}\|_{\mathrm{op}}$*-smooth in* $\mathbb{R}^d$*;*

(iii) For Poisson P_i*'s,* ℓ *is* $\|\widehat{\Sigma}\|_{\mathrm{op}} \exp(b \cdot \max_{1 \le i \le n} \|x_i\|_2)$*-smooth in* $\mathsf{B}_d(b)$ *for any* $b > 0$*.*

Moreover, consider following optimization algorithm with step size $\eta \in (0, 1/L_{\mathrm{Dist}}]$*:*

(i)-(ii) Gradient descent with iterates of (1)*;*

(iii) Projected gradient descent on $\mathsf{B}_d(b)$ *with iterates of* (11)*.*

Define λ^*_{gd} *for (i)-(iii) respectively as*

$$\lambda^*_{\mathrm{gd}} = \frac{\sigma_{\mathrm{Dist}}}{b\sqrt{n}} \sqrt{\mathrm{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F \sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\mathrm{op}} \delta}.$$

Suppose the following stopped-time T *is an integer:*

$$T = \frac{1}{\eta \lambda^*_{\mathrm{gd}}}.$$

Then, for any $\delta > 0$ *and* $b > 0$*, the following holds with the probability at least* $1 - e^{-\delta}$ *for (i)-(ii), or* $1 - 1/n - e^{-\delta}$ *for (iii),*

$$\mathrm{Risk}(\theta_T^{(\mathrm{gd})}) - \inf_{\theta:\ \|\theta\|_2 \le b} \mathrm{Risk}(\theta) \le \frac{b\sigma_{\mathrm{Dist}}}{\sqrt{n}} \sqrt{\mathrm{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F \sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\mathrm{op}} \delta}.$$

In general, for $T = \lceil \frac{1}{\eta \lambda^*_{\mathrm{gd}}} \rceil$*, i.e.,* $T = \arg\min_{t \in \mathbb{N}} \{\lambda_t \le \lambda^*_{\mathrm{gd}}\}$*, the same bound holds as above with an additional discretization error term on the right-hand side, same as Proposition 4.*

Theorem 4 shows that early-stopped (projected) gradient descent achieves essentially the same oracle risk bound as optimally tuned ridge regression, stated in Theorem 3, up to the small discretization error.

Comparison with existing literature. A tighter analysis has been done in linear regression. Ali et al. (2019) compares ridge regression and gradient flow, a continuous-time version of gradient flow, leveraging their closed-form solutions. They establish that the out-of-sample prediction risk bound of gradient flow at time t is within a factor of 1.7 of the risk from the ridge solution with $\lambda = 1/t$. Furthermore, the ratio of their minimum Bayes risks is tightly bounded.

In the context of overparametrized logistic regression, where data is more likely to be linearly separable, Wu et al. (2025) derive high-probability upper bounds on the excess risk for early-stopped gradient descent. They prove the existence of a stopping time that achieves rates of $O(d/n)$ in the well-specified case and $O(\sqrt{d/n})$ in the misspecified case. Notably, their definition of excess risk does not reflect the size of interested domain, b. They also suggests a connection between the gradient descent and ridge regularization, with respect to the angle between their estimators.

While our results only cover small step sizes bounded by $1/L$, Wu et al. (2024) study gradient descent with large step size for logistic regression on linearly separable data. They identify three phases of training with gradient descent which consequently leads to a monotonic decrease of of the loss.

5 Application: GLMs with exponentiated gradient descent

In this section, we explore early-stopped mirror descent with basic inequalities developed in Section 2.2, and its corresponding explicit regularization, named Bregman-divergence-penalized regularization. As a key application of our general theory, we focus on the exponentiated gradient descent for GLMs and Kullback–Leibler (KL) divergence penalty. These results are connected to stacking or model aggregation, which are discussed in detail later this section.

5.1 Risk analysis: KL-penalized GLM

Let's recall related definitions and notations from previous sections: the training data (X, Y) is generated as Assumption As4 and the prediction risk of an estimator follows Definition 2; the GLM loss function is of the form $\ell(\theta) = \frac{1}{n}(-Y^\top X\theta + A(X\theta))$. Then, we define *Bregman-divergence-penalized GLM* loss function and estimator as the following:

$$\hat{\theta}_{\lambda,\phi,z} := \operatorname*{arg\,min}_{\theta\in\mathcal{K}} \ell_{\lambda,\phi,z}(\theta) \quad \text{where} \quad \ell_{\lambda,\phi,z}(\theta) := \ell(\theta) + \lambda D_\phi(\theta, z),$$

where $\lambda \geq 0$ is the regularization parameter, a set $\mathcal{K} \subseteq \mathbb{R}^d$ is closed and convex, a function $\phi : \mathcal{K} \to \mathbb{R}$ is convex, $z \in \mathcal{K}$, and $D_\phi(\theta, z)$ is a Bregman divergence defined in Section 2.2.

While general theoretical results and discussion for Bregman-divergence-penalized GLMs can be found in Appendix F, our focus is an instance of this, named *KL-penalized GLM* estimator. Our interest is to find an estimator lying in the d-dimensional simplex,

$$\Delta_d := \Big\{\theta \in \mathbb{R}^d : \theta_i \geq 0, \sum\nolimits_{i=1}^d \theta_i = 1\Big\},$$

and thus we choose $\mathcal{K} = \Delta_d$. One popular example of such setting is *stacking* or *model aggregation*, where base predictors $\{h_i\}_{i=1}^d$ are given, and we construct an aggregated predictor $h_\theta = \sum_{i=1}^d \theta_i h_i$ under certain risk criterion (Wolpert, 1992; Breiman, 1996). The vector of weights $\theta = (\theta_1, \cdots, \theta_d)^\top$ is the parameter to be learned from the data, and is typically constrained in Δ_d as it represents a convex combination of the base predictors. The base predictors $\{h_i\}_{i=1}^d$ are used as benchmarks for evaluating the aggregation method, and will not be updated during the learning procedure.

In our setting, the task is to learn $\theta \in \Delta_d$ with a small prediction risk, subject to certain KL-divergence budget constraint. Thus, we choose ϕ to be the negative entropy function $\phi(\theta) = \sum_{i=1}^d \theta_i \log \theta_i$ for $\theta \in \Delta_d$, so that D_ϕ becomes the KL divergence $D_{\mathrm{KL}}(a, b) = \sum_{i=1}^d a_i \log(a_i/b_i)$, as mentioned in Section 3.1. Also, we measure the KL divergence with respect to the uniform distribution $\pi = (1/d, \ldots, 1/d)^\top \in \Delta_d$. In conclusion, we consider following KL-penalized GLM loss function and estimator:

$$\hat{\theta}_\lambda := \operatorname*{arg\,min}_{\theta\in\Delta_d} \ell_\lambda(\theta) \quad \text{where} \quad \ell_\lambda(\theta) := \ell(\theta) + \lambda D_{\mathrm{KL}}(\theta, \pi). \tag{14}$$

Then, Proposition 5, which is analogous to Proposition 1 in ridge GLM, holds as a special case of general result for Bregman-divergence-penalized GLM (see Appendix F). Consequently, we can establish oracle risk bound on the prediction risk of $\hat{\theta}_\lambda$, under a sub-Gaussian noise assumption and fixed design similar to Proposition 2. When we relax the assumption to $\max_{1\leq j\leq d} \|X_{\cdot j}\|_2 \leq C_d\sqrt{n}$ where C_d is a constant only

depends on d, the risk bound from the proposition is also just C_d folded and become $\tilde{O}(\sigma C_d\sqrt{(b\log d)/n})$. Then, finally, we can establish oracle risk bounds for specific GLMs based on Proposition 6, whose result and proof is parallel to Theorem 3. The proofs can be found in Appenxis F.

Proposition 5 (Risk bound for KL-penalized GLM estimator)**.** *For the KL-penalized GLM estimator $\hat{\theta}_\lambda$, its prediction risk is bounded by:*

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

Proposition 6 (Oracle risk bound with sub-Gaussian noise for KL-penalized GLM estimator)**.** *Assume that each ϵ_i in Assumption As4 is sub-Gaussian with parameter σ_i^2. Further assume that $\max_{1\le j\le d}\|X_{\cdot j}\|_2 \le \sqrt{n}$ where $X_{\cdot j}$ denotes the j-th column of X. Write $\sigma := \max(\sigma_1, \ldots, \sigma_n)$. Then, for any $\delta > 0$ and $b > 0$, by choosing*

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

the following holds with probability at least $1 - e^{-\delta}$:

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \inf_{\theta:\, D_{\mathrm{KL}}(\theta,\pi)\le b} \mathrm{Risk}(\theta) \le 4\sigma\sqrt{\frac{b(\log(2d) + \delta)}{n}}.$$

Theorem 5 (Specific cases of KL-penalized GLM estimator)**.** *Under Assumption As4 and the setting of Proposition 6, consider the KL-penalized GLM with the parameter λ, i.e., with the loss function of ℓ_λ in (14), where ℓ uses (i) Gaussian, (ii) Bernoulli, or (iii) Poisson distributions. Further assume, for each (i)-(iii), the distribution P_i and sub-Gaussian parameter σ_{Dist} are as given in Theorem 3. Then, for any $\delta > 0$ and $b > 0$, if we choose*

$$\lambda = \sigma_{\mathrm{Dist}}\sqrt{\frac{\log(2d) + \delta}{nb}},$$

then with probability at least $1 - e^{-\delta}$ for (i)-(ii), or $1 - 1/n - e^{-\delta}$ for (iii),

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \inf_{\theta:\, D_{\mathrm{KL}}(\theta,\pi)\le b} \mathrm{Risk}(\theta) \le 4\sigma_{\mathrm{Dist}}\sqrt{\frac{b(\log(2d) + \delta)}{n}}.$$

5.2 Risk analysis: Early-stopped exponentiated gradient descent GLM estimator

Now we consider the explicit regularization counterpart of KL-penalized GLM, that is, early stopping exponentiated gradient descent on the unpenalized GLM loss function $\ell(\theta)$. As before, broader observation for early-stopped mirror descent corresponding to Bregman-divergence-penalized GLM is possible, and can be found in Appendix F.

Recall exponentiated gradient descent iterates in (3) from Section 3.1. Just for clarity, we denote the T-th iterate as $\theta_T^{(\mathrm{egd})}$. Let the initialization is the uniform distribution, i.e., $\theta_0 = \pi \in \Delta_d$, and define

$$\lambda_T = \frac{1}{\eta T}$$

as the effective regularization parameter. Then, we can obtain general and oracle risk bounds for $\theta_T^{(\mathrm{egd})}$, similar to those of KL-penalized GLM, in Proposition 7 and 8. Remark that, in the bounds, λ_T essentially plays the role of λ in KL-penalized GLM. In conclusion, we obtain the risk bounds of $\theta_T^{(\mathrm{egd})}$ for specific GLMs in Theorem 6 as in Theorem 5 of KL-penalized GLMs. The proofs are deferred to Appendix F.

Proposition 7 (Risk bound for early-stopped exponentiated gradient descent GLM estimator)**.** *Under Assumptions As2 and As3, consider exponentiated gradient descent iterates $\theta_T^{(\mathrm{egd})}$ with a constant step size satisfying $\eta \in (0, 1/L]$ and the initialization $\theta_0 = \pi = (1/d, \ldots, 1/d)^\top \in \mathcal{K}$. Then, for any $T \in \mathbb{N}$ and $\theta \in \mathcal{K}$,*

$$\mathrm{Risk}(\theta_T^{(\mathrm{egd})}) - \mathrm{Risk}(\theta) \le \frac{1}{2\lambda_T}\left\|\frac{X^\top \epsilon}{n}\right\|_\infty^2 + \lambda_T D_{\mathrm{KL}}(\theta, \pi).$$

Proposition 8 (Oracle risk bound with sub-Gaussian noise for early-stopped exponentitated gradient descent GLM estimator)**.** *Under the same setting as in Proposition 7, suppose that each ϵ_i in Assumption As4 is sub-Gaussian with parameter σ_i^2. Further assume that $\max_{1\le j\le d}\|X_{\cdot j}\|_2 \le \sqrt{n}$ where $X_{\cdot j}$ denotes the j-th column of X. Write $\sigma := \max(\sigma_1,\dots,\sigma_n)$. For any $\delta > 0$ and $b > 0$, define*

$$\lambda^*_{\mathrm{egd}} = \sigma\sqrt{\frac{\log(2d)+\delta}{nb}}.$$

Suppose the following stopped-time T is an integer:

$$T = \frac{1}{\lambda^*_{\mathrm{egd}}\eta}.$$

Then, the following holds with probability at least $1-e^{-\delta}$:

$$\mathrm{Risk}(\theta_T^{(\mathrm{egd})}) - \inf_{\theta:\, D_{\mathrm{KL}}(\theta,\pi)\le b}\mathrm{Risk}(\theta) \le 2\sigma\sqrt{\frac{b(\log(2d)+\delta)}{n}}.$$

In general, for $T = \lceil 1/(\lambda^*_{\mathrm{egd}}\eta)\rceil$, i.e., $T = \arg\min_{t\in\mathbb{N}}\{\lambda_t \le \lambda^*_{\mathrm{egd}}\}$, the same bound holds as above, with an additional discretization error term of $\eta^2\sigma^3\cdot(\log(2d)+\delta)^{3/2}/(n^{3/2}b^{1/2})$.

Theorem 6 (Specific cases of early-stopped exponentiated gradient descent GLM estimator)**.** *Under Assumption As4 and the setting of Proposition 8, consider the GLM with the loss function ℓ in (4), using the distribution (i) to (iii) from Theorem 3. For each (i)-(iii), assume that the distribution P_i and sub-Gaussian parameter σ_{Dist} are as given in Theorem 3, and the loss ℓ is L_{Dist}-smooth on Δ_d with respect to $\|\cdot\|_1$:*

(i) For Gaussian distribution: $L_{\mathrm{Dist}} = \|\widehat{\Sigma}\|_{1\to\infty} = \frac{1}{n}\max_{j\in[d]}\|X_{\cdot j}\|_2^2 \le 1$;

(ii) For Bernoulli distribution: $L_{\mathrm{Dist}} = \frac{1}{4}\|\widehat{\Sigma}\|_{1\to\infty} = \frac{1}{4n}\max_{j\in[d]}\|X_{\cdot j}\|_2^2 \le \frac{1}{4}$;

(iii) For Poission distribution: $L_{\mathrm{Dist}} = \frac{1}{n}\max_{j\in[d]}\sum_{i=1}^n \exp(\|x_i\|_\infty)x_{ij}^2$.

Respectively for (i) to (iii), consider exponentiated gradient descent updates with a constant stepsize $\eta = 1/L_{\mathrm{Dist}}$, and define λ^*_{egd} as

$$\lambda^*_{\mathrm{egd}} = \sigma_{\mathrm{Dist}}\sqrt{\frac{\log(2d)+\delta}{nb}}.$$

Suppose the following stopped-time T is an integer:

$$T = \frac{1}{\eta\lambda^*_{\mathrm{egd}}} = \frac{L_{\mathrm{Dist}}}{\lambda^*_{\mathrm{egd}}}.$$

Then, the following holds with probability $1-e^{-\delta}$ for (i) and (iii), or $1-1/n-e^{-\delta}$ for (iii):

$$\mathrm{Risk}(\theta_T^{(\mathrm{egd})}) - \inf_{\theta:\, D_{\mathrm{KL}}(\theta,\pi)\le b}\mathrm{Risk}(\theta) \le 2\sigma_{\mathrm{Dist}}\sqrt{\frac{b(\log(2d)+\delta)}{n}}.$$

In general, for $T = \lceil 1/(\eta\lambda^*_{\mathrm{egd}})\rceil$, the same bound holds as above with an additional discretization error of $(\sigma^3_{\mathrm{Dist}}/L^2_{\mathrm{Dist}})\cdot(\log(2d)+\delta)^{3/2}/(n^{3/2}b^{1/2})$.

Comparison with existing literature. Extensive research has focused on investigating the theoretical properties of the optimal aggregated predictor and designing efficient algorithms for its computation. In this context, let $\hat{\theta}$ denote the aggregation weights output by a statistical procedure (e.g., empirical risk minimization) or an efficient algorithm (e.g., mirror descent). The central object of interest in such analysis is the excess risk,

$$R(\hat{\theta}) - \inf_{1\le i\le d} R(h_i),$$

where R denotes the risk function. This excess risk measures how well $\hat{\theta}$ performs compared to the best base predictor.

