Journal of Machine Learning Research 22 (2021) 1-62 Submitted 6/19; Revised 12/20; Published 6/21

An algorithmic view of ℓ_2 regularization and some path-following algorithms

Yunzhang Zhu zhu.219@osu.edu
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA

Renxiong Liu liu.6732@buckeyemail.osu.edu
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA

Editor: Rina Foygel Barber

Abstract

We establish an equivalence between the ℓ_2-regularized solution path for a convex loss function, and the solution of an ordinary differentiable equation (ODE). Importantly, this equivalence reveals that the solution path can be viewed as the flow of a hybrid of gradient descent and Newton method applying to the empirical loss, which is similar to a widely used optimization technique called trust region method. This provides an interesting algorithmic view of ℓ_2 regularization, and is in contrast to the conventional view that the ℓ_2 regularization solution path is similar to the gradient flow of the empirical loss. New path-following algorithms based on homotopy methods and numerical ODE solvers are proposed to numerically approximate the solution path. In particular, we consider respectively Newton method and gradient descent method as the basis algorithm for the homotopy method, and establish their approximation error rates over the solution path. Importantly, our theory suggests novel schemes to choose grid points that guarantee an arbitrarily small suboptimality for the solution path. In terms of computational cost, we prove that in order to achieve an ϵ-suboptimality for the entire solution path, the number of Newton steps required for the Newton method is $\mathcal{O}(\epsilon^{-1/2})$, while the number of gradient steps required for the gradient descent method is $\mathcal{O}\left(\epsilon^{-1}\ln(\epsilon^{-1})\right)$. Finally, we use ℓ_2-regularized logistic regression as an illustrating example to demonstrate the effectiveness of the proposed path-following algorithms.

Keywords: ℓ_2 regularization, path-following algorithms, Newton method, gradient descent method, convergence rate analysis.

1. Introduction

It is of great interest to study statistical procedures from a computational perspective. Many regularization techniques can be understood as iterative algorithmic procedures, providing an interesting algorithmic view of regularization. For instance, Friedman and Popescu (2004) studied variants of gradient descent and showed that they closely correspond to those induced by commonly used regularization methods. Building on the works by Efron et al. (2004); Hastie et al. (2007), M. Freund et al. (2017) showed that the classic boosting

©2021 Yunzhang Zhu and Renxiong Liu.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v22/19-477.html`.

algorithm in linear regression can be viewed as the iterates generated by applying subgradient descent algorithm to the loss function defined as the maximum absolute correlation between the features and residuals.

Tikhonov (or ℓ_2) regularization (Tikhonov and Arsenin, 1977) is ubiquitously used in many modeling procedures, and in the statistical literature it traces back to Hoerl and Kennard (1970), where it is often referred to as ridge regression. It is natural to seek an algorithmic view of ℓ_2 regularization, that is, what algorithm would produce a sequence of iterates that is identical to the ℓ_2-regularized solutions. Surprisingly, this has not been formally established for a general convex loss function, with most related works focusing on least squares error loss. For example, Fleming (1990) showed an equivalence between ℓ_2 regularization and the iterates of certain optimization algorithms. More recently, Suggala et al. (2018) formally studied the connection between ℓ_2 regularization path and the iterates of gradient descent algorithm. They established a pointwise bound between these two paths and use this to establish the risk bound of the iterates of gradient descent algorithm. Neu and Rosasco (2018) proposed a weighting scheme for gradient descent iterates so that it is exactly equal to some ℓ_2-regularized solution. Another related work is by Ali et al. (2019), which compares the risk of gradient flow to that of ℓ_2-regularized solutions in the context of least squares regression.

Another line of work has focused on the similarity between algorithmic approaches and explicit regularization approaches in terms of their statistical performance. Earlier work includes (Frank and Friedman, 1993), who pointed out a similarity between ridge regression and partial least squares regression, where the latter has been shown to be equivalent to conjugate gradient descent with squared-error loss (Wold et al., 1984). More recently, Yao et al. (2007) considered nonparametric regression in a reproducing kernel Hilbert space (RKHS) and provided some theoretical justification for early stopping of gradient descent algorithm. Raskutti et al. (2014) proposed a data-dependent and easily computable stopping rule for gradient descent, and showed that it can achieve similar risk bounds as that of the ridge regression.

In this article, we establish an algorithmic view of ridge regression for a general convex loss function. We first establish an equivalence between ℓ_2-regularized solution path for a convex loss function, and the solution of an ODE. This reveals a formal equivalence between ℓ_2 regularization solution path and the iterates produced by a hybrid of gradient descent and Newton algorithm when the step size tends to 0. This equivalence has been previously discovered by Suggala et al. (2018) (see proof of Theorem 1 in Suggala et al. (2018)). However, a rigorous proof was not provided by Suggala et al. (2018).

More formally, denote by $L_n(\theta)$ some convex empirical loss function, where $\theta \in \mathbb{R}^p$ is the parameter. Let $C(\cdot)$ be a differentiable increasing function with $C(0) = 0$ and $\lim_{t\to\infty} C(t) = \infty$. We consider the solution path of an ℓ_2-regularized convex optimization problem:

$$\theta(t) = \arg\min_{\theta\in\mathbb{R}^p} \left(C(t) \cdot L_n(\theta) + \frac{1}{2}\|\theta\|_2^2 \right) . \tag{1}$$

Note that as t varies from 0 to ∞, the solution $\theta(t)$ varies from $\mathbf{0}$ to a minimizer of $L_n(\theta)$. Throughout this article, we focus on ℓ_2 regularization, although some of the results in this article can be easily extended to the case where the regularizer is a general quadratic function.

Our first main result is that under some smoothness condition on $L_n(\theta)$, the solution path defined by (1) coincides with the *global solution* to the following *ordinary differential equation (ODE)*,

$$\theta'(t) = -C'(t)\left[C(t) \cdot \nabla^2 L_n(\theta(t)) + I\right]^{-1} \nabla L_n(\theta(t)), t \geq 0\,, \tag{2}$$

with an initial condition $\theta(0) = \mathbf{0}$. More specifically, under the assumption that $L_n(\theta)$ is convex and has continuous Hessian, we show that $\theta(t)$ is differentiable in t and the solution to the above ODE is also a solution path to the original optimization problem (1).

To better interpret the ODE formulation in (2), we consider a special choice of $C(t) = \exp(t)-1$ throughout this article. In fact, based on the aforementioned equivalence, it is easy to see that the choice of $C(t)$ is not essential, because the solution to the ODE can always be viewed as the solution path to (1) regardless of the choice of $C(t)$. In another words, different choices of $C(t)$ produce the same path $\theta(t)$ in $\mathbb{R}^p$—they just “travel” at different speeds to the minimum ℓ_2 norm minimizer of $L_n(\theta)$ as t varies from 0 to ∞. Another rationale for choosing $C(t) = \exp(t) - 1$ is that this agrees with the common practice of picking grid points on a log scale for tuning parameter selection.

Plugging $C(t) = \exp(t) - 1$ into (2), the ODE becomes

$$\theta'(t) = -\left[(1 - e^{-t}) \cdot \nabla^2 L_n(\theta(t)) + e^{-t} \cdot I\right]^{-1} \nabla L_n(\theta(t)), t \geq 0 \text{ with } \theta(0) = \mathbf{0}\,. \tag{3}$$

The left hand side $\theta'(t)$ can be viewed as the local direction of the solution path at time t. Interestingly, we can see from (3) that the search direction can be thought of as certain hybrid of gradient descent search direction $-\nabla L_n(\theta(t))$ and Newton direction $-\left[\nabla^2 L_n(\theta(t))\right]^{-1} \nabla L_n(\theta(t))$. Moreover, the search direction is closer to gradient search direction when t is small, and closer to Newton direction when t is large. This provides an interesting algorithmic perspective of ℓ_2 regularization, and partially confirms previous belief that the ℓ_2 regularization path is closely related to the solution path generated by the gradient descent method. In particular, the ODE update direction (3) resembles to that of the trust region algorithm or its precursor the Levenberg–Marquardt algorithm (Levenberg, 1944; Conn et al., 2000). Both algorithms produce similar types of hybrid of gradient descent and Newton direction:

$$\theta_{k+1} = \theta_k - \left[\nabla^2 L_n(\theta_k) + \lambda I\right]^{-1} \nabla L_n(\theta_k)\,, \tag{4}$$

where $\lambda > 0$ is often adaptively chosen or determined by the size of the trust region. Although these optimization algorithms have very similar update directions, they are designed with the goal of finding a minimizer of the unregularized loss $L_n(\theta)$ reliably and efficiently. By contrast, our focus here is to provide an algorithmic interpretation of the ℓ_2-regularized solution path, and to design numerical procedures to approximate the entire solution path for the regularized problem. A more detailed discussion of this connection is provided in Section 3.

Aside from providing a conceptual connection between the ℓ_2-regularized solution path and solutions to an ODE, the ODE formulation also opens up avenues for designing algorithms to approximate the entire ℓ_2-regularized solution path or the minimum norm minimizer of $L_n(\theta)$. In particular, the ODE formulation (2) is known as the *initial-value* *problems* in the numerical ODE literature (see, e.g., Butcher, 2016). Many effective numerical ODE solvers such as the Euler's method and Runge-Kutta method (see Chapter 2 and 3 of Butcher, 2016) can be used to approximately solve the ODE over a discrete set of grid points.

In addition to ODE solvers, we also propose two new path-following (homotopy) methods based on Newton method and gradient descent method as their "working horse" algorithms to approximate the solution path $\theta(t)$ over a given region $[0, t_{\max})$, where $0 < t_{\max} \leq \infty$. An approximate solution path $\tilde{\theta}(t)$ is constructed through linearly interpolating the approximate solutions at the selected grid points (see Section 3.1 for its formal definition). Theoretically, we bound the global approximation error of the entire solution path in terms of $\sup_{0\leq t\leq t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\}$ for both the Newton method and gradient descent method (c.f. Theorem 4 and 8), where $f_t(\theta) = (1-e^{-t})L_n(\theta) + (e^{-t}/2)\cdot\|\theta\|_2^2$ is a scaled version of the regularized objective function. These bounds reveal an important interplay between the choice of grid points and accuracy of the solutions at the selected grid points. In particular, they allow us to design novel schemes to select grid points $t_1, \ldots, t_N$ so that the overall computations required to achieve a prespecified suboptimality is minimized.

Using the newly proposed grid point selection schemes, we further derive upper bounds on the total number of steps required to achieve an ϵ-suboptimality, i.e.,

$$\sup_{0\leq t\leq t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \lesssim \epsilon\,, \tag{5}$$

where $\epsilon > 0$. In particular, the number of Newton steps and gradient steps required to achieve ϵ suboptimality are at most $\mathcal{O}(\epsilon^{-1/2})$ and $\mathcal{O}\left(\epsilon^{-1}\ln(\epsilon^{-1})\right)$, respectively. To the best of our knowledge, these complexity results are new, and parallel to existing complexity results for the Newton method and gradient descent method when applied to solving a single optimization problem (i.e., the problem corresponding to $t = t_{\max}$). Moreover, the new complexity results also suggest that Newton method, being more expensive at each iteration, requires less number of iterations as compared with the gradient descent method. Numerical experiments on a ℓ_2-regularized logistic regression corroborate with the theoretical results in that the Newton method tends to perform better than the gradient descent method for small to medium scale problems, while the gradient descent method is more efficient for large-scale problems.

In optimization, homotopy techniques have been used in many algorithms including the interior point algorithm (Nesterov and Nemirovskii, 1993). For example, the solution to a constrained convex optimization problem can be viewed as the limit of the solutions to a family of unconstrained optimization by introducing a barrier (or penalty) function. However, the focus of these methods is the recovery of the limit of the path, rather than the entire solution path. That said, the idea of the warm-start strategy has been well developed, which consists of the so-called "working horse" algorithm and the policy for updating the penalty parameter (see, e.g., Chapter 1.3 of Nesterov and Nemirovskii, 1993). Typically, Newton method is used as the "working horse" for the modern path-following interior point methods. In statistical learning literature, Osborne (1992) and Osborne et al. (2000) applied the homotopy technique to generate piecewise linear trajectories in quantile regression and LASSO, respectively. Later Efron et al. (2004), Hastie et al. (2004), and Rosset and Zhu (2007) exploited the homotopy path-following methods to generate an entire solution path for a family of regularization problems. Subsequent developments include Friedman et al. (2007); Hoefling (2010); Arnold and Tibshirani (2016), among others. These works often leverage the piecewise linearity of the solution path so that an exact path-following algorithm can be explicitly derived. For situations where the solution paths are not piecewise linear, approaches based on ODE solvers were considered in Wu (2011); Zhou and Wu (2014) and a path-following algorithm based on Newton method was considered in Rosset (2004). In particular, Rosset (2004) also proposed to use one-step Newton update to generate the solution path, and is the most relevant to our work. However, it used a constant step size scheme and only established the pointwise closeness to the solution path. To the best of our knowledge, our work is the first to theoretically analyze the global approximation error of the entire solution path.

To summarize, our key contributions are that

- we provide an algorithmic view of ℓ_2 regularization through establishing a formal equivalence to the solution of an ODE, which further reveals an interesting connection to the trust region algorithm and Levenberg-Marquardt algorithm;
- we propose two path-following algorithms based on Newton update and gradient descent update, and establish global approximation-error bounds for the solution paths generated by both algorithms;
- we also consider various numerical ODE solvers to approximate the ℓ_2-regularized solution path.

The rest of the paper is organized as follows. Section 2 discusses the properties of the solution path, and provides a proof of the equivalence to the ODE solution. Section 3 introduces the linear interpolation scheme and discusses various approaches to approximate the regularized solution path. In Section 4, global approximation-error bounds for two path-following algorithms are established. New grid point selection schemes and the associated computational complexities are derived. In Section 5, we compare the proposed methods with some competing methods through a simulated study using ℓ_2-regularized logistic regression. We close with some remarks in Section 6.

2. Properties of the solution path

In this section, we first start with an informal derivation of the ODE (2) using the optimality condition of the ℓ_2-regularized solutions. We then rigorously establish the differentiability of $\theta(t)$, which turns out to be the key ingredient in establishing the equivalence between (1) and (2). Note that the optimality condition of (1) at time t is

$$C(t)\nabla L_n(\theta(t)) + \theta(t) = 0\,. \tag{6}$$

If we assume for now that $\theta(t)$ is differentiable in t, by taking derivative with respect to t, we obtain that

$$C'(t)\nabla L_n(\theta(t)) + C(t)\nabla^2 L_n(\theta(t))\theta'(t) + \theta'(t) = 0\,,$$

which implies that

$$\theta'(t) = -C'(t)\left(C(t)\nabla^2 L_n(\theta(t)) + I\right)^{-1}\nabla L_n(\theta(t))\,.$$

It is easy to see that $\theta(0) = \mathbf{0}$. Thus, it follows that the ℓ_2-regularized solution path must be a solution to the ODE (2).

Next, we make the above argument rigorous. The missing piece of the above argument is the differentiability of the solution path $\theta(t)$ in t. To formally establish this, we impose convexity and smoothness conditions on $L_n(\theta)$, and show that the solution path $\theta(t)$ is differentiable in t. The assumptions needed on the loss function $L_n(\theta)$ are described below in Assumption (A0).

Assumption (A0). Suppose that $L_n(\theta)$ is convex and has continuous second derivative, with $\mathbf{0} \in \mathbf{dom}\, L_n$, where $\mathbf{dom}\, L_n$ denotes the domain of $L_n(\theta)$.

Theorem 1 *Assume that $C(t)$ is a strictly increasing and differentiable function, with $C(0) = 0$ and $\lim_{t\to\infty} C(t) = \infty$. Under Assumption (A0), the solution path $\theta(t)$ defined by* (1) *is differentiable in t, and it is the unique solution to the ODE* (2).

Some remarks are in order. The above result has been informally presented in Suggala et al. (2018) without rigorously proving the differentiability of the regularization path $\theta(t)$. The connection to the ODE suggests that the choice of $C(t)$ is not essential. If we choose $C(t) = \exp(t) - 1$, then it follows from Theorem 1 that the search direction at time t is

$$-\left((1 - e^{-t})\nabla^2 L_n(\theta(t)) + e^{-t} I\right)^{-1} \nabla L_n(\theta(t)),$$

which is similar to the search direction of Levenberg-Marquardt algorithm (Levenberg, 1944). Interestingly, the direction is a hybrid of Newton and gradient direction, and it is close to gradient direction when t is small and close to the Newton direction when t is large.

It is also worth pointing out that Efron et al. (2004) shows that ℓ_1-regularized solution path is piecewise linear. By contrast, here we show that the ℓ_2-regularized solution path is more smooth in the sense that it is differentiable everywhere. Moreover, without the smoothness assumption, the differentiability of the solution path can not be established in general. Examples include ℓ_2-regularized quantile regression and support vector machine, both of which have nonsmooth loss functions and their solution paths were shown to be nondifferentiable in t by Osborne (1992) and Hastie et al. (2004), respectively. In this sense, the smoothness assumption for the loss function is necessary.

Next we present some properties of the ℓ_2-regularized solution path, which may be interesting on their own. In particular, it shows that the ℓ_2 norm of the solutions $\|\theta(t)\|_2$ is nondecreasing, while $\|\theta(t)\|_2 / C(t)$ is nonincreasing. Moreover, it is shown that the solution $\theta(t)$ converges to the minimum ℓ_2 norm minimizer of $L_n(\theta)$ as t goes to infinity if it is finite.

Corollary 1 *Assume that $C(t)$ is a strictly increasing and differentiable function, with $C(0) = 0$ and $\lim_{t\to\infty} C(t) = \infty$. Then*

(i) $\|\theta(t)\|_2$ is nondecreasing in t and $L_n(\theta(t))$ is nonincreasing in t;

(ii) $\|\theta(t)\|_2 / C(t)$ is nonincreasing in t;

(iii) if $L_n(\theta)$ is a continuous, closed proper convex function and the minimum ℓ_2 norm minimizer of $L_n(\theta)$, denoted as $\theta^\star$, is finite, then $\lim_{t\to\infty} \theta(t) = \theta^\star$.

We remark that the convergence of $\theta(t)$ to the minimum ℓ_2 norm minimizer has already been established in Theorem 8 of Suggala et al. (2018). Moreover, the monotonicity property of the solution path and the loss function is also probably well-known as folklore. We include them here to make the paper largely self-contained. Also, as pointed in Suggala et al. (2018), results of similar flavor have also been obtained recently for various types of optimization algorithms (see, e.g., Soudry et al., 2017; Gunasekar et al., 2017, 2018). Moreover, it is noted that the smoothness assumption on the loss function is not necessary for establishing monotonicity or convergence to the minimum ℓ_2 norm solution. As such, this result is applicable to nonsmooth loss functions such as support vector machine and quantile regression.

Both Theorem 1 and Corollary 1 can be extended to handle general quadratic regularizers. More specifically, it can be shown that Theorem 1 and Corollary 1 continue to hold if a general quadratic regularization function $\frac{1}{2}(\theta-\theta_0)^\top Q(\theta-\theta_0)$ is used, where Q is a positive definite matrix and θ_0 is some starting point. The corresponding ODE becomes

$$\theta'(t) = -C'(t)\left[C(t)\cdot\nabla^2 L_n(\theta(t)) + Q\right]^{-1}\nabla L_n(\theta(t)), t \geq 0 \text{ with } \theta(0) = \theta_0 \,.$$

For Corollary 1, the limit of $\theta(t)$ would be the minimizer of $L_n(\theta)$ that is closest to θ_0 with distance induced by $\|\cdot\|_Q$-norm.

3. Approximation of the solution path

When the solution path $\theta(t)$ is not piecewise linear, typically only an approximate solution path can be obtained. There are in general two types of approaches to obtain an approximate solution path. One is based on the idea of homotopy method (Osborne, 1992; Nesterov and Nemirovskii, 1993; Rosset, 2004), and the other one is based on numerical ODE methods (see, e.g., Wu, 2011; Zhou and Wu, 2014). In this section, we study these two types of approximation schemes. Specifically, for homotopy methods, we use Newton update and gradient descent update as the basis, and derive the corresponding path-following algorithms. We also consider numerical ODE solvers based on the explicit forward Euler method and the (second-order) Runge-Kutta method (Butcher, 2016).

Note that although the focus of typical homotopy algorithms is to find a single solution at the limit, here we use the idea of homotopy algorithm with the goal of approximating the entire solution path (see, e.g., Friedman et al., 2007) through linear interpolation. More specifically, given the approximate solutions $\{\theta_k\}_{k=1}^N$ at a set of prespecified grid points $0 < t_1 < \cdots < t_N < \infty$, we propose an approximate solution path through linearly interpolating these solutions. This produces a continuous approximate solution path for $\theta(t)$. Throughout this section, we assume that $C(t) = \exp(t) - 1$ and consider (3) instead of (2), because they generate the same solution path.

3.1 Approximate solution path through linear interpolation

Suppose that the goal is to approximate the solution path $\theta(t)$ over a given interval $[0, t_{\max})$ for some $t_{\max} \in (0, \infty]$, where we allow $t_{\max} = \infty$. Given a set of grid points $0 < t_1 < \cdots < t_N < \infty$, and the approximate solutions $\{\theta_k\}_{k=1}^N$ at these grid points, a natural way to produce an approximate solution path over $[0, t_{\max})$ is by linear interpolation. In particular, we define a piecewise linear function $\tilde{\theta}(t)$ as the approximate solution path through linearly interpolating the solutions at each grid point:

$$\tilde{\theta}(t) = \frac{t_{k+1} - t}{t_{k+1} - t_k}\theta_k + \frac{t - t_k}{t_{k+1} - t_k}\theta_{k+1} \text{ for any } t \in [t_k, t_{k+1}], k = 0, 1, \ldots, N-1\,,$$
$$\tilde{\theta}(t) = \theta_N \text{ for any } t_N < t \leq t_{\max} \text{ if } t_N < t_{\max}\,,$$

where $t_0 = 0$ and $\theta_0 = \mathbf{0}$. This defines an approximate solution path $\tilde{\theta}(t)$ for any $t \in [0, t_{\max})$. In view of this definition, we may also assume that $t_{N-1} \leq t_{\max}$, because we do not need $\tilde{\theta}(t)$ over $t \in [t_{N-1}, t_N]$ if $t_{N-1} > t_{\max}$. We also remark that the above interpolation scheme allows two possible approaches to approximating the solution path around $t_{\max}$. The first approach is to specify all grid points from $[0, t_{\max})$ and use a constant path $\tilde{\theta}(t) = \theta_N$ to approximate $\theta(t)$ when $t_N < t \leq t_{\max}$. The other approach is to allow $t_N > t_{\max}$ but $t_{N-1} < t_{\max}$ when $t_{\max} < \infty$, and use a linear interpolation of θ_{N-1} and θ_N to approximate $\theta(t)$ when $t_{N-1} < t < t_{\max}$.

To obtain the approximate solution path $\tilde{\theta}(t)$ as constructed above, one also needs to choose the grid points $t_1, \ldots, t_N$ and a numerical algorithm to generate the approximate solutions at these grid points, both of which will likely have an impact on how well the solution path approximates the true path $\theta(t)$. For the rest of this section, we first discuss some path following algorithms that can produce solutions at a given set of grid points. Given a path following algorithm, the issue of how to optimally design its grid points to minimize the overall computations will be investigated later in Section 4.

3.2 Path following algorithm: Newton

In this subsection, we propose a path following algorithm based on Newton update over a set of grid points. A special version of this algorithm was considered in Rosset (2004) with $C(t) = 1/t$. The Newton method is constructed based on taking one-step Newton steps at each grid point to obtain an approximate solution at the next grid point. More specifically, we consider an one-step Newton update at t_{k+1} using θ_k as the initial solution, which can be shown to have the following form

$$\theta_{k+1} = \theta_k - \left((1 - e^{-t_{k+1}})\nabla^2 L_n(\theta_k) + e^{-t_{k+1}} I\right)^{-1} \left((1 - e^{-t_{k+1}})\nabla L_n(\theta_k) + e^{-t_{k+1}}\theta_k\right)\,. \quad (7)$$

To facilitate a comparison to the update of the Euler's method to be presented later in (11), we present an alternative updating formula. Let $g_k = (1 - e^{-t_k})\nabla L_n(\theta_k) + e^{-t_k}\theta_k$ denote the scaled gradient at θ_k. By substituting θ_k with $e^{t_k} g_k - (e^{t_k} - 1)\nabla L_n(\theta_k)$ in (7), we obtain an alternative expression for the Newton update,

$$\theta_{k+1} = \theta_k - \left((1 - e^{-t_{k+1}})\nabla^2 L_n(\theta_k) + e^{-t_{k+1}} I\right)^{-1} \left((1 - e^{-\alpha_{k+1}})\nabla L_n(\theta_k) + e^{-\alpha_{k+1}} g_k\right)\,, \quad (8)$$

where $\alpha_{k+1} = t_{k+1} - t_k$. It will be shown later that the iterates generated by the Newton method are all "close" to the true solution path $\theta(t)$ in some sense. (c.f. Theorem 5). Moreover, it will be theoretically justified later that only one Newton step is needed at each grid point as the overall approximation error would not improve further if more Newton steps are taken (c.f. Theorem 4). We also establish that the linearly interpolated solution path based on Newton algorithm can achieve ϵ-suboptimality after taking at most $\mathcal{O}(\epsilon^{-1/2})$ Newton iterations (c.f. Theorem 6).

3.3 Path following algorithm: gradient descent

In this subsection, we consider the gradient descent algorithm as the basis algorithm for the path following scheme. More specifically, at time t_{k+1}, we perform n_{k+1} gradient descent steps to minimize $f_{t_{k+1}}(\theta) = (1 - e^{-t_{k+1}})L_n(\theta) + (e^{-t_{k+1}}/2) \cdot \|\theta\|_2^2$ starting from θ_k. The update can be written down explicitly as

$$\theta_{k+1} = \circ^{n_{k+1}}(I - \eta_{k+1}\nabla f_{t_{k+1}})\theta_k \,, \tag{9}$$

where η_{k+1} is the gradient step size chosen at step $k+1$, and $\circ^l h$ denotes l function compositions of h. In practice, a varying gradient step size can be implemented using a line search. As suggested by subsequent theoretical analysis (see Theorem 7), multiple gradient descent steps are needed to ensure a small approximation error and convergence. This is in contrast to the Newton method, for which one step is sufficient to achieve good approximation and convergence. Moreover, the search direction at each step can be thought of as a "damped" gradient descent search direction as we have that

$$\nabla f_{t_k}(\theta) = (1 - e^{-t_k})\nabla L_n(\theta) + e^{-t_k}\theta \,, \tag{10}$$

which becomes closer and closer to the gradient search direction $\nabla L_n(\theta)$ as t_k increases.

In practice, the gradient descent method has the advantage that it is typically cheaper to compute as compared to the Newton method, although multiple steps need to be taken in order for it to enjoy a good approximation-error bound (c.f. Theorem 7 and 8). We also establish that the linearly interpolated solution path using the gradient descent iterates can achieve ϵ-suboptimality after taking at most $\mathcal{O}(\epsilon^{-1}\ln(\epsilon^{-1}))$ gradient descent iterations (c.f. Theorem 10).

3.4 Numerical ODE methods

In view of Theorem 1, the solution path of (1) is also the unique solution of the ODE (3). Hence, any numerical methods that approximately solve (3) with initial condition $\theta(0) = \mathbf{0}$ would also produce an approximate solution path for (1). In the numerical ODE literature, the ODE (3) is often referred to as the *initial value problem* and standard solvers are available to find an approximate solution. In this subsection, we consider two popular approaches: the explicit forward Euler method and the second-order Runge-Kutta method (Butcher, 2016).

The explicit forward Euler method leads to the following updating scheme:

$$\theta_{k+1} = \theta_k - \alpha_{k+1}\left[(1 - e^{-t_k}) \cdot \nabla^2 L_n(\theta_k) + e^{-t_k} I\right]^{-1} \nabla L_n(\theta_k) \,. \tag{11}$$

Note that if we choose a constant step size $\alpha_k = \alpha$, then $t_k = k\alpha$. Again this update is similar to the Levenberg–Marquardt algorithm. The difference here is that the iterates are close to the true path as $\alpha \to 0$, while in the Levenberg-Marquardt algorithm, the goal is to recover the unregularized solution as $k \to \infty$. Euler's method has been known to have bad approximation error, and is referred to as first-order method as the approximation error $\|\theta_k - \theta(t_k)\|$ is typically of order $\mathcal{O}(\alpha)$ when $\alpha_k = \alpha$ for all $k \geq 1$.

Higher order approximation can be achieved using more sophisticated approximation schemes. Runge-Kutta method is such a scheme whose global approximation error is

$\|\theta_k - \theta(t_k)\| = \mathcal{O}(\alpha^m)$ with $m \geq 2$ when $\alpha_k = \alpha$ for all $k \geq 1$ (see Chapter 3 of Butcher, 2016). Although it can achieve higher-order approximation accuracy compared to the Euler's method, it does require higher computational cost at each step. For example, the second-order Runge-Kutta method considers the following update

$$\theta_{k+1} = \theta_k + \frac{\alpha_{k+1}}{2}\left(J(\theta_k, t_k) + J\left(\theta_k + \alpha_{k+1} J(\theta_k, t_k), t_{k+1}\right)\right), \tag{12}$$

where $J(\theta, t) = -\left((1-e^{-t})\nabla^2 L_n(\theta(t)) + e^{-t} I\right)^{-1} \nabla L_n(\theta(t))$. It can be immediately seen that, compared to the Euler's method and the Newton method, it requires solving two linear systems as opposed to just one for the Euler method and Newton method. Therefore, there is an apparent trade-off between approximation error and per-iteration cost here. Another popular choice is the fourth-order Runge-Kutta method, which achieves a fourth-order approximation accuracy, but again requires solving four linear systems at each iteration. Empirically, it will be demonstrated in Section 5 that the first-order ODE method generally performs much worse than the Newton method, while the second-order ODE method performs slightly worse than the Newton method.