The existing literature has established a 'fast rate' of convergence for the excess prediction risk of some specific aggregated predictors in model aggregation. Seminal works, including Dalalyan and Salmon (2012) and Lecué and Mendelson (2013), demonstrate that for i.i.d. data $\{(x_i, y_i)\}_{i=1}^n$ and certain bounded or quadratic loss functions, the celebrated *exponentially weighted aggregate* algorithm achieves a rate of $O(\log d/n)$. This result was subsequently extended by Juditsky et al. (2008) to a broader class of loss functions satisfying a key structural assumption known as *exponential concavity*, using mirror averaging algorithms based on online mirror descent. Later, Lecué and Rigollet (2014) propose the Q-aggregation procedure, whose loss function is a mixture of those used in model aggregation and randomized model selection (see Section 6), penalized by a weighted ℓ^1-norm. The authors establish that Q-aggregation achieves the fast rate of $O(\log d/n)$ if the loss function is strongly convex and Lipschitz on a finite interval. As we can see, a critical aspect of these fast rate results is that they rely on the exponential concavity (or strong convexity) of the loss function, which is a stronger condition than standard convexity. When the loss function is only assumed to be convex, a standard parametric rate $O(1/\sqrt{n})$ is expected, cf. Theorem 2 of Lecué (2007).

In contrast, our analysis requires neither exponential concavity assumption nor i.i.d. data points. Since $\sup_{\theta\in\Delta_d} D_{\mathrm{KL}}(\theta, \pi) \le \log d$ (see Appendix B.3), our risk bounds imply

$$\mathrm{Risk}(\hat{\theta}) - \inf_{\theta\in\Delta_d} \mathrm{Risk}(\theta) = O\left(\frac{\log d}{\sqrt{n}}\right).$$

An important remark is that the above infimum is over all possible convex combinations of the base learners, instead of only the base learners themselves. As we can see, our rate is slower than the fast rate by a factor of $\sqrt{n}$. We conjecture that it may be possible to recover the fast rate for GLMs by fully exploiting the strong convexity of the loss function $\ell(\cdot)$ over Δ_d. However, this refinement is beyond the scope of the present paper and we leave it future work.

6 Application: Risk of randomized predictors

This section focuses on *random model selection*, another type of approach for constructing a meta-learner from a collection of base learners. This method is distinguished from model aggregation in Section 5, since model aggregation outputs a composite prediction as a convex combination of outputs from the base learners. Formally, given a set of candidate models $\mathcal{B}$, we will randomly select one model $\beta \in \mathcal{B}$ according to a probability distribution θ over $\mathcal{B}$. If $\mathcal{B}$ is finite, then θ becomes a probability vector in the simplex $\Delta_{|\mathcal{B}|}$. Our goal is to find a distribution $\hat{\theta}$ so that a randomly selected model $\hat{\beta} \in \mathcal{B}$ according to $\hat{\theta}$ behaves nicely.

A model is an any form of function $\beta : \mathcal{X} \to \mathcal{Y}$. Given a loss function $r : \mathcal{Y}^2 \to \mathbb{R}$ and n observed data $\{(x_i, y_i)\}_{i=1}^n$, we consider population risk R and empirical risk $\widehat{R}_n$ of a model β as following:

$$R(\beta) = \mathbb{E}_{(X,Y)}\big[r(\beta(X), Y)\big] \text{ or } \mathbb{E}_{Y|X}\big[r(\beta(X), Y)\big]; \quad \text{and} \quad \widehat{R}_n(\beta) = \frac{1}{n}\sum\nolimits_{i=1}^n r(\beta(x_i), y_i).$$

For instance, but not limited to, for the GLMs considered in earlier sections, one may take $r(y', y) = -yy' + A(y')$, and thus, using β to represent the parameter of the GLM model,

$$\widehat{R}_n(\beta) = \ell(\beta) = \frac{1}{n}\sum\nolimits_{i=1}^n \Big(- y_i x_i^\top \beta + A(x_i^\top \beta)\Big).$$

One approach to construct a random model selector θ is exponential weighting based on the empirical risk evaluated on each model. Namely, we can use

$$\hat{\theta}_\lambda(\mathrm{d}\beta) \propto \exp(-\widehat{R}_n(\beta)/\lambda) \cdot \pi(\mathrm{d}\beta), \tag{15}$$

where π is a base measure and λ is a tuning parameter. For example, when $\mathcal{B}$ is finite, one possible base measure is the uniform base measure $\pi(\beta) = 1/|\mathcal{B}|$. The above $\hat{\theta}_\lambda$ is called "Gibbs posterior" in Bayesian statistics literature, and can be equivalently defined via the following KL-divergence-penalized random model selection problem (Alquier, 2024):

$$\hat{\theta}_\lambda = \operatorname*{arg\,min}_{\theta\in\mathcal{P}(\mathcal{B})} \Big\{ \mathbb{E}_{\beta\sim\theta}\Big[\widehat{R}_n(\beta)\Big] + \lambda D_{\mathrm{KL}}(\theta, \pi)\Big\}, \tag{16}$$

where $\mathcal{P}(\mathcal{B})$ is the set of all probability measures on $\mathcal{B}$, and the expectation $\mathbb{E}_{\beta\sim\theta}[\widehat{R}_n(\beta)]$ is taken only with respect to the randomness over β, i.e., $\mathbb{E}_{\beta\sim\theta}[\widehat{R}_n(\beta)] = \int_{\mathcal{B}} \widehat{R}_n(\beta)\theta(\mathrm{d}\beta)$. The optimization problem (16) is sometimes referred to as *information risk minimization* in literature that uses information-theoretic tools to derive generalization bounds for learning algorithms (Zhang, 2006; Xu and Raginsky, 2017).

Let $\hat{\beta}_\lambda$ be a random model drawn by the distribution $\hat{\theta}_\lambda$. Using basic inequalities, we can upper bound expected population risk of random $\hat{\beta}_\lambda$ in Proposition 9. In fact, our bounds presented below apply to any pair of $(\widehat{R}_n, R)$ with minimal assumptions, and only depend on

$$\left\|\widehat{R}_n - R\right\|_{L^\infty(\mathcal{B})} := \sup_{\beta\in\mathcal{B}} \left|\widehat{R}_n(\beta) - R(\beta)\right|,$$

which is related to the rate of uniform convergence of $\widehat{R}_n$ to R over the model space $\mathcal{B}$. Note that $\mathbb{E}_{\beta\sim\theta}[R(\beta)]$ in the proposition serves a similar role of Risk(β) in Section 4 and 5. The proof is in Appendix G.

Proposition 9 (Expected population risk of model sampled via $\hat{\theta}_\lambda$)**.** *Let a probability distribution $\hat{\theta}_\lambda$ be as defined in Equation (16). Then, the following bounds holds for expected population risk of a model randomly selected via $\hat{\theta}_\lambda$: for any distribution $\theta \in \mathcal{P}(\mathcal{B})$,*

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}\left[R(\beta)\right] - \mathbb{E}_{\beta\sim\theta}\left[R(\beta)\right] \le \frac{1}{\lambda}\left\|\widehat{R}_n - R\right\|^2_{L^\infty(\Theta)} + 2\lambda D_{\mathrm{KL}}(\theta,\pi).$$

We should consider not just an explicit regularized estimator $\hat{\theta}_\lambda$, but also an implicit regularized estimator. Let $\theta_T^{(\mathrm{egd})}$ be the T-th exponentiated gradient descent iterate (3) for optimizing $f(\theta) := \mathbb{E}_{\beta\sim\theta}[\widehat{R}_n(\beta)]$, with the initialization $z \in \mathcal{P}(\mathcal{B})$ and a constant step size $\eta > 0$. We can derive a similar excess risk bound for $\theta_T^{(\mathrm{egd})}$ using basic inequality for exponentiated gradient descent in Proposition 10. Note that the proposition allows any arbitrarily large step size η. This is because the loss function $f(\theta)$ is linear in θ, and therefore L-smooth for any $L > 0$. The proof is deferred to Appendix G.

Proposition 10 (Expected population risk of model sampled via $\theta_T^{(\mathrm{egd})}$)**.** *Denote $\lambda_T = 1/\eta T$. Then, the following bound holds for expected population risk of a model randomly sampled via $\theta_T^{(\mathrm{egd})}$: for any stopped-time $T \in \mathbb{N}$ and distribution $\theta \in \mathcal{P}(\mathcal{B})$,*

$$\mathbb{E}_{\beta\sim\theta_T^{(\mathrm{egd})}}\left[R(\beta)\right] - \mathbb{E}_{\beta\sim\theta}\left[R(\beta)\right] \le \frac{1}{2\lambda_T}\left\|\widehat{R}_n - R\right\|^2_{L^\infty(\Theta)} + \lambda_T D_{\mathrm{KL}}(\theta,\pi).$$

An interesting fact is that $\theta_T^{(\mathrm{egd})}$ and $\hat{\theta}_\lambda$ are actually the same estimator in this scenario. Note that the gradient of f is constant: $\nabla f(\theta) = (\widehat{R}_n(\beta))_{\beta\in\mathcal{B}}$ for any $\theta \in \mathcal{B}$. Therefore, the $\theta_T^{(\mathrm{egd})}$ has the following form by mathematical induction:

$$\theta_T^{(\mathrm{egd})}(\mathrm{d}\beta) = \frac{\exp(-\eta T\widehat{R}_n(\beta))\cdot\pi(\mathrm{d}\beta)}{\int_{\mathcal{B}}\exp(-\eta T\widehat{R}_n(\beta))\cdot\pi(\mathrm{d}\beta)} = \frac{\exp(-\widehat{R}_n(\beta)/\lambda_T)\cdot\pi(\mathrm{d}\beta)}{\int_{\mathcal{B}}\exp(-\widehat{R}_n(\beta)/\lambda_T)\cdot\pi(\mathrm{d}\beta)}. \tag{17}$$

This is identical to $\hat{\theta}_\lambda$ in (15), and more interestingly, in Proposition 9 and 10, we were able to observe their equivalence without looking into the closed form solution of $\hat{\theta}_\lambda$ and $\theta_T^{(\mathrm{egd})}$. Such an equivalence between the implicit regularization of exponentiated gradient descent and explicit regularization via KL penalty holds for general linear loss functions f.

Discussion and comparison with Alquier (2024). The main results on the population risk of $\hat{\beta}_\lambda$ in Alquier (2024) are obtained under the assumptions that the observed data $\{Z_i\}_{i=1}^n$ are i.i.d., and a non-negative loss is bounded by an absolute constant $C > 0$. Under this setting, they define $R(\beta) = \mathbb{E}_Z[r(\beta, Z)]$ and $\widehat{R}_n(\beta) = \frac{1}{n}\sum_{i=1}^n r(\beta, Z_i)$. Then, the following holds with probability at least $1 - e^{-\delta}$: for any $\theta \in \mathcal{P}(\mathcal{B})$,

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

Unfortunately, our upper bound in Proposition 9 is not always comparable to the above bound, since it is impossible to upper bound the term $\|\widehat{R}_n - R\|_{L^\infty(\mathcal{B})}$ only under the bounded loss assumption. If we further assume that $\mathcal{B}$ is finite, then applying Hoeffding's inequality and a simple union bound yields

$$\mathbb{P}\left(\left\|\widehat{R}_n - R\right\|_{L^\infty(\mathcal{B})} \ge C\sqrt{\frac{\log(2|\mathcal{B}|) + \delta}{2n}}\right) \le e^{-\delta}.$$

Then Proposition 9 implies that, with probability at least $1 - e^{-\delta}$,

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}\left[R(\beta)\right] - \mathbb{E}_{\beta\sim\theta}\left[R(\beta)\right] \leq \frac{C^2(\log(2|\mathcal{B}|)+\delta)}{2n\lambda} + 2\lambda D_{\mathrm{KL}}(\theta,\pi).$$

This bound has an additional $\log|\mathcal{B}|$ factor compared to the bound by Alquier (2024). However, it may be possible to remove this extra factor via a more refined analysis (in which Young's inequality, used in the proof of Proposition 10, is replaced by a more sophisticated bound, such as Donsker-Vadharan, which is critical in the analysis in Alquier (2024)).

7 Other iterative algorithms and basic inequalities

The basic inequalities can be derived for other iterative algorithms as well. Here we introduce several of them which were not used in the three applications in the paper (Section 3, 4, and 5).

7.1 Proximal gradient descent

While gradient descent corresponds to the forward Euler method for solving ordinary differential equations, proximal gradient descent is its backward (implicit) Euler method analog, used to minimize composite function $f = g + h$ with convex differentiable g, and convex but potentially non-differentiable h. Given initialization θ_0 and step sizes $(\eta_t)_{t=0}^{\infty}$, proximal gradient descent iterates via

$$\theta_{t+1} = \mathrm{Prox}_{\eta_t h}(\theta_t - \eta_t \nabla g(\theta_t)), \quad \text{where} \quad \mathrm{Prox}_h(\theta) := \operatorname*{arg\,min}_{z\in\mathbb{R}^d} \frac{1}{2}\|\theta - z\|_2^2 + h(z).$$

The operator Prox_h is called the proximal operator. The proximal gradient descent iterates can be equivalently expressed as

$$\theta_{t+1} = \theta_t - \eta_t G_{\eta_t}(\theta_t), \quad \text{where} \quad G_\eta(\theta) := \frac{1}{\eta}\Big(\theta - \mathrm{Prox}_{\eta h}(\theta - \eta\nabla g(\theta))\Big). \tag{18}$$

Proximal gradient descent encompasses several well-known algorithms as special cases, and some of them will be discussed later. A key requirement of this algorithm is that the proximal operator must be computable in closed form (or efficiently approximable), as it defines the core of each iterate.

Assumption As5 (Proximal gradient descent setting)**.** *$f : \mathbb{R}^d \to \mathbb{R}$ is a convex function of the form $f = g + h$, where g is convex and differentiable, and h is convex but possibly non-differentiable. The proximal operator Prox_h is computable.*

Theorem 7 (Basic inequality for proximal gradient descent)**.** *Under Assumption As5, consider proximal gradient descent with iterates (18). Suppose one of the following holds:*

(i) g is L-smooth in a convex set $C \subseteq \mathbb{R}^d$, with step sizes $\eta_t \in (0, 1/L]$, and $\theta_t \in C$ for any $t \geq 0$;

(ii) g is zero (i.e., $f = h$), with no constraint on $\eta_t > 0$.

Then, for any reference point $z \in \mathbb{R}^d$ and any stopped-time $T \in \mathbb{N}$, it holds that

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

Proof of Theorem 1. The proof parallels the gradient descent case, but relies on standard inequalities involving the proximal operator, Prox_h.
Step 1: Bounding the proximity difference at t and $t+1$. We measure proximity via the Euclidean distance. By the definition of θ_{t+1} in (18),

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

Step 2: Bounding the criterion difference $f(\theta_t) - f(z)$. By generalized descending lemma (Lemma H2) for proximal gradient descent, for any $z \in \mathbb{R}^d$, and under the respective conditions of (i) and (ii), we have:

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

Substituting into result from Step 1, we obtain in both (i) and (ii):

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

Additionally, by applying the same lemma with $z \leftarrow \theta_t$, we deduce that $f(\theta_{t+1}) \le f(\theta_t)$. Therefore,

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

Step 3: Aggregating bounds over $t = 0, \ldots, T-1$. Summing both sides of the result from Step 2 over $t < T$ results in a telescoping cancellation of squared norm terms, yielding

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

□

Notable example: Projected gradient descent. For a closed convex set $\mathcal{K} \subseteq \mathbb{R}^d$, choosing

$$h(x) = I_C(x) := \begin{cases} 0 & : x \in C \\ \infty & : x \notin \mathcal{K} \end{cases}$$

reduces the proximal gradient descent to the projected gradient descent over $\mathcal{K}$, whose iterates follow (11). As $\theta_{t+1} \in \mathcal{K}$ by definition, if g is L-smooth in $\mathsf{B}_d(b)$, Theorem 7 holds, equivalently written as, for any $z \in \mathcal{K}$,

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

Note that this is a same conclusion as in Section 4.2, where we viewed the projected gradient descent as a special case of mirror descent.

Notable example: ISTA (Iterative soft-thresholding algorithm). Another popular use case of the proximal gradient descent is ISTA for lasso penalty. Consider $f(\theta) = \frac{1}{2n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1$ and write $g(\theta) = \frac{1}{2n}\|Y - X\theta\|_2^2$ and $h(\theta) = \lambda\|\theta\|_1$. The reason why ISTA is popular is due to its closed form solution for the proximal update:

$$\mathrm{Prox}_{\eta h}(\theta) = \operatorname*{arg\,min}_{z\in\mathbb{R}^d} \frac{1}{2\eta}\|\theta - z\|_2^2 + \lambda\|z\|_1 = S_{\eta\lambda}(z) \quad \text{where } \forall i \in [d],\ [S_\gamma(z)]_i = \begin{cases} z_i - \gamma & : z_i > \gamma \\ 0 & : z_i \in [-\gamma, \gamma] \\ z_i + \gamma & : z_i < \gamma \end{cases}.$$

$S_\gamma(\cdot)$ is called a soft-thresholding operator. Note that $g(\theta)$ is $\|\widehat{\Sigma}\|_{\mathrm{op}}$-smooth in $\mathbb{R}^d$. Therefore, with an appropriate step sizes, Theorem 7 implies

$$\frac{1}{2n}\|Y - X\theta_T\|_2^2 + \lambda\|\theta_T\|_1 + \frac{1}{2\eta T}\|\theta_T\|_2^2 \le \frac{1}{2n}\|Y - Xz\|_2^2 + \lambda\|z\|_1 + \frac{1}{2\eta T}\|z\|_2^2$$

for any $z \in \mathbb{R}^d$. This suggests an interesting connection between ISTA and the elastic net (Zou and Hastie, 2005), but it is not exactly matched with the popular use of elastic net in practice, which use 'alpha' and 'ℓ_1-ratio' to control the penalty terms.

7.2 NoLips algorithm

We next consider the NoLips algorithm proposed by Bauschke et al. (2017). While its iterates are the same as those of mirror descent in (2), the NoLips variant relaxes the strong convexity of ϕ and the smoothness of f, replacing them with another condition, as explained in Assumption As6. This algorithm can be viewed as an instance of mirror descent but operates under a slightly different set of assumptions.

Assumption As6 (NoLips setting)**.** *Let $\mathcal{K}$ and Ω be closed convex sets in $\mathbb{R}^d$ such that $\mathcal{K} \subseteq \Omega$, whose interiors are not empty. A function $f : \Omega \to \mathbb{R}$ is convex on $\mathcal{K}$, and it is differentiable on $\operatorname{int}(\Omega)$. A function $\phi : \Omega \to \mathbb{R}$ is of Legendre type, and it is continuous on Ω. Furthermore, there exists a constant $L > 0$ such that $L\phi - f$ is convex on $\mathcal{K} \cap \operatorname{int}(\Omega)$, called the 'Lipschitz-like convexity condition'.*

Theorem 8 (Basic inequality; Last iterate of NoLips algorithm)**.** *Under Assumption As6, consider NoLips iterates which has the same update as (2) with an initialization $\theta_0 \in \operatorname{int}(\Omega)$ and step sizes $\eta_t \in (0, 1/L]$. Then, for any reference point $z \in \mathcal{K}$ and stopped-time $T \in \mathbb{N}$ it holds that*

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

In particular, for a constant step size $\eta_t = \eta$, this simplifies to

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

Proof of Theorem 8. The proof proceeds similarly to Theorem 2, with careful use of Assumption As6.
Step 1: Bounding the proximity difference at t and $t+1$. We measure proximity via the Bregman divergence. Note that $\theta_t \in \mathcal{K} \cap \operatorname{int}(\Omega)$ for any t, due to Lemma A2. Following the same Step 1 of Theorem 2, we have

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

Step 2: Bounding the criterion difference $f(\theta_t) - f(z)$. Note that this step is an adaptation of Lemma 5 and Theorem 1.i in Bauschke et al. (2017). Since $L\phi - f$ is convex on $\mathcal{K} \cap \operatorname{int}(\Omega)$ and $\theta_t, \theta_{t+1} \in \mathcal{K} \cap \operatorname{int}(\Omega)$, by Lemma H3, for any t,

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

Moreover, since f is convex on $\mathcal{K}$, we know $f(\theta_t) \leq f(z) + \langle \nabla f(\theta_t), \theta_t - z \rangle$ for any $z \in \mathcal{K}$. Therefore, we have

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

Combining this with the result from Step 1, we obtain

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

As $\eta_t \leq 1/L$, this implies

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

Moreover, we know that $f(\theta_{t+1}) - f(\theta_t)$ for any t, by plugging $z \leftarrow \theta_t$ to the above inequality. Thus, we have

$$f(\theta_T) - f(z) \leq D_\phi(z, \theta_0) - D_\phi(z, \theta_{t+1}).$$

Step 3: Aggregating bounds over $t = 0, \ldots, T-1$. Summing both sides of the inequality from Step 2 over $t < T$ results in a telescoping cancellation of the Bregman divergence terms, yielding

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

This completes the proof. □

8 Experiments

This section presents empirical results that corroborate our theoretical findings on the relationship between implicit and corresponding explicit regularization. Three main results are presented: two aspects of optimization dynamics in training time; and the prediction risk of the estimators for the test time, as defined in (5). The Python code to reproduce our experiments is available at https://github.com/100shpaik/basicineq.

Optimization tasks. We consider two iterative algorithms: gradient descent (GD), initialized at $0 \in \mathbb{R}^d$; and exponentiated gradient descent (EGD) initialized at the uniform distribution $\pi = (1/d, \dots, 1/d) \in \Delta_d$. They are applied to three GLMs (logistic, linear, and Poisson regression) in both underparametrized and overparametrized ($n < d$ and $n > d$) regimes. Their explicit regularization counterparts are solved as well: ridge-regularization for GD; and KL-regularization for EGD.

Notation. We denote the estimators from GD and EGD at iteration T by θ_T, and the estimators from explicit regularizations with the parameter λ by $\hat{\theta}_\lambda$. For the exact definition of each estimator, please recall (1), (3),(6), and (14). The total elapsed time for the iteration T is defined as $\tau = \tau_T := \sum_{t=0}^{T-1} \eta_t$, which corresponds to the time in the associated continuous flow for iterative algorithms.

Elapsed time τ and regularization parameter λ. The range of τ and $1/\lambda$ covers $[10^{-4}, 10^3]$ for GD, and $[10^{-4}, 10^4]$ for EGD. Throughout, the x-axis of any figure represents the total elapsed time τ but in $\log_{10}$ scale. Further details about the optimization, including learning rate schedules $\{\eta_t\}_{t=0}^{\infty}$ and numerical solvers, can be found in Appendix I.

Data distributions. Training data $(X, Y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ are generated as follows. The entries of the design matrix X are independently sampled from $\mathcal{N}(0, 1)$, and thus the population covariance is $\Sigma = I$. Y is generated from a well-specified model with a true parameter θ_{true}. The components of θ_{true} for GD were independently sampled from $\text{Unif}[-1, 1]$; for EGD they were independently sampled from $\text{Unif}[0, 1]$ and normalized to have a unit $\|\cdot\|_1$-norm, implying $\theta_{\text{true}} \in \Delta_d$. We introduce additional parameter $\gamma > 0$ which controls the signal-to-noise ratio of Y. Then y_i for $i \in [n]$ independently as the following:

- For linear regression, $y_i = x_i^\top \theta_{\text{true}} + \gamma \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 1)$;
- For logistic regression, $y_i \sim \text{Bernoulli}(p_i)$ where $p_i = 1/(1 + \exp(-\gamma x_i^\top \theta_{\text{true}}))$;
- For Poisson regression, $y_i \sim \text{Pois}(\mu_i)$ where $\mu_i = \gamma x_i^\top \theta_{\text{true}}$.

The specific values of (n, d) and γ are summarized in Table 1. The values of γ were selected to effectively show the non-monotonic prediction risk curves, as too small or large γ typically leads to monotone curves.

GLM	**GD**		**EGD**	
	underparam. $(n, d) = (200, 20)$	overparam. $(n, d) = (100, 200)$	underparam. $(n, d) = (200, 20)$	overparam. $(n, d) = (30, 60)$
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	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 1.2$	$\gamma = 3.5$

Table 1: (n, d) and γ values used in the experiments.

8.1 Training-time envelope functions

We first examine the optimization dynamics of the estimators in relation to the theoretical bounds derived in Corollary 1(a) and 3. The results are plotted in Figure 1. For each subfigure, three rows represent different GLM tasks, and two columns represent (n, d)-regimes.

Figure 1(a) displays the results for GD. The red line shows the quantity $f(\theta_T) + \|\theta_T\|_2^2/(4\tau)$ for the GD iterates. This is bounded by the blue and green lines for the ridge-regularized estimators, whose values are $f(\hat{\theta}_\lambda) + \lambda \|\hat{\theta}_\lambda\|_2^2$ where $\lambda = 1/\tau$ or $\lambda = 1/(4\tau)$ respectively, as suggested in Corollary 1. We can also observe that the red line closely follows the green line, suggesting more resemblance of the GD and $\lambda = 1/(4\tau)$ in the

prediction risk figure as well. Meanwhile, it is well known that $f(\hat{\theta}_\lambda) + \lambda\|\hat{\theta}_\lambda\|_2^2$ is a decreasing function of $\lambda > 0$; see Lemma I1.

Figure 1(b) displays similar but more nuanced results for EGD. Again the red line is for the EGD estimator, whose value is $f(\theta_T) + \|\theta_T - \pi\|_1^2/(4\tau)$. The upper bound from Corollary 3 is the orange line with the value of $f(\hat{\theta}_\lambda) + \lambda\|\hat{\theta}_\lambda - \pi\|_1^2$ with $\lambda = (d+1)/(2\tau)$. However, this bound is much looser than what we have seen in the GD subfigure. Plotting the blue and green line, which displays the same value as the orange line but with $\lambda = 1/\tau$ and $\lambda = 1/(4\tau)$, we can empirically see tighter envelope functions.

Figure 1(c) is replacing the $\|\theta - \pi\|_1^2$ penalty term from the subfigure (b) to $D_{\mathrm{KL}}(\theta, \pi)$, which is more natural to the KL-regularized estimator. Figure 1(c) is visually very similar compared to Figure 1(b), with a couple deviation. Precisely speaking, the EGD trajectory (red) is almost unchanged compared to Figure 1(b), yet the envelopes values from the KL-regularized estimator (blue, green, orange) slightly decreased. With this more natural penalty, the red line is now more centrally located between the blue and green envelopes. This finding will be revisited in the prediction risk figure.

Meanwhile, alignments of these curves for very small or large τ are intuitive. When $\tau \to 0$, equivalently $\lambda \to \infty$, both estimators are close to the initialization 0, because there was no update for θ_T, and $\hat{\theta}_\lambda$ should be near 0 as the penalty is significantly large otherwise. On the other hand, when $\tau \to \infty$ or $\lambda \to 0$), both estimators achieves the infimum of the original loss function without any penalty.

8.2 Prediction risk

We now compare the prediction risk, defined in (5), using the same estimators plotted in the previous figure. Figure 2 displays the prediction risk curves against total elapsed time τ. The color scheme of red, blue, and green is consistent with Figure 1.

Figure 2(a) shows the GD results. The red line for the GD estimator tracks the green line for the EGD estimator with $\lambda = 1/(4\tau)$ than the blue line. This is consistent with an observation from Figure 1(a), a closer resemblance of the red and green line in the training envelope functions.

Figure 2(b) for EGD tells a different story, which aligns with the observation in Figure 1(c). The EGD curve (red) initially follows the $\lambda = 1/\tau$ curve (blue) for small τ, and their minimum risks are also achieved in similar locations. However, in larger τ, the red line goes in between of the blue and green.

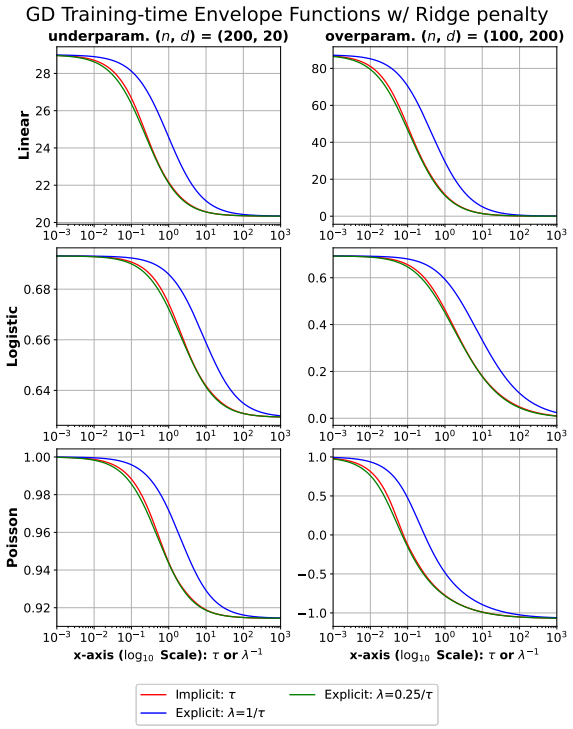
Regarding the minimum prediction risk, while the implicit and explicit regularizations indeed obtain similar minimum values, they can both be better or worse than the other, depending on the GLM tasks, (n, d)-regimes, and randomness on the training data.

8.3 Solution path

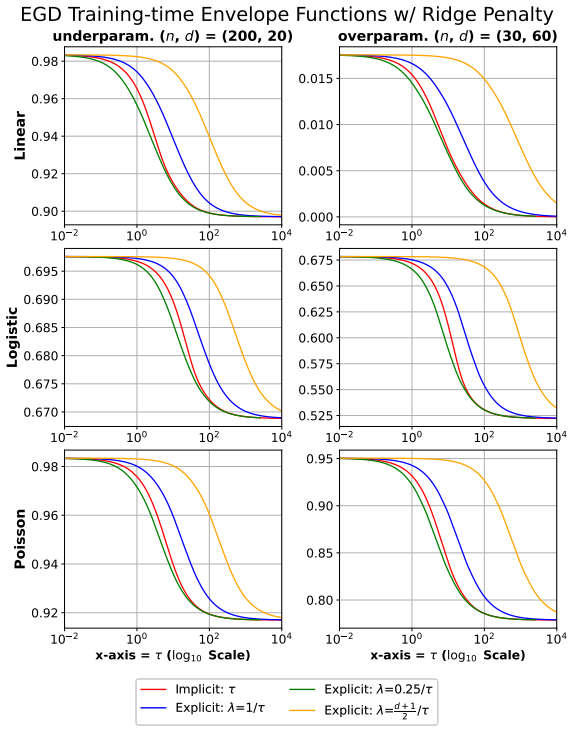
Providing more granular comparison, we visualize the solution paths during the training, i.e., the evolution of each component of estimators. Figure 1 compares the solution paths of θ_T and $\hat{\theta}_\lambda$ side-to-side (columns), for three GLMs (rows) in two (n, d)-regimes (columns). The underparametrized regime displays all $d = 20$ components, while the overparametrized regime only displays the first 40 components. The $\log_{10}$ scale x-axis represents τ for the iterative regularization and $1/\lambda$ for the explicit regularization.

The solution paths of two estimators look strikingly similar. Yet the equivalence of the solution paths was not formally proved in the paper, this provides more compelling visual evidence for the deep connection between the implicit and corresponding explicit regularization. Notably, for EGD and KL-regularization, many components converge to zero, so that induces sparsity even though the true parameter θ_0 is dense.

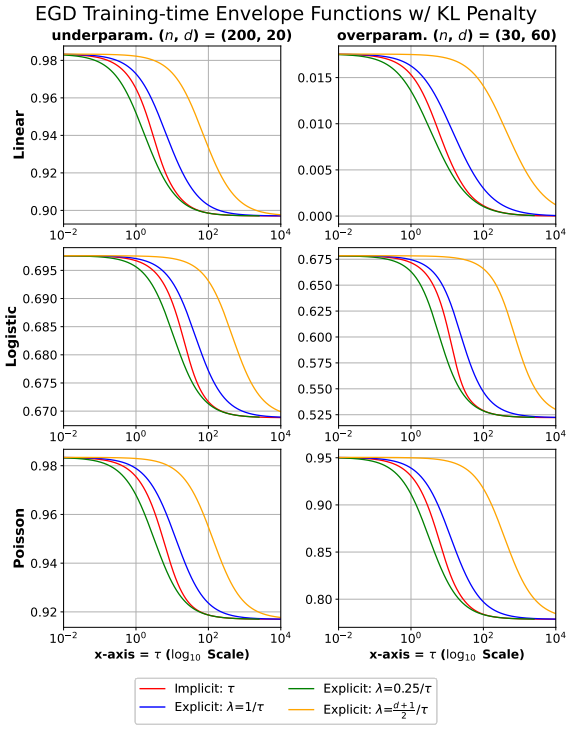
In the overparametrized regime for GD and ridge-regularization, the logistic and Poisson regression estimators diverge as τ increases. This is consistent with the known results. For logistic regression on linearly separable data, the GD solution diverges while its direction converges to the max-margin direction (Soudry et al., 2018), and linear separability is more likely to happen in an overparmetrized regime. For Poisson regression, a solution exists if any only if there exists $\delta \in \mathrm{Null}(X^\top)$ such that $y_i + \delta_i > 0$ for $i \in [n]$. In overparametrized regime, as $X^\top \in \mathbb{R}^{d \times n}$ and $n < d$, it is more likely to be $\mathrm{Null}(X) = \emptyset$, thus a solution only exists when $y_i > 0$ for any i.



(a) GD and ridge-regularization. Use $\|\theta\|_2^2$ penalty for plotting. $\tau \in [10^{-3}, 10^3]$.