3.5 Discussion and connections

The two types of updates are derived from two different perspectives. The numerical ODE approach tries to approximate the solutions to the corresponding ODE, while the homotopy methods are based on applying path-following optimization algorithms with warm-start. Moreover, it is worth pointing out that the updating formulas of the Euler's method and Newton method, are very similar. In fact, the only difference is the presence of an extra gradient term in the Newton update (8). If we ignore the gradient term in the Newton update (8), we have that

$$\begin{aligned}
\theta_{k+1} &= \theta_k - (1-e^{-\alpha_{k+1}})\left((1-e^{-t_{k+1}})\nabla^2 L_n(\theta_k) + e^{-t_{k+1}} I\right)^{-1} \nabla L_n(\theta_k) \\
&\approx \theta_k - \alpha_{k+1}\left[(1-e^{-t_k}) \cdot \nabla^2 L_n(\theta_k) + e^{-t_k} I\right]^{-1} \nabla L_n(\theta_k),
\end{aligned}$$

where the right hand side is the Euler's update (11). In practice, however, we will show that the Newton method work much better than the Euler method in terms of approximation accuracy.

In terms of computational cost and ease of implementation, the gradient descent update has the smallest per-iteration cost, but it requires running more steps at each grid point, especially when t_k is large (c.f. Theorem 9). By contrast, the Newton method and the ODE solver have higher per-iteration cost, but only requires one update at each grid point. We also remark that other optimization algorithms could also be used in the path following algorithm. For example, glmnet (Friedman et al., 2010) uses coordinate descent algorithm in the path following algorithm to get an approximate solution path. Other viable choices include accelerated gradient descent or conjugate gradient descent algorithm. Hybrid approaches that mix two types of algorithms can also be considered. We shall investigate these alternative approaches in the future.

4. Solution path approximation-error bounds

In this section, we derive approximation-error bounds for the solution path over $[0, t_{\max})$ generated by the Newton method and gradient descent method. The bounds for the ODE solvers have been extensively studied in the numerical ODE literature, but are less satisfactory in that most results are proved for generic ODE problems. We present one such version in Appendix B.

We aim to bound the function-value suboptimality of an approximate solution path $\tilde{\theta}(t)$ measured by $\sup_{0\le t\le t_{\max}}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$, where $f_t(\theta) := (1-e^{-t})L_n(\theta) + (e^{-t}/2)\|\theta\|_2^2$ is a scaled version of the objective function. Given the definition of $\theta(t)$, this is a natural performance metric that captures the accuracy of the approximate solution path. In what follows, we call $\sup_{0\le t\le t_{\max}}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$ the global approximation error for $\tilde{\theta}(t)$. Our analysis proceeds in two steps: (i) we first relate the global approximation error to approximation errors at the selected grid points measured by the size of the gradients $\|g_k\|_2$, where $g_k := \nabla f_{t_k}(\theta_k) = (1-e^{-t_k})\nabla L_n(\theta_k) + e^{-t_k}\theta_k$; (ii) we then bound $\|g_k\|_2$ for the Newton method and gradient descent method proposed in Section 3.

For step (i), we have the following result.

Theorem 2 *For any $0 < t_1 < t_2 < \cdots < t_N < \infty$, we have that*

$$\sup_{t\in[0,t_1]}\left\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\right\} \le \max\left(e^{t_1}\|g_1\|_2^2,\, \|\theta_1\|_2^2\right) + \frac{e^{t_1}(1-e^{-t_1})^2}{2}\|\nabla L_n(\mathbf{0})\|_2^2\,, \tag{13}$$

$$\begin{aligned}\sup_{t\in[t_k,t_{k+1}]}\left\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\right\} &\le e^{t_{k+1}}\max\left\{\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2\|g_k\|_2^2,\, \|g_{k+1}\|_2^2\right\} \\ &\quad + (e^{-t_k} - e^{-t_{k+1}})^2\max\left\{\frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2},\, \frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\right\}\end{aligned} \tag{14}$$

for any $k = 1, \dots, N-1$. If we further assume that $\|\theta(t_{\max})\|_2 < \infty$, then we have that

$$\sup_{t_N<t\le t_{\max}}\left\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\right\} \le \frac{e^{t_N}(1-e^{-t_{\max}})}{1-e^{-t_N}}\|g_N\|_2^2 + \frac{3}{2(e^{t_N}-1)}\|\theta(t_{\max})\|_2^2 \tag{15}$$

when $t_N < t_{\max}$.

We can see that the upper bounds consist of two parts, with the first part (depending on g_k) being algorithm-specific and the other part stemming from interpolation over the selected grid points. We call them *optimization error* and *interpolation error*, respectively. Note that the optimization error depends on the size of the gradient at time t_k and is roughly of order $e^{t_k}\|g_k\|_2^2$, while the interpolation error is essentially independent of the choice of optimization algorithm as it only depends on how finely we choose the grid points and the norm of the solutions along the solution path (typically $\|\theta_k\|_2 = \mathcal{O}(\|\theta(t_k)\|_2)$, c.f., Lemma 2). In other words, given a specific set of grid points, the interpolation error is irreducible for any optimization algorithms. The optimization error, however, does depend on the optimization algorithms, and can be pushed to be arbitrarily small if we run the algorithm long enough at the selected grid points. In this sense, if the goal is to approximate the solution path, then both the grid points and the optimization algorithm should be designed carefully to strike a balance between these two types of errors to save the overall computation. For instance, it would be wasteful to have the optimization error much smaller than the interpolation error, because the additional computations would not improve the overall approximation error in terms of order.

We next derive bounds on $\|g_k\|_2$ for the Newton method and gradient descent method to obtain an overall approximation-error bound for $\sup_{0\le t\le t_{\max}}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$. Using the bounds on $\|g_k\|_2$, we then investigate how many Newton steps or gradient steps are needed so that the optimization error can be dominated by the interpolation error. Moreover, novel grid point schemes will be constructed to control both the overall approximation error and amount of computation. This allows us to derive upper bounds on the total number of iterations to achieve a prespecified suboptimality over the entire path for both methods.

4.1 Newton method

In this subsection, we show that by taking only one Newton step at each grid point, the optimization error is comparable to the interpolation error, under some conditions on the grid points. To show this, we first bound $\|g_k\|_2$ for the Newton method. The following local Lipschitz Hessian condition on $L_n(\theta)$ is assumed.

Assumption (A1). Assume that $L_n(\theta)$ is a proper, closed, convex function, and there exists constants $\beta > 0$, $0 \le \gamma_1 < 2$, and $0 \le \gamma_2 < 2$ such that the second-order derivative of $L_n(\theta)$ exists and satisfies a local Lipschitz condition

$$\|\nabla L_n(\theta+\delta) - \nabla L_n(\theta) - \nabla^2 L_n(\theta)\delta\|_2 \le \beta\delta^\top \left[\nabla^2 L_n(\theta)\right]^{\gamma_1}\delta \tag{16}$$

for any $\theta \in \mathbf{dom}\, L_n$ and δ satisfying $\theta+\delta \in \mathbf{dom}\, L_n$ and

$$\delta^\top[\nabla^2 L_n(\theta)]^{\gamma_2}\delta \le \beta^{-2}\,. \tag{17}$$

Assumption (A1) can be thought of as a local version of Lipschitz Hessian condition, and is similar to the (generalized) self-concordant condition imposed for the convergence analysis of second-order method (see, e.g., Nesterov and Nemirovskii, 1993; Sun and Tran-Dinh, 2017). This avoids making the assumption that $L_n(\theta)$ is strongly convex. It will later be verified that many commonly used loss functions satisfy Assumption (A1) (see Table 1). The following result provides bound on $\|g_k\|_2$ under Assumption (A1) and some conditions on the step sizes $\{\alpha_k\}_{k=1}^\infty$.

Theorem 3 *Suppose that Assumption (A1) holds for some constants $\beta > 0$, $0 \le \gamma_1 < 2$, and $0 \le \gamma_2 < 2$. We further assume that the step sizes $\{\alpha_k\}_{k=1}^\infty$ satisfy*

$$C_1\beta(e^{\alpha_1}-1)^{\min(2-\gamma_1,1-\gamma_2/2)}\|\nabla L_n(\mathbf{0})\|_2 \le 1,\ \alpha_k < \ln(2),\ 2^{-1}\alpha_k \le \alpha_{k+1} \le 2\alpha_k; \tag{18a}$$

$$C_2\beta e^{t_{k+1}}\max\left((e^{t_{k+1}}-1)^{-\gamma_1}, (e^{t_{k+1}}-1)^{-1-\gamma_2/2}\right)(e^{\alpha_{k+1}}-1)\|\theta_k\|_2 \le 1 \tag{18b}$$

for any $k \ge 1$, where $C_1 = 15\mathbb{I}(\gamma_1 \le 1) + 15\min\left(\nu^{\gamma_1-1}, \nu^{\gamma_1}e^{-\alpha_1}(1-e^{-\alpha_1})\right)\mathbb{I}(\gamma_1 > 1)$, $C_2 = 442$, and ν denotes the maximum eigenvalue of $\nabla^2 L_n(\mathbf{0})$. Then, the scaled gradients g_k evaluated at the iterates θ_k generated by the Newton method in (7) satisfy

$$\|g_k\|_2 \le \frac{\|\theta(t_k)\|_2}{2(e^{t_k}-1)}(1-e^{-\alpha_k}) \text{ for every } k \ge 1\,. \tag{19}$$

Some remarks are in order. First, fixing t_k, the upper bound for $\|g_k\|_2$ decreases as the step size α_k decreases. In other words, smaller step size generally leads to a small upper bound. Moreover, the first term in the upper bound decreases as k increases, because $\|\theta(t_k)\|_2/(e^{t_k}-1)$ is a nonincreasing function of t_k (c.f. part (ii) of Corollary 1). Second, the existence of step sizes that satisfy (18) is not obvious. A novel step size scheme will be proposed later so that it satisfies (18) and at the same time leads to fast exploration of the solution path. Finally, we remark that the dependence of C_1 on the largest eigenvalue of $\nabla^2 L_n(\mathbf{0})$ is to ensure that the bound (19) holds for $\|g_1\|_2$, and such dependence can be eliminated if multiple Newton steps are taken at $t = t_1$ to ensure (19) for $\|g_1\|_2$.

To facilitate a comparison to the theoretical analysis of Rosset (2004) and second-order Runge-Kutta method, an alternative bound on $\|g_k\|_2$ is presented below, which can be derived using some partial results obtained in the proof of Theorem 3.

Corollary 2 *Under the assumptions in Theorem 3, we have that the gradients g_k evaluated at the iterates θ_k generated by the Newton method in (7) satisfy*

$$\|g_1\|_2 \le \frac{C_1}{15}\beta\|\nabla L_n(\mathbf{0})\|_2^2(e^{\alpha_1}-1)^{\max(2,3-\gamma_1)}\,, \tag{20a}$$

$$\|g_k\|_2 \le 30\beta\left(\|\theta(t_{k-1})\|_2 + \|\nabla L_n(\mathbf{0})\|_2\right)^2 \frac{e^{-\gamma_1 t_k}(e^{\alpha_k}-1)^2}{(1-e^{-t_k})^{\gamma_1-1}} \tag{20b}$$

for any $k \ge 2$.

The above corollary can be viewed as an extension of Theorem 1 in Rosset (2004), which established that $\|\theta_k - \theta(t_k)\|_2 \lesssim \alpha^2$ when $t_k = t_0 + k\alpha$ is equally spaced over a bounded interval $[t_0, t_{\max}]$ with $C(t) = 1/t$. In particular, we can see from (20) that $\|g_k\|_2 \lesssim \alpha_k^2$ when $\gamma_1 \le 1$ or $t_k = \mathcal{O}(1)$. In other words, when $\gamma_1 \le 1$, we have $\|g_k\|_2 \lesssim \alpha_k^2$ for all $k \ge 1$; and when $\gamma_1 > 1$, we have $\|g_k\|_2 \lesssim \alpha_k^2$ when t_k is large enough. This suggests that the precision at the selected grid points for the Newton method is often of order $\mathcal{O}(\alpha_k^2)$. This rate is comparable to that derived in Rosset (2004) and that of the second-order Runge-Kutta method (see Chapter 3 of Butcher, 2016) if a constant step size scheme is taken $\alpha_k = \alpha$.

Combining the bounds for $\|g_k\|_2$ in Theorem 3 with Theorem 2, we show that for the Newton method, the optimization error is comparable to the interpolation error. Moreover, we can also obtain an approximation-error bound for the Newton solution path in terms of function-value suboptimality. This is summarized below.

Theorem 4 *Under the assumptions in Theorem 2 and 3, we have that the approximate solution path $\tilde{\theta}(t)$ generated by the Newton method satisfies*

$$\sup_{0\le t\le t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \le 8\max\left\{ (e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2,\ \max_{1\le k\le N} e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2 \|\theta_k\|_2^2 \right\} \tag{21}$$

when $t_{N-1} \le t_{\max} < t_N$ *for some* $N \ge 1$*; and*

$$\sup_{0\le t\le t_{\max}} \left\{ f_t(\tilde\theta(t)) - f_t(\theta(t)) \right\} \le \max \Bigg\{ 8(e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2,$$
$$8 \max_{1\le k\le N-1} e^{-t_k} \left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2 \|\theta_k\|_2^2\,, \frac{2\max(\|\theta(t_{\max})\|_2^2, \|\theta_N\|_2^2)}{(e^{t_N}-1)} \Bigg\} \tag{22}$$

when $0 < t_N < t_{\max}$ *for some* $N \ge 2$ *and* $\|\theta(t_{\max})\|_2 < \infty$*.*

In the proof of the above theorem, it is shown that taking just one Newton step at each grid point can ensure that the optimization error is comparable to the interpolation error. Specifically, it is shown in the proof of Theorem 4 that for all $k \ge 1$,

$$\underbrace{e^{t_{k+1}} \max\left\{ \left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2 \|g_k\|_2^2,\, \|g_{k+1}\|_2^2 \right\}}_{\text{optimization error}}$$
$$\lesssim \underbrace{(e^{-t_k} - e^{-t_{k+1}})^2 \max\left\{ \frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2}, \frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2} \right\}}_{\text{interpolation error}}, \tag{23}$$

where the LHS is the optimization error and the RHS is the interpolation error in the bounds in Theorem 2. In this sense, it is wasteful to take more than one Newton step at each grid point.

Another important consequence of Theorem 3 and 4 is that a principled scheme of choosing the step sizes (or equivalently the grid points) can be designed to ensure any prespecified level of suboptimality while minimizing the overall computations. More specifically, for any $\epsilon > 0$ and $t_{\max} > 0$, suppose that our goal is to design a step size scheme that satisfies all the conditions in (18) and at the same time ensures that $\sup_{0\le t\le t_{\max}} \left\{ f_t(\tilde\theta(t)) - f_t(\theta(t)) \right\} \lesssim \epsilon$. In view of (22) in Theorem 4, this amounts to running the Newton method by choosing a sequence of step sizes $\{\alpha_k\}$ satisfying both (18) and

$$(e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2 \lesssim \epsilon\,, e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2 \|\theta_k\|_2^2 \lesssim (e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2\,, \tag{24}$$

and terminating the Newton method at $k+1 = N$ when

$$t_N > t_{\max} \text{ or } \frac{2\|\theta_N\|_2^2}{(e^{t_N}-1)} \le (e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2\,. \tag{25}$$

If such a sequence of step sizes $\{\alpha_k\}$ exists, then by its construction and (22) of Theorem 4, we have that

$$\sup_{0\le t\le t_{\max}} \left\{ f_t(\tilde\theta(t)) - f_t(\theta(t)) \right\} \lesssim (e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2 \lesssim \epsilon\,. \tag{26}$$

Therefore, it remains to prove the existence of such a sequence satisfying all the conditions in (18) and (24) for any $\epsilon > 0$, and that the Newton method must terminate within finite steps. This is shown in the theorem below.

Theorem 5 *Suppose that $\|\theta(t_{\max})\|_2 < \infty$ with $t_{\max} \in (0, \infty]$. For any $\epsilon > 0$ and $0 < \alpha_{\max} \leq 10^{-1}$, using the step sizes defined below*

$$\begin{aligned}\alpha_1 \leq \min\Bigg\{&\alpha_{\max}, \ln\left(1 + \frac{\sqrt{\epsilon}}{\|\nabla L_n(\mathbf{0})\|_2}\right), \\ &\ln\left(1 + (\max(C_1, \sqrt{2}C_2)\beta\|\nabla L_n(\mathbf{0})\|_2)^{-\min(2-\gamma_1, 1-\gamma_2/2)}\right)\Bigg\} \text{ and} \end{aligned} \tag{27}$$

$$\begin{aligned}\alpha_{k+1} = \min\Bigg\{&\alpha_{\max}, 2\alpha_k, \ln\left(1 + \frac{e^{t_k/2}(e^{\alpha_1} - 1)\|\nabla L_n(\mathbf{0})\|_2(1 - e^{-t_k})}{\|\theta_k\|_2}\right), \\ &\ln\left(1 + \left(C_2\beta e^{t_k}\max\left((e^{t_k} - 1)^{-\gamma_1}, (e^{t_k} - 1)^{-1-\gamma_2/2}\right)\|\theta_k\|_2\right)^{-1}\right)\Bigg\}; k \geq 1\end{aligned} \tag{28}$$

and the termination criterion in (25)*, the Newton method terminates after a finite number of iterations, and when terminated, the generated solution path $\tilde{\theta}(t)$ satisfies*

$$\sup_{0 \leq t \leq t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \lesssim \epsilon\,. \tag{29}$$

This result confirms the existence of a step size sequence that ensures any prespecified suboptimality for the solution path generated by the Newton method. The step size choices in (27) and (28) are motivated by (24). Moreover, as we can see from (27) and (29), the suboptimality ϵ is controlled by the initial step size α_1. Indeed, for small enough ϵ, we can see that $\epsilon = (e^{\alpha_1} - 1)^2\|\nabla L_n(\mathbf{0})\|_2^2$, which implies that

$$\sup_{0 \leq t \leq t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \lesssim (e^{\alpha_1} - 1)^2\|\nabla L_n(\mathbf{0})\|_2^2\,. \tag{30}$$

Importantly, the above result suggests that even when $t_{\max} = \infty$, we can achieve arbitrarily small suboptimality for the entire path using a finite number of grid points. To the best of our knowledge, this type of theoretical analysis is new in the literature for path following algorithms.

We next investigate how fast the Newton method explores the solution path by deriving its computational complexity. As we can see from the step size scheme in Theorem 5, both the value of γ_1 and the speed that $\|\theta(t)\|_2$ grows as a function of t will likely have a big impact on how aggressively we can choose the step sizes α_k. In particular, if $\gamma_1 \geq 1$ and $\|\theta_k\|_2$ is bounded (e.g., when $\theta^\star$ is finite), then the last term in the min function of (28) is at least of order $\mathcal{O}(1)$, while the third term in the min function is increasing. Therefore, aggressive step sizes can be taken in this case until it reaches $\mathcal{O}(1)$, which will likely lead to a fast exploration of the solution path. On the other hand, if $\gamma_1 < 1$ or $\|\theta_k\|_2$ grows quite quickly to infinity as k increases, then the last term in the min function goes to zero as $k \to \infty$. This means that the step sizes need to decrease to zero eventually, leading to a slower exploration of the solution path. The following result gives an upper bound on the number of Newton steps needed for the Newton method when $\gamma_1 \geq 1$ and $\|\theta(t_{\max})\|_2$ is treated as a finite constant.

Theorem 6 *Suppose that $\|\theta(t_{\max})\|_2 < \infty$ and $\gamma_1 \geq 1$. For any $\epsilon > 0$, using the step sizes defined in* (27) *and* (28)*, and the termination criterion in* (25)*, the number of Newton steps required to achieve ϵ-suboptimality* (29) *is at most $\mathcal{O}(\epsilon^{-1/2})$.*

We remark the above result holds even when $t_{\max} = \infty$. Moreover, it is rather difficult to theoretically bound the number of Newton steps when $\gamma_1 < 1$. Fortunately, for many commonly used loss functions, Assumption (A1) holds with some $\gamma_1 \geq 1$, as demonstrated by the following proposition.

Proposition 1 *The Assumption (A1) holds for*

- *log-barrier function $L_n(\theta) = -\ln(\theta)$ with $\gamma_1 = \frac{3}{2}$ and $\gamma_2 = 1$;*
- *entropy-barrier function $L_n(\theta) = \theta \ln(\theta) - \ln(\theta)$ with $\gamma_1 = \frac{3}{2}$ and $\gamma_2 = 1$;*
- *logistic function $L_n(\theta) = \ln(1 + e^{-\theta})$ with $\gamma_1 = 1$ and $\gamma_2 = 0$;*
- *exponential function $L_n(\theta) = e^{-\theta}$ with $\gamma_1 = 1$ and $\gamma_2 = 0$.*
- *square function $L_n(\theta) = \theta^2$ with any $\gamma_1 \in [0, 2)$ and $\gamma_2 \in [0, 2)$.*

We summarize these results in Table 1. A detailed proof is provided in the Appendix.

Function	form of $L_n(\theta)$	$\mathbf{dom}(L_n)$	γ_1	γ_2	Application
Log-barrier	$-\ln(\theta)$	$\mathbb{R}^{++}$	$\frac{3}{2}$	1	Poisson regression
Entropy-barrier	$\theta \ln(\theta) - \ln(\theta)$	$\mathbb{R}^{++}$	$\frac{3}{2}$	1	Interior-point
Logistic	$\ln(1 + e^{-\theta})$	$\mathbb{R}$	1	0	Logistic regression
Exponential	$e^{-\theta}$	$\mathbb{R}$	1	0	Boosting
Square	θ^2	$\mathbb{R}$	[0,2)	[0,2)	Least square regression

Table 1: Some commonly used loss functions that satisfy Assumption (A1).

As such, Theorem 6 applies to all losses listed in Table 1 since Assumption (A1) is satisfied with $\gamma_1 \geq 1$ for all losses. Thus the total number of Newton steps required to ensure (29) is at most $\mathcal{O}(\epsilon^{-1/2})$ for these loss functions.

Finally, we note that our theoretical results for the Newton method are widely applicable to a large class of functions. It can even include loss functions that are not self-concordant, which is a typical condition imposed to establish complexity bound for the classical Newton method (see, e.g., Nesterov and Nemirovskii, 1993) without making strong convexity assumptions. For example, among the losses in Table 1, the logistic regression loss function $\log(1+e^{-\theta})$ and the exponential loss $e^{-\theta}$ are not self-concordant. Indeed, a separate rate of convergence analysis is needed for the Newton method when applied to logistic regression problems (see, e.g., Bach et al., 2010). On the other hand, the generality of our analysis likely will lead to conservative rates and step size choices for problems with better conditioning. For instance, we expect that some of the above results can be improved and a better step size scheme can be constructed if we assume that the loss function $L_n(\theta)$ is strongly convex or "locally" strongly convex along the solution path $\theta(t)$. Due to space limit, we leave this for future investigation.

4.2 Gradient descent method

We next bound $\|g_k\|_2$ for the gradient descent method proposed in Section 3. We then use the bound to derive conditions on the number of gradient steps needed to ensure that the optimization error is comparable to the interpolation error. For gradient descent method, we impose the following Lipschitz gradient assumption on $L_n(\theta)$.

Assumption (A2). Assume that $L_n(\theta)$ has L-Lipschitz continuous gradient:

$$\|\nabla L_n(\theta_1) - \nabla L_n(\theta_2)\|_2 \le L\|\theta_1 - \theta_2\|_2\,. \tag{31}$$

Theorem 7 *Let $m_k = m(1-e^{-t_k}) + e^{-t_k}$, $L_k = L(1-e^{-t_k}) + e^{-t_k}$, where $m \ge 0$ is the strong convexity parameter for $L_n(\theta)$. Under Assumption (A2) and the condition that*

$$n_1 \ge \frac{\log(10m_1)}{-\log\left(1-\frac{2m_1L_1}{m_1+L_1}\eta_1\right)},\; n_{k+1} \ge \frac{\log(24)}{-\log\left(1-\frac{2m_{k+1}L_{k+1}}{m_{k+1}+L_{k+1}}\eta_{k+1}\right)} \text{ and } \eta_k \le \frac{2}{m_k+L_k} \tag{32}$$

for any $k \ge 1$, the iterates generated by the gradient descent method (defined by (9)) satisfies

$$\|g_k\|_2 \le \frac{2L_k}{m_{k-1}}\left(1-\frac{2m_kL_k}{m_k+L_k}\eta_k\right)^{n_k}\frac{(e^{\alpha_k}-1)\|\theta(t_k)\|_2}{(e^{t_k}-1)} \tag{33}$$

for any $k \ge 1$ and step sizes $\alpha_k \le \ln(2)$ satisfying $2^{-1}\alpha_k \le \alpha_{k+1} \le 2\alpha_k; k \ge 1$.

As we can see from the condition on n_k in (32), the number of gradient steps needed at each grid point is likely to be more than one to ensure (33). This is in contrast to the Newton method, for which only one Newton step is taken at each iteration. It will be shown later that taking multiple gradient steps is necessary to ensure that the optimization error is comparable to the interpolation error. Moreover, we can see that when $m > 0$, that is, when $L_n(\theta)$ is m-strongly convex with $m > 0$, then the lower bound on n_k behaves like a constant. When $m = 0$, however, then $m_k = e^{-t_k}$ and the number of gradient steps n_k scales as $\mathcal{O}\left(e^{t_k}\right)$ in the worst case, suggesting that the number of gradient steps needed should increase as k increases.

Interestingly, unlike the Newton method, the optimization error bound for gradient descent method may not be dominated by the interpolation error. In order for the optimization error to be comparable to the interpolation error, more gradient steps need to be taken beyond what is required in (32). The following theorem derives conditions on n_k under which the the optimization error is dominated by the interpolation error, and establishes an approximation-error bound for the solution path generated by the gradient descent method building on Theorem 2 and 7.

Theorem 8 *Under the assumptions in Theorem 7 with (32) replaced by*

$$n_1 \ge \frac{\log(10m_1L_1)}{-\log\left(1-\frac{2m_1L_1}{m_1+L_1}\eta_1\right)},\; n_{k+1} \ge \frac{\log(24)+\max(0,\log(L_{k+1}/m_k))}{-\log\left(1-\frac{2m_{k+1}L_{k+1}}{m_{k+1}+L_{k+1}}\eta_{k+1}\right)},\; \eta_k \le \frac{2}{m_k+L_k} \tag{34}$$

for $k \geq 1$, the approximate solution path $\tilde{\theta}(t)$ generated by the gradient descent method satisfies

$$\sup_{0\leq t\leq t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\}$$
$$\leq 2\max\left((e^{\alpha_1}-1)^2 \|\nabla L_n(\mathbf{0})\|_2^2 \,, \max_{1\leq k\leq N-1} e^{-t_k} \left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}} \right)^2 \|\theta_k\|_2^2 \right). \quad (35)$$

when $t_N \geq t_{\max}$ for some N; and

$$\sup_{0\leq t\leq t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \leq 2\max\Bigg((e^{\alpha_1}-1)^2 \|\nabla L_n(\mathbf{0})\|_2^2 \,,$$
$$\max_{1\leq k\leq N-1} e^{-t_k} \left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}} \right)^2 \|\theta_k\|_2^2 \,, \frac{\|\theta(t_{\max})\|_2^2}{(e^{t_N}-1)} \Bigg) \quad (36)$$

when $t_N \leq t_{\max}$ and $\|\theta(t_{\max})\|_2 < \infty$.

Compared to Theorem 7, the conditions (34) on n_k in the above theorem are stronger than the conditions (32), which is to ensure that the optimization error is dominated by the interpolation error. Moreover, it is unnecessary to run more than those required by the conditions in (34), as taking beyond this many gradient steps would not improve the overall approximation error for the entire path (at least in terms of order).

Similar to the Newton method, a novel step size scheme can be designed to ensure that the approximation error is small for all $k \geq 1$. In particular, we choose

$$\alpha_1 \leq \min\left\{ \alpha_{\max},\, \ln\left(1 + \frac{\epsilon^{1/2}}{\|\nabla L_n(\mathbf{0})\|_2}\right) \right\} \text{ and} \quad (37)$$
$$\alpha_{k+1} = \min\left\{ \alpha_{\max}, 2\alpha_k,\, \ln\left(1 + \frac{\epsilon^{1/2} e^{t_k/2}(1-e^{-t_k})}{\|\theta_k\|_2}\right) \right\}; k \geq 1\,, \quad (38)$$

where $\alpha_{\max} = \ln(2)$, and terminate the algorithm at $k+1 = N$ when

$$t_N > t_{\max} \text{ or } \frac{2\|\theta_N\|_2^2}{(e^{t_N}-1)} \leq \epsilon\,. \quad (39)$$

Similar to the Newton method, we show that the solution path generated by the gradient descent method using the above step size scheme and termination criterion achieves ϵ-suboptimality (up to a multiplicative constant). This is summarized in the following theorem.