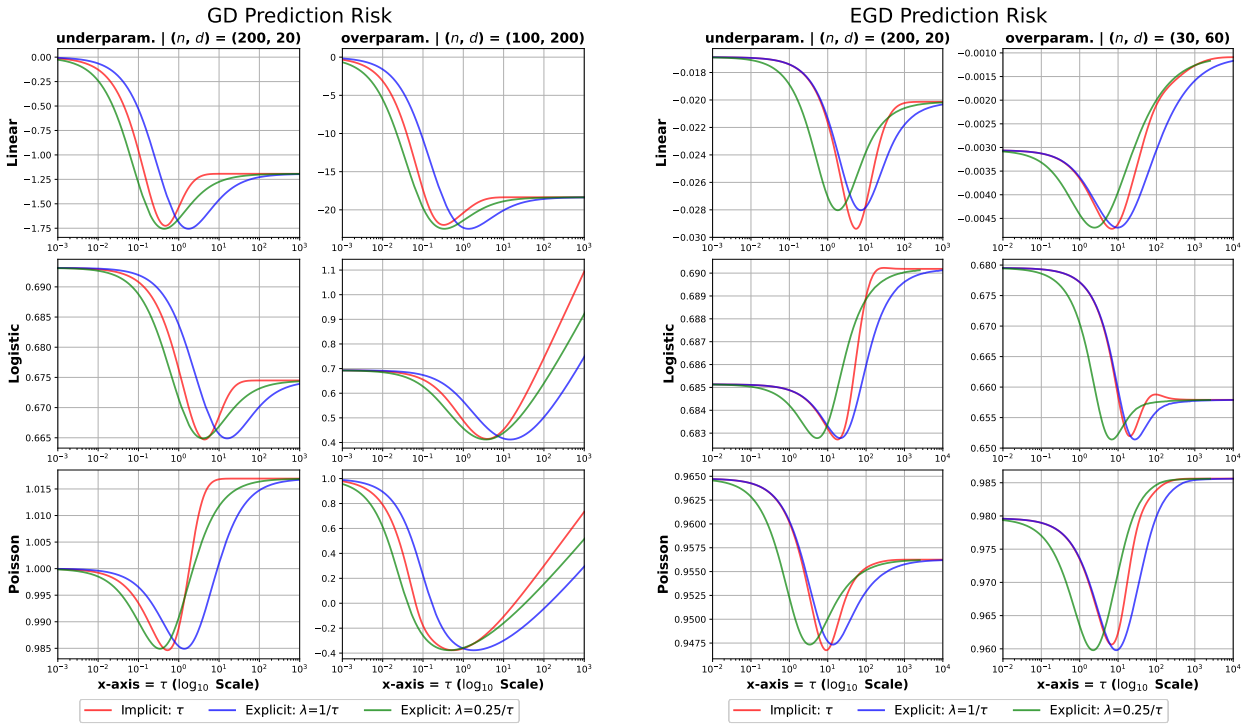


(b) EGD and KL-regularization. Use $\|\theta - \pi\|_1^2$ penalty for plotting. $\tau \in [10^{-2}, 10^4]$.



(c) EGD and KL-regularization. Use $D_{\mathrm{KL}}(\theta, \pi)$ penalty for plotting. $\tau \in [10^{-2}, 10^4]$.

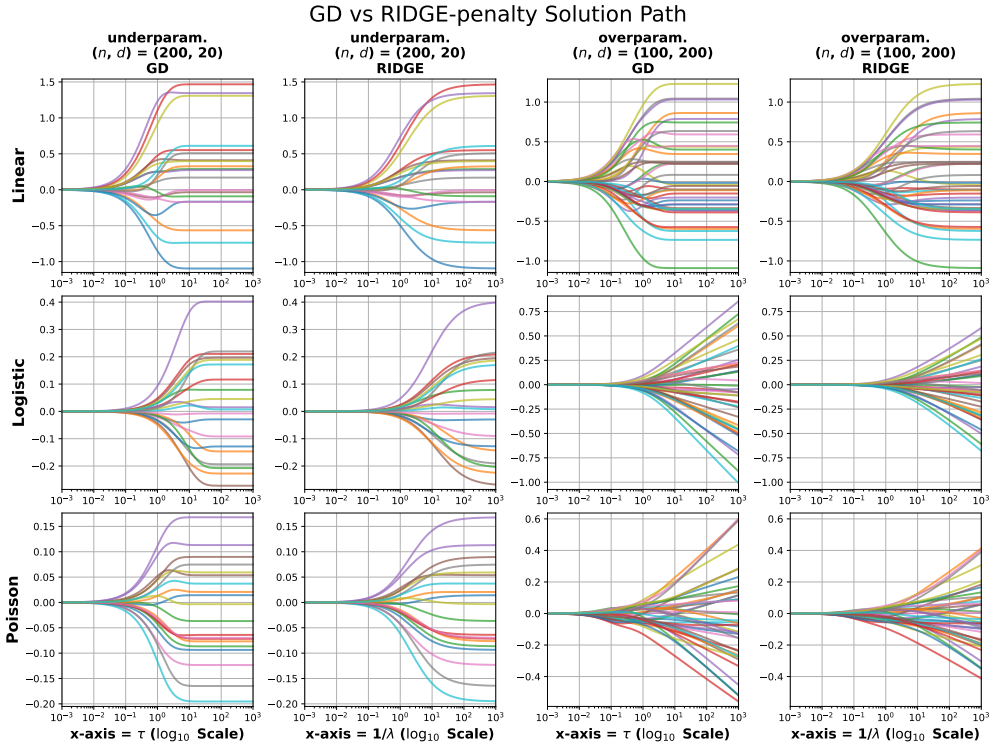
Figure 1: Training loss plus penalty trajectories (y-axis) for total elapsed time τ (x-axis; $\log_{10}$-scale). Each subfigure has six plots, corresponding to three GLMs in two (n, d)-regimes. The red color represents θ_T and the other color represents $\hat{\theta}_\lambda$ with different λ's as a function of τ.



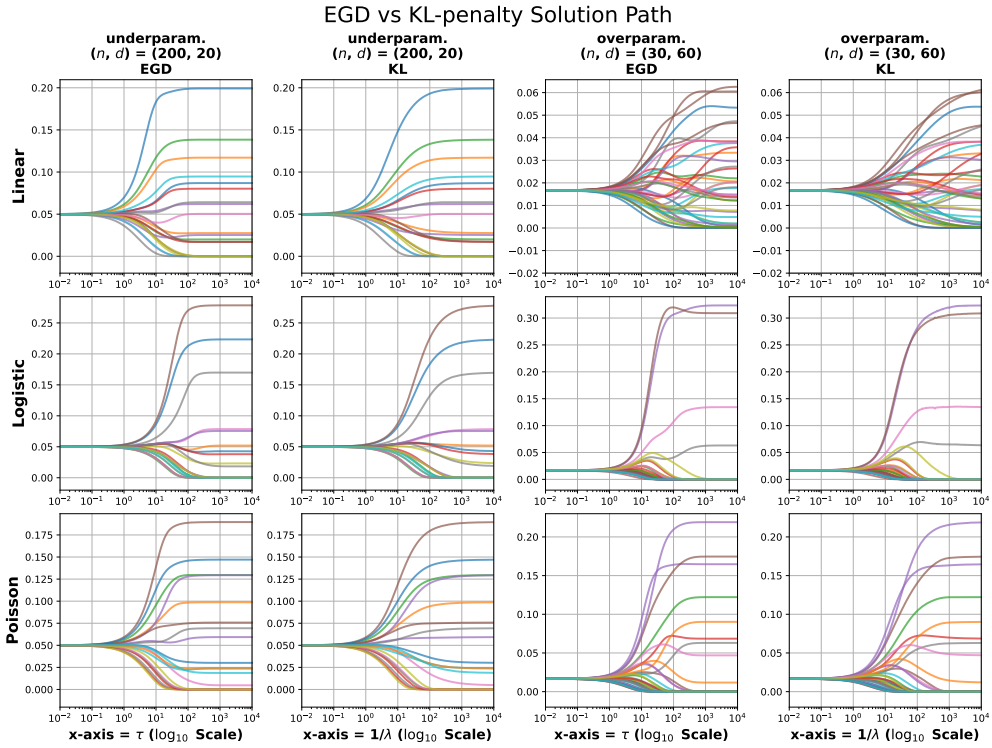
(a) GD versus ridge-regularization. $\tau \in [10^{-3}, 10^{3}]$.

(b) EGD versus KL-regularization. $\tau \in [10^{-2}, 10^{4}]$.

Figure 2: Prediction risk (y-axis) for total elapsed time τ (x-axis; $\log_{10}$-scale). Each subfigure has six plots, corresponding to three GLMs in two (n, d)-regimes. The red color represents θ_T and the other color represents $\hat{\theta}_\lambda$ with different λ's as a function of τ.



(a) GD versus ridge-regularization.



(b) EGD versus KL-regularization.

Figure 3: Solution path (y-axis) for τ or $1/\lambda$ (x-axis; $\log_{10}$-scale). The underparametrized regimes visualized all $d = 20$ components of the estimators, but the overparametrized regimes only shows the first 40 components. The color scheme in each plot pair of the implicit and explicit regularization is the same, as they are matched to the ordering of the components.

9 Discussion

We introduced basic inequalities for iterative algorithms as a unifying framework for the optimization and statistical analysis. This framework allows a comparison between the implicit regularization of the iterative algorithms and the corresponding explicit regularization in the original optimization problem. We demonstrated the broad utility of this framework through training dynamics and prediction risk analysis. However, our results also reveal a trade-off: the generality of basic inequalities and the cost of tightness, as the resulting bounds are not always as tight as those from some algorithm- or problem-specific observations in the literature.

This trade-off opens several interesting directions for future research. First, using stronger assumptions on loss functions, such as strong convexity, may lead to tighter basic inequalities that can refine the analysis. Conversely, we may extend the applicability of this framework by relaxing assumptions on losses, such as non-smoothness or non-convexity, as well as using modern models such as deep neural networks and transformers. These extensions would cover larger loss and algorithms classes from which we can get insights via basic inequalities.

References

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pages 1370–1378. PMLR, 2019.

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

Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. In *International Conference on Learning Representations*, 2019.

Francis Bach. Self-concordant analysis for logistic regression. 2010.

Peter Bartlett, Michael Collins, Ben Taskar, and David McAllester. Exponentiated gradient algorithms for large-margin structured classification. *Advances in neural information processing systems*, 17, 2004.

Heinz H Bauschke, Jonathan M Borwein, et al. 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.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

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

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

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

Michael Collins, Amir Globerson, Terry Koo, Xavier Carreras Pérez, and Peter Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin markov networks. *Journal of Machine Learning Research*, 9:1775–1822, 2008.

Arnak S Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. 2012.

Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–451, 2004.

Leonhard Euler. *Institutiones calculi integralis*, volume 1. impensis Academiae imperialis scientiarum, 1792.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

David Helmbold, Jyrki Kivinen, and Manfred KK Warmuth. Worst-case loss bounds for single neurons. *Advances in neural information processing systems*, 8, 1995.

David P Helmbold, Robert E Schapire, Yoram Singer, and Manfred K Warmuth. On-line portfolio selection using multiplicative updates. *Mathematical Finance*, 8(4):325–347, 1998.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pages 1772–1798. PMLR, 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*, pages 2109–2136. PMLR, 2020.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B Tsybakov. Learning by mirror averaging. 2008.

Anatoli B Juditsky, Alexander V Nazin, Alexandre B Tsybakov, and Nicolas Vayatis. Recursive aggregation of estimators by the mirror descent algorithm with averaging. *Problems of Information Transmission*, 41 (4):368–384, 2005.

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and computation*, 132(1):1–63, 1997.

Guillaume Lecué. Optimal rates of aggregation in classification under low noise assumption. 2007.

Guillaume Lecué and Shahar Mendelson. On the optimality of the aggregate with exponential weights for low temperatures. 2013.

Guillaume Lecué and Philippe Rigollet. Optimal learning with q-aggregation. 2014.

B Lemaire. An asymptotical variational principle associated with the steepest descent method for a convex function. *Journal of Convex Analysis*, 3:63–70, 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 L Sharpnack, Alessandro Rinaldo, and Ryan J Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. *Advances in neural information processing systems*, 30, 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.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 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.

David Pollard. A few good inequalities. 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*, pages 55–69. Springer, 2002.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

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

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 Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. *Advances in Neural Information Processing Systems*, 31, 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.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Ryan Tibshirani. High-dimensional regression: Ridge, 2023.

Sara A van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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 *The Thirty Seventh Annual Conference on Learning Theory*, pages 5019–5073. PMLR, 2024.

Jingfeng Wu, Peter 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. *Advances in neural information processing systems*, 30, 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*, pages 1538–1579, 2005.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A Regarding Section 2: Complements for Theorem 1 and 2

A.1 Gradient descent

Lemma A1 (Descent lemma for gradient descent)**.** *Under Assumption As1, consider gradient descent with iterates (1) and step sizes $\eta_t \in (0, 2/L]$. Then, for any $t \geq 0$,*

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

Proof of Lemma A1. The L-smoothness of f implies that

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

Since $\theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t)$ as in (1), it completes the proof. □

A.2 Bregman divergence and ϕ

Lemma A2 (Well-definedness of the Bregman divergence induced by ϕ)**.** *Under Assumption As2, a function $D(\cdot, v) : \Omega \to \mathbb{R}^d$ is well-defined if and only if $v \in \mathrm{int}(\Omega)$.*

Proof. By the definition of Legendre type, ϕ is essentially smooth. Also, as ϕ is a proper continuous convex function over a closed domain Ω, by Theorem 7.1 in Rockafellar (1997), ϕ is a closed function. Therefore, by Theorem 26.1 in Rockafellar (1997), we know $\partial\phi(v) = \emptyset$ if $v \notin \mathrm{int}(\Omega)$, and $\partial\phi(v) = \{\nabla\phi(v)\}$ if $v \in \mathrm{int}(\Omega)$. □

Lemma A3 (Strong convexity and lower bound of Bregman divergence)**.** *Under Assumption As2 and As3, for any $u \in \mathcal{K}$ and $v \in \mathcal{K} \cap \mathrm{int}(\Omega)$,*

$$D_\phi(u, v) \geq \frac{\alpha}{2}\|u - v\|^2.$$

Proof of Lemma A3. Note that $D_\phi(\cdot, v)$ is well-defined by Lemma A2. The desired bound directly follows from the α-strong convexity of ϕ in $\mathcal{K}$: $D_\phi(u, v) = \phi(u) - \phi(v) - \langle\nabla\phi(v), u - v\rangle \geq \frac{\alpha}{2}\|u - v\|^2$. □

A.3 Mirror descent

Lemma A4 (Well-definedness of $D_\phi(\cdot, \theta_t)$ in mirror descent)**.** *Under Assumption As2, consider mirror descent updates with iterates (2). Then, if $\theta_0 \in \mathcal{K} \cap \mathrm{int}(\Omega)$ then $\theta_t \in \mathcal{K} \cap \mathrm{int}(\Omega)$ for any $t \geq 0$. In other words,*

$$D_\phi(\cdot, \theta_t) : \Omega \to \mathbb{R} \text{ is well-defined for any } t \geq 0.$$

In particular, when $\mathcal{K} = \Omega$, this is equivalent to: if $\theta_0 \in \mathrm{int}(\mathcal{K})$ then $\theta_t \in \mathrm{int}(\mathcal{K})$ for any $t \geq 0$.

Proof. Assume $\theta_t \in \mathcal{K} \cap \mathrm{int}(\Omega)$, then $D_\phi(\cdot, \theta_t)$ is well-defined by Lemma A2. Recall θ_{t+1} in (2):

$$\theta_{t+1} = \arg\min_{\theta \in \mathcal{K}} \eta_t\langle\nabla f(\theta_t), \theta\rangle + D_\phi(\theta, \theta_t).$$

For brevity, define a function $F : \mathcal{K} \to \mathbb{R}$ as $F(\theta) = \eta_t\langle\nabla f(\theta_t), \theta\rangle + D_\phi(\theta, \theta_t)$. Due to the first order optimality condition, there exists a subgradient $g \in \partial F(\theta_{t+1})$ such that $\langle g, \theta - \theta_{t+1}\rangle \geq 0$ for all $\theta \in \mathcal{K}$. Meanwhile, since $\mathcal{K} \subseteq \Omega$, by Theorem 23.8 in Rockafellar (1997), for any $\theta \in \mathcal{K}$,

$$\partial F(\theta) = \partial\Big(\eta_t\langle\nabla f(\theta_t), \theta\rangle\Big) + \partial\Big(D_\phi(\theta, \theta_t)\Big) = \eta_t\nabla f(\theta_t) - \partial\phi(\theta) - \nabla\phi(\theta_t).$$

Suppose $\theta_{t+1} \notin \mathrm{int}(\Omega)$, then $\partial\phi(\theta_{t+1}) = \emptyset$ since ϕ is of Legendre type, as observed in Lemma A2. Therefore, $\partial F(\theta_{t+1})$ is also empty, which is contradictory to the existence of the aforementioned subgradient g. Therefore, $\theta_{t+1} \in \mathrm{int}(\Omega)$, and thus $D_\phi(\cdot, \theta_{t+1}) : \Omega \to \mathbb{R}$ is well-defined. The proof is completed by mathematical induction. □

Lemma A5 (Three-point inequality for mirror descent)**.** *Under Assumption As2, consider mirror descent updates with iterates (2). Then, for any $z \in \mathcal{K}$ and $t \geq 0$,*

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

Proof of Lemma A5. For brevity, define $F(\theta) = \eta_t \langle \nabla f(\theta_t), \theta \rangle + D_\phi(\theta, \theta_t)$. By Lemma A2 and the first order optimality condition for θ_{t+1},

$$\nabla F(\theta_{t+1}) = \eta_t \nabla f(\theta_t) + \nabla \phi(\theta_{t+1}) - \nabla \phi(\theta_t) \quad \text{and} \quad 0 \leq \langle \nabla F(\theta_{t+1}), z - \theta_{t+1} \rangle \text{ for any } z \in \mathcal{K}.$$

Rearranging the terms, this is equivalent to

$$\eta_t \langle \nabla f(\theta_t), \theta_{t+1} - z \rangle \leq \langle \nabla \phi(\theta_{t+1}) - \nabla \phi(\theta_t), z - \theta_{t+1} \rangle \text{ for any } z \in \mathcal{K}.$$

Since $\nabla \phi(\theta_t)$ and $\nabla \phi(\theta_{t+1})$ are well-defined, the standard three-point identity for D_ϕ suggests that

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

which concludes the proof. □

Lemma A6. *Under Assumption As2, consider mirror descent updates in (2). Then, for any $t \in \mathbb{N}_0$ and $z \in \mathcal{K}$,*

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

Proof of Lemma A6. Since f is convex in $\mathcal{K}$, we have

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

Multiplying both sides by η_t,

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

Applying Lemma A5 completes the proof. □

Lemma A7 (Descent lemma for mirror descent)**.** *Under Assumption As2 and As3, consider mirror descent updates in (2) with the step sizes $\eta_t \in (0, 2\alpha/L]$. Then, for any $t \in \mathbb{N}_0$,*

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

Proof of Lemma A7. Since f is L-smooth with respect to $\|\cdot\|$ in $\mathcal{K} \cap \operatorname{int}(\Omega)$, it follows that

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

Meanwhile, Lemma A5 with $z = \theta_t$ yields

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

Combining these two inequalities, we obtain that

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

Since both $D_\phi(\theta_t, \theta_{t+1})$ and $D_\phi(\theta_{t+1}, \theta_t)$ are lower bounded by $\frac{\alpha}{2} \|\theta_{t+1} - \theta_t\|^2$ by Lemma A3, the proof is completed by noting that $\eta_t \leq \alpha / L$. □

B Regarding Section 3: Complements for Corollary 1 to 4

B.1 Proof for part (d) and (e) of Corollary 1

(d). Note that the projection map $\text{Proj}_S(u)$ is well-defined as S is closed and convex. Let $s \in S$. From part (c), the sequence $\{\|\theta_t - s\|_2\}_{t=0}^{\infty}$ is non-increasing and bounded below by 0, hence it converges. This implies the sequence $\{\theta_t\}_{t=0}^{\infty}$ is bounded. By the Bolzano-Weierstrass theorem, there exists a subsequence $\{\theta_{t_i}\}_{i=1}^{\infty}$ that converges to a limit point $\theta_\infty := \lim_{i\to\infty} \theta_{t_i}$. By the continuity of f and the result from part (b), we have $f(\theta_\infty) = \lim_{i\to\infty} f(\theta_{t_i}) = \inf f$. Thus, $\theta_\infty \in S$.

Now we show the entire sequence converges to θ_∞. Since $\theta_\infty \in S$, part (c) implies that for any $t \geq t_i$, $\|\theta_t - \theta_\infty\|_2 \leq \|\theta_{t_i} - \theta_\infty\|_2$. As $i \to \infty$, the right-hand side converges to 0. Therefore, $\lim_{t\to\infty} \|\theta_t - \theta_\infty\|_2 = 0$, which means $\lim_{t\to\infty} \theta_t = \theta_\infty$.

Finally, by the continuity of the norm, for any $s \in S$, $\|\theta_\infty - s\|_2 = \lim_{t\to\infty} \|\theta_t - s\|_2 \leq \|\theta_0 - s\|_2$. Choosing $s = \text{Proj}_S(\theta_0)$ gives $\|\theta_\infty - \text{Proj}_S(\theta_0)\|_2 \leq \|\theta_0 - \text{Proj}_S(\theta_0)\|_2 = \text{Dist}_S(\theta_0)$.

(e). Let $P := \text{Proj}_S(\theta_0)$ and $v = P - \theta_\infty$. For any $c \geq 0$, define $\beta_c := P + c \cdot \text{Dist}_S(\theta_0) \cdot (v/\|v\|_2) \in S$. Since $\beta_c \in S$, due to part (c) and (d), we must have $\|\theta_\infty - \beta_c\|_2 \leq \|\theta_0 - \beta_c\|_2$. Since the points θ_∞, P, and β_c are collinear by construction,

$$\|\theta_\infty - \beta_c\|_2 = \|\theta_\infty - P\|_2 + \|P - \beta_c\|_2 = \|v\|_2 + c \cdot \text{Dist}_S(\theta_0).$$

By the Pythagorean theorem,

$$\|\theta_0 - \beta_c\|_2^2 = \|\theta_0 - P\|_2^2 + \|P - \beta_c\|_2^2 = \text{Dist}_S(\theta_0)^2 + (c \cdot \text{Dist}_S(\theta_0))^2 = (1 + c^2) \cdot \text{Dist}_S(\theta_0)^2.$$

Substituting these expressions into the inequality $\|\theta_\infty - \beta_c\|_2 \leq \|\theta_0 - \beta_c\|_2$ gives us that

$$\|v\|_2 + c \cdot \text{Dist}_S(\theta_0) \leq \sqrt{1 + c^2} \cdot \text{Dist}_S(\theta_0).$$

As $c \to \infty$, the term $\sqrt{1 + c^2} - c \to 0$. Therefore, we must have $\|v\|_2 \leq 0$, implying $v = 0$, and thus, $P = \theta_\infty$.

B.2 Proof for part (d) and (e) of Corollary 2

(d). Fix $s \in S$. From part (c), the sequence $\{D_\phi(s, \theta_t)\}_{t=0}^{\infty}$ is non-increasing. The lower bound $D_\phi(s, \theta_t) \geq \frac{\alpha}{2}\|s - \theta_t\|^2$ implies that the sequence $\{\theta_t\}_{t=0}^{\infty}$ is bounded with respect to $\|\cdot\|$, and thus bounded with respect to $\|\cdot\|_2$. Since the sequence lies within the closed set $\mathcal{K}$, it has a convergent subsequence $\{\theta_{t_i}\}_{i=1}^{\infty}$ with limit $\theta_\infty := \lim_{i\to\infty} \theta_{t_i} \in \mathcal{K}$. By continuity of f on $\mathcal{K}$ and the result of part (b), we have $f(\theta_\infty) = \inf f$, therefore

$$\theta_\infty \in S.$$

We now show that the entire sequence of $\{\theta_t\}_{t=0}^{\infty}$ converges to θ_∞. Suppose this is not true. Then there exists a subsequence $\{\theta_{t_j}\}_{j=1}^{\infty}$ and $\delta > 0$ such that $\|\theta_{t_j} - \theta_\infty\| \geq \delta$ for all j. With the same argument we did earlier, there exists a sub-subsequence $\{\theta_{t_{j_k}}\}_{k=1}^{\infty}$ that converges to another limit point $\tilde{\theta}_\infty \in S$ such that $\|\tilde{\theta}_\infty - \theta_\infty\| \geq \delta$. Considering two different assumptions given in the theorem statement, we can prove

$$D_\phi(\theta_\infty, \theta_{t_i}) \to 0 \quad \text{as} \quad i \to \infty :$$

(i) Suppose $S \cap \text{int}(\Omega) \neq \emptyset$. Choose any $s \in S \cap \text{int}(\Omega)$. Since $\{D_\phi(s, \theta_{t_i})\}_{i=0}^{\infty}$ is bounded, by Theorem 3.8(ii) in Bauschke et al. (1997), we know that $\theta_\infty \in \text{int}(\Omega)$ and $D_\phi(\theta_\infty, \theta_{t_i}) \to 0$ as $i \to \infty$.

(ii) Suppose for any $y \in \Omega$ and for any sequence $\{y_n\}_{n=1}^{\infty} \subset \text{int}(\Omega)$ converging to y, $D_\phi(y, y_n) \to 0$. Then we know $D_\phi(\theta_\infty, \theta_{t_i}) \to 0$ as $i \to \infty$.

Then, by its decreasing nature proved in part (b), we conclude that $D_\phi(\theta_\infty, \theta_t) \to 0$ as $t \to \infty$. Similarly we know that $D_\phi(\tilde{\theta}_\infty, \theta_t) \to 0$. However, these give a contradiction:

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

Thus, the entire sequence $\{\theta_t\}_{t=0}^{\infty}$ must converge, and we conclude

$$\lim_{t\to\infty} \theta_t = \theta_\infty.$$

(e). Suppose S is a non-empty affine subspace with $S \subset \mathrm{int}(\Omega)$. For brevity, write $P := \mathrm{BregProj}_S(\theta_0) \in S$. Let $v := P - \theta_\infty \neq 0$, then $P + cv \in S$ for any $c \in \mathbb{R}$ since S is affine. Meanwhile, note that $\nabla\phi(P)$ and $\nabla\phi(\theta_\infty)$ are well-defined since $P, \theta_\infty \in S \subset \mathrm{int}(\Omega)$.

We know three inequalities about P, θ_∞ and $P + cv$. First, the three-point identity of 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 the result of part (c), we know that $D_\phi(P + cv, \theta_\infty) \leq D_\phi(P + cv, \theta_0)$. 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 inequalities, we have

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

Since the above inequality holds for arbitrary $c \in \mathbb{R}$, we conclude $\nabla\phi(P) = \nabla\phi(\theta_\infty)$. This implies

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

where the last inequality holds by the alternative definition of α-strong convexity. Thus $v = 0$ and $\theta_\infty = P$.

B.3 Proof of Corollary 3

First we will prove (i) in the corollary statement using Corollary 2. Note that the negative entropy function ϕ is 1-strongly convex with respect to $\|\cdot\|_1$ due to Pinsker's inequality. Also, it is well known that the Bregman divergence generated by the negative entropy function is the KL divergence: for $a = (a_1, \ldots, a_d)^\top$ and $b = (b_1, \ldots, b_d)^\top$,

$$D_\phi(a, b) = \sum\nolimits_{j=1}^{d} a_j \log(a_j / b_j).$$

Let $s_j := z_j - 1/d$. Then $\sum_{j=1}^d s_j = 0$ and $\|z - \pi\|_1 = \sum_{j=1}^d |s_j|$. Observe that

$$\begin{aligned}
D_\phi(z, \pi) = \sum_{j=1}^{d} z_j \log(dz_j) &= \sum_{j=1}^{d} (s_j + 1/d) \log(1 + ds_j) \\
&\leq \sum_{j=1}^{d} (s_j + 1/d) ds_j = d \sum_{j=1}^{d} s_j^2 \overset{(*)}{\leq} \frac{d}{2} \Big(\sum_{j=1}^{d} |s_j| \Big)^2 = \frac{d}{2} \|z - \pi\|_1^2,
\end{aligned}$$

where the first inequality holds from $\log(1 + x) \leq x$ for $x > -1$ (and when $ds_j = -1$, the first inequality holds with equality where both sides are zero due to $(s_j + 1/d)$-term). The inequality $(*)$ holds due to follow reasoning. Let $C = \sum_i s_i \cdot \mathbb{1}(s_i \geq 0)$, then $\sum_i s_i \cdot \mathbb{1}(s_i < 0) = C$ and $\sum_i |s_i| = 2C$ since $\sum_i s_i = 0$. Therefore,

$$\begin{aligned}
\sum_i s_i^2 &= \sum_i s_i^2 \cdot \mathbb{1}(s_i \geq 0) + \sum_i s_i^2 \cdot \mathbb{1}(s_i < 0) \\
&\leq \Big(\sum_i |s_i| \mathbb{1}(s_i \geq 0) \Big)^2 + \Big(\sum_i |s_i| \cdot \mathbb{1}(s_i < 0) \Big)^2 = 2C^2 = (\sum_i |s_i|)^2 / 2.
\end{aligned}$$

Thus, as we derived that $D_{\mathrm{KL}}(z, \pi) \leq (d/2)\|z - \pi\|_1^2$ for any $z \in \mathcal{K}$ where π is uniform, Corollary 2(a) completes the first half of the proof.

To prove (ii) in the corollary statement, we need a different upper bound on $D_\phi(z, \pi)$, then following similar steps as in the proof of Corollary 2(a). By Theorem 1 of Sason (2015), we know that for any $z \in \mathcal{K}$,

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

Also, as we have already observed in Corollary 2(a), $D_{\mathrm{KL}}(z, \theta_T) \geq \frac{1}{4}\|\theta_0 - \theta_T\|^2 - \frac{1}{2}\|z - \theta_0\|^2$. Therefore,

$$D_{\mathrm{KL}}(z, \pi) - D_{\mathrm{KL}}(z, \theta_T) \leq \frac{\log d}{2}\|z - \theta_0\|_1 - \frac{1}{4}\|\theta_0 - \theta_T\|^2 - \frac{1}{2}\|z - \theta_0\|^2.$$

Finally, the basic inequality in Theorem 2 completes the proof.

B.4 Lemma for Corollary 4

Lemma B1 (The solution set of the GLM is affine)**.** *Consider the GLM loss and estimator in Definition 1. Suppose that the search space is $\{\theta : \theta \in \mathcal{K}\}$ with an affine set $\mathcal{K}$ (e.g., $\mathbb{R}^d$ or unbounded simplex), and A is a strictly convex function. Then the solution set of GLM is either empty or an affine set.*

Proof. Say the solution set is S. Suppose S is not empty, and $s \in S$. Define a set $U := \mathcal{K} \cap (\{s\} + \{v \in \mathbb{R}^d : Xv = 0\})$, which is an affine set. Then, it is enough to show that $S = U$. Clearly, $S \supseteq U$ since the GLM loss function $\ell(\theta)$ depends on θ only via $X\theta$.

To show $S \subseteq U$, take any element u in U, and write $v = u - s$. Since $\ell(s) = \ell(u) = \min_{\theta \in \mathcal{K}} \ell(\theta)$ and $\ell(\cdot)$ is a convex function (see Lemma D1), we know that $\ell(s + cv) = \ell(s)$ for any $c \in (0, 1)$. Therefore, differentiating twice with respect to c, we get that

$$\begin{aligned} 0 = \frac{\mathrm{d}^2}{\mathrm{d}c^2}\ell(s + cv) = \frac{\mathrm{d}}{\mathrm{d}c}\langle v, \nabla\ell(s + cv)\rangle = \langle v, \nabla^2\ell(s + cv)v\rangle &= (Xv)^\top \mathrm{diag}\Big(\ddot{A}\big(X(s + cv)_{i\in[n]}\big)\Big)Xv \\ &= \sum_{i=1}^{n}(Xv)_i^2\ddot{A}\big((Xs + cXv)_i\big). \end{aligned}$$

Since $\ddot{A}(\cdot) > 0$ due to strict convexity, we have $Xv = 0$. In conclusion, $u \in (\{s\} + \{v \in \mathbb{R}^d : Xv = 0\})$, which implies $S \subseteq U$. □

C Generalized Linear Models

Generalized linear models (GLMs) refer to a broader model class related to diverse distributions within the exponential family, moving beyond the linear regression which is naturally related to the Gaussian distribution. An univariate exponential family distribution, in its canonical form, models the density or mass function $p(z|\xi)$ proportional to $\exp(\xi S(z) - A(\xi))$. Here, $\xi \in \mathbb{R}$ is the natural parameter, $S : \mathbb{R} \to \mathbb{R}$ is the sufficient statistic (often, and in our focus, $S(z) = z$), and $A : \mathbb{R} \to \mathbb{R}$ is the cumulant function. Key properties derived from A are $\mathbb{E}[S(Z)] = \dot{A}(\xi)$ and $\mathrm{Var}(S(Z)) = \ddot{A}(\xi)$, where dots denote differentiation. Familiar examples include: Gaussian distribution with $N(\mu, \sigma^2)$ with fixed σ^2 has $\xi = \mu$, $S(z) = z$, and $A(\xi) = \xi^2/2$; Bernoulli(p) has $\xi = \log(p/(1-p))$ which is called the logit link, $S(z) = z$, and $A(\xi) = \log(1 + e^\xi)$; and Poisson(μ) has $\xi = \log(\mu)$ which is called the log link, $S(z) = z$, and $A(\xi) = e^\xi$.