Theorem 9 *Suppose that $\|\theta(t_{\max})\|_2 < \infty$ with $t_{\max} \in (0, \infty]$, and (34) in Theorem 8 is satisfied. For any $\epsilon > 0$, using the step sizes and the termination criterion specified above in (37), (38), and (39), the gradient descent method terminates after a finite number of iterations, and when terminated, the generated solution path $\tilde{\theta}(t)$ satisfies*

$$\sup_{0\leq t\leq t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \lesssim \epsilon\,. \quad (40)$$

Next, we derive the computational complexity of the gradient descent method. To make it directly comparable to the Newton method, we consider the case $m = 0$. In this case, in order for the optimization error to be comparable to the interpolation error, n_{k+1} must satisfy (34), which can be shown to be equivalent to $n_{k+1} \geq \mathcal{O}(e^{t_k}(t_k+1))$. Building on this, an upper bound on the number of gradient steps needed can be derived when $\|\theta(t_{\max})\|_2$ is treated as a finite constant. This is summarized in the following theorem.

Theorem 10 *Suppose that $\|\theta(t_{\max})\|_2 < \infty$, and the assumptions in Theorem 8 are met with $m = 0$. Using $\eta_k = \mathcal{O}(\min(1, L^{-1}))$, the step sizes defined in* (37) *and* (38)*, and the termination criterion in* (39)*, the number of gradient steps required to achieve ϵ-suboptimality* (40) *is at most $\mathcal{O}\left(\epsilon^{-1}\ln(\epsilon^{-1})\right)$.*

Compared with the Newton method that requires $\mathcal{O}(\epsilon^{-1/2})$ number of Newton steps, gradient descent method requires substantially more updates. Of course, since the per-iteration cost of the gradient descent method is much lower than that of Newton method, an overall computational-complexity comparison depends on how problem dimension scales with suboptimality ϵ. In general, we expect that the Newton method may be more suitable for small to medium scale problems or when a small suboptimality is desired, whereas gradient descent method may be more suitable for large scale problems with medium accuracy. This will also be confirmed through some numerical experiments in Section 5. As a side remark, a hybrid approach combining the gradient descent method and the Newton method is likely to work better than either one. Due to space limit, we choose to investigate this strategy in the future.

Moreover, for the unregularized problem, it is well-known that the number of gradient steps required for the regular gradient descent method to achieve an ϵ-suboptimality (i.e., $L_n(\theta_k) - L_n(\theta^\star) < \epsilon$) is $\mathcal{O}(\epsilon^{-1})$ when $m = 0$. In view of this and the above result, one can essentially claim that for the gradient descent method starting from $\theta_0 = \mathbf{0}$, computing the entire solution path for the ℓ_2-regularized problem requires roughly the same amount of computation as compared to computing a single unregularized solution (up to a logarithm term $\ln(\epsilon^{-1})$).

The implementation of the gradient descent method requires the specification of n_k and η_k, both of which depend on unknown problem-specific parameters m and L (see (34)). In practice, we implement the gradient method using a backtracking line search (Boyd and Vandenberghe, 2004) and terminates the gradient descent method at t_k when

$$\|g_k\|_2 \leq \frac{e^{\alpha_k} - 1}{C_0(e^{t_k} - 1)} \|\theta_k\|_2 \,, \tag{41}$$

for some absolute constant C_0. In the proof of Theorem 8, it is shown that if n_k and η_k satisfy the conditions in (34), then (41) holds for $C_0 = 12$. Here if we use (41) directly as a termination criterion for the gradient descent method at t_k, we can still establish the approximation-error bound in Theorem 8 and 9.

Corollary 3 *Suppose that $\|\theta(t_{\max})\|_2 < \infty$ with $t_{\max} \in (0, \infty]$. Moreover, we assume that at each t_k, we run the gradient descent method with backtracking line search until* (41) *is satisfied for some absolute constant C_0. Then for any $\epsilon > 0$, using the step sizes and the termination criterion specified in* (37), (38), *and* (39)*, the gradient descent method* *terminates after a finite number of iterations, and when terminated, the generated solution path* $\tilde{\theta}(t)$ *satisfies*

$$\sup_{0 \le t \le t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \lesssim \epsilon \,. \tag{42}$$

Again, the advantage of using the backtracking line search and the termination criterion (41) for the gradient descent is that it avoids having to specify n_k and η_k, both of which may depend on unknown problem-specific parameters m and L.

5. Numerical studies

In this section, we use ℓ_2-regularized logistic regression as an illustrating example to study the operating characteristics of the various proposed methods. Let $X = (X_1, \ldots, X_n)^\top$ and $Y = (Y_1, \ldots, Y_n)^\top$ denote the design matrix and the binary response vector, where $X_i \in \mathbb{R}^p$ and $Y_i \in \{+1, -1\}$; $i = 1, \ldots, n$. The empirical loss function for logistic regression is

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-Y_i X_i^\top \theta}) \,. \tag{43}$$

We first verify that the above loss function satisfies Assumption (A1).

Proposition 2 *The logistic regression loss function* $L_n(\theta)$ *defined in* (43) *satisfies Assumption (A1) with* $\gamma_1 = 1, \gamma_2 = 0$, *and* $\beta = 2 \max_{1 \le i \le n} \|X_i\|_2$.

In view of the above results, Theorem 6 can be applied to logistic regression if $\|\theta(t_{\max})\|_2 < \infty$. We note that for logistic regression, the MLE could be at the "infinity" i.e., $\|\theta^\star\|_2 = \infty$, when the two classes are separable (see, e.g., Geyer, 2009).

In our numerical experiments, we consider six methods: Euler method, second-order Runge-Kutta method, Newton method, the method proposed by Rosset (2004), gradient descent method, and glmnet (Friedman et al., 2010). The first four methods are "second-order" algorithms in the sense that they all involve solving linear systems. Gradient descent method only requires gradient evaluations, and glmnet uses warm start strategies and cyclical coordinate descent method to compute an approximate solution path. We implement all methods in R using Rcpp (Eddelbuettel et al., 2011; Eddelbuettel, 2013), except for glmnet for which we use the R package glmnet. We remark that the method of Rosset (2004) is also a path-following algorithm based on Newton updates. Compared with our proposed Newton method, it considers equally-spaced grid points using $C(t) = 1/t$ and starts with an initial solution $\theta(t_{\max})$ at $t_{\max}$. As will be demonstrated later, this makes it less efficient compared with the proposed Newton method. Finally, we point out that the proposed Newton method and gradient descent method can be applied to the case $t_{\max} = \infty$ for the nonseparable case, while all the other four methods can be only applied to the case $t_{\max} < \infty$. Throughout, we use $t_{\max} = 10$ in all of the numerical experiments. Increasing $t_{\max}$ further will make the proposed methods even more competitive in the comparisons.

We first compare all methods in terms of runtime and suboptimality. Two scenarios will be considered depending on whether the two classes are separable or not. For the non-separable case, we sample the components of the response vector $Y \in \mathbb{R}^n$ from a Bernoulli distribution, where $\mathbb{P}(Y_i = +1) = 1/2$ and $\mathbb{P}(Y_i = -1) = 1/2$ for $i = 1, 2, \ldots, n$. Conditioned on Y_i, we generate X_i's independently from $N_p(Y_i\mu, \sigma^2 I_{p\times p})$, where $\mu \in \mathbb{R}^p$ and $\sigma^2 > 0$. Note that μ and σ^2 controls the Bayes risk, which is $\Phi(-\|\mu\|_2/\sigma)$ under the 0/1 loss, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Here we choose $\mu = (1/\sqrt{p}, \ldots, 1/\sqrt{p})$ and $\sigma^2 = 1$ so that the Bayes risk is $\Phi(-1) \approx 0.15$. For the separable case, we generate X_i's independently from $N_p(Y_i\mu, I_{p\times p})$ where $\mu = (1/\sqrt{p}, \ldots, 1/\sqrt{p})$ until $Y_i\mu^\top X_i > 1$, which makes the two classes linearly separable. In fact, the two classes can be separated by the hyperplane $\mu^\top X_i = 0$. For both scenarios, three choices of problem dimensions are considered: $(n, p) = (1000, 500)$, $(n, p) = (1000, 1000)$, and $(n, p) = (1000, 2000)$.

To assess the accuracy for the approximate solution path $\tilde{\theta}(t)$ generated by each method, we use the global approximation error $\sup_{0\le t\le t_{\max}}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$, where $\tilde{\theta}(t)$ is the linear interpolation of the iterates θ_k generated by each method. To approximate the global approximation error, we sample N points $s_1, \ldots, s_N$ uniformly from $(0, t_{\max})$ and use $\max_{1\le i\le N}\{f_{s_i}(\tilde{\theta}(s_i)) - f_{s_i}(\theta(s_i))\}$ as an approximation of $\sup_{0\le t\le t_{\max}}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$. Here the exact solutions $\theta(s_i)$ at s_i's are calculated using the CVX solver (Grant and Boyd, 2014, 2008). In all simulations, we use $N = 100$.

We first compare the four "second-order" methods: Newton, Euler, Runge-Kutta, and the method of Rosset (2004) as they all involve solving linear systems. In order to make a fair comparison among these four methods, we design our experiments so that their runtime are about the same. This can be achieved by controlling the step sizes in these methods to ensure that they all take the same number of Newton steps. Specifically, for any particular choice of initial step size, we first run the proposed Newton method, record the number of Newton steps taken (denoted as N_{Newton}), and define $\alpha = t_{\max}/N_{\text{Newton}}$. Then, for the Euler method and the second-order Runge-Kutta method, we use a constant step scheme with $\alpha_k = \alpha$ and $\alpha_k = 2\alpha$. For the method of Rosset (2004), we choose the N_{Newton} grid points equally spaced with $C(t) = 1/t$. This is to ensure that all four methods have identical computational complexity. We also consider two initial step sizes: $\alpha_1 = 0.01, 0.1$ for the Newton method to see the impact of α_1 on the suboptimality.

Figure 1 and 2 present the global approximation errors (on a log scale) of the aforementioned four second-order methods for nonseparable and separable cases, respectively. Note that two initial step sizes $\alpha_1 = .01, .01$ are used for the proposed Newton method, and the other methods use the corresponding initial step sizes so that the overall computations are comparable to that of the Newton method. Among the four methods, the proposed Newton method performs the best, followed by the second-order Runge-Kutta method, the Euler method, and the method of Rosset (2004). The method of Rosset (2004) is much worse compared to other methods due to the way it chooses the grid points.

Next we compare the Newton method and gradient descent method against glmnet in terms of both runtime and approximation error. In this case, it is difficult to control the initial step sizes so that they have similar runtime. As such, we choose to look at the trade-off curve of runtime and approximation error for these three methods. Figure 3 presents plots of runtime versus approximation error based on 100 simulations, as we vary the initial step size for each method. We can see from Figure 3 that the proposed Newton method runs the fastest when the desired suboptimality is small (high precision), especially when the problem dimension is small. Also, as expected, the gradient method runs the slowest when the desired suboptimality is small. Interestingly, the glmnet performs better than the gradient descent method in most cases, but worse than the Newton method when the desired

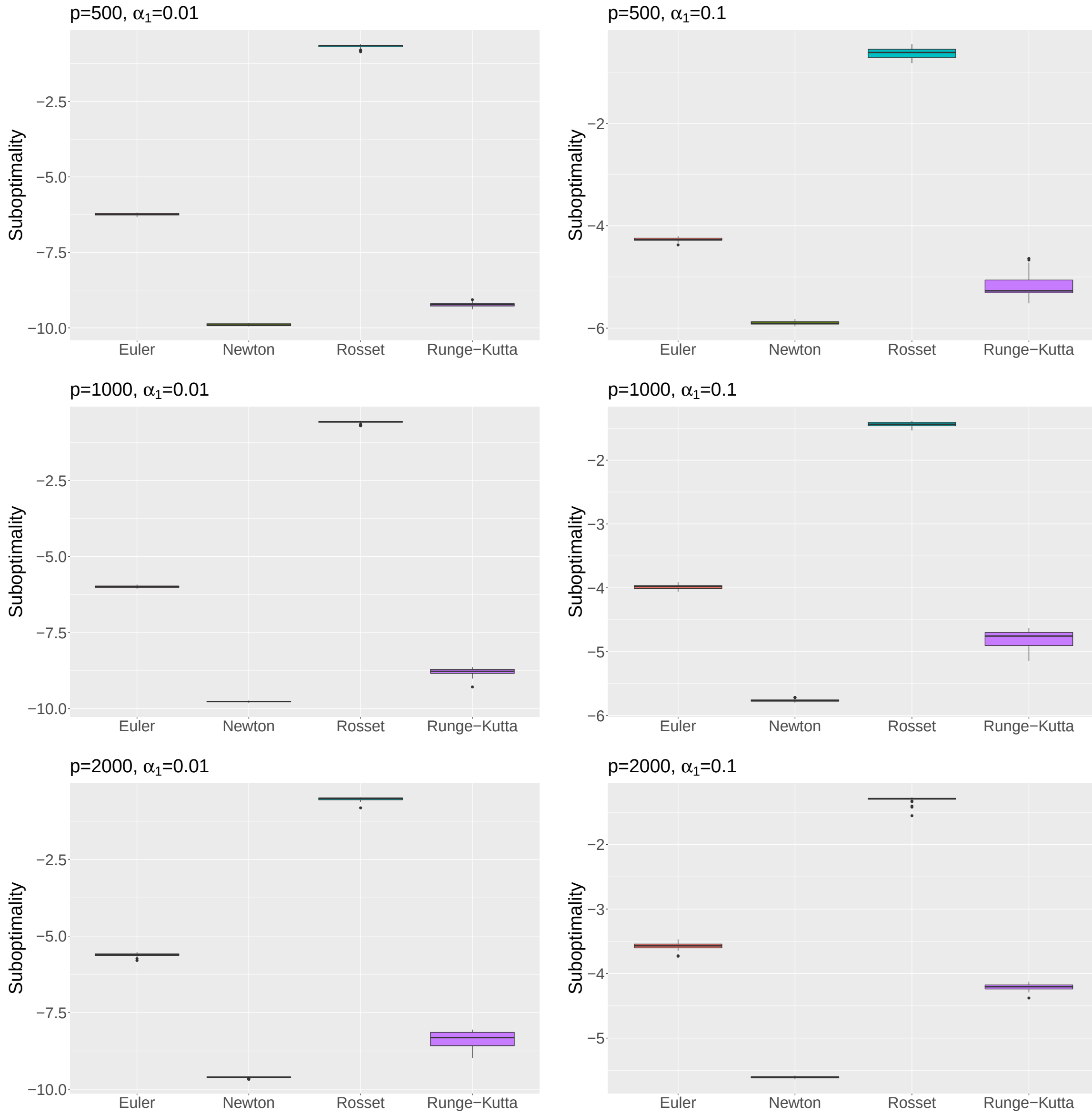

Figure 1: Suboptimalities $\sup_{0 \le t \le 10}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$ (in log scale) of the approximate solution paths generated by the proposed Newton method (Newton), the second-order Runge-Kutta method (Runge-Kutta), the Euler method (Euler), and the method of Rosset (2004) (Rosset) for ℓ_2-regularized logistic regression when the data is nonseparable.

suboptimality is small. This could be partially explained by the fact that the coordinate descent algorithms can usually be viewed as a type of methods that is between "first-order" and "second-order" method.

In summary, in terms of approximation error and computational efficiency, the Newton method and the second-order Runge-Kutta method both work quite well when the problem dimension is not too large or the desired suboptimality is small. For large-scale problems,

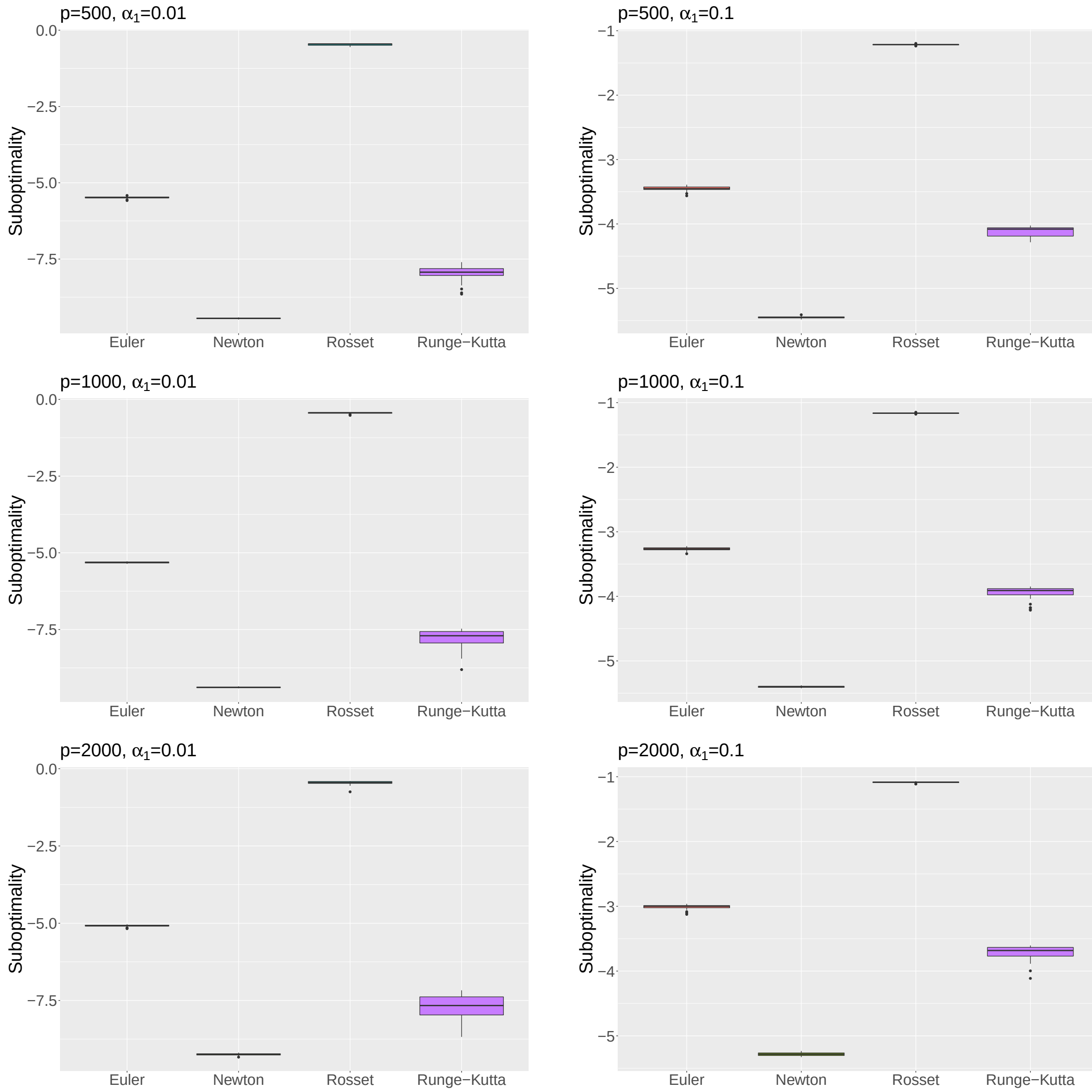

Figure 2: Suboptimalities $\sup_{0\le t\le 10}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$ (in log scale) of the approximate solution paths generated by the proposed Newton method (Newton), the second-order Runge-Kutta method (Runge-Kutta), the Euler method (Euler), and the method of Rosset (2004) (Rosset) for ℓ_2-regularized logistic regression when the data is separable.

however, gradient descent method and glmnet seem to be more scalable, although glmnet produces solution paths with better suboptimality.

Lastly, we investigate how the initial step size of various solution path algorithms would affect their statistical performances. As we have argued before, the initial step size determines the approximation error. To assess the accuracy of the approximation to the true statistical risk, we consider a generative model for logistic regression. Specifically, we first generate the predictors $X_1, \ldots, X_n \in \mathbb{R}^p$ from normal distribution $N_p(0, I_{p\times p})$. Given pre-

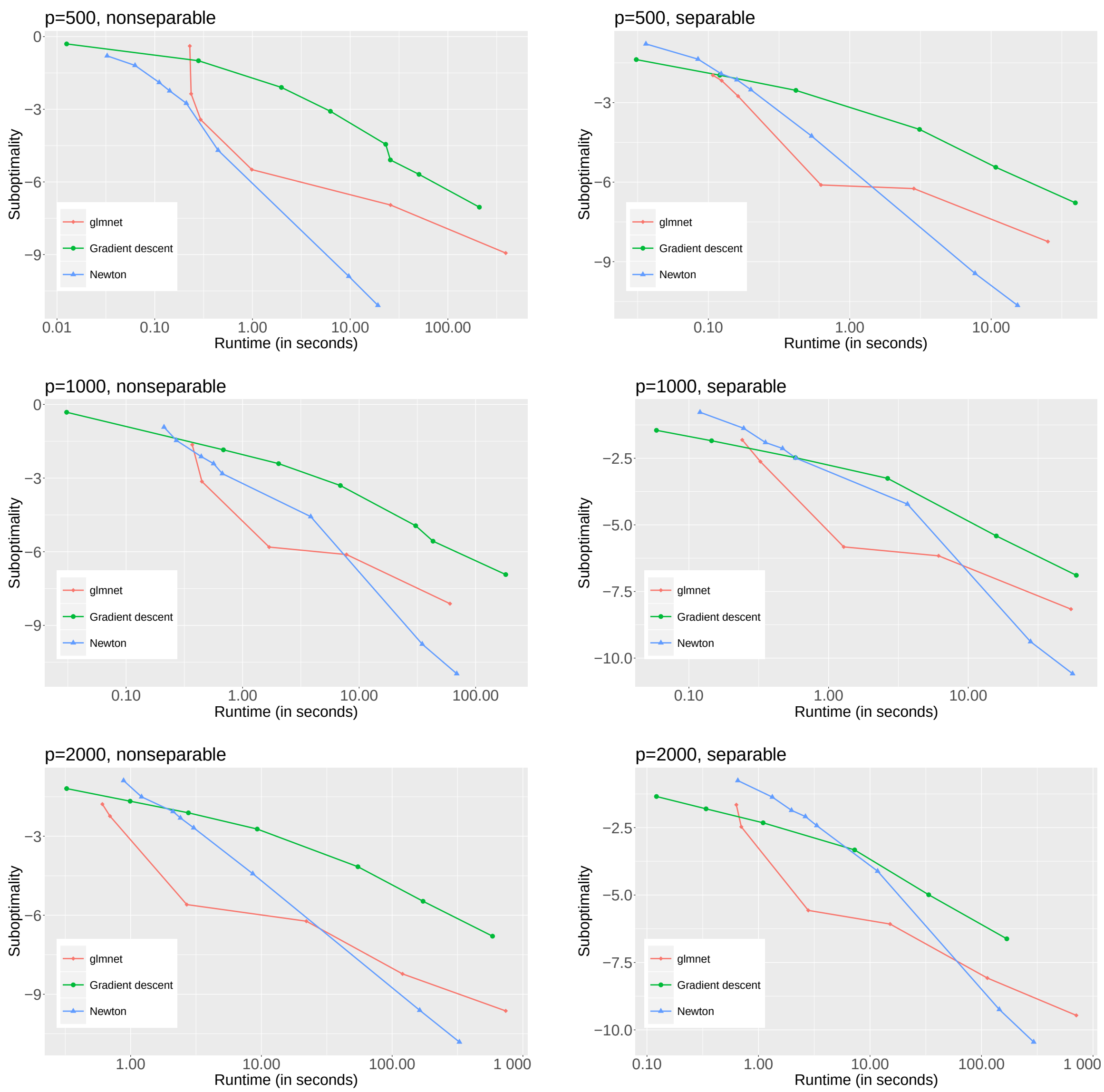

Figure 3: Runtime v.s. suboptimality for the proposed Newton method, gradient descent method, and glmnet under six different scenarios, when applied to ℓ_2-regularized logistic regression.

dictor X_i, we draw the binary response $Y_i \in \{-1, +1\}$ from Bernoulli distributions with $\mathbb{P}_\theta(Y_i = +1) = \exp(X_i^\top \theta)/(1 + \exp(X_i^\top \theta))$ for $i = 1, 2 \ldots, n$, where the true regression coefficient $\theta \in \mathbb{R}^p$ is drawn from $N_p(0, (16/p) \cdot I_{p \times p})$. Three choices of problem dimensions $(n, p) = (500, 100)$, $(n, p) = (500, 500)$, and $(n, p) = (500, 1000)$ will be considered. The statistical risk of an approximate solution path $\tilde{\theta}(t)$ is quantified by the Kullback–Leibler divergence:

$$R(\tilde{\theta}(t), \theta) = \mathbb{E}_\theta \log(1 + \exp(-Y X^\top \tilde{\theta}(t))) - \mathbb{E}_\theta \log(1 + \exp(-Y X^\top \theta)) \,.$$

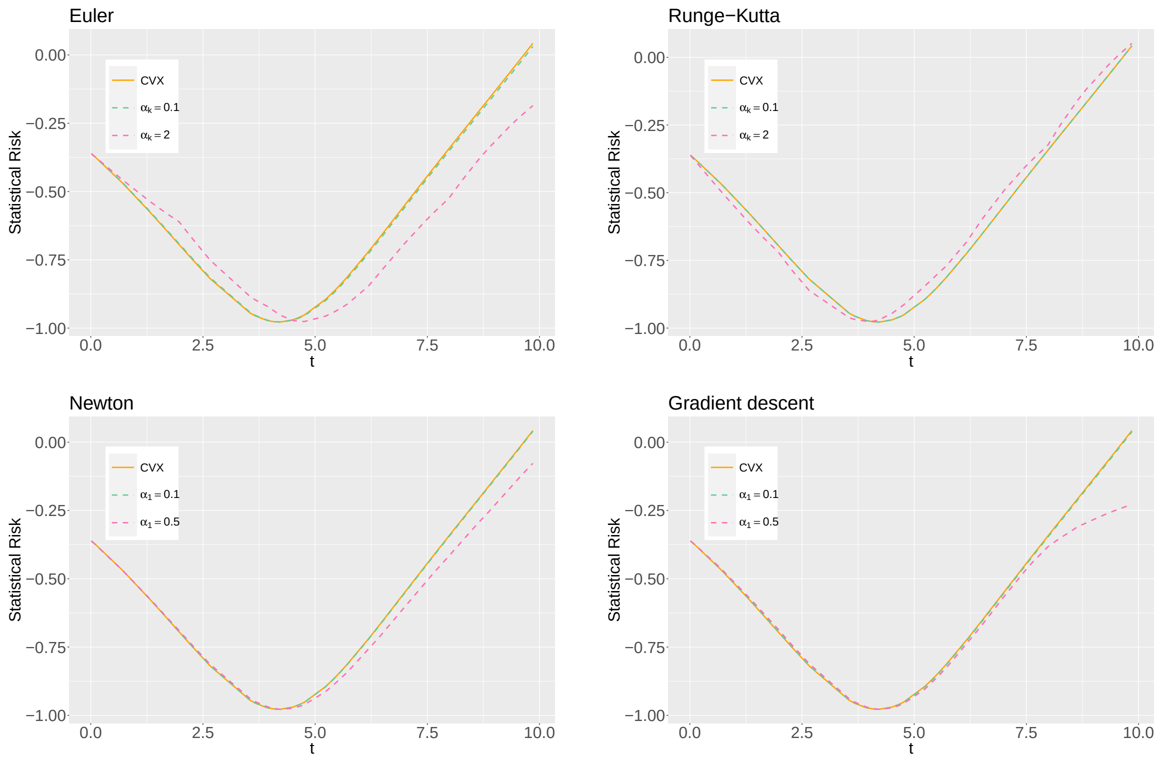

Figure 4: Approximate risk curve $\log_{10}(R(\tilde{\theta}(t), \theta))$ of the proposed algorithms applied to ℓ_2-regularized logistic regression when problem dimension is $(n, p) = (500, 100)$. The CVX (orange) curve denotes the true risk curve $\log_{10}(R(\theta(t), \theta))$ with $\theta(t)$ computed using the CVX solver. For algorithms with constant step size (Euler and Runge-Kutta), α_k denotes the step size; while α_1 denotes the initial step size for Newton and gradient descent method.

Note that the statistical risk for the exact solution path $\theta(t)$ is $R(\theta(t), \theta)$, which we refer to as the true risk curve (as a function of t). Here, we calculate the exact solution path $\theta(t)$ using CVX (Grant and Boyd, 2014, 2008). Again, the goal is to see the impact of the initial step size on how close the approximate risk curve $R(\tilde{\theta}(t), \theta)$ is to the true risk curve $R(\theta(t), \theta)$.

Figures 4–6 plot the approximate risk curve $R(\tilde{\theta}(t), \theta)$ against the true risk curve (on a log scale) by varying the initial step sizes for the proposed methods. Note that under all scenarios, when the initial step size is 0.1 (i.e., $\alpha_1 = 0.1$), the approximate risk curves approximate the true risk curve quite well for all four methods. This seems to suggest that good approximation error leads to good approximation of the risk curve. As the initial step size increases, interestingly, we observe that Runge-Kutta continues to provide reasonable good results, suggesting that they are more tolerant of a large initial step size (see the results when $\alpha_k = 2$ for Runge-Kutta methods on Figures 4–6). On the other hand, the Newton method and the gradient descent method requires the initial step sizes to be much smaller to obtain reasonable risk curve approximation. That says, this does not necessarily imply that the Newton method is less efficient than the ODE-based methods, because the

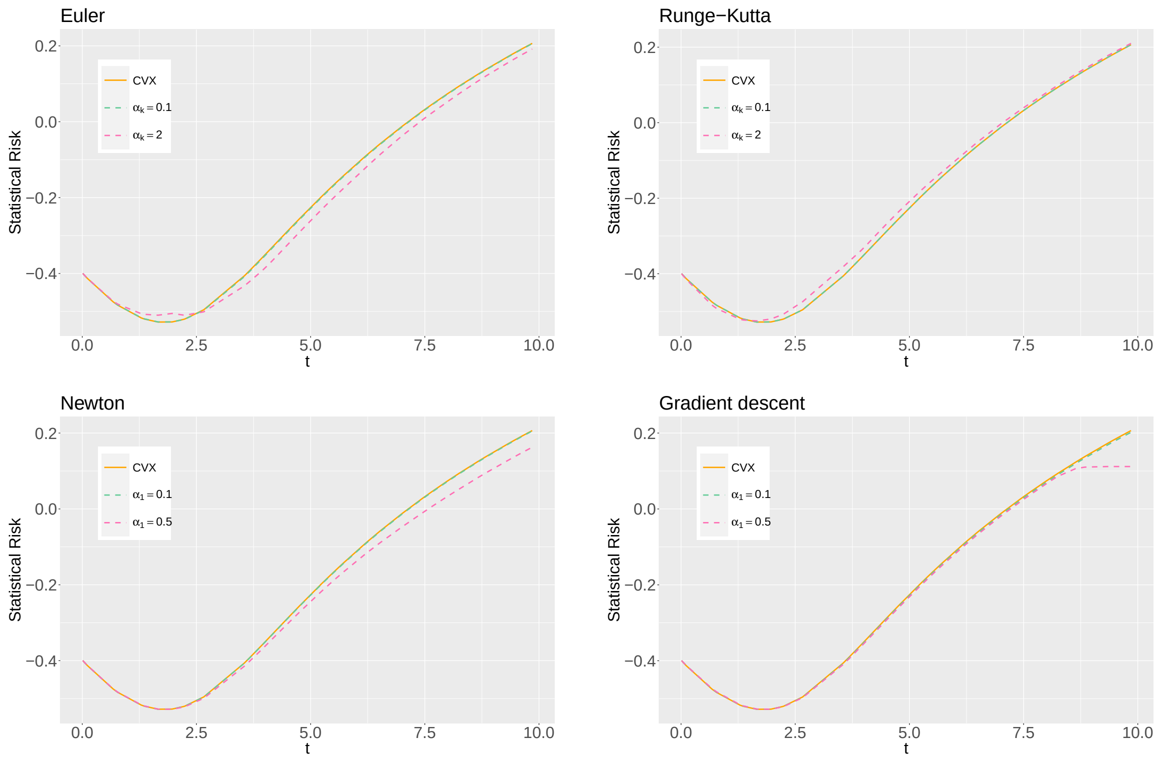

Figure 5: Approximate risk curve $\log_{10}(R(\tilde{\theta}(t), \theta))$ of the proposed algorithms applied to ℓ_2-regularized logistic regression when problem dimension is $(n, p) = (500, 500)$. The CVX (orange) curve denotes the true risk curve $\log_{10}(R(\theta(t), \theta))$ with $\theta(t)$ computed using the CVX solver. For algorithms with constant step size (Euler and Runge-Kutta), α_k denotes the step size; while α_1 denotes the initial step size for Newton and gradient descent method.

Newton method will adaptively increase step sizes while the ODE-based methods always fix their step sizes.

6. Discussion

In this article, we established a formal connection between ℓ_2-regularized solution path and the solution of an ODE. This connection provides an interesting algorithmic view of ℓ_2 regularization. In particular, the solution path turns out to be similar to the iterates of a hybrid algorithm that combines the gradient descent update and the Newton update. Moreover, we proposed various new path-following algorithms to approximate the ℓ_2-regularized solution path. Global approximation-error bounds for these methods are also derived, which in turn suggest some interesting schemes for choosing the grid points. Computational complexities are also derived using the proposed grid point schemes.

One important aspect we did not touch on is the statistical properties of ℓ_2-regularized solution path, which has been studied extensively in the literature (see, e.g., Dobriban and Wager, 2018, and references therein). Interestingly, Ali et al. (2019), in the context of least squares regression, connects the statistical properties of gradient descent iterates to that

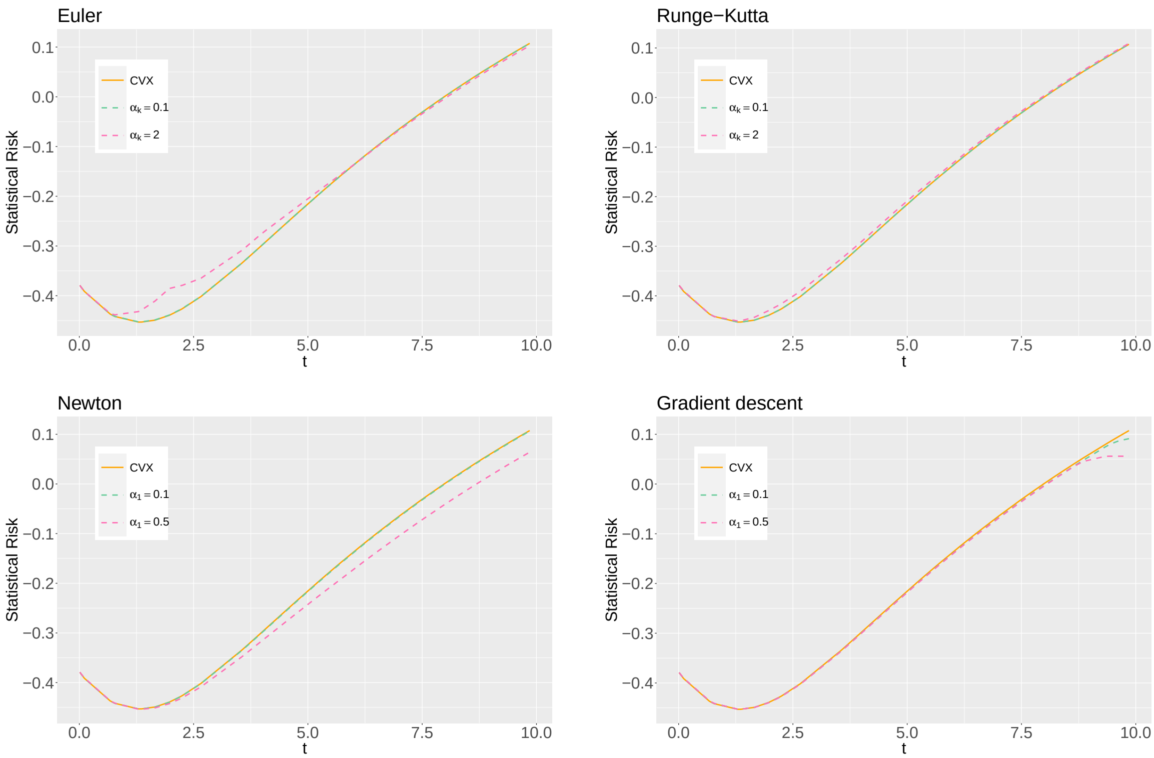

Figure 6: Approximate risk curve $\log_{10}(R(\tilde{\theta}(t), \theta))$ of the proposed algorithms applied to ℓ_2-regularized logistic regression when problem dimension is $(n, p) = (500, 1000)$. The CVX (orange) curve denotes the true risk curve $\log_{10}(R(\theta(t), \theta))$ with $\theta(t)$ computed using the CVX solver. For algorithms with constant step size (Euler and Runge-Kutta), α_k denotes the step size; while α_1 denotes the initial step size for Newton and gradient descent method.

of ridge regression solution path. In particular, they show that the statistical risk of the gradient descent path is no more than 1.69 times that of ridge regression, along the entire path. Motivated by our proposed homotopy method based on damped gradient descent updates (9), it would be interesting to investigate whether a damped version of gradient descent algorithm would enjoy a more favorable statistical risk compared to regular gradient descent. Further investigation is necessary.

Acknowledgments

We would like to thank the Associate Editor and reviewers for their insightful comments and encouragement to revise our paper. The feedback substantially improved the paper. We would also like to acknowledge support for this project from the National Science Foundation (DMS-17-12580, DMS-17-21445 and DMS-20-15490).

Appendix A. Proofs of main results

This section collects the proofs of Theorem 1–9, Corollary 1–3, and Proposition 1 and 2. Throughout this section, some standard results for m-strongly convex functions will be repeatedly used in the proofs, which are stated below. We omit their proofs as all of them can be found in standard convex analysis textbooks (see, e.g., Boyd and Vandenberghe, 2004).

Suppose that $f(\cdot)$ is a m-strongly convex function with minimizer $x^\star$. Then for any x and y,

$$m\|x-y\|_2^2 \le \langle \nabla f(x) - \nabla f(y)\,, x-y\rangle \le \frac{1}{m}\|\nabla f(x) - \nabla f(y)\|_2^2\,, \tag{44}$$

$$\frac{m}{2}\|x-x^\star\|_2^2 \le f(x) - f(x^\star) \le \frac{1}{2m}\|\nabla f(x)\|_2^2 \text{ and } \|x-x^\star\|_2 \le \frac{1}{m}\|\nabla f(x)\|_2\,. \tag{45}$$

Proof of Theorem 1. We first show differentiability of $\theta(t)$ at $t=0$. By the optimality of $\theta(t)$ and strong convexity of the objective function, we have that for any $t \ge 0$,

$$\|\theta(t) - \mathbf{0}\|_2^2 \le (\theta(t) - \mathbf{0})^\top (\mathbf{0} - C(t)\nabla L_n(\mathbf{0})) \le C(t)\|\theta(t)\|_2 \left\|\nabla L_n(\mathbf{0})\right\|_2\,,$$

where we have used (44). This implies that $\|\theta(t)\|_2 \le C(t)\left\|\nabla L_n(\mathbf{0})\right\|_2$. Thus $\theta(t)$ is continuous at $t=0$ since $\lim_{t\to 0} C(t) = C(0) = 0$ and $\mathbf{0} \in \mathbf{dom}\, L_n$. Moreover,

$$\frac{\theta(t)}{t} = -\frac{C(t)}{t}\nabla L_n(\theta(t)) \to -C'(0)\nabla L_n(\mathbf{0}) \text{ as } t \to 0\,, \tag{46}$$

where we have used the continuity of $\nabla L_n(\theta)$ and $\theta(t)$ at $\theta = \mathbf{0}$ and $t=0$, respectively. Therefore, $\theta(t)$ is differentiable at $t=0$.

Next we show the differentiability of $\theta(t)$ for $t>0$. Denote by $f_t(\theta) = C(t)L_n(\theta) + \frac{1}{2}\|\theta\|_2^2$. Since $f_t(\cdot)$ is 1-strongly convex for all $t \ge 0$, by using (44) and the fact that $\nabla f_{t'}(\theta(t')) = \nabla f_t(\theta(t)) = \mathbf{0}$, we have that for any $t > 0$

$$\begin{aligned}
\|\theta(t') - \theta(t)\|_2^2 &\le \langle \nabla f_{t'}(\theta(t')) - \nabla f_{t'}(\theta(t))\,, \theta(t') - \theta(t)\rangle = -\langle \nabla f_{t'}(\theta(t))\,, \theta(t') - \theta(t)\rangle \\
&= -\langle C(t')\nabla L_n(\theta(t)) + \theta(t)\,, \theta(t') - \theta(t)\rangle \\
&= -\langle -C(t')\theta(t)/C(t) + \theta(t)\,, \theta(t') - \theta(t)\rangle \\
&\le \frac{|C(t') - C(t)|}{C(t)}\|\theta(t)\|_2\|\theta(t') - \theta(t)\|_2\,,
\end{aligned} \tag{47}$$

which implies that

$$\|\theta(t') - \theta(t)\|_2 \le \frac{|C(t') - C(t)|}{C(t)}\|\theta(t)\|_2\,, \tag{48}$$

when $t>0$. This gives a bound on how fast can $\theta(t)$ can vary as t increases. Next, we use this to establish differentiability of $\theta(t)$. Note that for any $t, t' \ge 0$

$$C(t)\nabla L_n(\theta(t)) + \theta(t) = 0 \text{ and } C(t')\nabla L_n(\theta(t')) + \theta(t') = 0\,. \tag{49}$$

Taking the difference, we obtain that

$$\theta(t') - \theta(t) = -(C(t') - C(t))\nabla L_n(\theta(t)) - C(t')\nabla^2 L_n(\theta(t))(\theta(t') - \theta(t)) - C(t')\Delta\,, \tag{50}$$

where $\Delta = \nabla L_n(\theta(t')) - \nabla L_n(\theta(t)) - \nabla^2 L_n(\theta(t))(\theta(t') - \theta(t))$. Rearranging and dividing both sides by $t' - t$, we obtain that

$$\frac{\theta(t') - \theta(t)}{t' - t} = -\left(C(t')\nabla^2 L_n(\theta(t)) + I\right)^{-1} \left(\frac{C(t') - C(t)}{t' - t} \nabla L_n(\theta(t)) + C(t')\frac{\Delta}{t' - t}\right), \quad (51)$$

where the matrix $C(t')\nabla^2 L_n(\theta(t)) + I$ is invertible because $\nabla^2 L_n(\theta(t)) \succeq 0$. Since $C(t)$ is differentiable, it remains to show that $\|C(t')\frac{\Delta}{t'-t}\|_2 \to 0$ as $t' \to t$. By Assumption (A0) and (48), we obtain that

$$\begin{aligned}
&\left\|C(t')\frac{\Delta}{t' - t}\right\|_2 \\
\leq\ & \frac{C(t')}{|t' - t|} \int_0^1 \left\| \left[\nabla^2 L_n(\theta(t) + \tau(\theta(t') - \theta(t))) - \nabla^2 L_n(\theta(t))\right] (\theta(t') - \theta(t)) \right\|_2 d\tau \\
\leq\ & \frac{C(t')}{|t' - t|} \sup_{0 \leq \tau \leq 1} \left|\rho\left(\nabla^2 L_n(\theta(t) + \tau(\theta(t') - \theta(t))) - \nabla^2 L_n(\theta(t))\right)\right| \|\theta(t') - \theta(t)\|_2 \\
\leq\ & \frac{C(t')}{C(t)} \frac{|C(t') - C(t)|}{|t' - t|} \|\theta(t)\|_2 \sup_{0 \leq \tau \leq 1} \left|\rho\left(\nabla^2 L_n(\theta(t) + \tau(\theta(t') - \theta(t))) - \nabla^2 L_n(\theta(t))\right)\right| \to 0
\end{aligned}$$

as $t' \to t$, where $\rho(A)$ denotes the spectral norm of a matrix A, and we have used the fact that

$$\sup_{0 \leq \tau \leq 1} \left|\rho\left(\nabla^2 L_n(\theta(t) + \tau(\theta(t') - \theta(t))) - \nabla^2 L_n(\theta(t))\right)\right| \to 0 \text{ as } t' \to t \quad (52)$$

by Assumption (A0) and $\frac{C(t') - C(t)}{t' - t} \to C'(t)$ as $t' \to t$ since $C(t)$ is differentiable. Combining this with (51), it follows that

$$\theta'(t) = \lim_{t' \to t} \frac{\theta(t') - \theta(t)}{t' - t} = -C'(t)\left(C(t)\nabla^2 L_n(\theta(t)) + I\right)^{-1} \nabla L_n(\theta(t)).$$

This completes the proof of Theorem 1.

Proof of Corollary 1. To prove (i), rearranging terms in (47), we obtain that

$$(C(t') - C(t))\left(\|\theta(t')\|_2^2 - \|\theta(t)\|_2^2\right) \geq (C(t) + C(t'))\|\theta(t) - \theta(t')\|_2^2 \geq 0,$$

which implies that $\|\theta(t)\|_2$ is nondecreasing in t. For nonincreasingness of $L_n(\theta(t))$, note that

$$\begin{aligned}
C(t')L_n(\theta(t')) + \frac{1}{2}\|\theta(t')\|_2^2 \ &\leq\ C(t')L_n(\theta(t)) + \frac{1}{2}\|\theta(t)\|_2^2 \\
&\leq\ (C(t') - C(t))L_n(\theta(t)) + C(t)L_n(\theta(t')) + \frac{1}{2}\|\theta(t')\|_2^2,
\end{aligned}$$

which implies that $(C(t') - C(t))(L_n(\theta(t')) - L_n(\theta(t))) \leq 0$. Hence, if $C(t') - C(t) > 0$ then $L_n(\theta(t')) \leq L_n(\theta(t))$, which proves that $L_n(\theta(t))$ is nonincreasing in t.

To prove (ii), we modify the proof of (48) to accommodate the case where $L_n(\theta)$ might not be differentiable. Note that for any $t > 0$ and any $g_{t'} \in \partial L_n(\theta(t'))$ and $g_t \in \partial L_n(\theta(t))$, we have $\langle g_{t'} - g_t \,, \theta(t') - \theta(t)\rangle \geq 0$, where $\partial L_n(\theta)$ denotes the *subdifferential* of $L_n(\cdot)$ at θ. Hence, for any $h_{t'} \in \partial f_{t'}(\theta(t')), h_t \in \partial f_{t'}(\theta(t))$

$$\langle h_{t'} - h_t \,, \theta(t') - \theta(t)\rangle \geq \|\theta(t') - \theta(t)\|_2^2 \tag{53}$$

Since $\mathbf{0} \in \partial f_{t'}(\theta(t'))$ and $-C(t')\theta(t)/C(t) + \theta(t) \in \partial f_{t'}(\theta(t))$, substituting $h_{t'}$ with $\mathbf{0}$ and h_t with $-C(t')\theta(t)/C(t) + \theta(t)$, we obtain that

$$\langle C(t')\theta(t)/C(t) - \theta(t) \,, \theta(t') - \theta(t)\rangle \geq \|\theta(t') - \theta(t)\|_2^2 \,, \tag{54}$$

which implies that

$$\|\theta(t') - \theta(t)\|_2^2 \leq \langle C(t')\theta(t)/C(t) - \theta(t) \,, \theta(t') - \theta(t)\rangle \leq |C(t')/C(t) - 1| \|\theta(t)\|_2 \|\theta(t') - \theta(t)\|_2 \,,$$

which proves (48) when $L_n(\theta)$ might not be differentiable. Using this, we have that for any $t > t'$ and $\theta(t) \neq \theta(t')$,

$$\|\theta(t)\|_2 - \|\theta(t')\|_2 \leq \|\theta(t') - \theta(t)\|_2 \leq (C(t) - C(t'))\|\theta(t)\|_2 / C(t) \,, \tag{55}$$

which implies that

$$\|\theta(t)\|_2 / C(t) \leq \|\theta(t')\|_2 / C(t') \,. \tag{56}$$

This also holds when $\theta(t) = \theta(t')$ because $C(t)$ is an increasing function. This proves that part (ii).

Lastly, we prove part (iii). Denote by $\theta^\star$ the minimum ℓ_2 norm minimizer of $L_n(\theta)$. Next, we show that $\theta(t)$ converges to $\theta^\star$ as $t \to \infty$ if $\theta^\star$ is finite. Note that $\mathbf{0} \in \partial L_n(\theta^\star)$ and $\mathbf{0} \in C(t)\partial L_n(\theta(t)) + \theta(t)$. As a result,

$$\mathbf{0} \in C(t) \left(\partial L_n(\theta(t)) - \partial L_n(\theta^\star)\right) + \theta(t) \,,$$

where $A - B$ denotes the set $\{a - b : a \in A \text{ and } b \in B\}$. Multiplying $\theta(t) - \theta^\star$ on both sides, we obtain that

$$(\theta(t) - \theta^\star)^\top \theta(t) \in -C(t)(\theta(t) - \theta^\star)^\top \left(\partial L_n(\theta(t)) - \partial L_n(\theta^\star)\right) \,,$$

which implies that $(\theta(t) - \theta^\star)^\top \theta(t) \leq 0$. Therefore, $\|\theta(t)\|_2^2 \leq (\theta^\star)^\top \theta(t) \leq \|\theta^\star\|_2 \|\theta(t)\|_2$, which implies that $\|\theta(t)\|_2 \leq \|\theta^\star\|_2 < \infty$ for any $t \geq 0$. Denote by $\bar{\theta}$ the limit of any converging subsequence $\theta(t_k)$, that is, $\bar{\theta} = \lim_{k\to\infty} \theta(t_k)$ for some $t_k \to \infty$. Then, $\|\bar{\theta}\|_2 = \lim_{k\to\infty} \|\theta(t_k)\|_2 \leq \|\theta^\star\|_2$. Next, we show that $\bar{\theta}$ must also be a minimizer of $L_n(\theta)$. To this end, note that $L_n(\bar{\theta}) = \lim_{k\to\infty} L_n(\theta(t_k))$ by using the continuity of $L_n(\theta)$ in θ. Moreover, by optimality of $\theta(t_k)$,

$$L_n(\theta(t_k)) \leq L_n(\theta(t_k)) + \frac{1}{2C(t_k)} \|\theta(t_k)\|_2^2 \leq L_n(\theta^\star) + \frac{1}{2C(t_k)} \|\theta^\star\|_2^2 \,. \tag{57}$$

By letting $k \to \infty$ and using the fact that $L_n(\bar{\theta}) = \lim_{k\to\infty} L_n(\theta(t_k))$ due to continuity of $L_n(\theta)$, we have that

$$L_n(\bar{\theta}) = \lim_{k\to\infty} L_n(\theta(t_k)) \leq \lim_{k\to\infty} \left(L_n(\theta^\star) + \frac{1}{2C(t_k)} \|\theta^\star\|_2^2 \right) = L_n(\theta^\star) \,,$$

where the last step uses the assumption that $\|\theta^\star\|_2 < \infty$. This proves that $\bar{\theta}$ must also be a minimizer of $L_n(\theta)$.

Now if $\bar{\theta} \neq \theta^\star$, then their convex combination $\frac{1}{2}(\bar{\theta} + \theta^\star)$ must also be a minimizer of $L_n(\theta)$ due to the convexity of $L_n(\theta)$. On the other hand, the convex combination has strictly smaller norm than that of $\theta^\star$, because $\|\frac{1}{2}(\bar{\theta} + \theta^\star)\|_2 < \frac{1}{2}(\|\bar{\theta}\| + \|\theta^\star\|_2) \leq \|\theta^\star\|_2$. This contradicts with the definition of $\theta^\star$. Hence, we must have $\lim_{k\to\infty} \theta(t_k) = \bar{\theta} = \theta^\star$ for every converging subsequence $\theta(t_k)$. Consequently, the sequence $\theta(t)$ must converge to $\theta^\star$. This completes the proof of Corollary 1.

Proof of Theorem 2. For any $t \in [t_k, t_{k+1}]$, we let $w_k = \frac{t_{k+1}-t}{t_{k+1}-t_k}$, for $k = 0, 1, \ldots, N-1$. Then $\tilde{\theta}(t) = w_k\theta_k + (1-w_k)\theta_{k+1}$. By convexity of $f_t(\cdot)$, we have $f_t(\tilde{\theta}(t)) \leq w_k f_t(\theta_k) + (1 - w_k)f_t(\theta_{k+1})$. Thus,

$$f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \leq w_k(f_t(\theta_k) - f_t(\theta(t))) + (1-w_k)(f_t(\theta_{k+1}) - f_t(\theta(t)))\,. \tag{58}$$

For any $k = 1, \ldots, N-1$, the term $f_t(\theta_k) - f_t(\theta(t))$ in (58) can be bounded as follows:

$$\begin{aligned} f_t(\theta_k) - f_t(\theta(t)) &\leq \frac{1}{2e^{-t}}\|\nabla f_t(\theta_k)\|_2^2 = \frac{e^t}{2}\left\|\frac{1-e^{-t}}{1-e^{-t_k}}\nabla f_{t_k}(\theta_k) + \frac{e^{-t}-e^{-t_k}}{1-e^{-t_k}}\theta_k\right\|_2^2 \\ &\leq e^t\left(\frac{1-e^{-t}}{1-e^{-t_k}}\right)^2\|\nabla f_{t_k}(\theta_k)\|_2^2 + e^t\left(\frac{e^{-t}-e^{-t_k}}{1-e^{-t_k}}\right)^2\|\theta_k\|_2^2 \\ &= e^t\left(\frac{1-e^{-t}}{1-e^{-t_k}}\right)^2\|g_k\|_2^2 + e^t\left(\frac{e^{-t}-e^{-t_k}}{1-e^{-t_k}}\right)^2\|\theta_k\|_2^2\,, \end{aligned}$$

where the first inequality uses the fact that $f_t(\cdot)$ is e^{-t}-strongly convex and (45). Similarly, we can bound the term $f_t(\theta_{k+1}) - f_t(\theta(t))$ by

$$e^t\left(\frac{1-e^{-t}}{1-e^{-t_{k+1}}}\right)^2\|g_{k+1}\|_2^2 + e^t\left(\frac{e^{-t}-e^{-t_{k+1}}}{1-e^{-t_{k+1}}}\right)^2\|\theta_{k+1}\|_2^2$$

for any $k = 0, 1, \ldots, N-1$. Combining these two bounds, we have that

$$\begin{aligned} &w_k(f_t(\theta_k) - f_t(\theta(t))) + (1-w_k)(f_t(\theta_{k+1}) - f_t(\theta(t))) \\ &\leq e^{t_{k+1}}\max\left\{\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2\|g_k\|_2^2,\ \|g_{k+1}\|_2^2\right\} \\ &\quad + (e^{-t_k} - e^{-t_{k+1}})^2\max\left\{\frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2},\ \frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\right\}\,, \end{aligned}$$

for any $k = 1, \ldots, N-1$. This proves (14).

When $k = 0$, the term $f_t(\theta_k) - f_t(\theta(t))$ in (58) can be bounded as follows

$$f_t(\theta_0) - f_t(\theta(t)) = f_t(\mathbf{0}) - f_t(\theta(t)) \leq \frac{1}{2e^{-t}}\|\nabla f_t(\mathbf{0})\|_2^2 = \frac{e^t(1-e^{-t})^2}{2}\|\nabla L_n(\mathbf{0})\|_2^2$$

for any $0 \leq t < t_1$, where we have used (45) in the above inequality. Following a similar argument as before, we obtain that

$$\begin{aligned} &w_0(f_t(\theta_0) - f_t(\theta(t))) + (1-w_0)(f_t(\theta_1) - f_t(\theta(t))) \\ \leq\ &\frac{e^{t_1}(1-e^{-t_1})^2}{2}\|\nabla L_n(\mathbf{0})\|_2^2 + \max\left(e^{t_1}\|g_1\|_2^2,\ \|\theta_1\|_2^2\right) \end{aligned}$$

for any $t \in [0, t_1]$. This proves (13).

Now we bound $f_t(\tilde{\theta}(t)) - f_t(\theta(t))$ when $t_N < t \le t_{\max}$. Toward this end, notice that

$$\begin{aligned}
&f_t(\tilde{\theta}(t)) - f_t(\theta(t)) = f_t(\theta_N) - f_t(\theta(t)) \\
=&\frac{1-e^{-t}}{1-e^{-t_N}}(f_{t_N}(\theta_N) - f_{t_N}(\theta(t_N))) + \frac{e^{-t_N}-e^{-t}}{2(1-e^{-t_N})}(\|\theta(t_N)\|_2^2 - \|\theta_N\|_2^2) + f_t(\theta(t_N)) - f_t(\theta(t))\,.
\end{aligned}$$

Next, we bound these three terms separately. For the first term, by using (45), we have that $f_{t_N}(\theta_N) - f_{t_N}(\theta(t_N)) \le 2^{-1}e^{t_N}\|g_N\|_2^2$. Using this, we obtain that

$$\frac{1-e^{-t}}{1-e^{-t_N}}(f_{t_N}(\theta_N) - f_{t_N}(\theta(t_N))) \le \frac{(1-e^{-t})e^{t_N}}{2(1-e^{-t_N})}\|g_N\|_2^2\,. \tag{59}$$

For the second term, note that

$$\begin{aligned}
&\|\theta(t_N)\|_2^2 - \|\theta_N\|_2^2 = 2(\theta(t_N)-\theta_N)^\top\theta(t_N) - \|\theta(t_N)-\theta_N\|_2^2 \\
&\le (2\|\theta(t_N)\|_2 - \|\theta(t_N)-\theta_N\|_2)\|\theta(t_N)-\theta_N\|_2 \le 2\|\theta(t_N)\|_2\|\theta(t_N)-\theta_N\|_2 \\
&\le 2\|\theta(t_N)\|_2 e^{t_N}\|g_N\|_2 = 2e^{t_N}\|\theta(t_N)\|_2\|g_N\|_2\,.
\end{aligned}$$

Thus, the second term can be bounded by $\frac{1-e^{-(t-t_N)}}{1-e^{-t_N}}\|\theta(t_N)\|_2\|g_N\|_2$, which can be further bounded using the Cauchy–Schwarz inequality:

$$\frac{1-e^{-(t-t_N)}}{1-e^{-t_N}}\|\theta(t_N)\|_2\|g_N\|_2 \le \frac{1}{2}\left(\frac{(1-e^{-t})e^{t_N}}{1-e^{-t_N}}\|g_N\|_2^2 + \frac{(1-e^{-(t-t_N)})^2}{(1-e^{-t_N})(1-e^{-t})e^{t_N}}\|\theta(t_N)\|_2^2\right)\,.$$

To bound the third term, by optimality of $\theta(t_N)$, we have $f_{t_N}(\theta(t_N)) \le f_{t_N}(\theta(t))$, which in turn implies that $L_n(\theta(t_N)) - L_n(\theta(t)) \le .5(e^{t_N}-1)^{-1}\left(\|\theta(t)\|_2^2 - \|\theta(t_N)\|_2^2\right)$. Using this, the third term can be bounded as follows

$$\begin{aligned}
&f_t(\theta(t_N)) - f_t(\theta(t)) = (1-e^{-t})(L_n(\theta(t_N)) - L_n(\theta(t))) + \frac{e^{-t}}{2}\left(\|\theta(t_N)\|_2^2 - \|\theta(t)\|_2^2\right) \\
&\le \frac{1-e^{-t}}{2(e^{t_N}-1)}\left(\|\theta(t)\|_2^2 - \|\theta(t_N)\|_2^2\right) + \frac{e^{-t}}{2}\left(\|\theta(t_N)\|_2^2 - \|\theta(t)\|_2^2\right) \\
&= \frac{1-e^{-(t-t_N)}}{2(e^{t_N}-1)}\left(\|\theta(t)\|_2^2 - \|\theta(t_N)\|_2^2\right)\,.
\end{aligned}$$

Combining the three bounds and using the fact that

$$\frac{1-e^{-(t-t_N)}}{2(e^{t_N}-1)} = \frac{(1-e^{-(t-t_N)})^2}{2(1-e^{-t_N})(1-e^{-t})e^{t_N}} + \frac{1-e^{-(t-t_N)}}{2(e^t-1)}\,,$$

we obtain that

$$\begin{aligned}
\sup_{t_N<t\le t_{\max}}\left\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\right\} &\le \sup_{t_N<t\le t_{\max}}\left\{\frac{(1-e^{-t})e^{t_N}}{1-e^{-t_N}}\|g_N\|_2^2 + \frac{1-e^{-(t-t_N)}}{2(e^{t_N}-1)}\|\theta(t)\|_2^2\right\} \\
&\le \frac{e^{t_N}(1-e^{-t_{\max}})}{1-e^{-t_N}}\|g_N\|_2^2 + \sup_{t_N<t\le t_{\max}}\frac{(1-e^{-(t-t_N)})^2}{2(1-e^{-t_N})(1-e^{-t})e^{t_N}}\|\theta(t)\|_2^2 \\
&\quad + \sup_{t_N<t\le t_{\max}}\frac{1-e^{-(t-t_N)}}{2(e^t-1)}\left(\|\theta(t)\|_2^2 - \|\theta(t_N)\|_2^2\right)\,.
\end{aligned}$$

Moreover, by (48), we have that

$$\|\theta(t)\|_2^2 - \|\theta(t_N)\|_2^2 \le \|\theta(t) - \theta(t_N)\|_2(\|\theta(t)\|_2 + \|\theta(t_N)\|_2)$$
$$\le \frac{e^t - e^{t_N}}{e^t - 1}\|\theta(t)\|_2(\|\theta(t)\|_2 + \|\theta(t_N)\|_2) \le 2\frac{e^t - e^{t_N}}{e^t - 1}\|\theta(t)\|_2^2$$

Combining the above two inequalities and using the fact that

$$\sup_{t_N < t \le t_{\max}} \frac{(1 - e^{-(t - t_N)})^2}{1 - e^{-t}} = \frac{(1 - e^{-(t_{\max} - t_N)})^2}{1 - e^{-t_{\max}}},$$

we obtain that

$$\sup_{t_N < t \le t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \le \frac{e^{t_N}(1 - e^{-t_{\max}})}{1 - e^{-t_N}}\|g_N\|_2^2 + \frac{3(1 - e^{-(t_{\max} - t_N)})^2}{2(e^{t_N} - 1)(1 - e^{-t_{\max}})}\|\theta(t_{\max})\|_2^2,$$

which implies (15). This completes the proof of Theorem 2.