In a GLM, the natural parameter ξ is assumed to be linearly related to a predictor vector $x \in \mathbb{R}$ via $\xi = x^\top\theta$, where $\theta \in \mathbb{R}^d$. Under this assumption, estimation of θ is performed by maximum likelihood for the chosen exponential family.

The formal definition of the GLM loss function and estimator are defined as the following. Given data $(x_i, y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ and an exponential family characterized by (S, A), assume that $\xi_i = x_i^\top\theta$ for $i = 1, \ldots, n$. The maximum likelihood estimator for θ is

$$\hat{\theta}_0 := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} -\log p_{S,A}(y_i|x_i^\top\theta) = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\Big(-S(y_i)x_i^\top\theta + A(x_i^\top\theta)\Big). \tag{19}$$

The main article assumes that the sufficient statistic is the identity map, i.e., $S(y) = y$. However, our analytical framework can be readily extended to general S by substituting y_i with $S(y_i)$ for where appropriate. For instance, the vector Y can be simply replaced with $(S(y_1), \ldots, S(y_n))^\top$ in (4).

D Regarding Section 4.1

D.1 Lemmas

Lemma D1 (Convexity of ℓ and strong convexity of ℓ_λ)**.** *Recall $\ell(\theta)$ and $\ell_\lambda(\theta)$ defined in (4) and (6). $\ell(\theta)$ is a convex function with respect to $\theta \in \mathbb{R}^d$, and moreover, for any $\lambda > 0$, $\ell_\lambda(\theta)$ is a 2λ-strongly convex function with respect to $\theta \in \mathbb{R}^d$.*

Proof of Lemma D1. Easily observe that $\nabla \ell(\theta) = -\frac{1}{n} X^\top (Y - \nabla A(X\theta))$ and $\nabla^2 \ell(\theta) = \frac{1}{n} X^\top \nabla^2 A(X\theta) X$. Note that $\nabla^2 A(v) = \mathrm{diag}((\frac{\mathrm{d}^2}{\mathrm{d}u^2} A(u)|_{v_i})_{i=1}^n) \in \mathbb{R}^{n \times n}$. Since it is a well-known property of exponential families that A is convex, we know $\nabla^2 A(v) \succeq 0$. This implies $\nabla^2 \ell(\theta) \succeq 0$, and thus, $\ell(\theta)$ is convex. Moreover, $\ell_\lambda(\theta)$ is a 2λ-strongly convex function for $\lambda > 0$ because $\nabla^2 \ell_\lambda(\theta) = \nabla^2 \ell(\theta) + 2\lambda I_n \succeq 2\lambda I_n$. □

D.2 Proof of Proposition 1

For any $\omega \in \mathbb{R}_{\geq 0}$ and $u, v \in \mathbb{R}^d$ such that $L_\omega(u) \leq L_\omega(v)$, this can be rewritten as

$$\frac{1}{n}\Big(A(Xu) - A(Xv)\Big) \leq Y^\top \frac{X}{n}(u - v) + \omega\Big(\|v\|_2^2 - \|u\|_2^2\Big).$$

By combining this with Definition 2 of the prediction risk, we have

$$\begin{aligned}\mathrm{Risk}(u) - \mathrm{Risk}(v) &= \frac{1}{n}\Big(A(Xu) - A(Xv) - \mu_0^\top Xu + \mu_0^\top Xv\Big) \\ &\leq Y^\top \frac{X}{n}(u - v) + \omega\Big(\|v\|_2^2 - \|u\|_2^2\Big) - \mu_0^\top \frac{X}{n}(u - v) = \epsilon^\top \frac{X}{n}(u - v) + \omega\Big(\|v\|_2^2 - \|u\|_2^2\Big),\end{aligned}$$

which is equivalent to

$$\mathrm{Risk}(u) - \mathrm{Risk}(v) \leq \Big\langle \frac{X^\top \epsilon}{n}, u - v \Big\rangle + \omega\Big(\|v\|_2^2 - \|u\|_2^2\Big).$$

Therefore, as we know $L_\lambda(\hat{\theta}_\lambda) \leq L_\lambda(\theta)$ for any β by the definition of $\hat{\theta}_\lambda$, the above inequality suggests

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \mathrm{Risk}(\theta) \leq \epsilon^\top \frac{X}{n}(\hat{\theta} - \theta) + \lambda\Big(\|\theta\|_2^2 - \|\hat{\theta}\|_2^2\Big). \tag{20}$$

Finally, we can prove that

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

from following observation: Using Young's inequality, i.e., $2ab \leq ca^2 + b^2/c$ for any $c > 0$, we have

$$\begin{aligned}2\epsilon^\top \frac{X}{n}\Big(\hat{\theta}_\lambda - \theta\Big) + 2\lambda\Big(\|\beta\|_2^2 - \|\hat{\theta}_\lambda\|_2^2\Big) &= 2\Big(\frac{X}{n}^\top \epsilon\Big)^\top (\hat{\theta}_\lambda - \theta) + 2\lambda\Big(\|\theta\|_2^2 - \|\hat{\theta}_\lambda\|_2^2\Big) \\ &\leq \frac{1}{\lambda}\Big\|\frac{X}{n}^\top \epsilon\Big\|_2^2 + \lambda\|\hat{\theta}_\lambda - \theta\|_2^2 + 2\lambda\Big(\|\theta\|_2^2 - \|\hat{\theta}_\lambda\|_2^2\Big) \\ &\leq \frac{1}{\lambda}\Big\|\frac{X}{n}^\top \epsilon\Big\|_2^2 + \lambda\Big(2\|\hat{\theta}_\lambda\|_2^2 + 2\|\theta\|_2^2\Big) + 2\lambda\Big(\|\theta\|_2^2 - \|\hat{\theta}_\lambda\|_2^2\Big) \\ &= \frac{1}{\lambda}\Big\|\frac{X}{n}^\top \epsilon\Big\|_2^2 + 4\lambda\|\theta\|_2^2.\end{aligned}$$

This completes the proof.

D.3 Proof of Proposition 2

Since $\epsilon_i \sim \mathrm{sG}(\sigma)$, due to the Remark 1 of Hsu et al. (2012), we know

$$\mathbb{P}\left(\left\|\frac{X^\top \epsilon}{n}\right\|_2^2 > \frac{\sigma^2}{n}\left[\mathrm{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F\sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\mathrm{op}}\delta\right]\right) \le e^{-\delta}, \tag{21}$$

with the fact that $\|XX^\top\|_F = \|X^\top X\|_F$ and $\mathrm{tr}(XX^\top) = \mathrm{tr}(X^\top X)$. Applying this concentration inequality to Proposition 1 gives us that, with probability at least $1 - e^{-\delta}$,

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \inf_{\theta:\ \|\theta\|_2 \le b} \mathrm{Risk}(\theta) \le \frac{1}{2\lambda} C_{\mathrm{sG}} + 2\lambda b^2,$$

where we define

$$C_{\mathrm{sG}} := \frac{\sigma^2}{n}\left[\mathrm{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F\sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\mathrm{op}}\delta\right] \tag{22}$$

for brevity. Therefore, when we choose $\lambda = \sqrt{C_{\mathrm{sG}}}/(2b)$, i.e., λ as (7), the following holds with probability at least $1 - e^{-\delta}$:

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \inf_{\theta:\ \|\theta\|_2 \le b} \mathrm{Risk}(\theta) \le 2b\sqrt{C_{\mathrm{sG}}}.$$

D.4 Proof of Theorem 3

Recall that Assumption As4 says $y_i | x_i \overset{ind.}{\sim} P_i$.

Ridge-penalized GLM with Gaussian distribution. Since $\epsilon_i = y_i - \mu_i \sim \mathcal{N}(0, \sigma_i)$, clearly $\epsilon_i \sim \mathrm{sG}(\sigma_i^2)$. Thus $\epsilon_i \sim \mathrm{sG}(\sigma_{\mathrm{Dist}}^2)$ with $\sigma_{\mathrm{Dist}} = \max_{i\in[n]} \sigma_i$. Then the result directly comes from Proposition 2.

Ridge-penalized GLM with Bernoulli distribution. Since $\epsilon_i = y_i - \mu_i \in [-\mu_i, 1 - \mu_i]$, we know $\epsilon_i \sim \mathrm{sG}(1/4)$. Then the result directly comes from Proposition 2.

Ridge-penalized GLM with Poisson distribution. In Poisson regression case, we need an additional observation about a high-probability bound of Poisson random variables before we jump into the main proof. For such an upper bound, we follow a similar process as Appendix A.4 of Lin et al. (2017). Define an event

$$\mathcal{E} := \{\epsilon_i < D \text{ for all } 1 \le i \le n\} = \{y_i - \mu_i < D \text{ for all } 1 \le i \le n\} \quad \text{where} \quad D = 4\big(\|\mu\|_\infty + 1/3\big)\log n.$$

Note that $D > 1$ for $n \ge 3$. Then we observe $\mathbb{P}(\mathcal{E}^c) \le 1/n$ from following:

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}(\exists i, y_i - \mu_i \ge D) \le \sum\nolimits_{i=1}^n \mathbb{P}(y_i - \mu_i \ge D) \overset{(*)}{\le} n \times 1/n^2 = 1/n.$$

The inequality $(*)$ is elaborated more here. By the Poisson concentration result from Pollard (2017), for $X \sim \mathrm{Pois}(\mu)$,

$$\mathbb{P}(X - \mu \ge x) \le \exp\left(-\frac{x^2}{2\mu}\psi_{\mathrm{Benn}}\left(\frac{x}{\mu}\right)\right) \ \forall x > 0, \quad \text{where} \quad \psi_{\mathrm{Benn}}(x) = \frac{(1+x)\log(1+x) - x}{x^2/2}.$$

Moreover, when $x \ge 1$, *(Really? Need to check this by my own, at lease once.)*

$$\frac{x^2}{2\mu}\psi_{\mathrm{Benn}}\left(\frac{x}{\mu}\right) \ge \frac{1/2}{\mu + 1/3}x.$$

Therefore, we have the following for any $1 \le i \le n$ with $n \ge 3$, which completes the proof of inequality $(*)$:

$$\mathbb{P}(\epsilon_i \ge D) \le \exp\Big(-\frac{1/2}{\mu_i + 1/3}D\Big) \le \exp\Big(-\frac{1/2}{\|\mu\|_\infty + 1/3}D\Big) = n^{-2}.$$

Now we are ready to finish the main proof. For any $\delta > 0$, define an event

$$\mathcal{S}_\delta := \left\{ \left\| \frac{X^\top \epsilon}{n} \right\|_2^2 > \frac{\sigma_{\text{Pois}}^2}{n} \left[\text{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F \sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\text{op}} \delta \right] \right\} \quad \text{where} \quad \sigma_{\text{Pois}} := \frac{D + \|\mu\|_\infty}{2}.$$

Since we have observed $\mathbb{P}(\mathcal{E}^c) \le 1/n$, we can upper bound $\mathbb{P}(\mathcal{S}_\delta)$ as

$$\mathbb{P}(\mathcal{S}_t) = \mathbb{P}(\mathcal{S}_\delta \cap \mathcal{E}^c) + \mathbb{P}(\mathcal{S}_\delta \cap \mathcal{E}) \le \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{S}_\delta | \mathcal{E}) \mathbb{P}(\mathcal{E}) \le 1/n + \mathbb{P}(\mathcal{S}_\delta | \mathcal{E}) \overset{(**)}{\le} 1/n + e^{-\delta},$$

where the inequality $(**)$ holds by following reasoning. Under the event $\mathcal{E}$, we know that $\{\epsilon_i\}_{i=1}^n$ are still mutually independent and $\epsilon_i \in [-\mu_i, D)$. In other words, $\epsilon_i \sim \text{sG}(\sigma_i^2)$ with $\sigma_i = (D + \mu_i)/2$ are mutually independent under $\mathcal{E}$. Then $\epsilon_i \sim \text{sG}(\sigma_{\text{Pois}}^2)$ holds for all $1 \le i \le n$ since $\sigma_{\text{Pois}} \ge \sigma_i$. Therefore, due to Hsu et al. (2012) as stated in (21), we have $\mathbb{P}(\mathcal{S}_\delta | \mathcal{E}) \le e^{-\delta}$, and this implies $(**)$.

Finally, due to an upper bound of $\mathbb{P}(\mathcal{S}_\delta)$ and Proposition 1, with probability at least $1 - 1/n - e^{-\delta}$,

$$\text{Risk}(\hat{\theta}_\lambda) \le \inf_{\theta:\, \|\theta\|_2 \le b} \text{Risk}(\theta) + \frac{1}{2\lambda} C_{\text{Pois}} + 2\lambda b^2$$

where we define $C_{\text{Pois}} := (\sigma_{\text{Pois}}^2/n)[\text{tr}\big(\widehat{\Sigma}\big) + 2\big\|\widehat{\Sigma}\big\|_F \sqrt{\delta} + 2\big\|\widehat{\Sigma}\big\|_{\text{op}} \delta]$ for brevity. Hence choosing $\lambda = \sqrt{C_{\text{Pois}}}/(2b)$ gives us that, with probability at least $1 - 1/n - e^{-\delta}$,

$$\text{Risk}(\hat{\theta}_\lambda) \le \inf_{\theta:\, \|\theta\|_2 \le b} \text{Risk}(\theta) + 2b\sqrt{C_{\text{Pois}}}.$$

D.5 Analysis on linear regression with closed form solution

This is regarding the analysis of prediction risk for linear regression solution $\hat{\theta} = (X^\top X)^{-1} X^\top Y$, when it is well-defined. The risk in linear regression, following the definition (5), is: $\text{Risk}(\theta) = \frac{1}{n}(-\mu^\top X\theta + \frac{1}{2}\|X\theta\|_2^2)$. Meanwhile, since $Y = X\theta_0 + \epsilon$, we know that $\hat{\theta} = \theta_0 + (X^\top X)^{-1} X^\top \epsilon$, and thus $X\hat{\theta} = X\theta_0 + \epsilon$ where $H = X(X^\top X)^{-1} X^\top$ is the projection matrix. Therefore,

$$\text{Risk}(\hat{\theta}) - \text{Risk}(\theta_0) = \frac{1}{n}\Big(-\mu^\top X(\hat{\theta} - \theta_0) + \frac{1}{2}\|X\hat{\theta}\|_2^2 - \frac{1}{2}\|X\theta_0\|_2^2 \Big) = \frac{1}{2n}\|H\epsilon\|_2^2.$$

By Remark 1 in Hsu et al. (2012), and since $H^\top = H$ and $H^2 = H$, we know

$$\mathbb{P}\Big(\|H\epsilon\|_2^2/\sigma^2 > \text{tr}(H) + 2\sqrt{\text{tr}(H)\delta} + 2\|H\|_{\text{op}}\delta \Big) \le e^{-\delta}.$$

We know that $\text{tr}(H) = \text{tr}((X^\top X)^{-1} X^\top X) = d$ and $\|H\|_{\text{op}} = 1$ since H is idempotent. Thus we finally obtain $O(\sigma^2 d/n)$ high-probability bound: with probability at least $1 - e^{-\delta}$,

$$\text{Risk}(\hat{\beta}) - \text{Risk}(\beta_0) \le \frac{\sigma^2}{2n}\Big(d + 2\sqrt{d\delta} + 2\delta \Big).$$

E Regarding Section 4.2

E.1 Proof of Proposition 3

Recall the basic inequalities (i) for gradient descent in Theorem 1 and (ii) for projected gradient descent that follows the same form in Theorem 7, as explained in Section 4.2. These basic inequalities suggest that (i) for any $\theta \in \mathbb{R}^d$ in gradient descent, or (ii) for any $\theta \in \mathsf{B}_d(b)$ in projected gradient decent, the following holds:

$$\ell(\theta_T^{(\text{gd})}) + \frac{\lambda_T}{2}\|\theta_T^{(\text{gd})} - \theta\|_2^2 \le \ell(\theta) + \frac{\lambda_T}{2}\|\theta\|_2^2$$

since the initialization is set as $\theta_0^{(\text{gd})} = 0 \in \mathbb{R}^d$. By the definition of ℓ in (4), this is equivalent to

$$\frac{1}{n}\Big(A(X\theta_T^{(\text{gd})}) - A(X\theta) \Big) \le Y^\top \frac{X}{n}(\theta_T^{(\text{gd})} - \theta) + \frac{\lambda_T}{2}\Big(\|\theta\|_2^2 - \|\theta_T^{(\text{gd})} - \theta\|_2^2 \Big).$$

Following the same calculation as in the proof of Proposition 1,

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

Since $2\epsilon^\top \frac{X}{n}(\theta_T^{(\mathrm{gd})} - \theta) \le \frac{1}{\lambda_T}\|\frac{1}{n}X^\top\epsilon\|_2^2 + \lambda_T\|\theta_T^{(\mathrm{gd})} - \theta\|_2^2$ by Young's inequality, we conclude

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

E.2 Proof of Proposition 4

We need following auxiliary lemma for simpler computation.