Next, we present a supporting lemma for the proof of Theorem 3.

Lemma 1 *Under Assumption (A0), we have that*

$$\|\theta(t)\|_2 \le (e^t - 1)\|\nabla L_n(\mathbf{0})\|_2 \text{ and } \|\theta(t)\|_2 \le (1 - e^{-t})\left(\|\nabla L_n(\mathbf{0})\|_2 + \|\theta(t')\|_2\right) \tag{60}$$

for any $t' \ge t > 0$.

Proof of Lemma 1. Since $f_t(\cdot)$ is e^{-t} strongly convex, using (45), we have

$$\|\theta(t) - \mathbf{0}\|_2 \le e^t\|\nabla f_t(\mathbf{0})\|_2 = (e^t - 1)\|\nabla L_n(\mathbf{0})\|_2,$$

which proves the first inequality $\|\theta(t)\|_2 \le (e^t - 1)\|\nabla L_n(\mathbf{0})\|_2$. Combining this with the fact that $\|\theta(t)\|_2 \le \|\theta(t')\|_2$, we have that

$$\|\theta(t)\|_2 \le \min\left((e^t - 1)\|\nabla L_n(\mathbf{0})\|_2, \|\theta(t')\|_2\right) \le (1 - e^{-t})\left(\|\nabla L_n(\mathbf{0})\|_2 + \|\theta(t')\|_2\right),$$

which proves the second inequality in (60). This completes the proof of Lemma 1.

Proof of Theorem 3. Note that

$$\begin{aligned} g_{k+1} &= (1 - e^{-t_{k+1}})\nabla L_n(\theta_{k+1}) + e^{-t_{k+1}}\theta_{k+1} \\ &= \underbrace{(1 - e^{-t_{k+1}})\left(\nabla L_n(\theta_{k+1}) - \nabla L_n(\theta_k) - \nabla^2 L_n(\theta_k)(\theta_{k+1} - \theta_k)\right)}_{\text{Part I}} + \\ &\quad \underbrace{(1 - e^{-t_{k+1}})\left(\nabla L_n(\theta_k) + \nabla^2 L_n(\theta_k)(\theta_{k+1} - \theta_k)\right) + e^{-t_{k+1}}\theta_{k+1}}_{\text{Part II}}. \end{aligned}$$

Moreover, based on the definition of θ_{k+1}, we have that

$$\left((1 - e^{-t_{k+1}})\nabla^2 L_n(\theta_k) + e^{-t_{k+1}}I\right)(\theta_{k+1} - \theta_k) + (1 - e^{-\alpha_{k+1}})\nabla L_n(\theta_k) + e^{-\alpha_{k+1}}g_k = 0. \tag{61}$$

Combining this with the fact that $g_k = (1 - e^{-t_k})\nabla L_n(\theta_k) + e^{-t_k}\theta_k$, we obtain that

$$\begin{aligned}\text{Part II} &= \big((1 - e^{-t_{k+1}})\nabla^2 L_n(\theta_k) + e^{-t_{k+1}} I\big)(\theta_{k+1} - \theta_k) + (1 - e^{-t_{k+1}})\nabla L_n(\theta_k) + e^{-t_{k+1}}\theta_k \\ &= (1 - e^{-t_{k+1}})\nabla L_n(\theta_k) + e^{-t_{k+1}}\theta_k - (1 - e^{-\alpha_{k+1}})\nabla L_n(\theta_k) - e^{-\alpha_{k+1}} g_k \\ &= e^{-\alpha_{k+1}}\big((1 - e^{-t_k})\nabla L_n(\theta_k) + e^{-t_k}\theta_k - g_k\big) = \mathbf{0}\,.\end{aligned}$$

Hence, we have that

$$\begin{aligned}\|g_{k+1}\|_2 &= (1 - e^{-t_{k+1}})\left\|\nabla L_n(\theta_{k+1}) - \nabla L_n(\theta_k) - \nabla^2 L_n(\theta_k)(\theta_{k+1} - \theta_k)\right\|_2 \\ &\leq \beta(1 - e^{-t_{k+1}})(\theta_{k+1} - \theta_k)^\top \left[\nabla^2 L_n(\theta_k)\right]^{\gamma_1}(\theta_{k+1} - \theta_k)\,, \end{aligned} \tag{62}$$

where the last inequality uses (16) in Assumption (A1), provided that

$$(\theta_{k+1} - \theta_k)^\top[\nabla^2 L_n(\theta_k)]^{\gamma_2}(\theta_{k+1} - \theta_k) \leq \beta^{-2}\,,$$

which is to be verified later by induction. Next, we define

$$H_{k+1} = (1 - e^{-t_{k+1}})\nabla^2 L_n(\theta_k) + e^{-t_{k+1}} I \text{ and } J_{k+1} = H_{k+1}^{-1}\left[\nabla^2 L_n(\theta_k)\right]^{\gamma_1} H_{k+1}^{-1}\,. \tag{63}$$

Notice that $\theta_{k+1} - \theta_k = -H_{k+1}^{-1}((1 - e^{-\alpha_{k+1}})\nabla L_n(\theta_k) + e^{-\alpha_{k+1}} g_k)$. Combining this with $g_k = (1 - e^{-t_k})\nabla L_n(\theta_k) + e^{-t_k}\theta_k$, we obtain that,

$$\theta_1 = -(1 - e^{-\alpha_1})H_1^{-1}\nabla L_n(\mathbf{0}), \tag{64}$$

$$\begin{aligned}\theta_{k+1} - \theta_k &= -H_{k+1}^{-1}\big((1 - e^{-\alpha_{k+1}})\nabla L_n(\theta_k) + e^{-\alpha_{k+1}} g_k\big) \\ &= -H_{k+1}^{-1}\left(\frac{1 - e^{-t_{k+1}}}{1 - e^{-t_k}} g_k - \frac{e^{-t_k} - e^{-t_{k+1}}}{1 - e^{-t_k}}\theta_k\right) \text{ for any } k \geq 1\,. \end{aligned} \tag{65}$$

Combining (62) and (65), and using the fact that $\|a + b\|_2^2 \leq 2(\|a\|_2^2 + \|b\|_2^2)$ for any vectors a and b, we have that

$$\|g_1\|_2 \leq \beta(1 - e^{-\alpha_1})\theta_1^\top[\nabla^2 L_n(\mathbf{0})]^{\gamma_1}\theta_1 \tag{66}$$

$$\|g_{k+1}\|_2 \leq 2\beta\lambda_{\max}(J_{k+1})(1 - e^{-t_{k+1}})\left\{\frac{(1 - e^{-t_{k+1}})^2}{(1 - e^{-t_k})^2}\|g_k\|_2^2 + \frac{(e^{-t_k} - e^{-t_{k+1}})^2}{(1 - e^{-t_k})^2}\|\theta_k\|_2^2\right\} \tag{67}$$

for any $k \geq 1$. Using (67) and the fact that

$$\lambda_{\max}(J_{k+1}) \leq \sup_{\lambda:\lambda\geq 0}\frac{\lambda^{\gamma_1}}{((1 - e^{-t_{k+1}})\lambda + e^{-t_{k+1}})^2} = \frac{e^{(2-\gamma_1)t_{k+1}}}{4(1 - e^{-t_{k+1}})^{\gamma_1}}(2 - \gamma_1)^{2-\gamma_1}\gamma_1^{\gamma_1} \tag{68}$$

for any $0 \leq \gamma_1 \leq 2$ and $k \geq 0$, we obtain that for any $k \geq 1$,

$$\|g_{k+1}\|_2 \leq \frac{\beta h(\gamma_1)e^{(2-\gamma_1)t_{k+1}}}{2(1 - e^{-t_{k+1}})^{\gamma_1 - 1}}\left\{\frac{(1 - e^{-t_{k+1}})^2}{(1 - e^{-t_k})^2}\|g_k\|_2^2 + \frac{(e^{-t_k} - e^{-t_{k+1}})^2}{(1 - e^{-t_k})^2}\|\theta_k\|_2^2\right\}, \tag{69}$$

where $h(\gamma_1) = (2 - \gamma_1)^{2-\gamma_1}\gamma_1^{\gamma_1}$. Throughout the proof, we shall treat $h(\gamma)$ as an absolute constant as $1 \leq h(\gamma) \leq 4$ for $0 \leq \gamma \leq 2$. We then use induction to show that

$$\|g_{k+1}\|_2 \leq \frac{c_0\beta h(\gamma_1)e^{(2-\gamma_1)t_{k+1}}}{2(1 - e^{-t_{k+1}})^{\gamma_1 - 1}}\frac{(e^{-t_k} - e^{-t_{k+1}})^2}{(1 - e^{-t_k})^2}\|\theta_k\|_2^2 \tag{70}$$

for any $k \geq 1$ and $c_0 = 26/25$. To this end, in view of (69), we only need to show that

$$\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\|g_k\|_2 \leq (c_0-1)^{1/2}\frac{(e^{-t_k}-e^{-t_{k+1}})}{(1-e^{-t_k})}\|\theta_k\|_2\,,$$

or equivalently,

$$\|g_k\|_2 \leq (c_0-1)^{1/2}\frac{(e^{-t_k}-e^{-t_{k+1}})}{(1-e^{-t_{k+1}})}\|\theta_k\|_2 \tag{71}$$

for any $k \geq 1$. To verify this, our plan is to show that (i) inequality (71) holds for $k=1$ using the bound in (66) for $\|g_1\|_2$; and (ii) inequality (71) holds for $\|g_k\|_2$ if the bound (70) holds for $\|g_k\|_2$ and inequality (71) holds for $\|g_{k-1}\|_2$.

When $k=1$, we note that

$$\begin{aligned}
\left\|[\nabla^2 L_n(\mathbf{0})]^{\gamma_1}\theta_1\right\|_2 &= (1-e^{-\alpha_1})\left\|[\nabla^2 L_n(\mathbf{0})]^{\gamma_1}H_1^{-1}\nabla L_n(\mathbf{0})\right\|_2\\
&= (1-e^{-\alpha_1})\sqrt{[\nabla L_n(\mathbf{0})]^\top H_1^{-1}[\nabla^2 L_n(\mathbf{0})]^{2\gamma_1}H_1^{-1}\nabla L_n(\mathbf{0})}\\
&\leq \begin{cases} 2^{-1}\sqrt{h(2\gamma_1)}e^{(1-\gamma_1)\alpha_1}(1-e^{-\alpha_1})^{1-\gamma_1}\|\nabla L_n(\mathbf{0})\|_2 & \text{if } \gamma_1 \leq 1;\\ \min\left(\nu^{\gamma_1-1},\, \nu^{\gamma_1}e^{-\alpha_1}(1-e^{-\alpha_1})\right)\|\nabla L_n(\mathbf{0})\|_2 & \text{if } \gamma_1 > 1\,. \end{cases}
\end{aligned}$$

where ν denotes the largest eigenvalue of $\nabla^2 L_n(\mathbf{0})$, and we have used (68).

Combining this with (66), we obtain that

$$\begin{aligned}
\|g_1\|_2 &\leq \beta(1-e^{-\alpha_1})\theta_1^\top[\nabla^2 L_n(\mathbf{0})]^{\gamma_1}\theta_1 \leq \beta(1-e^{-\alpha_1})\left\|[\nabla^2 L_n(\mathbf{0})]^{\gamma_1}\theta_1\right\|_2\|\theta_1\|_2\\
&\leq \begin{cases} \beta e^{(1-\gamma_1)\alpha_1}(1-e^{-\alpha_1})^{2-\gamma_1}\|\nabla L_n(\mathbf{0})\|_2\|\theta_1\|_2 & \text{if } \gamma_1 \leq 1;\\ \beta\min\left(\nu^{\gamma_1-1},\, \nu^{\gamma_1}e^{-\alpha_1}(1-e^{-\alpha_1})\right)(1-e^{-\alpha_1})\|\nabla L_n(\mathbf{0})\|_2\|\theta_1\|_2 & \text{if } \gamma_1 > 1 \end{cases}\\
&= \frac{C_1}{15}\beta e^{-\alpha_1}(e^{\alpha_1}-1)^{\max(2-\gamma_1,1)}\|\nabla L_n(\mathbf{0})\|_2\|\theta_1\|_2\,,
\end{aligned} \tag{72}$$

which can be upper bounded by

$$(c_0-1)^{1/2}\frac{(e^{-t_1}-e^{-t_2})}{(1-e^{-t_2})}\|\theta_1\|_2 = \frac{(e^{-t_1}-e^{-t_2})}{5(1-e^{-t_2})}\|\theta_1\|_2\,, \tag{73}$$

if we choose $c_0 = 26/25$, because (18a), where we have used the fact that $e^{t_1}(e^{-t_1}-e^{-t_2})(1-e^{-t_2})^{-1} \geq 3^{-1}$.

When $k \geq 2$, we next verify (71) when the bound in (70) holds for $\|g_k\|_2$ and inequality (71) holds for $\|g_{k-1}\|_2$. First using (65), we have that

$$\begin{aligned}
\|\theta_{k-1}\|_2 &\leq \|\theta_k\|_2 + \frac{1-e^{-t_k}}{1-e^{-t_{k-1}}}\left(e^{-t_k}+\frac{e^{-t_{k-1}}-e^{-t_k}}{1-e^{-t_{k-1}}}\right)^{-1}\|g_{k-1}\|_2\\
&\leq \|\theta_k\|_2 + \frac{1-e^{-t_k}}{1-e^{-t_{k-1}}}\left(e^{-t_k}+\frac{e^{-t_{k-1}}-e^{-t_k}}{1-e^{-t_{k-1}}}\right)^{-1}(c_0-1)^{1/2}\frac{(e^{-t_{k-1}}-e^{-t_k})}{(1-e^{-t_k})}\|\theta_{k-1}\|_2\\
&= \|\theta_k\|_2 + (c_0-1)^{1/2}\frac{1-e^{-\alpha_k}}{1-e^{-t_k}}\|\theta_{k-1}\|_2\,,
\end{aligned}$$

which implies that

$$\|\theta_{k-1}\|_2 \le \frac{\|\theta_k\|_2}{1-(c_0-1)^{1/2}\frac{1-e^{-\alpha_k}}{1-e^{-t_k}}}\,. \tag{74}$$

Using this, we have that

$$\begin{aligned}
\|g_k\|_2 &\le \frac{c_0\beta h(\gamma_1)e^{(2-\gamma_1)t_k}}{2(1-e^{-t_k})^{\gamma_1-1}}\frac{(e^{-t_{k-1}}-e^{-t_k})^2}{(1-e^{-t_{k-1}})^2}\|\theta_{k-1}\|_2^2\\
&\le \frac{c_0\beta h(\gamma_1)e^{(2-\gamma_1)t_k}}{2(1-e^{-t_k})^{\gamma_1-1}}\frac{(e^{-t_{k-1}}-e^{-t_k})^2}{(1-e^{-t_{k-1}})^2}\|\theta_{k-1}\|_2\frac{\|\theta_k\|_2}{1-(c_0-1)^{1/2}\frac{1-e^{-\alpha_k}}{1-e^{-t_k}}}\\
&\le (c_0-1)^{1/2}\frac{(e^{-t_k}-e^{-t_{k+1}})}{(1-e^{-t_{k+1}})}\|\theta_k\|_2\,,
\end{aligned}$$

provided that

$$12\sqrt{2}\beta h(\gamma_1)\frac{c_0(c_0-1)^{-1/2}}{1-(c_0-1)^{1/2}\frac{1-e^{-\alpha_k}}{1-e^{-t_k}}}\frac{e^{t_k}(e^{\alpha_k}-1)}{(e^{t_k}-1)^{\gamma_1}}\|\theta_{k-1}\|_2 \le 1 \tag{75}$$

for any $k \ge 2$, where we have used the fact that

$$\frac{(1-e^{-t_{k+1}})(1-e^{-t_k})(e^{\alpha_k}-1)}{(1-e^{-t_{k-1}})^2(1-e^{-\alpha_{k+1}})} \le 24\sqrt{2}\,,$$

because $\alpha_{k+1} \ge \alpha_k/2$, $\alpha_k \le 2\alpha_{k-1}$, and $\alpha_{k-1} \le \ln(2)$ for any $k \ge 1$ by assumption. Note that (75) can be ensured by the second inequality in (18) if we choose $c_0 = 26/25$.

It remains to check Assumption (A1) through bounding $(\theta_{k+1}-\theta_k)^\top[\nabla^2 L_n(\theta)]^{\gamma_2}(\theta_{k+1}-\theta_k)$, which can be achieved through similar arguments used in the derivations of (69) ignoring the term $\beta(1-e^{-t_{k+1}})$. Recall that for any $0 \le \gamma_2 \le 2$

$$\lambda_{\max}(H_{k+1}^{-1}[\nabla^2 L_n(\theta_k)]^{\gamma_2}H_{k+1}^{-1}) \le \sup_{\lambda:\lambda\ge 0}\frac{\lambda^{\gamma_2}}{((1-e^{-t_{k+1}})\lambda+e^{-t_{k+1}})^2} = \frac{h(\gamma_2)e^{(2-\gamma_2)t_{k+1}}}{4(1-e^{-t_{k+1}})^{\gamma_2}}\,. \tag{76}$$

Then replacing $\lambda_{\max}(J_{k+1})$ in (69) with the above bound, we obtain that for any $k \ge 1$

$$\begin{aligned}
&(\theta_{k+1}-\theta_k)^\top[\nabla^2 L_n(\theta_k)]^{\gamma_2}(\theta_{k+1}-\theta_k)\\
&\le \frac{h(\gamma_2)e^{(2-\gamma_2)t_{k+1}}}{2(1-e^{-t_{k+1}})^{\gamma_2}}\left(\frac{(1-e^{-t_{k+1}})^2}{(1-e^{-t_k})^2}\|g_k\|_2^2+\frac{(e^{-t_k}-e^{-t_{k+1}})^2}{(1-e^{-t_k})^2}\|\theta_k\|_2^2\right)\\
&\le \frac{h(\gamma_2)e^{(2-\gamma_2)t_{k+1}}}{(1-e^{-t_{k+1}})^{\gamma_2}}\frac{c_0(1-e^{-\alpha_{k+1}})^2\|\theta_k\|_2^2}{(e^{t_k}-1)^2} \le \frac{h(\gamma_2)}{4\beta^2h^2(\gamma_1)} \le \frac{1}{\beta^2}\,,
\end{aligned}$$

provided that

$$6\sqrt{c_0}\beta h(\gamma_1)\frac{e^{t_{k+1}}(e^{\alpha_{k+1}}-1)}{(e^{t_{k+1}}-1)^{1+\gamma_2/2}}\|\theta_k\|_2 \le 1 \tag{77}$$

for any $k \ge 1$, which can be ensured by the second inequality in (18). Here we have used the fact that $(1-e^{-t_{k+1}})/(1-e^{-t_k}) \le 3$ if $\alpha_{k+1} \le 2\alpha_k; k \ge 2$.

Moreover, when $k = 1$, using the first equation in (65) and the eigenvalue bound in (76), we obtain that

$$
\begin{aligned}
&(\theta_{k+1} - \theta_k)^\top [\nabla^2 L_n(\theta_k)]^{\gamma_2} (\theta_{k+1} - \theta_k) = \theta_1^\top [\nabla^2 L_n(\mathbf{0})]^{\gamma_2} \theta_1 \\
&= (1 - e^{-\alpha_1})^2 (\nabla L_n(\mathbf{0}))^\top H_1^{-1} [\nabla^2 L_n(\mathbf{0})]^{\gamma_2} H_1^{-1} \nabla L_n(\mathbf{0}) \le \frac{h(\gamma_2) e^{(2-\gamma_2)\alpha_1}}{4(1 - e^{-\alpha_1})^{\gamma_2}} (1 - e^{-\alpha_1})^2 \|\nabla L_n(\mathbf{0})\|_2^2 \\
&= 4^{-1} h(\gamma_2)(e^{\alpha_1} - 1)^{2-\gamma_2} \|\nabla L_n(\mathbf{0})\|_2^2 \le \frac{1}{\beta^2}\,,
\end{aligned}
$$

provided that $\beta(e^{\alpha_1} - 1)^{1-\gamma_2/2} \|\nabla L_n(\mathbf{0})\|_2 \le 1$, which can be ensured by the first condition in (18). This completes the proof of (70). Finally, the bound in (19) follows from (71) and Lemma 2,

$$
\begin{aligned}
\|g_k\|_2 &\le (c_0 - 1)^{1/2} \frac{(e^{-t_k} - e^{-t_{k+1}})}{(1 - e^{-t_k})} \|\theta_k\|_2 \le \frac{e^{-t_k} - e^{-t_k - 2\alpha_k}}{5(1 - e^{-t_k})} \|\theta_k\|_2 \le \frac{2(1 - e^{-\alpha_k})}{5(e^{t_k} - 1)} \|\theta_k\|_2 \\
&\le \frac{(1 - e^{-\alpha_k})}{2(e^{t_k} - 1)} \|\theta(t_k)\|_2\,,
\end{aligned}
$$

where we have used the fact that $\alpha_{k+1} \le 2\alpha_k; k \ge 1$. This completes the proof of Theorem 3.

Proof of Corollary 2. By using (64), we have that $\|\theta_1\|_2 = (1 - e^{-\alpha_1}) \|H_1^{-1} \nabla L_n(\mathbf{0})\|_2 \le (e^{\alpha_1} - 1) \|\nabla L_n(\mathbf{0})\|_2$. Using this and (72), it follows that,

$$
\|g_1\|_2 \le \frac{C_1}{15} \beta e^{-\alpha_1} (e^{\alpha_1} - 1)^{\max(1, 2-\gamma_1)} \|\nabla L_n(\mathbf{0})\|_2 \|\theta_1\|_2 \le \frac{C_1}{15} \beta (e^{\alpha_1} - 1)^{\max(2, 3-\gamma_1)} \|\nabla L_n(\mathbf{0})\|_2^2\,,
$$

which proves the first bound in (20).

Next, we turn to the proof of the second bound in (20). Using (70) with $c_0 = 26/25$, (82) in Lemma 2, and the fact that $(1 - e^{-t_{k+1}})/(1 - e^{-t_k}) \le 3$ when $\alpha_{k+1} \le 2\alpha_k; k \ge 1$, we obtain that

$$
\begin{aligned}
\|g_{k+1}\|_2 &\le \frac{c_0 \beta h(\gamma_1) e^{(2-\gamma_1)t_{k+1}}}{2(1 - e^{-t_{k+1}})^{\gamma_1 - 1}} \frac{(e^{-t_k} - e^{-t_{k+1}})^2}{(1 - e^{-t_k})^2} \|\theta_k\|_2^2 \\
&\le \frac{52 \beta e^{(2-\gamma_1)t_{k+1}}}{25(1 - e^{-t_{k+1}})^{\gamma_1 - 1}} \frac{3^2 (e^{-t_k} - e^{-t_{k+1}})^2}{(1 - e^{-t_{k+1}})^2} \left(\frac{5}{4}\right)^2 \|\theta(t_k)\|_2^2 \\
&\le \frac{30 \beta e^{-\gamma_1 t_{k+1}} (e^{\alpha_{k+1}} - 1)^2}{(1 - e^{-t_{k+1}})^{\gamma_1 - 1}} \frac{\|\theta(t_k)\|_2^2}{(1 - e^{-t_{k+1}})^2} \\
&\le \frac{30 \beta e^{-\gamma_1 t_{k+1}} (e^{\alpha_{k+1}} - 1)^2}{(1 - e^{-t_{k+1}})^{\gamma_1 - 1}} \left(\|\theta(t_k)\|_2 + \|\nabla L_n(\mathbf{0})\|_2\right)^2
\end{aligned}
$$

for any $k \ge 1$, where the last inequality uses Lemma 1. This proves the second bound in (20). This completes the proof of Corollary 2.

Proof of Theorem 4. Using (13) and (71), we have that

$$
\sup_{t \in [0, t_1]} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \le \max\left((e^{\alpha_1} - 1)^2 \|\nabla L_n(\mathbf{0})\|_2^2,\ \|\theta_1\|_2^2 \right)\,. \tag{78}
$$

Using (71), we have that

$$e^{t_{k+1}}\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2\|g_k\|_2^2 \le e^{t_{k+1}}\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2\left((c_0-1)^{1/2}\frac{(e^{-t_k}-e^{-t_{k+1}})}{(1-e^{-t_{k+1}})}\|\theta_k\|_2\right)^2$$
$$\le (e^{-t_k}-e^{-t_{k+1}})^2\frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2}$$

and

$$\begin{aligned}e^{t_{k+1}}\|g_{k+1}\|_2^2 &\le e^{t_{k+1}}\left((c_0-1)^{1/2}\frac{(e^{-t_{k+1}}-e^{-t_{k+2}})}{(1-e^{-t_{k+2}})}\|\theta_{k+1}\|_2\right)^2\\ &\le 4^{-1}\frac{(e^{\alpha_{k+2}}-1)^2}{(e^{\alpha_{k+1}}-1)^2}e^{-2\alpha_{k+2}+\alpha_{k+1}}(e^{-t_k}-e^{-t_{k+1}})^2\frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\\ &\le (e^{-t_k}-e^{-t_{k+1}})^2\frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\,,\end{aligned}$$

where we have used the fact that $\alpha_{k+2}\le 2\alpha_{k+1}$ and

$$\frac{(e^{\alpha_{k+2}}-1)^2}{(e^{\alpha_{k+1}}-1)^2}e^{-2\alpha_{k+2}+\alpha_{k+1}} \le (e^{\alpha_{k+1}}+1)^2 e^{-3\alpha_{k+1}} \le 4\,.$$

Combining, we obtain that

$$e^{t_{k+1}}\max\left\{\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2\|g_k\|_2^2,\ \|g_{k+1}\|_2^2\right\} \le (e^{-t_k}-e^{-t_{k+1}})^2\max\left\{\frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2},\ \frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\right\}$$

Combining this with (14) in Theorem 2, it follows that

$$\sup_{t\in[t_k,t_{k+1}]}\left\{f_t(\tilde\theta(t))-f_t(\theta(t))\right\} \le 2(e^{-t_k}-e^{-t_{k+1}})^2\max\left\{\frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2},\ \frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\right\}\,.$$

Therefore,

$$\begin{aligned}&\max_{1\le k\le N-1}\sup_{t\in[t_k,t_{k+1}]}\left\{f_t(\tilde\theta(t))-f_t(\theta(t))\right\}\\ &\qquad\le \max_{1\le k\le N-1}2(e^{-t_k}-e^{-t_{k+1}})^2\max\left\{\frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2},\ \frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\right\}\\ &\qquad\le 2\max_{1\le k\le N}\left\{\frac{e^{-t_k}\|\theta_k\|_2^2}{(1-e^{-t_k})^2}\max\left(e^{-\alpha_{k+1}}(e^{\alpha_{k+1}}-1)^2, e^{-\alpha_k}(e^{\alpha_k}-1)^2\right)\right\}\\ &\qquad\le 8\max_{1\le k\le N}\left\{\frac{e^{-t_k}(e^{\alpha_{k+1}}-1)^2\|\theta_k\|_2^2}{(1-e^{-t_k})^2}\right\}\,. \qquad (79)\end{aligned}$$

Combining this with (78), we have that

$$\begin{aligned}\sup_{0\le t\le t_{\max}}\left\{f_t(\tilde\theta(t))-f_t(\theta(t))\right\} &\le \sup_{0\le t\le t_N}\left\{f_t(\tilde\theta(t))-f_t(\theta(t))\right\}\\ &\le 8\max\left\{(e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2,\ \max_{1\le k\le N}\left(\frac{e^{-t_k}(e^{\alpha_{k+1}}-1)^2\|\theta_k\|_2^2}{(1-e^{-t_k})^2}\right)\right\}\,,\end{aligned}$$

if $t_{N-1} \le t_{\max} \le t_N$ for some $N \ge 1$. This proves (21).

Lastly, when $0 < t_N \le t_{\max}$, using (71), we obtain that

$$\begin{aligned}\frac{e^{t_N}\|g_N\|_2^2}{1-e^{-t_N}} &\le \frac{e^{t_N}}{1-e^{-t_N}}\left((c_0-1)^{1/2}\frac{(e^{-t_N}-e^{-t_{N+1}})}{(1-e^{-t_{N+1}})}\|\theta_N\|_2\right)^2 \\ &\le \frac{(1-e^{-\alpha_{N+1}})^2}{4(1-e^{-t_{N+1}})^2(e^{t_N}-1)}\|\theta_N\|_2^2 \le \frac{\|\theta_N\|_2^2}{4(e^{t_N}-1)}\,.\end{aligned}$$

Combining this with (15) in Theorem 2, we obtain that when $t_N \le t_{\max}$,

$$\sup_{t_N < t \le t_{\max}}\left\{f_t(\tilde\theta(t)) - f_t(\theta(t))\right\} \le \frac{\|\theta_N\|_2^2}{4(e^{t_N}-1)} + \frac{3\|\theta(t_{\max})\|_2^2}{2(e^{t_N}-1)}\,. \tag{80}$$

Combining (79) and (80), we obtain that when $t_{\max} = \infty$

$$\begin{aligned}\sup_{0\le t\le t_{\max}}\left\{f_t(\tilde\theta(t)) - f_t(\theta(t))\right\} \le \max\Bigg\{&8(e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2, \\ 8\max_{1\le k\le N-1} e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2\|\theta_k\|_2^2,\ &\frac{2\max(\|\theta(t_{\max})\|_2^2, \|\theta_N\|_2^2)}{(e^{t_N}-1)}\Bigg\},\end{aligned}$$

which implies (22). This completes the proof of Theorem 4.

Next, we present a supporting lemma to be used in the proof of Theorem 5.

Lemma 2 *Under the assumptions in Theorem 3, we have that*

$$\|\theta_1\|_2 \le (e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2 \text{ and} \tag{81}$$

$$\frac{1}{1+(c_0-1)^{1/2}} \le \frac{\|\theta_k\|_2}{\|\theta(t_k)\|_2} \le \frac{1}{1-(c_0-1)^{1/2}}\,, \tag{82}$$

where $c_0 = 26/25$.

Proof of Lemma 2. For the first inequality, using (64), we have that

$$\begin{aligned}\|\theta_1\|_2 = \|(1-e^{-\alpha_1})H_1^{-1}\nabla L_n(\mathbf{0})\|_2 &\le \lambda_{\max}(H_1^{-1})(1-e^{-\alpha_1})\|\nabla L_n(\mathbf{0})\|_2 \\ &\le e^{\alpha_1}(1-e^{-\alpha_1})\|\nabla L_n(\mathbf{0})\|_2 = (e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2,\end{aligned}$$

which proves (81).

We next prove (82). Using (71) and (45), we have that $\|\theta_k - \theta(t_k)\|_2 \le e^{t_k}\|g_k\|_2$ and

$$\begin{aligned}\|\theta_k\|_2 &\le \|\theta(t_k)\|_2 + \|\theta(t_k) - \theta_k\|_2 \le \|\theta(t_k)\|_2 + e^{t_k}\|g_k\|_2 \\ &\le \|\theta(t_k)\|_2 + (c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}\|\theta_k\|_2\,,\end{aligned}$$

which implies that

$$\|\theta_k\|_2 \le \frac{\|\theta(t_k)\|_2}{1-(c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}}\,.$$

Similarly, we also have

$$\|\theta(t_k)\|_2 \le \|\theta_k\|_2 + \|\theta(t_k) - \theta_k\|_2 \le \|\theta_k\|_2 + (c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}\|\theta_k\|_2\,.$$

Combining, we obtain that

$$\frac{1}{1+(c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}} \le \frac{\|\theta_k\|_2}{\|\theta(t_k)\|_2} \le \frac{1}{1-(c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}}\,,$$

which implies (82). This completes the proof of Lemma 2.

Proof of Theorem 5. We first verify α_1 and α_{k+1} satisfy the conditions in (18). By the definitions of α_1 and α_{k+1}, we only need to prove that $\alpha_{k+1} \ge \alpha_k/2$.

If $\alpha_{k+1} = \alpha_{\max} = 1/10$ or $2\alpha_k$, then trivially $\alpha_{k+1} \ge \alpha_k/2$. Hence, we only need to consider the case $\alpha_{k+1} = A_k$ or B_k, where

$$\begin{aligned}
A_k &:= \ln\left(1+\frac{e^{t_k/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_k})}{\|\theta_k\|_2}\right)\,,\\
B_k &:= \ln\left(1+\left(C_2\beta e^{t_k}\max\left((e^{t_k}-1)^{-\gamma_1},(e^{t_k}-1)^{-1-\gamma_2/2}\right)\|\theta_k\|_2\right)^{-1}\right)\,.
\end{aligned}$$

It is easy to check that $A_k \le B_k$ if and only if

$$C_2\beta\|\nabla L_n(\mathbf{0})\|_2(e^{\alpha_1}-1)e^{t_k/2}\max\left((e^{t_k}-1)^{1-\gamma_1},(e^{t_k}-1)^{-\gamma_2/2}\right) \le 1\,. \tag{83}$$

We first show that $A_k \le B_k$ when $e^{t_k} \le 2$. Equivalently, we need to show that the above inequality (83) holds if $e^{t_k} \le 2$. To this end, we consider two cases: (i) $\gamma_1 \ge 1$; and (ii) $\gamma_1 < 1$. It is easy to see that for any $s \ge 0$, function $x/(x^2-1)^s$ with $e^{\alpha_1/2} \le x \le \sqrt{2}$ achieves its maximum at the two boundary points, that is,

$$x/(x^2-1)^s \le \max\left(\sqrt{2}, e^{\alpha_1/2}/(e^{\alpha_1}-1)^s\right)\,. \tag{84}$$

For case (i), note that when $e^{t_k} \le 2$,

$$\begin{aligned}
e^{t_k/2}\max\left((e^{t_k}-1)^{1-\gamma_1},(e^{t_k}-1)^{-\gamma_2/2}\right) &= \frac{e^{t_k/2}}{(e^{t_k}-1)^{\max(\gamma_1-1,\gamma_2/2)}}\\
&\le \max\left(\sqrt{2}, e^{\alpha_1/2}/(e^{\alpha_1}-1)^{\max(\gamma_1-1,\gamma_2/2)}\right)
\end{aligned}$$

where we have used (84). Hence, inequality (83) holds if

$$C_2 e^{\alpha_1/2}\beta\|\nabla L_n(\mathbf{0})\|_2(e^{\alpha_1}-1)^{\min(2-\gamma_1,1-\gamma_2/2)} \le 1 \text{ and } \sqrt{2}C_2\beta\|\nabla L_n(\mathbf{0})\|_2(e^{\alpha_1}-1) \le 1\,,$$

both of which can be ensured by (27).

Similarly, for case (ii), note that when $e^{t_k} \le 2$,

$$e^{t_k/2}\max\left((e^{t_k}-1)^{1-\gamma_1},(e^{t_k}-1)^{-\gamma_2/2}\right) = \frac{e^{t_k/2}}{(e^{t_k}-1)^{\gamma_2/2}} \le \max\left(\sqrt{2}, e^{\alpha_1/2}/(e^{\alpha_1}-1)^{\gamma_2/2}\right)\,,$$

where we have used (84). Hence, inequality (83) holds if

$$C_2 e^{\alpha_1/2}\beta\|\nabla L_n(\mathbf{0})\|_2(e^{\alpha_1}-1)^{1-\gamma_2/2}\leq 1 \text{ and } \sqrt{2}C_2\beta\|\nabla L_n(\mathbf{0})\|_2(e^{\alpha_1}-1)\leq 1\,,$$

both of which can be ensured by (27). This completes the proof that $A_k \leq B_k$ when $e^{t_k}\leq 2$.

Since $e^{\alpha_1}\leq 2$, we only need to consider the case $\alpha_2 = A_1$ when $k=1$. To show that $e^{\alpha_2}-1\geq e^{\alpha_1/2}$, we note that

$$\begin{aligned}\frac{e^{\alpha_2}-1}{e^{\alpha_1/2}-1} &= \frac{e^{A_1}-1}{e^{\alpha_1/2}-1} = \frac{e^{t_1/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_1})}{\|\theta_1\|_2(e^{\alpha_1/2}-1)}\\ &\geq \frac{e^{\alpha_1/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_1})}{(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(e^{\alpha_1/2}-1)} = e^{-\alpha_1/2}(e^{\alpha_1/2}+1) > 1\,,\end{aligned}$$

where we have used (81) in Lemma 2. This proves that $\alpha_2 \geq \alpha_1/2$.

For $k\geq 2$, we consider two cases (i) $\alpha_{k+1}=A_k$; and (ii) $\alpha_{k+1}=B_k$. For case (i), using the fact that $\alpha_k \leq A_{k-1}$, we have

$$\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k/2}-1} = (e^{\alpha_k/2}+1)\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k}-1} \geq (e^{\alpha_k/2}+1)\frac{e^{A_k}-1}{e^{A_{k-1}}-1} \tag{85}$$

$$\begin{aligned}&= (e^{\alpha_k/2}+1)e^{\alpha_k/2}\frac{(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_k})}{\|\theta_k\|_2}\frac{\|\theta_{k-1}\|_2}{(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_{k-1}})}\\ &= (e^{-\alpha_k/2}+1)\frac{(e^{t_k}-1)\|\theta_{k-1}\|_2}{(e^{t_{k-1}}-1)\|\theta_k\|_2}\,.\end{aligned} \tag{86}$$

Applying Lemma 2, we have that

$$\begin{aligned}\frac{(e^{t_k}-1)\|\theta_{k-1}\|_2}{(e^{t_{k-1}}-1)\|\theta_k\|_2} &\geq \frac{(e^{t_k}-1)\|\theta(t_{k-1})\|_2}{(e^{t_{k-1}}-1)\|\theta(t_k)\|_2}\frac{1-(c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}}{1+(c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}}\\ &\geq \frac{1-(c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}}{1+(c_0-1)^{1/2}\frac{(1-e^{-\alpha_{k+1}})}{(1-e^{-t_{k+1}})}} \geq \frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}}\,,\end{aligned} \tag{87}$$

where the last inequality uses part (ii) of Corollary 1. Combining this with (86), we obtain that

$$\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k/2}-1} \geq (e^{-\alpha_k/2}+1)\frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}} \geq 1\,, \tag{88}$$

because $\alpha_k \leq 10^{-1}$ and $c_0 = 26/25$. This proves case (i).

For case (ii), we have $\alpha_{k+1}=B_k$ and $\alpha_k\leq B_{k-1}$. Since we have shown that $\alpha_{k+1}=A_k$ when $e^{t_k}>2$, we must have that $e^{t_k}\leq 2$ and $e^{t_{k-1}} > 2e^{-\alpha_k}$. Using these, we have that

$$\begin{aligned}\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k/2}-1} &= (e^{\alpha_k/2}+1)\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k}-1} \geq (e^{\alpha_k/2}+1)\frac{e^{B_k}-1}{e^{B_{k-1}}-1}\\ &= (e^{\alpha_k/2}+1)\frac{e^{t_{k-1}}\max\left((e^{t_{k-1}}-1)^{-\gamma_1},(e^{t_{k-1}}-1)^{-1-\gamma_2/2}\right)\|\theta_{k-1}\|_2}{e^{t_k}\max\left((e^{t_k}-1)^{-\gamma_1},(e^{t_k}-1)^{-1-\gamma_2/2}\right)\|\theta_k\|_2}\\ &\geq (e^{\alpha_k/2}+1)\frac{\max\left((e^{t_{k-1}}-1)^{1-\gamma_1},(e^{t_{k-1}}-1)^{-\gamma_2/2}\right)}{e^{\alpha_k}\max\left((e^{t_k}-1)^{1-\gamma_1},(e^{t_k}-1)^{-\gamma_2/2}\right)}\frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}}\,.\end{aligned}$$

where we have used (87). Now if $e^{t_{k-1}} \geq 2$, then

$$\begin{aligned}
\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k/2}-1} &\geq (e^{\alpha_k/2}+1)\frac{(e^{t_{k-1}}-1)^{\max(1-\gamma_1,-\gamma_2/2)}}{e^{\alpha_k}(e^{t_k}-1)^{\max(1-\gamma_1,-\gamma_2/2)}}\frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}} \\
&\geq (e^{\alpha_k/2}+1)\frac{(e^{t_{k-1}}-1)}{e^{\alpha_k}(e^{t_k}-1)}\frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}} \\
&\geq \frac{(e^{\alpha_k/2}+1)e^{-\alpha_k}}{e^{\alpha_k}+(e^{\alpha_k}-1)/(e^{t_{k-1}}-1)}\frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}} \\
&\geq \frac{(e^{\alpha_k/2}+1)e^{-\alpha_k}}{2e^{\alpha_k}-1}\frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}} \geq 1\,,
\end{aligned}$$

where we have used the fact that the last inequality holds when $\alpha_k \leq 10^{-1}$ and $c_0 = 26/25$. If $e^{t_{k-1}} < 2$, we have $e^{t_k} < 2e^{\alpha_k}$ and

$$\begin{aligned}
\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k/2}-1} &\geq (e^{\alpha_k/2}+1)\frac{(e^{t_{k-1}}-1)^{\min(1-\gamma_1,-\gamma_2/2)}}{e^{\alpha_k}(e^{t_k}-1)^{\max(1-\gamma_1,-\gamma_2/2)}}\frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}} \\
&\geq (e^{\alpha_k/2}+1)\frac{(2e^{-\alpha_k}-1)}{e^{\alpha_k}(2e^{\alpha_k}-1)}\frac{1-(c_0-1)^{1/2}}{1+(c_0-1)^{1/2}} \geq 1\,,
\end{aligned}$$

where we have used the fact that the last inequality holds when $\alpha_k \leq 10^{-1}$ and $c_0 = 26/25$. This completes the proof of $\alpha_{k+1} \geq \alpha_k/2$.

Next, we show that the algorithm terminates after a finite number of steps. We first show that t_k diverges. To this end, using (60) and (82), we have

$$\begin{aligned}
&\frac{e^{t_k/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_k})}{\|\theta_k\|_2} \geq \frac{e^{t_k/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_k})}{\|\theta(t_k)\|_2/(1-(c_0-1)^{1/2})} \\
&\geq \frac{e^{t_k/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_k})(1-(c_0-1)^{1/2})}{(e^{t_k}-1)\|\nabla L_n(\mathbf{0})\|_2} \geq 2^{-1}(e^{\alpha_1}-1)e^{-t_k/2}\,,
\end{aligned}$$

which implies that

$$A_k \geq \ln(1+2^{-1}(e^{\alpha_1}-1)e^{-t_k/2}) \geq \frac{(e^{\alpha_1}-1)e^{-t_k/2}}{2+(e^{\alpha_1}-1)e^{-t_k/2}} = \frac{(e^{\alpha_1}-1)}{2e^{t_k/2}+(e^{\alpha_1}-1)}\,. \tag{89}$$

Moreover, $B_k \geq A_k$ when $e^{t_k} < 2$, and when $e^{t_k} \geq 2$, we have

$$\begin{aligned}
&e^{t_k}\max\Big((e^{t_k}-1)^{-\gamma_1},(e^{t_k}-1)^{-1-\gamma_2/2}\Big)\|\theta_k\|_2 \\
&\leq 2e^{t_k}\max\Big((e^{t_k}-1)^{1-\gamma_1},(e^{t_k}-1)^{-\gamma_2/2}\Big)\|\nabla L_n(\mathbf{0})\|_2 \\
&\leq 2e^{2t_k}\|\nabla L_n(\mathbf{0})\|_2\,,
\end{aligned}$$

which implies that when $e^{t_k} \geq 2$

$$\begin{aligned}
B_k &\geq \ln(1+(C_1\beta e^{2t_k}\|\nabla L_n(\mathbf{0})\|_2)^{-1}) \geq \frac{(2C_1\beta e^{2t_k}\|\nabla L_n(\mathbf{0})\|_2)^{-1}}{1+(2C_1\beta e^{2t_k}\|\nabla L_n(\mathbf{0})\|_2)^{-1}} \\
&= \frac{1}{1+(2C_1\beta e^{2t_k}\|\nabla L_n(\mathbf{0})\|_2)}\,.
\end{aligned}$$

Thus,

$$B_k \geq \min\left\{\frac{(e^{\alpha_1}-1)}{2e^{t_k/2}+(e^{\alpha_1}-1)}, \frac{1}{1+(2C_1\beta e^{2t_k}\|\nabla L_n(\mathbf{0})\|_2)}\right\}.$$

Combining we have that

$$\alpha_{k+1} \geq \min\left(\alpha_{\max}, 2\alpha_k, \frac{(e^{\alpha_1}-1)}{2e^{t_k/2}+(e^{\alpha_1}-1)}, \frac{1}{1+2C_1\beta e^{2t_k}\|\nabla L_n(\mathbf{0})\|_2}\right). \tag{90}$$

Now we prove the divergence of t_k by contradiction. Suppose that t_k does not diverge. Then there must exist a constant T such that $t_k < T$ for all k. However, now we have

$$\alpha_{k+1} \geq \min\left(\alpha_{\max}, 2\alpha_k, \frac{(e^{\alpha_1}-1)}{2e^{T/2}+(e^{\alpha_1}-1)}, \frac{1}{1+2C_1\beta e^{2T}\|\nabla L_n(\mathbf{0})\|_2}\right),$$

which implies that α_k is lower bounded by a positive constant when k is large enough, implying that t_k should diverge. This is a contradiction. Hence, t_k diverges.

Now we are ready to show that the algorithm must terminate after a finite number of iterations. If $t_{\max} < \infty$, then $t_k \geq t_{\max}$ must hold for large enough k as t_k diverges. If $t_{\max} = \infty$, then we have that $\theta(t_{\max}) = \theta^\star$ is finite by assumption. Therefore, the termination criterion in (25) should also be met when N is large enough, because t_N diverges and

$$\frac{\max(\|\theta(t_{\max})\|_2^2, \|\theta_N\|_2^2)}{(e^{t_N}-1)} \leq \frac{2\max(\|\theta(t_{\max})\|_2^2, \|\theta(t_N)\|_2^2)}{(e^{t_N}-1)} \leq \frac{2\|\theta^\star\|_2^2}{(e^{t_N}-1)} \to 0 \text{ as } N \to \infty.$$

Finally, we are ready to prove (29) after the algorithm is terminated. Upon termination when $k+1 = N$, we have one of the two conditions in (25) must hold. If $t_N > t_{\max}$, it is easy to see that the step sizes defined in (27) and (28) satisfy all the assumptions in Theorem 4, by using (21) in Theorem 4 and the definition of α_k, we have that

$$\sup_{0\leq t\leq t_{\max}} \left\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\right\} \lesssim (e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2 \leq \epsilon.$$

If the second inequality in (25) holds, then

$$\begin{aligned}\frac{2\max(\|\theta(t_{\max})\|_2^2, \|\theta_N\|_2^2)}{(e^{t_N}-1)} &\leq \frac{2\|\theta(t_{\max})\|_2^2}{(e^{t_N}-1)} \leq \frac{2\|\theta_N\|_2^2}{(e^{t_N}-1)}\frac{\|\theta(t_{\max})\|_2^2}{\|\theta_N\|_2^2} \\ &\leq \frac{2\|\theta(t_{\max})\|_2^2}{\|\theta(t_N)\|_2^2}(e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2.\end{aligned}$$

Combining this with (22) in Theorem 4, we obtain that

$$\sup_{0\leq t\leq t_{\max}} \left\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\right\} \lesssim \frac{\|\theta(t_{\max})\|_2^2}{\|\theta(t_N)\|_2^2}(e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2 \leq \epsilon.$$

This completes the proof of Theorem 5.

We next present a supporting lemma for the proof of Theorem 6 and 10.

Lemma 3 *For any $\epsilon \in (0,1]$, let $(e^{\alpha_1}-1)^2 = c_1^2\epsilon$ and $\alpha_{k+1} = \ln(1+c_2 e^{t_k/2}(e^{\alpha_1}-1))$, where $c_1, c_2 > 0$ are constants and $t_k = \sum_{i=1}^k \alpha_i$. Define $k^\star := \max\{k : t_k \le \ln(\epsilon^{-1})\}$. Then,*

$$k^\star < \frac{1+c_1c_2+\sqrt{1+c_1c_2}}{c_1c_2\sqrt{\epsilon}\sqrt{1+c_1\sqrt{\epsilon}}}, \sum_{k=1}^{k^\star} e^{t_k}(t_k+1) \le \frac{(1+2c_1c_2+\sqrt{1+c_1c_2})(1+\ln(\epsilon^{-1}))}{c_1c_2\epsilon}. \quad (91)$$

Proof of Lemma 3. It is easy to see that α_k is strictly increasing for $k \ge 2$ and by definition of $k^\star$, we have that

$$\alpha_{k+1} \le \ln(1+c_1c_2) \text{ for any } 1 \le k \le k^\star, \quad (92)$$

which implies that $e^{\alpha_{k+1}/2} \le (1+c_1c_2)^{1/2}$ for any $1 \le k \le k^\star$. Moreover, since $e^{\alpha_{k+1}} = 1 + c_1c_2\sqrt{\epsilon}e^{t_k/2}$, we have that

$$e^{-t_k/2} - e^{-t_{k+1}/2} = \frac{e^{\alpha_{k+1}/2}-1}{e^{t_{k+1}/2}} = \frac{e^{-t_{k+1}/2}(e^{\alpha_{k+1}}-1)}{e^{\alpha_{k+1}/2}+1} = \frac{e^{-\alpha_{k+1}/2}c_1c_2\sqrt{\epsilon}}{e^{\alpha_{k+1}/2}+1} = \frac{c_1c_2\sqrt{\epsilon}}{e^{\alpha_{k+1}}+e^{\alpha_{k+1}/2}}$$

for any $k \ge 1$. Therefore, for any $1 \le k \le k^\star$,

$$\begin{aligned} e^{-t_1/2} - e^{-t_{k^\star+1}/2} &= \sum_{i=1}^{k^\star}\left(e^{-t_i/2} - e^{-t_{i+1}/2}\right) = \sum_{i=1}^{k^\star} \frac{c_1c_2\sqrt{\epsilon}}{e^{\alpha_{i+1}}+e^{\alpha_{i+1}/2}} \ge \frac{k^\star c_1c_2\sqrt{\epsilon}}{1+c_1c_2+\sqrt{1+c_1c_2}}, \\ e^{-t_k/2} - e^{-t_{k^\star}/2} &= \sum_{i=k}^{k^\star-1}\left(e^{-t_i/2} - e^{-t_{i+1}/2}\right) = \sum_{i=k}^{k^\star-1} \frac{c_1c_2\sqrt{\epsilon}}{e^{\alpha_{i+1}}+e^{\alpha_{i+1}/2}} \ge \frac{(k^\star-k)c_1c_2\sqrt{\epsilon}}{1+c_1c_2+\sqrt{1+c_1c_2}}, \end{aligned}$$

which implies the first inequality in (91) and

$$e^{-t_k/2} \ge e^{-t_{k^\star}/2} + \frac{(k^\star-k)c_1c_2\sqrt{\epsilon}}{1+c_1c_2+\sqrt{1+c_1c_2}} \ge \sqrt{\epsilon} + \frac{c_1c_2\sqrt{\epsilon}}{1+c_1c_2+\sqrt{1+c_1c_2}}(k^\star-k) \quad (93)$$

by using the fact that $t_{k^\star} \le \ln(\epsilon^{-1})$. Therefore, for any $1 \le k \le k^\star$,

$$e^{t_k} \le \frac{\epsilon^{-1}}{(1+C(k^\star-k))^2}, \quad (94)$$

where $C = \frac{c_1c_2}{1+c_1c_2+\sqrt{1+c_1c_2}} < 1$. Now we are ready to prove the second inequality in (91). By (94) and the fact that $t_{k^\star} \le \ln(\epsilon^{-1})$, it follows that

$$\begin{aligned} \sum_{k=1}^{k^\star} e^{t_k}t_k &\le e^{t_{k^\star}}t_{k^\star} + \sum_{k=1}^{k^\star-1} \epsilon^{-1}\frac{1}{(1+C(k^\star-k))^2}\ln\left(\epsilon^{-1}\frac{1}{(1+C(k^\star-k))^2}\right) \\ &\le \epsilon^{-1}\ln(\epsilon^{-1}) + \int_1^{k^\star} \epsilon^{-1}\frac{1}{(1+C(k^\star-x))^2}\ln\left(\epsilon^{-1}\frac{1}{(1+C(k^\star-x))^2}\right)dx \\ &= \epsilon^{-1}\ln(\epsilon^{-1}) + \frac{1}{2C\sqrt{\epsilon}}\int_a^b te^{t/2}dt = \epsilon^{-1}\ln(\epsilon^{-1}) + \frac{1}{C\sqrt{\epsilon}}((b-2)e^{b/2}-(a-2)e^{a/2}) \end{aligned}$$

where $a = \ln(\epsilon^{-1}) - 2\ln(1 + C(k^\star - 1))$ and $b = \ln(\epsilon^{-1})$. Using (94) with $k = 1$ and the fact that function $(t-2)e^{t/2}$ is increasing over $t \geq 0$, we have that

$$a = \ln(\epsilon^{-1}) - 2\ln(1 + C(k^\star - 1)) \geq t_1 > 0 \text{ and } (a-2)e^{a/2} \geq -2\,.$$

Hence,

$$\frac{(b-2)e^{b/2} - (a-2)e^{a/2}}{C\sqrt{\epsilon}} \leq \frac{(b-2)e^{b/2}}{C\sqrt{\epsilon}} + \frac{2}{C\sqrt{\epsilon}} = \frac{\epsilon^{-1}(\ln(\epsilon^{-1}) - 2)}{C} + \frac{2}{C\sqrt{\epsilon}} \leq \frac{\epsilon^{-1}\ln(\epsilon^{-1})}{C}\,,$$

provided that $\epsilon < 1$. Combining, we obtain that

$$\sum_{k=1}^{k^\star} e^{t_k} t_k \leq \epsilon^{-1}\ln(\epsilon^{-1}) + \frac{\epsilon^{-1}\ln(\epsilon^{-1})}{C} = \frac{(C+1)\epsilon^{-1}\ln(\epsilon^{-1})}{C}\,.$$

Similarly, by using (94), we have that

$$\begin{aligned}
\sum_{k=1}^{k^\star} e^{t_k} &\leq \epsilon^{-1} + \sum_{k=1}^{k^\star - 1} \epsilon^{-1} \frac{1}{(1 + C(k^\star - k))^2} \leq \epsilon^{-1} + \int_1^{k^\star} \frac{\epsilon^{-1}}{(1 + C(k^\star - x))^2}\, dx \\
&= \epsilon^{-1} + \frac{\epsilon^{-1}}{C}\left(1 - \frac{1}{1 + C(k^\star - 1)}\right) \leq \frac{(C+1)\epsilon^{-1}}{C}\,.
\end{aligned}$$

Consequently, we have that

$$\sum_{k=1}^{k^\star} e^{t_k}(t_k + 1) \leq \frac{(C+1)\epsilon^{-1}(1 + \ln(\epsilon^{-1}))}{C} = \frac{(1 + 2c_1c_2 + \sqrt{1 + c_1c_2})\epsilon^{-1}(1 + \ln(\epsilon^{-1}))}{c_1c_2}\,.$$

This completes the proof of Lemma 3.