Lemma E1. *Given a function $g(x) = \frac{a}{x} + bx$ defined in $(0,\infty)$ with $a, b > 0$. It is known that the function g obtains its minimum at $x^* = \sqrt{a/b}$. For $y \in (0,\infty)$ such that $1/y = 1/x^* + d$ with $d \ge 0$,*

$$g(y) - g(x^*) = ad^2y \le ad$$

Proof of Lemma E1. Note that $bx^* = a/x^*$.

$$g(y) - g(x^*) = a\Big(\frac{1}{y} - \frac{1}{x^*}\Big) + b(y - x^*) = ad - b\Big(\frac{1}{y} - \frac{1}{x^*}\Big)yx^* = d(a - byx^*) = d\Big(a - \frac{ay}{x^*}\Big) = ad^2y.$$

Also, $ad^2y \le ad$ since

$$dy = \Big(\frac{1}{y} - \frac{1}{x^*}\Big)y \le 1.$$

□

Now proceed to the main proof. Combining the high probability upper bound (21) with Proposition 3, we have

$$\mathrm{Risk}(\theta_T^{(\mathrm{gd})}) - \inf_{\theta:\ \|\theta\|_2 \le b} \mathrm{Risk}(\theta) \le \frac{1}{2\lambda_T}C_{\mathrm{sG}} + \frac{\lambda_T}{2}b^2.$$

with the probability at least $1 - e^{-\delta}$, with C_{sG} defined in (22). Observe that the right hand side achieves the minimum value of $b\sqrt{C_{\mathrm{sG}}}$ when $\lambda = \lambda_{\mathrm{gd}}^*$. Then we choose T as

$$T = \arg\min\{t \in \mathbb{N} : (\lambda_{\mathrm{gd}}^*)^{-1} \le \lambda_T^{-1}\} = \Big\lceil \frac{1}{\eta\lambda_{\mathrm{gd}}^*} \Big\rceil.$$

Note that the above minimum always exists since $\lambda_T^{-1} = \eta T \to \infty$ as $T \to \infty$. Meanwhile, there is additional discretization error due to the nature of gradient descent. Since it is clear that $0 \le 1/\lambda_T - 1/\lambda_{\mathrm{gd}}^* < \eta$, Lemma E1 says that

$$\Big(\frac{1}{2\lambda_T}C_{\mathrm{sG}} + \frac{\lambda_T}{2}b^2\Big) - \Big(\frac{1}{2\lambda_{\mathrm{gd}}^*}C_{\mathrm{sG}} + \frac{\lambda_{\mathrm{gd}}^*}{2}b^2\Big) \le \frac{C_{\mathrm{sG}}}{2}\Big(\frac{1}{\lambda_T} - \frac{1}{\lambda_{\mathrm{gd}}^*}\Big) < \frac{\eta C_{\mathrm{sG}}}{2}.$$

In conclusion, we proved that

$$\mathrm{Risk}(\theta_T^{(\mathrm{gd})}) \le \inf_{\theta:\ \|\theta\|_2 \le b} \mathrm{Risk}(\theta) + b\sqrt{C_{\mathrm{sG}}} + \frac{\eta C_{\mathrm{sG}}}{2}$$

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

E.3 Proof of Theorem 4

The theorem is proved straightforwardly from Proposition 4 once we prove L_{Dist} for each distribution. Note that $\nabla\ell(\theta) = \frac{1}{n}X^\top\nabla^2 A(X\theta)X$. We have listed $A(\eta)$ per distribution which was used in Section 4.

Gaussian distribution. $A(\xi) = \xi^2/2$ and $\ddot{A}(\xi) = 1$. Thus, $\ddot{A}(X\theta) = I_n$ and $\nabla^2\ell(\theta) = \widehat{\Sigma}$. Therefore ℓ is $\lambda_{\max}(\widehat{\Sigma})$–smooth.

Bernoulli distribution. $A(\xi) = \log(1+e^\xi)$ and $\ddot{A}(\xi) = e^\xi/(1+e^\xi)^2 \le 1/4$. Thus $\nabla^2 L(\theta) \preceq \frac{1}{4}\widehat{\Sigma}$, and hence ℓ is $\frac{1}{4}\lambda_{\max}(\widehat{\Sigma})$–smooth.

Poisson distribution. $A(\xi) = e^\xi$ and $\ddot{A}(\xi) = e^\xi$. For any $\theta \in \mathsf{B}_d(b)$, easily check $x_i^\top\theta \le \|x_i\|_2\|\theta\|_2 = b\|x_i\|_2$. Therefore $\nabla^2 A(X\theta) \preceq \exp(b \cdot \max_{1\le i\le n}\|x_i\|_2) I_n$, which implies that ℓ is L_{Pois}–smooth on $\mathsf{B}_d(b)$, where

$$L_{\mathrm{Pois}} = \exp(b \cdot \max_{1\le i\le n}\|x_i\|_2) \cdot \lambda_{\max}(\widehat{\Sigma}).$$

F Regarding Section 5: Model aggregation

F.1 General results: Bregman-divergence-regularization and mirror descent

Here we state general results for Bregman-divergence-penalized GLM and early-stopped mirror descent on GLM, not necessarily limited to KL-penalized GLM and early-stopped exponentiated gradient descent on GLM. In other words, the following two propositions are general versions of Proposition 5 and 7.

Proposition 11 is analogous to Proposition 1 in ridge GLM. Note that taking $\phi(\theta) = \|\theta\|_2^2/2$ and $z = 0$ indeed retrieves the conclusion of Proposition 1. Proposition 12 is analogous to Proposition 3 in gradient descent on GLM. Proposition 12 is analogous to Proposition 3 in early-stopped gradient descent on GLM.

Proposition 11 (Risk bound for Bregman-divergence-penalized GLM estimator)**.** *Assume that ϕ is α-strongly convex with respect to the norm $\|\cdot\|$ on $\mathcal{K}$ for some $\alpha > 0$. Denote $\|\cdot\|_*$ as the dual norm of $\|\cdot\|$. For any $\lambda > 0$, $z \in \mathcal{K}$, and a reference point $\theta \in \mathcal{K}$, the prediction risk of $\hat{\theta}_{\lambda,\phi,z}$ is bounded by:*

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

Proof of Proposition 11. By following the same step as the proof of Proposition 1, but using $u, v \in \mathbb{R}^d$ and $\omega \in \mathbb{R}_{\ge 0}$ such that $\ell_{\omega,\phi,z}(u) \le \ell_{\omega,\phi,z}(v)$ instead, we have

$$\frac{1}{n}\big(A(Xu) - A(Xv)\big) \le Y^\top\frac{X}{n}(u-v) + \omega\Big(D_\phi(v,z) - D_\phi(u,z)\Big)$$

and thus

$$\mathrm{Risk}(u) - \mathrm{Risk}(v) \le \Big\langle\frac{X^\top\epsilon}{n}, u - v\Big\rangle + \omega\Big(D_\phi(v,z) - D_\phi(u,z)\Big).$$

Therefore, as $\ell_{\lambda,\phi,z}(\hat{\theta}_{\lambda,\phi,z}) \le \ell_{\lambda,\phi,z}(\theta)$ for any θ due to the definition of $\hat{\theta}_{\lambda,\phi,z}$, we have

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

By the definition of dual norm and Young's inequality, we obtain that

$$\begin{aligned}
\Big\langle\frac{X^\top\epsilon}{n}, \hat{\theta}_{\lambda,\phi,z} - \theta\Big\rangle &\le \Big\|\frac{X^\top\epsilon}{n}\Big\|_*\|\hat{\theta}_{\lambda,\phi,z} - \theta\| \\
&\le \Big\|\frac{X^\top\epsilon}{n}\Big\|_*\|\hat{\theta}_{\lambda,\phi,z} - z\| + \Big\|\frac{X^\top\epsilon}{n}\Big\|_*\|\theta - z\| \\
&\le \frac{1}{2\lambda\alpha}\Big\|\frac{X^\top\epsilon}{n}\Big\|_*^2 + \frac{\lambda\alpha}{2}\|\hat{\theta}_{\lambda,\phi,z} - z\|^2 + \Big\|\frac{X^\top\epsilon}{n}\Big\|_*\|\theta - z\|,
\end{aligned}$$

Plugging this inequality to the above inequality about the risk, and also using $D_\phi(u,v) \ge (\alpha/2)\|u-v\|^2$ by α-strong convexity of ϕ, we have

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

Finally, by AM-GM inequality, we have

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

completing the proof. □

Proposition 12 (Risk bound for early-stopped mirror descent GLM estimator)**.** *Under Assumptions As2 and As3, consider mirror descent iterates (2) initialized at $z \in \mathcal{K}$ with a constant step size satisfying $\eta \in (0, \alpha/L]$. Let $\theta_T^{(\mathrm{md})}$ be the T-th iterate. Then, for any $T \in \mathbb{N}$ and $\theta \in \mathcal{K}$,*

$$\mathrm{Risk}(\theta_T^{(\mathrm{md})}) - \mathrm{Risk}(\theta) \le \frac{1}{2\lambda_T\alpha}\Big\|\frac{X^\top\epsilon}{n}\Big\|_*^2 + \lambda_T D_\phi(\theta, z). \tag{23}$$

Proof of Proposition 12. The proof is similar to that of Proposition 3. Applying the basic inequality of Theorem 2, we obtain that for any $\theta \in \mathcal{K}$:

$$\ell(\theta_T^{(\mathrm{md})}) - \ell(\theta) \le \lambda_T D_\phi(\theta, z) - \lambda_T D_\phi(\theta, \theta_T^{(\mathrm{md})}).$$

Similar to the proof of Proposition 11, we deduce that

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

completing the proof. □

F.2 Proof of Theorem 5

This is a special case of Proposition 11. Note that $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$ and KL divergence is 1-strongly convex with respect to $\|\cdot\|_1$ by Pinsker's inequality. Therefore, Proposition 11 suggests that, for any $\lambda > 0$ and any $\theta \in \Delta_d$,

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

F.3 Proof of Proposition 6

Recall that Proposition 5 gives an upper bound

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

Now consider sub-Gaussian noise $\epsilon_i \sim \mathrm{sG}(\sigma)$. Write $X = (x_{ij})_{i,j}$ and $v := X^\top\epsilon = (v_1, \ldots, v_d)^\top$. Then $v_j = \sum_{i=1}^n x_{ij}\epsilon_i$ and $\mathbb{E}[v_j] = 0$ due to the mean-zero property of ϵ_i. Check that $v_j \sim \mathrm{sG}(\sigma\|X_{\cdot j}\|_2)$ due to the following: for any $\alpha \ge 0$,

$$\mathbb{E}\Big[\exp(\alpha v_j)\Big] = \mathbb{E}\Big[\exp(\alpha\sum_{i=1}^n x_{ij}\,\epsilon_i)\Big] = \prod_{i=1}^n \mathbb{E}\Big[\exp(\alpha x_{ij}\epsilon_i)\Big] \le \prod_{i=1}^n \exp\Big(\frac{\alpha^2 x_{ij}^2}{2}\sigma^2\Big) = \exp\Big(\frac{\alpha^2}{2}\|X_{\cdot j}\|_2^2\sigma^2\Big)$$

where $X_{\cdot j}$ denotes the j-th column of X. Then, by the concentration inequality of the maximum of possibly dependent sub-Gaussian random variables, we know that

$$\mathbb{P}\Big(\|X^\top \epsilon\|_\infty = \max_{1\le j\le d} |v_j| \le \sigma \max_{1\le j\le d} \|X_{\cdot j}\|_2 \sqrt{2(\log(2d)+\delta)}\Big) \ge 1 - e^{-\delta}.$$

Since $\max_{1\le j\le d} \|X_{\cdot j}\|_2 \le \sqrt{n}$, we have

$$\mathbb{P}\left(\frac{1}{n}\|X^\top \epsilon\|_\infty \le \sigma\sqrt{\frac{2(\log(2d)+\delta)}{n}}\right) \ge 1 - e^{-\delta}. \tag{24}$$

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

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \inf_{\theta:\, D_{\mathrm{KL}}(\theta,\pi)\le b} \mathrm{Risk}(\theta) \le \frac{2\sigma^2(\log(2d)+\delta)}{n\lambda} + 2\lambda b.$$

Thus, when we choose

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

we know the following holds with probability at least $1 - e^{-\delta}$:

$$\mathrm{Risk}(\hat{\theta}_\lambda) - \inf_{\theta:\, D_{\mathrm{KL}}(\theta,\pi)\le b} \mathrm{Risk}(\theta) \le 4\sigma\sqrt{\frac{b(\log(2d)+\delta)}{n}}.$$

This concludes the proof of Proposition 6.

F.4 Proof of Theorem 5

σ_{Dist} has already been discussed in Theorem 3. Thus, the result directly comes from Proposition 6.

F.5 Proof of Proposition 7

This is a direct consequence of Proposition 12, since the KL divergence is 1-strongly convex with respect to the $\|\cdot\|_1$-norm, whose dual norm is $\|\cdot\|_\infty$.

F.6 Proof of Proposition 8

Recall that we have already assumed $\max_{1\le j\le d} \|X_{\cdot j}\|_2 \le \sqrt{n}$. Thus, we can use the high-probability upper bound for $\frac{1}{n}\|X^\top\epsilon\|_\infty$ obtained in (24). Plugging in this upper bound to Proposition 7, we have that, with probability at least $1 - e^{-\delta}$,

$$\mathrm{Risk}(\theta_T^{(\mathrm{egd})}) - \inf_{\theta:\, D_{\mathrm{KL}}(\theta,\pi)\le b} \mathrm{Risk}(\theta) \le \frac{\sigma^2(\log(2d)+\delta)}{n\lambda_T} + \lambda_T b. \tag{25}$$

Of course, if $T = 1/\lambda^*_{\mathrm{egd}}\eta$ is an integer, then $\lambda_T = \lambda^*_{\mathrm{egd}}$, and the desired excess risk bound follows. Otherwise, by taking $T = \lceil 1/(\lambda^*_{\mathrm{egd}}\eta)\rceil$ as in the proposition statement, we know that

$$\frac{1}{\lambda^*_{\mathrm{egd}}} \le \frac{1}{\lambda_T} \le \frac{1}{\lambda^*_{\mathrm{egd}}} + \eta.$$

This further implies the following bound on the discretization error

$$\begin{aligned}
&\frac{\sigma^2(\log(2d)+\delta)}{n\lambda_T} + \lambda_T b - 2\sigma\sqrt{\frac{b(\log(2d)+\delta)}{n}} \\
=& b\lambda^2_{\mathrm{egd}^*}\lambda_T\left(\frac{1}{\lambda_T} - \frac{1}{\lambda^*_{\mathrm{egd}}}\right)^2 \le b\eta^2\lambda^3_{\mathrm{egd}^*} \le b\frac{\lambda^3_{\mathrm{egd}^*}}{L^2_{\mathrm{Dist}}} \\
=& \frac{\sigma^3_{\mathrm{Dist}}}{L^2_{\mathrm{Dist}}}\cdot\frac{(\log(2d)+\delta)^{3/2}}{n^{3/2}b^{1/2}}.
\end{aligned}$$

This completes the proof of Proposition 8.

F.7 Proof of Theorem 6

Based on Proposition 8, it suffices to verify that for each of three distributions (i) to (iii), the loss function $\ell(\theta)$ is L_{Dist}-smooth with respect to $\|\cdot\|_1$ on Δ_d. Therefore, similar to the proof of Theorem 4, it suffices to verify that

$$\|\nabla^2\ell(\theta)\|_{1\to\infty} = \Big\|\frac{1}{n}\sum_{i=1}^n \ddot{A}(x_i^\top\theta)x_ix_i^\top\Big\|_{1\to\infty} \le L_{\mathrm{Dist}}, \quad \forall\theta\in\Delta_d.$$

Gaussian distribution. $\ddot{A}(\xi)=1$. Therefore,

$$\Big\|\frac{1}{n}\sum_{i=1}^n \ddot{A}(x_i^\top\theta)x_ix_i^\top\Big\|_{1\to\infty} = \|\widehat{\Sigma}\|_{1\to\infty} = \frac{1}{n}\max_{j\in[d]}\sum_{i=1}^n x_{ij}^2 = \frac{1}{n}\max_{j\in[d]}\|X_{\cdot j}\|_2^2 \le 1,$$

where the last inequality holds due to the assumption $\max_{1\le j\le d}\|X_{\cdot j}\|_2\le\sqrt{n}$.

Bernoulli distribution. $\ddot{A}(\xi)=e^\xi/(1+e^\xi)^2\in[0,1/4]$, hence

$$\Big\|\frac{1}{n}\sum_{i=1}^n \ddot{A}(x_i^\top\theta)x_ix_i^\top\Big\|_{1\to\infty} \le \frac{1}{4}\|\widehat{\Sigma}\|_{1\to\infty} = \frac{1}{4n}\max_{j\in[d]}\sum_{i=1}^n x_{ij}^2 = \frac{1}{4n}\max_{j\in[d]}\|X_{\cdot j}\|_2^2 \le \frac{1}{4}.$$

where the last inequality holds due to the assumption $\max_{1\le j\le d}\|X_{\cdot j}\|_2\le\sqrt{n}$.

Poisson distribution. $\ddot{A}(\xi)=e^\xi$. Therefore, for any $\theta\in\Delta_d$:

$$\Big\|\frac{1}{n}\sum_{i=1}^n \ddot{A}(x_i^\top\theta)x_ix_i^\top\Big\|_{1\to\infty} \le \Big\|\frac{1}{n}\sum_{i=1}^n \exp(\|x_i\|_\infty)x_ix_i^\top\Big\|_{1\to\infty} = \frac{1}{n}\max_{j\in[d]}\sum_{i=1}^n \exp(\|x_i\|_\infty)x_{ij}^2.$$

This completes the proof for (i) to (iii).

G Regarding Section 6: Random model selection

G.1 Proof of Proposition 9

By definition of $\hat{\theta}_\lambda$, for any $\theta\in\mathcal{P}(\mathcal{B})$:

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}[\widehat{R}_n(\beta)] - \mathbb{E}_{\beta\sim\theta}[\widehat{R}_n(\beta)] \le \lambda\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\hat{\theta}_\lambda,\pi)\big).$$

Meanwhile, defining a measure $\nu := \theta - \hat{\theta}_\lambda$ over $\mathcal{B}$,

$$\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}[R(\beta)] - \mathbb{E}_{\beta\sim\theta}[R(\beta)] = \int_{\mathcal{B}} -R(\beta)\nu(\mathrm{d}\beta) = \int_{\mathcal{B}}\big[\widehat{R}_n(\beta)-R(\beta)\big]\nu(\mathrm{d}\beta) + \int_{\mathcal{B}} -\widehat{R}_n(\beta)\nu(\mathrm{d}\beta).$$

Therefore, by the above inequality,

$$\begin{aligned}
\mathbb{E}_{\beta\sim\hat{\theta}_\lambda}[R(\beta)] - \mathbb{E}_{\beta\sim\theta}[R(\beta)] &\le \int_{\mathcal{B}}\big[\widehat{R}_n(\beta)-R(\beta)\big]\nu(\mathrm{d}\beta) + \lambda\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\hat{\theta},\pi)\big)\\
&\le \Big\|\widehat{R}_n - R\Big\|_{L^\infty(\mathcal{B})}\|\nu\|_{L^1(\mathcal{B})} + \lambda\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\hat{\theta}_\lambda,\pi)\big)\\
&\le \Big\|\widehat{R}_n - R\Big\|_{L^\infty(\mathcal{B})}\|\theta-\pi\|_{L^1(\mathcal{B})} + \Big\|\widehat{R}_n - R\Big\|_{L^\infty(\mathcal{B})}\|\pi-\hat{\theta}_\lambda\|_{L^1(\mathcal{B})}\\
&\quad + \lambda\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\hat{\theta}_\lambda,\pi)\big)\\
&\le \frac{1}{\lambda}\Big\|\widehat{R}_n - R\Big\|_{L^\infty(\mathcal{B})}^2 + \frac{\lambda}{2}\|\theta-\pi\|_{L^1(\mathcal{B})}^2 + \frac{\lambda}{2}\|\pi-\hat{\theta}_\lambda\|_{L^1(\mathcal{B})}^2\\
&\quad + \lambda\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\hat{\theta}_\lambda,\pi)\big)\\
&\le \frac{1}{\lambda}\Big\|\widehat{R}_n - R\Big\|_{L^\infty(\mathcal{B})}^2 + 2\lambda D_{\mathrm{KL}}(\theta,\pi),
\end{aligned}$$

where the penultimate inequality uses Young's inequality and the last inequality follows from Pinsker's inequality. This completes the proof.

G.2 Proof of Proposition 10

Recall that exponentiated gradient descent is a special case of mirror descent. Due to the basic inequality for mirror descent in Theorem 2, we have

$$\mathbb{E}_{\beta\sim\theta_T^{(\mathrm{egd})}}[\widehat{R}_n(\beta)] - \mathbb{E}_{\beta\sim\theta}[\widehat{R}_n(\beta)] \le \lambda_T\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\theta,\theta_T^{(\mathrm{egd})})\big).$$

Similar to the proof of Proposition 9, defining a measure $\nu := \theta - \theta_T^{(\mathrm{egd})}$ over $\mathcal{P}(\mathcal{B})$,

$$\mathbb{E}_{\beta\sim\theta_T^{(\mathrm{egd})}}[R(\beta)] - \mathbb{E}_{\beta\sim\theta}[R(\beta)] = \int_{\mathcal{B}} -R(\beta)\nu(\mathrm{d}\beta) = \int_{\mathcal{B}} \big[\widehat{R}_n(\beta) - R(\beta)\big]\nu(\mathrm{d}\beta) + \int_{\mathcal{B}} -\widehat{R}_n(\beta)\nu(\mathrm{d}\beta).$$

Therefore, by the above inequality,

$$\begin{aligned}
\mathbb{E}_{\beta\sim\theta_T^{(\mathrm{egd})}}[R(\beta)] - \mathbb{E}_{\beta\sim\theta}[R(\beta)] &\le \int_{\mathcal{B}} \big[\widehat{R}_n(\beta) - R(\beta)\big]\nu(\mathrm{d}\beta) + \lambda_T\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\theta,\theta_T^{(\mathrm{egd})})\big) \\
&\le \Big\|\widehat{R}_n - R\Big\|_{L^\infty(\mathcal{B})}\|\nu\|_{L^1(\mathcal{B})} + \lambda_T\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\theta,\theta_T^{(\mathrm{egd})})\big) \\
&\le \frac{1}{2\lambda_T}\Big\|\widehat{R}_n - R\Big\|^2_{L^\infty(\mathcal{B})} + \frac{\lambda_T}{2}\|\nu\|^2_{L^1(\Theta)} + \lambda_T\big(D_{\mathrm{KL}}(\theta,\pi) - D_{\mathrm{KL}}(\theta,\theta_T^{(\mathrm{egd})})\big) \\
&\le \frac{1}{2\lambda_T}\Big\|\widehat{R}_n - R\Big\|^2_{L^\infty(\mathcal{B})} + \lambda_T D_{\mathrm{KL}}(\theta,\pi),
\end{aligned}$$

which completes the proof.

H Regarding Section 7: Complements for Theorem 7 and 8

H.1 Proximal gradient descent

Lemma H1. *Consider proximal gradient descent with iterates (18). Then, for any $\theta \in \mathbb{R}^d$ and $\eta > 0$,*

$$G_\eta(\theta) \in \nabla g(\theta) + \partial h(\theta - \eta G(\theta))$$

where ∂h represents the subgradients of h.

Proof of Lemma H1. Recall the definition of the proximal operator and $G_\eta(\theta) = \frac{1}{\eta}(\theta - \mathrm{Prox}_{\eta h}(\theta - \eta\nabla g(\theta))$. By the first order optimality condition for the proximal operator,

$$0 \in \Big(\mathrm{Prox}_{\eta h}(\theta - \eta\nabla g(\theta)) - (\theta - \eta\nabla g(\theta))\Big) + \eta\partial h(\theta - \eta\nabla g(\theta)),$$

which means

$$0 \in \Big(\nabla g(\theta) - G_\eta(\theta)\Big) + \partial h(\theta - \eta\nabla g(\theta)).$$

This concludes the proof. □

Lemma H2. *For a composite function $f = g + h$ with convex differentiable g, and convex but potentially non-differentiable h, suppose one of the following holds:*

(i) g is L-smooth in a convex set $C \subseteq \mathbb{R}^d$, with step sizes $\eta_t \in (0, 1/L]$;

(ii) g is zero (i.e., $f = h$), with no constraint on $\eta > 0$.

Then, respectively, for any $z \in \mathbb{R}^d$ and

(i) for any $\theta \in C$ such that $\theta - \eta G_\eta(\theta) \in C$;

(ii) for any $\theta \in \mathbb{R}^d$,

the following holds, respectively:

$$(i): f(\theta - \eta G_\eta(\theta)) \le f(z) + \langle G_\eta(\theta), \theta - z\rangle - \frac{\eta}{2}\|G_\eta(\theta)\|_2^2,$$
$$(ii): f(\theta - \eta G_\eta(\theta)) \le f(z) + \langle G_\eta(\theta), \theta - z\rangle - \eta\|G_\eta(\theta)\|_2^2.$$

Proof of Lemma H2. (i) By the L-smoothness of g over C,

$$g(\theta - \eta G_\eta(\theta)) \le g(\theta) + \langle \nabla g(\theta), -\eta G_\eta(\theta)\rangle + \frac{L}{2}\|\eta G_\eta(\theta)\|_2^2.$$

Moreover, from the convexity of g over $\mathbb{R}^d$, we know $g(\theta) \le g(z) + \langle \nabla g(\theta), \theta - z\rangle$ for any $z \in \mathbb{R}^d$. Combining these two inequality gives that

$$\begin{aligned} g(\theta - \eta G_\eta(\theta)) &\le g(z) + \langle \nabla g(\theta), \theta - z\rangle + \langle \nabla g(\theta), -\eta G_\eta(\theta)\rangle + \frac{L}{2}\|\eta G_\eta(\theta)\|_2^2 \\ &\le g(z) + \langle \nabla g(\theta), \theta - z\rangle + \langle \nabla g(\theta), -\eta G_\eta(\theta)\rangle + \frac{\eta}{2}\|G_\eta(\theta)\|_2^2. \quad (\because \eta \le 1/L.) \end{aligned}$$

Meanwhile, from Lemma H1, we know $G_\eta(\theta) - \nabla g(\theta) \in \partial h(\theta - \eta G_\eta(\theta))$. Thus, by the definition of the subgradient,

$$\begin{aligned} h(\theta - \eta G_\eta(\theta)) &\le h(z) - \langle G_\eta(\theta) - \nabla g(\theta), z - (\theta - \eta G_\eta(\theta))\rangle \\ &= h(z) + \langle \nabla g(\theta), z - \theta + \eta G_\eta(\theta)\rangle + \langle G_\eta(\theta), \theta - z\rangle - \eta\|G_\eta(\theta)\|_2^2. \end{aligned}$$

Therefore, as $f(\theta - \eta G_\eta(\theta)) = g(\theta - \eta G_\eta(\theta)) + h(\theta - \eta G_\eta(\theta))$, combining the upper bounds on g and h gives us the following result, which concludes the proof for (i):

$$f(\theta - \eta G_\eta(\theta)) \le g(z) + h(z) + \langle G_\eta(\theta), \theta - z\rangle - \frac{\eta}{2}\|G_\eta(\theta)\|_2^2.$$

(ii) Note that $g \equiv 0$ implies that g is L-smooth over $\mathbb{R}^d$ with any $L > 0$. Recall the upper bound on $h(\theta - \eta G_\eta(\theta))$ in the proof of (i). As $g \equiv 0$, note that $\nabla g(\theta) = 0$ for any θ. Thus, the intermediate observation about $h(\theta - \eta G_\eta(\theta))$ from the proof of (i) gives that

$$h(\theta - \eta G_\eta(\theta)) \le h(z) + \langle G_\eta(\theta), \theta - z\rangle - \eta\|G_\eta(\theta)\|_2^2.$$

□

H.2 NoLips Mirror descent

Lemma H3 (Extended descent lemma for NoLips; Lemma 1 in Bauschke et al. (2017))**.** *Under Assumption As6, for any $u, v \in \mathcal{K} \cap \operatorname{int}(\Omega)$,*

$$f(u) \le f(v) + \langle \nabla f(v), u - v\rangle + LD_\phi(u, v).$$

Proof of Lemma H3. Due to the convexity of $L\phi - f$ on $\mathcal{K} \cap \operatorname{int}(\Omega)$, we know that

$$\langle L\nabla\phi(v) - \nabla f(v), u - v\rangle \le (L\phi(u) - f(u)) - (L\phi(v) - f(v)).$$

Rearranging this concludes the proof:

$$f(u) \le f(v) + \langle f(v), u - v\rangle + L\Big(\phi(u) - \phi(v) - \langle \phi(v), u - v\rangle\Big) = f(v) + \langle f(v), u - v\rangle + LD_\phi(u, v).$$

□

I Regarding Section 8: Complements for Experiments details

Lemma I1 (Regularized objective function evaluated at $\hat{\theta}_\lambda$)**.** *For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, define $\hat{\theta}_\lambda = \arg\min f(\theta) + \lambda\|\theta\|_2^2$ for $\lambda > 0$. Then a function $g : [0, \infty) \to \mathbb{R}$ where $g(\lambda) = f(\hat{\theta}_\lambda) + \lambda\|\hat{\theta}_\lambda\|_2^2$ is a non-decreasing function of $\lambda > 0$.*

Proof of Lemma I1. Choose any $\lambda_s < \lambda_b$ in $[0, \infty)$. By definition, $g(\lambda_s) = f(\hat{\theta}_{\lambda_s}) + \lambda_s\|\hat{\theta}_{\lambda_s}\|_2^2 \le f(\hat{\theta}_{\lambda_b}) + \lambda_s\|\hat{\theta}_{\lambda_b}\|_2^2$. Therefore,

$$g(\lambda_b) - g(\lambda_s) = f(\hat{\theta}_{\lambda_b}) + \lambda_b\|\hat{\theta}_{\lambda_b}\|_2^2 - g(\lambda_s) \ge (\lambda_b - \lambda_s)\|\hat{\theta}_{\lambda_b}\|_2^2 \ge 0,$$

whose equality holds if and only if $\hat{\theta}_{\lambda_b} = 0$. □

Optimization details: Implicit regularization. For iterative algorithms, learning rate schedules are used to cover small τ with high resolution and to reach large τ with less iterations at the same time. Table 2 and 3 summarize the learning rate schedules used in each combinations of GD and EGD for (i) three GLMs and (ii) underparamterized or overparametrized regime. The schedule $\{(\eta^{(i)}, T^{(i)})\}_{i=1}^k$ 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**	
	underparam. $(n, d) = (200, 20)$	overparam. $(n, d) = (100, 200)$
Linear	$\{(10^{-4}, 10^4),\ (10^{-3}, 10^5),\ (10^{-2}, 10^5)\}$	same as underparm.
Logistic	same as Linear	same as underparm.
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**	
	underparam. $(n, d) = (200, 20)$	overparam. $(n, d) = (30, 60)$
Linear	$\{(10^{-4}, 10^5), (10^{-3}, 10^5), (10^{-2}, 10^5), (10^{-1}, 10^5)]\}$	same as underparm.
Logistic	same as Linear	same as underparm.
Poisson	same as Linear	same as underparm.

Table 3: EGD learning rate schedules.

Optimization details: Explicit regularization. In both GD and EGD, for all (i) three GLMs and (ii) underparamterized or overparametrized regime, we solved 500 ridge- or KL-regularized optimization problems with different regularization parameter λ, where λ's are log-evenly spread through $[10^{-4}, 10^4]$. We used `scipy.optimize.minimize` function from `SciPy` library, where GD used `L-BFGS-B` solver and EGD used `SLSQP` solver. For the options for each solver, the GD always used (maxiter, ftol, gtol) $= (2 \times 10^4,\ 10^{-15},\ 10^{-8})$. The EGD used different options per GLM. Linear regression used (maxiter, ftol, eps) $= (4 \times 10^4,\ 2 \times 10^{-14},\ 2 \times 10^{-8})$ as a default, while used more conservative option of (maxiter, ftol, eps) $= (10^5,\ 10^{-14},\ 10^{-8})$ for $\lambda \in (10^{-2}, 10)$, as we observed the optimization does not converge in those λ. Logistic regression used (maxiter, ftol, eps) $= (4 \times 10^4,\ 10^{-13},\ 10^{-7})$. Poisson regression used (maxiter, ftol, eps) $= (6 \times 10^4,\ 10^{-14},\ 10^{-8})$.