Proof of Theorem 6. We first consider the case where $t_{\max} = \infty$ and $\gamma_1 \geq 1$. In view of the termination criterion (25), the algorithm will be terminated when $t_k = \mathcal{O}(\ln(\epsilon^{-1}))$. We define $N = \max\{k : t_k \leq \ln(\epsilon^{-1})\}$. By applying Lemma 1 and 2 with $c_0 = 26/25$, we have

$$\frac{\|\theta_k\|_2}{1 - e^{-t_k}} \leq \frac{5}{4}\frac{\|\theta(t_k)\|_2}{1 - e^{-t_k}} \leq \frac{5}{4}\left(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2\right)\,. \tag{95}$$

Therefore,

$$\begin{aligned}
e^{A_k} - 1 &= \frac{e^{t_k/2}(e^{\alpha_1} - 1)\|\nabla L_n(\mathbf{0})\|_2(1 - e^{-t_k})}{\|\theta_k\|_2} \geq \frac{4e^{t_k/2}(e^{\alpha_1} - 1)\|\nabla L_n(\mathbf{0})\|_2}{5(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2)}\,, \\
e^{B_k} - 1 &= \left(C_2\beta e^{t_k} \max\left((e^{t_k} - 1)^{-\gamma_1}, (e^{t_k} - 1)^{-1-\gamma_2/2}\right)\|\theta_k\|_2\right)^{-1} \\
&\geq \left(C_2\beta \max\left((e^{t_k} - 1)^{1-\gamma_1}, (e^{t_k} - 1)^{-\gamma_2/2}\right)(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2)\right)^{-1} \\
&\geq \left(C_2\beta(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2)\right)^{-1}\,,
\end{aligned}$$

which implies that,

$$\alpha_{k+1} \geq \ln(1 + \nu_1 e^{t_k/2}(e^{\alpha_1} - 1))$$

when $\alpha_{k+1} \leq \min(\alpha_{\max}, \ln(1+\nu_2))$, where we treat

$$\nu_1 = \frac{4\|\nabla L_n(\mathbf{0})\|_2}{5(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2)} \text{ and } \nu_2 = (C_2\beta(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2))^{-1}$$

as problem-dependent constants. Then, applying Lemma 3, we have that

$$N \leq \mathcal{O}\left(\epsilon^{-1/2} + \ln(\epsilon^{-1})/\min(\alpha_{\max}, \ln(1+\nu_2))\right) = \mathcal{O}\left(\epsilon^{-1/2}\right),$$

if we treat β, $\|\theta(t_{\max})\|_2$, and $\|\nabla L_n(\mathbf{0})\|_2$ as constants. When $t_{\max} < \infty$ and $\gamma_1 \geq 1$, the algorithm terminates at N if $t_N > t_{\max}$. Since α_k is increasing, it follows that $N \leq t_{\max}/\alpha_1 = \mathcal{O}\left(\epsilon^{-1/2}\right)$. This completes the proof of Theorem 6.

Proof of Theorem 7. It is easy to verify that $f_{t_k}(\theta)$ is m_k-strongly convex with L_k-Lipschitz gradient, where $m_k = m(1-e^{-t_k})+e^{-t_k}$ and $L_k = L(1-e^{-t_k})+e^{-t_k}$. By standard analysis of gradient descent for strongly convex and smooth functions (see, e.g., Theorem 2.1.14 of Nesterov, 1998), we have that

$$\|\theta_{k+1} - \theta(t_{k+1})\|_2 \quad \leq \quad \left(1 - \frac{2m_{k+1}L_{k+1}}{m_{k+1}+L_{k+1}}\eta_{k+1}\right)^{n_{k+1}} \|\theta_k - \theta(t_{k+1})\|_2 \,, \tag{96}$$

where $\eta_{k+1} \leq \frac{2}{m_{k+1}+L_{k+1}}$. Similar to the derivation of (48), we obtain that

$$\|\theta(t') - \theta(t)\|_2 \leq \frac{|C(t') - C(t)|}{C(t) + mC(t)C(t')}\|\theta(t)\|_2 \,, \tag{97}$$

for any $t' < t$, because $C(t')L_n(\theta) + \frac{1}{2}\|\theta\|_2^2$ is $(1 + C(t')m)$-strongly convex. Using this with $t' = t_k, t = t_{k+1}$, and applying the triangular inequality, we obtain that

$$\begin{aligned} & \|\theta_k - \theta(t_{k+1})\|_2 \leq \|\theta_k - \theta(t_k)\|_2 + \|\theta(t_k) - \theta(t_{k+1})\|_2 \\ = \quad & \|\theta_k - \theta(t_k)\|_2 + \frac{(e^{t_{k+1}} - e^{t_k})}{(e^{t_{k+1}} - 1)(1 + m(e^{t_k} - 1))}\|\theta(t_{k+1})\|_2 \\ \leq \quad & \|\theta_k - \theta(t_k)\|_2 + \frac{(e^{\alpha_{k+1}} - 1)}{(e^{t_{k+1}} - 1)m_k}\|\theta(t_{k+1})\|_2 \,. \end{aligned}$$

Combining this with (96), we get

$$\begin{aligned} \|\theta_{k+1} - \theta(t_{k+1})\|_2 \leq & \left(1 - \frac{2m_{k+1}L_{k+1}}{m_{k+1}+L_{k+1}}\eta_{k+1}\right)^{n_{k+1}} \\ & \left(\|\theta_k - \theta(t_k)\|_2 + \frac{(e^{\alpha_{k+1}} - 1)\|\theta(t_{k+1})\|_2}{(e^{t_{k+1}} - 1)m_k}\right) . \end{aligned} \tag{98}$$

Next we use induction to show that

$$\|\theta_k - \theta(t_k)\|_2 \leq 2\left(1 - \frac{2m_kL_k}{m_k+L_k}\eta_k\right)^{n_k} \frac{(e^{\alpha_k} - 1)\|\theta(t_k)\|_2}{(e^{t_k} - 1)m_{k-1}} \,. \tag{99}$$

Suppose that (99) holds for θ_k, then using (98) and (99), it follows that (99) holds for θ_{k+1} if

$$2\left(1 - \frac{2m_kL_k}{m_k+L_k}\eta_k\right)^{n_k} \frac{(e^{\alpha_k} - 1)\|\theta(t_k)\|_2}{(e^{t_k} - 1)m_{k-1}} \leq \frac{(e^{\alpha_{k+1}} - 1)\|\theta(t_{k+1})\|_2}{(e^{t_{k+1}} - 1)m_k} \tag{100}$$

for any $k \geq 1$. Next we show that (100) can be ensured by the conditions in (32). First, using the fact that

$$\frac{m_k}{m_{k-1}} = \frac{m(1-e^{-t_k}) + e^{-t_k}}{m(1-e^{-t_{k-1}}) + e^{-t_{k-1}}} \leq \begin{cases} m_1 & \text{when } k = 1 \\ \frac{1-e^{-t_k}}{1-e^{-t_{k-1}}} & \text{for any } k \geq 2\,, \end{cases} \tag{101}$$

and

$$\frac{(e^{t_{k+1}}-1)(e^{\alpha_k}-1)(1-e^{-t_k})}{(e^{t_k}-1)(e^{\alpha_{k+1}}-1)(1-e^{-t_{k-1}})} \leq 12 \text{ for any } k \geq 2,\ \frac{(e^{t_2}-1)(e^{\alpha_1}-1)}{(e^{t_1}-1)(e^{\alpha_2}-1)} \leq 5\,,$$

when $\alpha_k \leq 2\alpha_{k-1}$, $\alpha_{k+1} \geq \alpha_k/2$, and $\alpha_k \leq \ln(2)$ for $k \geq 1$, it follows that a sufficient condition for (100) is

$$n_1 \geq \frac{\log(10 m_1)}{-\log\left(1 - \frac{2m_1 L_1}{m_1 + L_1}\eta_1\right)} \text{ and } n_k \geq \frac{\log(24)}{-\log\left(1 - \frac{2m_k L_k}{m_k + L_k}\eta_k\right)}$$

for any $k \geq 2$.

Next, using the fact that $f_{t_k}(\cdot)$ has L_k-Lipschitz gradient, we have

$$\|g_k\|_2 \leq L_k\|\theta_k - \theta(t_k)\|_2 \leq 2L_k\left(1 - \frac{2m_k L_k}{m_k + L_k}\eta_k\right)^{n_k}\frac{(e^{\alpha_k}-1)\|\theta(t_k)\|_2}{(e^{t_k}-1)m_{k-1}}\,.$$

This completes the proof of Theorem 7.

Proof of Theorem 8. Since $C(t)L_n(\theta) + \frac{1}{2}\|\theta\|_2^2$ is $(1 + C(t)m)$-strongly convex, using (45), we have that

$$\|\theta(t) - \mathbf{0}\|_2 \leq \frac{\|C(t)\nabla L_n(\mathbf{0})\|_2}{1 + C(t)m}\,, \tag{102}$$

which implies that,

$$\|\theta(t_k)\|_2 \leq \frac{1 - e^{-t_k}}{m_k}\|\nabla L_n(\mathbf{0})\|_2\,. \tag{103}$$

By Theorem 7, since

$$n_k \geq \frac{\log(24) + \max(0, \log(L_k/m_{k-1}))}{-\log\left(1 - \frac{2m_k L_k}{m_k + L_k}\eta_k\right)}\,, \tag{104}$$

it follows that

$$\|g_k\|_2 \leq 2\left(1 - \frac{2m_k L_k}{m_k + L_k}\eta_k\right)^{n_k}\frac{(e^{\alpha_k}-1)\|\theta(t_k)\|_2}{(e^{t_k}-1)m_{k-1}} \leq (12)^{-1}\frac{(e^{\alpha_k}-1)\|\theta(t_k)\|_2}{(e^{t_k}-1)}\,. \tag{105}$$

Similar to (82), we can show that

$$\frac{1}{1 + (12)^{-1}} \leq \frac{\|\theta_k\|_2}{\|\theta(t_k)\|_2} \leq \frac{1}{1 - (12)^{-1}}\,. \tag{106}$$

Using this, Lemma 1, and (13), we have that

$$\sup_{t\in[0,t_1]}\left\{f_t(\tilde\theta(t))-f_t(\theta(t))\right\}\le \max\left(e^{t_1}\|g_1\|_2^2,\,\|\theta_1\|_2^2\right)+\frac{e^{t_1}(1-e^{-t_1})^2}{2}\|\nabla L_n(\mathbf{0})\|_2^2$$
$$\le \frac{e^{t_1}(1-e^{-t_1})^2}{2}\|\nabla L_n(\mathbf{0})\|_2^2+(3/2)\|\theta(t_1)\|_2^2\le 2(e^{\alpha_1}-1)^2\|\nabla L_n(\mathbf{0})\|_2^2\,.$$

Moreover, using (105), (106), and Corollary 1, we obtain that

$$e^{t_{k+1}}\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2\|g_k\|_2^2\le\frac{e^{t_{k+1}}}{144}\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2\left(\frac{(e^{\alpha_k}-1)\|\theta(t_k)\|_2}{(e^{t_k}-1)}\right)^2$$
$$\le\frac{e^{\alpha_{k+1}}}{144}\frac{(e^{\alpha_k}-1)^2(1-e^{-t_{k+1}})^2}{(e^{\alpha_{k+1}}-1)^2(1-e^{-t_k})^2}e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2\|\theta(t_k)\|_2^2\le(20)^{-1}e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2\|\theta_k\|_2^2$$
$$e^{t_{k+1}}\|g_{k+1}\|_2^2\le\frac{e^{t_{k+1}}}{144}\left(\frac{(e^{\alpha_{k+1}}-1)\|\theta(t_{k+1})\|_2}{(e^{t_{k+1}}-1)}\right)^2\le\frac{e^{\alpha_{k+1}}}{144}e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2\|\theta(t_k)\|_2^2$$
$$\le(50)^{-1}e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2\|\theta_k\|_2^2\,,$$

and

$$(e^{-t_k}-e^{-t_{k+1}})^2\max\left\{\frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2},\,\frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\right\}$$
$$\le\left(1-(12)^{-1}\right)^{-2}(e^{-t_k}-e^{-t_{k+1}})^2\max\left\{\frac{e^{t_{k+1}}\|\theta(t_k)\|_2^2}{(1-e^{-t_k})^2},\,\frac{e^{t_k}\|\theta(t_{k+1})\|_2^2}{(1-e^{-t_{k+1}})^2}\right\}$$
$$\le\left(1-(12)^{-1}\right)^{-2}e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2\|\theta(t_k)\|_2^2$$
$$\le\left(1-(12)^{-1}\right)^{-2}\left(1+(12)^{-1}\right)^2e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2\|\theta_k\|_2^2\,,$$

where we have used the fact that $e^{\alpha_{k+1}}(e^{\alpha_k}-1)(1-e^{-t_{k+1}})/(e^{\alpha_{k+1}}-1)(1-e^{-t_k})\le(\sqrt{2}+1)^2$. Combining these with (14), we have that

$$\sup_{t\in[t_k,t_{k+1}]}\left\{f_t(\tilde\theta(t))-f_t(\theta(t))\right\}\le e^{t_{k+1}}\max\left\{\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2\|g_k\|_2^2,\,\|g_{k+1}\|_2^2\right\}$$
$$+(e^{-t_k}-e^{-t_{k+1}})^2\max\left\{\frac{e^{t_{k+1}}\|\theta_k\|_2^2}{(1-e^{-t_k})^2},\,\frac{e^{t_k}\|\theta_{k+1}\|_2^2}{(1-e^{-t_{k+1}})^2}\right\}$$
$$\le\left((20)^{-1}+(1-(12)^{-1})^{-2}(1+(12)^{-1})^2\right)e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2\|\theta_k\|_2^2$$
$$\le 2e^{-t_k}\left(\frac{e^{t_{k+1}-t_k}-1}{1-e^{-t_k}}\right)^2\|\theta_k\|_2^2$$

for any $k\ge1$.

Lastly, using (15), (105), (80), and part (ii) of Corollary 1, we have

$$\begin{aligned}
\sup_{t_N < t \le t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} &\le \frac{e^{t_N}(1-e^{-t_{\max}})}{1-e^{-t_N}} \|g_N\|_2^2 + \frac{3}{2(e^{t_N}-1)} \|\theta(t_{\max})\|_2^2 \\
&\le \frac{e^{t_N}(1-e^{-t_{\max}})}{144(1-e^{-t_N})} \left(\frac{e^{\alpha_N}-1}{e^{t_N}-1} \right)^2 \|\theta(t_N)\|_2^2 + \frac{3}{2(e^{t_N}-1)} \|\theta(t_{\max})\|_2^2 \\
&\le \frac{2\|\theta(t_{\max})\|_2^2}{e^{t_N}-1} .
\end{aligned}$$

Combining the three bounds, we obtain that when $t_N \le t_{\max}$,

$$\begin{aligned}
\sup_{0 \le t \le t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \le 2 \max \Bigg(&(e^{\alpha_1}-1)^2 \|\nabla L_n(\mathbf{0})\|_2^2 , \\
&\max_{1 \le k \le N-1} e^{-t_k} \frac{(e^{\alpha_{k+1}}-1)^2}{(1-e^{-t_k})^2} \|\theta_k\|_2^2 , \frac{\|\theta(t_{\max})\|_2^2}{e^{t_N}-1} \Bigg) ,
\end{aligned}$$

and when $t_{N-1} \le t_{\max} < t_N$ for some $N \ge 1$.

$$\begin{aligned}
&\sup_{0 \le t \le t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \le \sup_{0 \le t \le t_N} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \\
&\le 2 \max \left((e^{\alpha_1}-1)^2 \|\nabla L_n(\mathbf{0})\|_2^2 , \max_{1 \le k \le N-1} e^{-t_k} \frac{(e^{\alpha_{k+1}}-1)^2}{(1-e^{-t_k})^2} \|\theta_k\|_2^2 \right) .
\end{aligned}$$

This completes the proof of Theorem 8.

Proof of Theorem 9. We first show that $\alpha_{k+1} \ge \alpha_k/2$. If $\alpha_{k+1} = 5^{-1}$ or $2\alpha_k$, then trivially $\alpha_{k+1} \ge \alpha_k/2$. Now we assume that $\alpha_{k+1} = A_k$, where

$$A_k := \ln\left(1 + \frac{\epsilon^{1/2} e^{t_k/2}(1-e^{-t_k})}{\|\theta_k\|_2} \right) .$$

First, using (105), (106), and the fact that $\|\theta_k - \theta(t_k)\|_2 \le e^{t_k}\|g_k\|_2$, we obtain that

$$\|\theta_k\|_2 \le \|\theta(t_k)\|_2 + \|\theta(t_k) - \theta_k\|_2 \le \|\theta(t_k)\|_2 + e^{t_k}\|g_k\|_2 \le \|\theta(t_k)\|_2 + e^{t_k} \frac{(e^{\alpha_k}-1)}{12(e^{t_k}-1)} \|\theta(t_k)\|_2 ,$$

which implies that

$$\|\theta_k\|_2 \le \|\theta(t_k)\|_2 \left(1 + \frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})} \right) . \tag{107}$$

When $k=1$, using (107) and the fact that $e^{\alpha_1}-1 \le \sqrt{\epsilon}/\|\nabla L_n(\mathbf{0})\|_2$, we obtain that

$$\begin{aligned}
\frac{e^{\alpha_2}-1}{e^{\alpha_1/2}-1} &= (e^{\alpha_1/2}+1) \frac{e^{\alpha_2}-1}{e^{\alpha_1}-1} \ge (e^{\alpha_1/2}+1) \frac{e^{A_1}-1}{e^{\alpha_1}-1} = (e^{\alpha_1/2}+1) e^{\alpha_1/2} \frac{\sqrt{\epsilon}(1-e^{-t_1})}{\|\theta_1\|_2 (e^{\alpha_1}-1)} \\
&\ge (e^{\alpha_1/2}+1) e^{\alpha_1/2} \frac{\sqrt{\epsilon}(1-e^{-t_1})}{\|\theta(t_1)\|_2 (e^{\alpha_1}-1)} \left(1 + \frac{(e^{\alpha_1}-1)}{12(1-e^{-t_1})} \right)^{-1} \\
&\ge \frac{(e^{-\alpha_1/2}+1)\sqrt{\epsilon}}{\|\nabla L_n(\mathbf{0})\|_2 (e^{\alpha_1}-1)} (1+e^{\alpha_1}/12)^{-1} \ge \frac{3}{4} (e^{-\alpha_1/2}+1)(1+e^{\alpha_1}/12)^{-1} \ge 1 ,
\end{aligned}$$

provided that $\alpha_1 \le \ln(2)$. This implies that $\alpha_2 \ge \alpha_1/2$.

When $k \ge 2$, note that $\alpha_k \le A_{k-1}$ and

$$\begin{aligned}\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k/2}-1} &= (e^{\alpha_k/2}+1)\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k}-1} \ge (e^{\alpha_k/2}+1)\frac{e^{A_k}-1}{e^{A_{k-1}}-1} \\ &= (e^{\alpha_k/2}+1)e^{\alpha_k/2}\frac{\epsilon^{1/2}(1-e^{-t_k})}{\|\theta_k\|_2}\frac{\|\theta_{k-1}\|_2}{\epsilon^{1/2}(1-e^{-t_{k-1}})} \\ &\ge (e^{-\alpha_k/2}+1)\frac{(e^{t_k}-1)\|\theta_{k-1}\|_2}{(e^{t_{k-1}}-1)\|\theta_k\|_2}\,. \end{aligned} \tag{108}$$

Now similar to (107), we obtain that

$$\|\theta(t_k)\|_2 \le \|\theta_k\|_2 + \|\theta(t_k)-\theta_k\|_2 \le \|\theta_k\|_2 + e^{t_k}\|g_k\|_2 \le \|\theta_k\|_2 + \frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}\|\theta(t_k)\|_2\,,$$

we have that

$$\left(1-\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}\right) \le \frac{\|\theta_k\|_2}{\|\theta(t_k)\|_2} \le \left(1+\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}\right)\,. \tag{109}$$

Hence,

$$\frac{(e^{t_k}-1)\|\theta_{k-1}\|_2}{(e^{t_{k-1}}-1)\|\theta_k\|_2} \ge \frac{(e^{t_k}-1)\|\theta(t_{k-1})\|_2}{(e^{t_{k-1}}-1)\|\theta(t_k)\|_2}\frac{1-\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}}{1+\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}} \ge \frac{1-\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}}{1+\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}}$$

where the last inequality uses part (ii) of Corollary 1. Combining this with (108), we obtain that

$$\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k/2}-1} \ge (e^{-\alpha_k/2}+1)\frac{1-\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}}{1+\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}}\,. \tag{110}$$

Therefore, to prove $\alpha_{k+1} \ge \alpha_k/2$, it suffices to show that the RHS of the above inequality is no smaller than 1. To this end, using the fact that $\alpha_k \le \ln(2)$ for all $k \ge 1$, we have

$$\frac{(e^{\alpha_k}-1)}{(1-e^{-t_k})} = \frac{e^{\alpha_k}(1-e^{-\alpha_k})}{1-e^{-t_k}} \le e^{\alpha_k} \le 2$$

for any $k \ge 1$. Combining this with (110), we have that

$$\frac{e^{\alpha_{k+1}}-1}{e^{\alpha_k/2}-1} \ge (e^{-\alpha_k/2}+1)\frac{1-\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}}{1+\frac{(e^{\alpha_k}-1)}{12(1-e^{-t_k})}} \ge (1/\sqrt{2}+1)\frac{1-6^{-1}}{1+6^{-1}} > 1\,,$$

which proves that $\alpha_{k+1} \ge \alpha_k/2$. Hence, α_k satisfy all the conditions in Theorem 8.

Next, we show that the algorithm will terminate in finite steps. We first show that t_k diverges. To this end, using (60) and (106), we have

$$\begin{aligned}\frac{\epsilon^{1/2}e^{t_k/2}(1-e^{-t_k})}{\|\theta_k\|_2} &\ge \frac{e^{t_k/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_k})}{\|\theta(t_k)\|_2/(1-(12)^{-1})} \\ &\ge \frac{e^{t_k/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_k})}{2(e^{t_k}-1)\|\nabla L_n(\mathbf{0})\|_2} \ge 2^{-1}(e^{\alpha_1}-1)e^{-t_k/2}\,,\end{aligned}$$

which implies that

$$A_k \geq \ln(1 + 2^{-1}(e^{\alpha_1} - 1)e^{-t_k/2}) \geq \frac{(e^{\alpha_1} - 1)e^{-t_k/2}}{2 + (e^{\alpha_1} - 1)e^{-t_k/2}} = \frac{(e^{\alpha_1} - 1)}{2e^{t_k/2} + (e^{\alpha_1} - 1)} \,. \tag{111}$$

Thus,

$$\alpha_{k+1} \geq \min\left(\alpha_{\max}, 2\alpha_k, \frac{(e^{\alpha_1} - 1)}{2e^{t_k/2} + (e^{\alpha_1} - 1)}\right) \,. \tag{112}$$

Now we prove the divergence of t_k by contradiction. Suppose that t_k does not diverge. Then there must exist a constant T such that $t_k < T$ for all k. However, now we have

$$\alpha_{k+1} \geq \min\left(\alpha_{\max}, 2\alpha_k, \frac{(e^{\alpha_1} - 1)}{2e^{T/2} + (e^{\alpha_1} - 1)}\right) \,,$$

which implies that α_k is lower bounded by a positive constant when k is large enough, implying that t_k should diverge. This is a contradiction. Hence, t_k diverges.

Now we are ready to show that the algorithm must terminate after a finite number of iterations. If $t_{\max} < \infty$, then $t_k \geq t_{\max}$ must hold for large enough k as t_k diverges. If $t_{\max} = \infty$, then we have that $\theta(t_{\max}) = \theta^\star$ is finite by assumption. Therefore, the termination criterion in (25) should also be met when N is large enough, because t_N diverges and

$$\frac{\max(\|\theta(t_{\max})\|_2^2, \|\theta_N\|_2^2)}{(e^{t_N} - 1)} \leq \frac{2\max(\|\theta(t_{\max})\|_2^2, \|\theta(t_N)\|_2^2)}{(e^{t_N} - 1)} \leq \frac{2\|\theta^\star\|_2^2}{(e^{t_N} - 1)} \to 0 \text{ as } N \to \infty \,.$$

Finally, we are ready to prove (40) upon termination. Using (35) and (36) in Theorem 8, and the definition of α_1 and α_{k+1}, we have that

$$\begin{aligned} &e^{-t_k}\left(\frac{e^{\alpha_{k+1}} - 1}{1 - e^{-t_k}}\right)^2 \|\theta_k\|_2^2 \leq (e^{\alpha_1} - 1)^2 \|\nabla L_n(\mathbf{0})\|_2^2 \leq \epsilon \text{ and} \\ &\frac{\|\theta(t_{\max})\|_2^2}{(e^{t_N} - 1)} \leq \frac{2\|\theta_N\|_2^2}{(e^{t_N} - 1)} \frac{\|\theta(t_{\max})\|_2^2}{2\|\theta_N\|_2^2} \leq \epsilon \frac{\|\theta(t_{\max})\|_2^2}{\|\theta(t_N)\|_2^2} \,. \end{aligned}$$

for any $k \geq 1$ when the algorithm is terminated, and after termination,

$$\sup_{0 \leq t \leq t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\} \leq \begin{cases} \frac{2\|\theta(t_{\max})\|_2^2}{\|\theta(t_N)\|_2^2}\epsilon & \text{when } t_{\max} \geq t_N \\ 2\epsilon & \text{when } t_{\max} < t_N \end{cases} \,,$$

which implies (40). This completes the proof of Theorem 9.

Proof of Theorem 10. We first bound the number of gradient steps at each iteration. Since $m = 0$, we have that $m_k = e^{-t_k}$. By assumption, we have that $\eta_k = \eta = \mathcal{O}(\min(1, L^{-1}))$. Then the upper bounds on n_1 and n_{k+1} in (34) can be bounded further as follows

$$\begin{aligned} &\frac{\log(24) + \max(0, \log(L_{k+1}/m_k))}{-\log\left(1 - \frac{2m_{k+1}L_{k+1}}{m_{k+1}+L_{k+1}}\eta_{k+1}\right)} \leq \frac{\log(24) + \max(0, \log(L_{k+1}/m_k))}{\frac{2m_{k+1}L_{k+1}}{m_{k+1}+L_{k+1}}\eta_{k+1}} \\ &\leq \frac{\log(24) + \max(0, \log(L_{k+1}/m_k))}{m_{k+1}\eta} \leq \frac{\log(24) + t_k + \log(L_{k+1})}{m_{k+1}\eta} = \mathcal{O}\left(e^{t_{k+1}}(t_{k+1} + 1)\right) \,, \end{aligned}$$

where we have used $L_k \geq m_k$ for any $k \geq 1$. Similarly, we also we have

$$\frac{\log(10m_1L_1)}{-\log\left(1-\frac{2m_1L_1}{m_1+L_1}\eta_1\right)} \lesssim e^{t_1},$$

where we have used the fact that $-\log(1-x) \geq x$, $L_k \leq \max(L,1)$, and treated L as a constant.

Next, By applying Lemma 1 and (109), we have

$$\frac{\|\theta_k\|_2}{1-e^{-t_k}} \leq \frac{7}{6}\frac{\|\theta(t_k)\|_2}{1-e^{-t_k}} \leq \frac{7}{6}\left(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2\right). \tag{113}$$

Therefore, when $\alpha_{k+1} \leq \ln(2)$, we have that

$$e^{\alpha_{k+1}} - 1 = \frac{e^{t_k/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2(1-e^{-t_k})}{\|\theta_k\|_2} \geq \frac{6e^{t_k/2}(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2}{7(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2)},$$

which implies that

$$\alpha_{k+1} \geq \ln(1+\nu_1 e^{t_k/2}(e^{\alpha_1}-1))$$

when $\alpha_{k+1} \leq \ln(2)$, where we treat

$$\nu_1 = \frac{6\|\nabla L_n(\mathbf{0})\|_2}{7(\|\theta(t_{\max})\|_2 + \|\nabla L_n(\mathbf{0})\|_2)}$$

as problem-dependent constants.

Now we are ready to derive the bound for the number of gradient steps. When $t_{\max} = \infty$, in view of the termination criterion (39), the algorithm will be terminated when $t_k = \mathcal{O}\left(\ln(\epsilon^{-1})\right)$. Therefore, the total number of gradient steps can be bounded as

$$\sum_{k=1}^{k^\star} n_k \lesssim \sum_{k=1}^{k^\star} e^{t_k}(t_k+1) \lesssim \epsilon^{-1}(1+\ln(\epsilon^{-1})),$$

where $k^\star = \max\{k : t_k \leq \ln(\epsilon^{-1})\}$ and we have used Lemma 3. Hence, the total number of gradient steps is at most $\mathcal{O}\left(\epsilon^{-1}\ln(\epsilon^{-1})\right)$.

When $t_{\max} < \infty$, then in view of the termination criterion (39), the algorithm will be terminated when $t_k > t_{\max}$. Therefore, the total number of gradient steps can be bounded as

$$\sum_{k=1}^{k^\star} n_k + \max(0, t_{\max} - t_{k^\star})/\ln(1+\nu_1) \lesssim \sum_{k=1}^{k^\star} e^{t_k}(t_k+1) + 1/\ln(1+\nu_1) \lesssim \epsilon^{-1}(1+\ln(\epsilon^{-1})),$$

where again we have used Lemma 3. In this case, the total number of gradient steps is also at most $\mathcal{O}\left(\epsilon^{-1}\ln(\epsilon^{-1})\right)$. This completes the proof of Theorem 10.

Proof of Corollary 3. We first show that the bounds in Theorem 8 continue to hold. The proof is similar to that of Theorem 8 with some slight modifications. In particular, upon termination of the gradient descent method at t_k, we have (41). Replacing the bound (105) in the proof of Theorem 8 by (41), and following a similar argument, we obtain that

$$\begin{aligned} &\sup_{0\le t\le t_{\max}} \left\{ f_t(\tilde\theta(t)) - f_t(\theta(t)) \right\} \\ &\qquad\lesssim \max\left((e^{\alpha_1}-1)^2 \|\nabla L_n(\mathbf{0})\|_2^2 \,,\, \max_{1\le k\le N-1} e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2 \|\theta_k\|_2^2 \right). \end{aligned}$$

when $t_N \ge t_{\max}$ for some N; and

$$\begin{aligned} \sup_{0\le t\le t_{\max}} \left\{ f_t(\tilde\theta(t)) - f_t(\theta(t)) \right\} \lesssim \max\Bigg(& (e^{\alpha_1}-1)^2 \|\nabla L_n(\mathbf{0})\|_2^2 \,, \\ & \max_{1\le k\le N-1} e^{-t_k}\left(\frac{e^{\alpha_{k+1}}-1}{1-e^{-t_k}}\right)^2 \|\theta_k\|_2^2 \,,\, \frac{\|\theta(t_{\max})\|_2^2}{(e^{t_N}-1)} \Bigg) \end{aligned}$$

when $t_N \le t_{\max}$ and $\|\theta(t_{\max})\|_2 < \infty$. Then the bound (42) follows from these bounds, following the proof of Theorem 9. This completes the proof of Corollary 3.

Proof of Proposition 1. We verify condition (16) for the functions in Table 1 separately.

Log-barrier function. Note that $L_n'(\theta) = -\theta^{-1}$ and $L_n''(\theta) = \theta^{-2}$. Consider $\gamma_1 = 3/2$, $\gamma_2 = 1$ and $\beta \ge 2$. Then (17) reduces to $|\delta| \le \theta\beta^{-1}$. It is easy to see the right hand side of (16) reduces to $\beta\delta^2\theta^{-3}$. Under the condition $|\delta| \le \theta\beta^{-1}$, the left hand side of (16) can be bounded as follows

$$\left| -\frac{1}{\theta+\delta} + \theta^{-1} - \frac{\delta}{\theta^2} \right| = \frac{\delta^2}{\theta^2}\left|\frac{1}{\theta+\delta}\right| = \frac{\delta^2}{\theta^3}\left|\frac{1}{1+\delta\theta^{-1}}\right| \le \frac{\delta^2}{\theta^3}\frac{\beta}{\beta-1} \le \beta\frac{\delta^2}{\theta^3}.$$

Thus, $-\ln(\theta)$ satisfies the condition with $\gamma_1 = \frac{3}{2}$ and $\gamma_2 = 1$.

Entropy-barrier function. Note that the first and second derivatives are $L_n'(\theta) = \ln(\theta) + 1 - \theta^{-1}$ and $L_n''(\theta) = \theta^{-1} + \theta^{-2}$, respectively. Consider $\gamma_1 = \frac{3}{2}$, $\gamma_2 = 1$ and any β satisfying $\frac{\beta}{(\beta-1)^2} \le 1$. Then (17) reduces to $|\delta| \le \theta\beta^{-1}$. The right hand side of (16) can be bounded as

$$\beta\delta^2(\theta^{-1}+\theta^{-2})^{3/2} = \beta\frac{\delta^2}{\theta^2}\frac{(1+\theta)^{3/2}}{\theta} \ge \beta\frac{\delta^2}{\theta^2}(\theta^{-1}+3/2).$$

By Taylor's expansion, there exists $s \in (0,1)$ such that

$$\begin{aligned} & \left| \ln(\theta+\delta) - \frac{1}{\theta+\delta} - \ln(\theta) + \theta^{-1} - (\theta^{-1}+\theta^{-2})\delta \right| \\ = \ & \left| \ln(1+\frac{\delta}{\theta}) + \frac{\delta}{\theta(\theta+\delta)} - \frac{\delta}{\theta}(1+\theta^{-1}) \right| = \left| -\frac{\delta^2}{2\theta^2(1+s\delta\theta^{-1})^2} - \frac{\delta}{\theta^2} + \frac{\delta}{\theta(\theta+\delta)} \right| \\ = \ & \frac{\delta^2}{\theta^2}\left(\frac{1}{\theta(1+\delta\theta^{-1})} + \frac{1}{2(1+s\delta\theta^{-1})^2} \right) \le \frac{\delta^2}{\theta^2}\left(\theta^{-1}\frac{1}{1-\beta^{-1}} + \frac{1}{2(1-\beta^{-1})^2} \right) \\ = \ & \frac{\delta^2}{\theta^2}(\frac{\beta}{\beta-1})^2\left(\theta^{-1}(1-\frac{1}{\beta}) + \frac{1}{2} \right) \le \beta\frac{\delta^2}{\theta^2}(\theta^{-1}+3/2)\frac{\beta}{(\beta-1)^2} \\ \le \ & \beta\frac{\delta^2}{\theta^2}(\theta^{-1}+3/2) \le \beta\delta^2(\theta^{-1}+\theta^{-2})^{3/2}, \end{aligned}$$

which implies that $\theta\ln(\theta)-\ln(\theta)$ satisfies (16) with $\gamma_1=3/2$ and $\gamma_2=1$.

Logistic function. Note that $L_n'(\theta)=-\frac{1}{1+e^\theta}$ and $L_n''(\theta)=\frac{e^\theta}{(1+e^\theta)^2}$. Consider $\gamma_1=1$, $\gamma_2=0$ and any β satisfying $e^{1/\beta}\le 2\beta$ (e.g., $\beta=2$). Then (17) reduces to $|\delta|\le\beta^{-1}$. The right hand side of (16) is $\beta\delta^2e^\theta(1+e^\theta)^{-2}$. By Taylor's expansion, there exists $s\in(0,1)$ such that

$$\left|-\frac{1}{1+e^{\theta+\delta}}+\frac{1}{1+e^\theta}-\frac{e^\theta}{(1+e^\theta)^2}\delta\right|=\frac{\delta^2}{2}\left|\frac{e^{\theta+s\delta}-1}{e^{\theta+s\delta}+1}\right|\frac{e^{\theta+s\delta}}{(e^{\theta+s\delta}+1)^2}\le\frac{\delta^2}{2}\frac{e^{\theta+s\delta}}{(e^{\theta+s\delta}+1)^2}\,.$$

Let $q=e^{s\delta}$. Since $|\delta|\le\beta^{-1}$, $s\in(0,1)$, we have $q\in(e^{-\frac{1}{\beta}},e^{\frac{1}{\beta}})$. Therefore,

$$\begin{aligned}\frac{qe^\theta}{(qe^\theta+1)^2} &= q\left(\frac{qe^\theta+1}{e^\theta+1}\right)^{-2}\frac{e^\theta}{(e^\theta+1)^2}=q\left(q+\frac{1-q}{e^\theta+1}\right)^{-2}\frac{e^\theta}{(e^\theta+1)^2}\\ &\le \max(q,q^{-1})\frac{e^\theta}{(e^\theta+1)^2}\le e^{\frac{1}{\beta}}\frac{e^\theta}{(e^\theta+1)^2}\,.\end{aligned}$$

Consequently, for any $|\delta|\le 1/\beta$ with β satisfying $e^{1/\beta}\le 2\beta$, we have

$$\frac{\delta^2}{2}\frac{e^{\theta+s\delta}}{(e^{\theta+s\delta}+1)^2}\le\frac{\delta^2}{2}e^{\frac{1}{\beta}}\frac{e^\theta}{(e^\theta+1)^2}\le\beta\delta^2\frac{e^\theta}{(1+e^\theta)^2}\,,$$

which implies that $\ln(1+e^{-\theta})$ satisfies (16) with $\gamma_1=1$ and $\gamma_2=0$.

Exponential function. Note that $L_n'(\theta)=-e^\theta$ and $L_n''(\theta)=e^{-\theta}$. Consider $\gamma_1=1$, $\gamma_2=0$, and $\beta\ge\frac{1}{2}$. Then (17) reduces to $|\delta|\le\beta^{-1}$. It is easy to see the right hand side of inequality (16) is $\beta e^{-\theta}\delta^2$. Note that $e^x\ge 1+x$ and $e^x-1-x-\beta x^2\le 0$ for any $\beta\ge 1/2$. Thus under the condition that $|\delta|\le\beta^{-1}$, the left hand side of inequality (16) satisfies

$$\left|-e^{-(\theta+\delta)}+e^{-\theta}-e^{-\theta}\delta\right|=e^\theta(e^{-\delta}-1+\delta)\le e^\theta\beta\delta^2\,,$$

which implies that e^θ satisfies the condition with $\gamma_1=1$ and $\gamma_2=0$.

Square function. Note that $L_n'(\theta)=2\theta$ and $L_n''(\theta)=2$. Since for any $\delta\in\mathbb{R}$, the left hand side of inequality (16) satisfies

$$\left|L_n'(\theta+\delta)-L_n'(\theta)-L_n''(\theta)\delta\right|=0$$

we conclude that θ^2 satisfies inequality (16) with any $0\le\gamma_1<2$, and $0\le\gamma_2<2$. This completes of the proof of Proposition 1.

Proof of Proposition 2. We first prove two claims to be used later in the proof.

Claim 1 *Suppose that function $f:\mathbb{R}\mapsto\mathbb{R}$ satisfies Assumption 1 with $\gamma_1=1$, $\gamma_2=0$ and β. Then for any p-dimensional vector a and any scalar b, the function $g(x):=f(a^\top x+b)$ also satisfies Assumption 1 with $\gamma_1=1$, $\gamma_2=0$, and $\beta\|a\|_2$.*

Proof of Claim 1. Suppose that Assumption (A1) holds for $f(\cdot)$ with $\gamma_1 = 1$, $\gamma_2 = 0$, and β. Then, for any $d \in \mathbb{R}^p$ with $\|d\|_2 \le (\beta\|a\|_2)^{-1}$, we have that $|a^\top d|^2 \le \|a\|_2^2\|d\|_2^2 \le \beta^{-2}$. By Assumption (A1), we have, for any $s \in \mathbb{R}$, that

$$\left|f'(s + a^\top d) - f'(s) - f''(s)a^\top d\right| \le \beta f''(s)(a^\top d)^2 .$$

Moreover, note that $\nabla g(x) = f'(a^\top x + b)a$ and $\nabla^2 g(x) = f''(a^\top x + b)aa^\top$. Consequently, for any $x \in \mathbb{R}^p$, we have that

$$\begin{aligned}
&\|\nabla g(x + d) - \nabla g(x) - \nabla^2 g(x)d\|_2 \\
&= \|(f'(a^\top x + b + a^\top d) - f'(a^\top x + b) - f''(a^\top x + b)a^\top d)a\|_2 \\
&= \left|f'(a^\top x + b + a^\top d) - f'(a^\top x + b) - f''(a^\top x + b)a^\top d\right| \cdot \|a\|_2 \\
&\le \beta f''(a^\top x + b)(a^\top d)^2\|a\|_2 = \beta\|a\|_2 d^\top \nabla^2 g(x)d ,
\end{aligned}$$

which implies that $g(\cdot)$ satisfies Assumption 1 with $\gamma_1 = 1$, $\gamma_2 = 0$, and $\beta\|a\|_2$.

Claim 2 *Suppose that functions f_i with $f_i : \mathbb{R}^p \mapsto \mathbb{R}$ satisfies Assumption (A1) with $\gamma_1 = 1$, $\gamma_2 = 0$, and $\beta = \beta_i$; $i = 1, \ldots, n$. Then $\alpha_1 f_1 + \alpha_2 f_2 \ldots + \alpha_n f_n$ also satisfies Assumption (A1) with $\gamma_1 = 1$, $\gamma_2 = 0$, and $\beta = \max_{1\le i\le n} \beta_i$, where $\alpha_1, \ldots, \alpha_n \ge 0$.*

Proof of Claim 2. For any d satisfying $\|d\| \le 1/\beta$, we have $\|d\| \le 1/\beta_i$ for all $i = 1, \ldots, n$. By Assumption (A1), this implies

$$\left\|\nabla f_i(s + d) - \nabla f_i(s) - \nabla^2 f_i(s)d\right\|_2 \le \beta_i d^\top \nabla^2 f_i(s)d \le \beta d^\top \nabla^2 f_i(s)d, \tag{114}$$

for any vector $s \in \mathbb{R}^p$ and $i = 1, \ldots, n$. Thus,

$$\begin{aligned}
&\left\|\sum_{i=1}^n \alpha_i \nabla f_i(s + d) - \sum_{i=1}^n \alpha_i \nabla f_i(s) - \sum_{i=1}^n \alpha_i \nabla^2 f_i(s)d\right\|_2 \\
&\le \sum_{i=1}^n \alpha_i \left\|\nabla f_i(s + d) - \nabla f_i(s) - \nabla^2 f_i(s)d\right\|_2 \\
&\le \sum_{i=1}^n \alpha_i \beta d^\top \nabla^2 f_i(s)d = \beta d^\top \left(\sum_{i=1}^n \alpha_i \nabla^2 f_i(s)\right) d .
\end{aligned}$$

Consequently, the function $\sum_{i=1}^n \alpha_i f_i$ also satisfies Assumption (A1) with $\gamma_1 = 1$, $\gamma_2 = 0$, and $\beta = \max_{1\le i\le n} \beta_i$.

Now we are ready to prove the main result. It has been shown in Proposition 1 that the logistic regression loss function satisfies Assumption (A1) with $\gamma_1 = 1$, $\gamma_2 = 0$ and $\beta = 2$. By using Claim 1, it follows that $\log(1 + e^{-Y_i X_i^\top \theta})$ also satisfies Assumption (A1) with $\gamma_1 = 1$, $\gamma_2 = 0$ and $\beta = 2\max_{1\le i\le n} \|X_i\|_2$ for $i = 1, \ldots, n$. Moreover, it follows from Claim 2 that the logistic regression empirical loss $n^{-1}\sum_{i=1}^n \log(1 + e^{-Y_i X_i^\top \theta})$ satisfies Assumption (A1) with $\gamma_1 = 1$, $\gamma_2 = 0$, and $\beta = 2\max_{1\le i\le n} \|X_i\|_2$. This completes the proof of Proposition 2.

Appendix B. Approximation-error bounds for the ODE methods

In this section, we follow the classical global approximation error analysis of ordinary differential equation, which studies the ODE $\theta'(t) = F(t, \theta(t))$. In particular, we focus on Euler's method and second-order Runge-Kutta method, which have been studied extensively in the numerical ODE literature (Hairer et al., 2008; Butcher, 2016). Both methods belong to the more general class of the so-called one-step method (Hairer et al., 2008), for which Lipschitz continuity of function $F(t, \theta)$ plays an important role in quantifying the approximation error. Here we present a result which is a direct application of Theorem 3.4 of Hairer et al. (2008).

Theorem 11 *(Theorem 3.4 of Hairer et al. (2008)) Assume that $L_n(\theta)$ is M-Lipschitz continuous and m-strongly convex. Moreover, assume that the gradient of $L_n(\theta)$ is L-Lipschitz continuous and the Hessian of $L_n(\theta)$ is S-Lipschitz continuous. We have that*

$$\|\theta_k - \theta(t_k)\|_2 \leq \frac{C\alpha^p}{L^\star}(e^{L^\star t_k} - 1)\,, \tag{115}$$

where α is the step size, $t_k = k\alpha$, C is an absolute constant, and $L^\star = \frac{MS}{\min\{m,1\}^2} + \frac{L}{\min\{m,1\}}$.

Note that the approximation error is a power function of the step size α. The power p is often referred to as the order the corresponding approximation method. As we can see from the above Theorem, Euler's method defined in (11) and the special case of Runge-Kutta method defined in (12) are first-order method and second-order method, respectively. In both cases, we can control global error in finite interval by adjusting step size α.

We also point out that the upper bound in (115) gets worse as $k \to \infty$, which is less desirable compared with the approximation error bounds derived for the other two path-following methods. This is likely due to the generality of problem class considered in Theorem 3.4 of Hairer et al. (2008). Indeed, some preliminary empirical studies suggest that the second-order Runge-Kutta is practically comparable to the Newton method in terms of approximation error. A more refined theoretical upper bound may hold for the particular ODE we consider here, although we choose not to pursue this due to space limit.

Proof of Theorem 11. Applying Theorem 3.4 of Hairer et al. (2008), it suffices to show that $F(t, \theta(t))$ is Lipschitz continuous with respect to $\theta(t)$, based on which we could bound the global error directly. For any $t > 0$, θ_1 and θ_2, let $\Delta F = F(t, \theta_1) - F(t, \theta_2)$. Note that

$$\begin{aligned}\|\Delta F\|_2 &= \left\|[(1-e^{-t})\nabla^2 L_n(\theta_1) + e^{-t}I]^{-1}\nabla L_n(\theta_1) - [(1-e^{-t})\nabla^2 L_n(\theta_2) + e^{-t}I]^{-1}\nabla L_n(\theta_2)\right\|_2 \\ &\leq \underbrace{\left\|\left[((1-e^{-t})\nabla^2 L_n(\theta_1) + e^{-t}I)^{-1} - ((1-e^{-t})\nabla^2 L_n(\theta_2) + e^{-t}I)^{-1}\right]\nabla L_n(\theta_2)\right\|_2}_{\text{Part I}} + \\ &\quad \underbrace{\left\|((1-e^{-t})\nabla^2 L_n(\theta_1) + e^{-t}I)^{-1}(\nabla L_n(\theta_1) - \nabla L_n(\theta_2))\right\|_2}_{\text{Part II}}\,.\end{aligned}$$

Since $L_n(\theta)$ is m-strongly convex and $\nabla L_n(\theta)$ is L-Lipschitz continuous, we have that for any θ, $\nabla^2 f_t(\theta) = (1-e^{-t})\nabla^2 L_n(\theta) + e^{-t}I$ satisfies that

$$[(1-e^{-t})\cdot L + e^{-t}]^{-1}I \preceq [\nabla^2 f_t(\theta)]^{-1} \preceq [(1-e^{-t})m + e^{-t}]^{-1}I\,.$$

Therefore,

$$\begin{aligned} \|(\nabla^2 f_t(\theta_1))^{-1} - (\nabla^2 f_t(\theta_2))^{-1}\|_2 &= \|(\nabla^2 f_t(\theta_1))^{-1}(\nabla^2 f_t(\theta_2) - \nabla^2 f_t(\theta_1))(\nabla^2 f_t(\theta_2))^{-1}\|_2 \\ &\leq \|(\nabla^2 f_t(\theta_1))^{-1}\|_2\|\nabla^2 f_t(\theta_2) - \nabla^2 f_t(\theta_1)\|_2\|(\nabla^2 f_t(\theta_2))^{-1}\|_2 \\ &\leq [(1-e^{-t})m + e^{-t}]^{-2}(1-e^{-t})\|\nabla^2 L_n(\theta_1) - \nabla^2 L_n(\theta_1)\|_2 \\ &\leq [(1-e^{-t})m + e^{-t}]^{-2}(1-e^{-t})S\|\theta_1 - \theta_2\|_2 \,. \end{aligned}$$

Moreover, since $L_n(\theta)$ is M-Lipschitz continuous and convex, we have that

$$\|\nabla L_n(\theta)\|_2^2 \leq |L_n(\theta + \nabla L_n(\theta)) - L_n(\theta)| \leq M\|\nabla L_n(\theta)\|_2 \,, \tag{116}$$

which implies that $\|\nabla L_n(\theta)\|_2 \leq M$. Consequently, we can bound Part I as follows

$$\begin{aligned} \text{Part I} &\leq \|[\nabla^2 f_t(\theta_1)]^{-1} - [\nabla^2 f_t(\theta_2)]^{-1}\|_2\|\nabla L_n(\theta_2)\|_2 \\ &\leq M\|[\nabla^2 f_t(\theta_1)]^{-1} - [\nabla^2 f_t(\theta_2)]^{-1}\|_2 \\ &\leq M[(1-e^{-t})m + e^{-t}]^{-2}(1-e^{-t})S\|\theta_1 - \theta_2\|_2 \,. \end{aligned}$$

For part II, we have that

$$\begin{aligned} \text{Part II} &\leq \|[(1-e^{-t})\nabla^2 L_n(\theta_1) + e^{-t}I]^{-1}\|_2\|\nabla L_n(\theta_1) - \nabla L_n(\theta_2)\|_2 \\ &\leq [(1-e^{-t})m + e^{-t}]^{-1}L\|\theta_1 - \theta_2\|_2 \,. \end{aligned}$$

Combining the above two bounds, it follows that

$$\|\Delta F\|_2 \leq \left(M[(1-e^{-t})m + e^{-t}]^{-2}(1-e^{-t})S + [(1-e^{-t})m + e^{-t}]^{-1}L\right)\|\theta_1 - \theta_2\|_2 \,.$$

Let $S^* = M[(1-e^{-t})m + e^{-t}]^{-2}(1-e^{-t})S + [(1-e^{-t})m + e^{-t}]^{-1}L$. It can be shown that

$$S^* \leq L^\star := \frac{MS}{\min\{m,1\}^2} + \frac{L}{\min\{m,1\}} \,.$$

Hence, we have that $\|\Delta F\|_2 \leq L^\star\|\theta_1 - \theta_2\|_2$ for any $m > 0$, which implies that $F(t,\theta)$ is $L^\star$-Lipschitz continuous with respect to θ. This completes the proof of Theorem 11.

Appendix C. Additional experiments

In this section, we provide some additional simulation results for ridge regression. In particular, we compare the proposed methods based on Newton and gradient descent updates against glmnet in terms of both runtime and approximation error, under the setting of ridge regression. In our simulation, the data $\{(X_i, Y_i)\}_{i=1}^n$ are generated from the usual linear regression model $Y_i = X_i^\top\theta^\star + \tilde{\epsilon}$, where $\tilde{\epsilon} \sim N(0, \sigma^2)$, $\theta^\star = (1/\sqrt{p}, \ldots, 1/\sqrt{p})^\top$, and $X_1, \ldots, X_n$ are IID samples from $N_p(0, I_{p\times p})$. We consider two different scenarios with $\sigma^2 = 1/4$ and $\sigma^2 = 4$. Moreover, for each scenario, we consider three different problem dimensions: $(n,p) = (1000, 500)$, $(n,p) = (1000, 1000)$, and $(n,p) = (1000, 2000)$.

Again, we use the global approximation error $\sup_{0\leq t\leq t_{\max}}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$ to assess the accuracy for the approximate solution path $\tilde{\theta}(t)$, where $\tilde{\theta}(t)$ is the linear interpolation of the iterates θ_k generated by each method. Moreover, we sample N points $s_1, \ldots, s_N$

uniformly from $(0, t_{\max})$ and use $\max_{1 \le i \le N}\{f_{s_i}(\tilde{\theta}(s_i)) - f_{s_i}(\theta(s_i))\}$ as an approximation of $\sup_{0 \le t \le t_{\max}}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$. Here $\theta(s_i)$ is the exact solution at s_i and can be computed explicitly. In our simulations, we use $N = 100$ and $t_{\max} = 10$.

Figure 7 plots runtime versus approximation error based on 100 simulations. Similar to Figure 3, we can see from Figure 7 that in all scenarios the proposed Newton method runs the fastest when the required accuracy is high (small suboptimality). Moreover, glmnet is no better than Newton method for smaller problems ($p = 500$ and 1000); while glmnet outperforms both Newton method and the gradient method when low accuracy solution is sufficient and problem dimension is large ($p = 2000$). Lastly, in all cases the gradient method runs faster than Newton method when the desired accuracy is low.

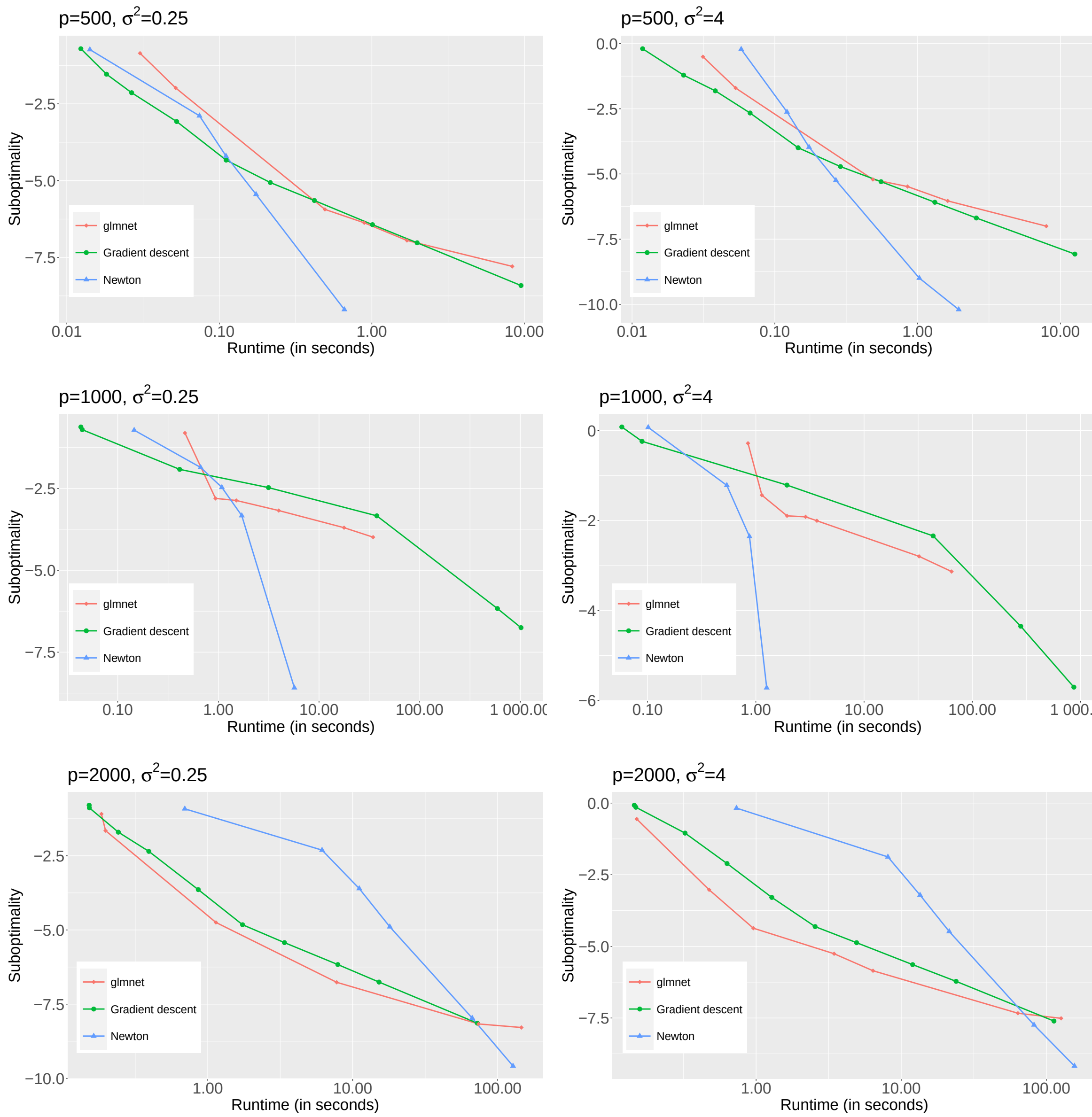

Figure 7: Runtime v.s. suboptimality for the proposed Newton method, gradient descent method, and glmnet under six different scenarios, when applied to ridge regression.

References

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares. In *International Conference on Artificial Intelligence and Statistics*, volume 22, 2019.

Taylor B. Arnold and Ryan J. Tibshirani. Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1): 1–27, 2016. doi: 10.1080/10618600.2015.1008638.

Francis Bach et al. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

Andrew R Conn, Nicholas IM Gould, and Ph L Toint. *Trust region methods*, volume 1. Siam, 2000.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.*, 46(1):247–279, 02 2018. doi: 10.1214/17-AOS1549.

Dirk Eddelbuettel. *Seamless R and C++ integration with Rcpp*. Springer, 2013.

Dirk Eddelbuettel, Romain François, J Allaire, Kevin Ushey, Qiang Kou, N Russel, John Chambers, and D Bates. Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

B. Efron, T. Hastie, I. Johnstone, and R. Tishirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004.

Henry E Fleming. Equivalence of regularization and truncated iteration in the solution of ill-posed image reconstruction problems. *Linear Algebra and its applications*, 130:133–150, 1990.

Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993. ISSN 00401706.

Jerome Friedman and Bogdan Popescu. Gradient directed regularization for linear regression and classiöcation. *Technical Report*, March 2004.

Jerome Friedman, Trevor Hastie, Holger Hofling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Charles J. Geyer. Likelihood inference in exponential families and directions of recession. *Electron. J. Statist.*, 3:259–289, 2009. doi: 10.1214/08-EJS349.

Michael Grant and Stephen Boyd. Cvx: Matlab software for disciplined convex programming, version 2.1, 2014.

Michael C Grant and Stephen P Boyd. Graph implementations for nonsmooth convex programs. In *Recent advances in learning and control*, pages 95–110. Springer, 2008.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6151–6159. Curran Associates, Inc., 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Ernst Hairer, Syvert P Nørsett, and Gerhard Wanner. *Solving ordinary differential equations I: nonstiff problems*, volume 8. Springer Science & Business Media, 2008.

T. Hastie, S. Rosset, R. Tishirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391 – 1415, 2004.

Trevor Hastie, Jonathan Taylor, Robert Tibshirani, Guenther Walther, et al. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010. doi: 10.1198/jcgs.2010.09208.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, pages 164–168, 1944.

Robert M. Freund, Paul Grigas, and Rahul Mazumder. A new perspective on boosting in linear regression via subgradient optimization and relatives. *Ann. Statist.*, 45(6):2328–2364, 2017. doi: 10.1214/16-AOS1505.

Yu. Nesterov. *Introductory Lectures on Convex Programming Volume I: Basic course*. 1998.

Yu. E. Nesterov and A. S. Nemirovskii. *Interior Point Polynomial Methods in Convex Programming: Theory and Algorithms*. SIAM Publications, 1993.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *COLT*, 2018.

MR Osborne. An effective method for computing regression quantiles. *IMA Journal of Numerical Analysis*, 12:151 – 166, 1992.

MR Osborne, B Presnell, and BA Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389 – 403, 2000.

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.

Saharon Rosset. Following curved regularized optimization solution paths. *Advances in Neural Information Processing Systems*, 17:1153–1160, 2004.

Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *Ann. Statist.*, 35 (3):1012–1030, 2007. doi: 10.1214/009053606000001370.

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10631–10641. Curran Associates, Inc., 2018.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, pages 1–69, 2017.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-Posed Problems.* Wiley, New York, 1977.

S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, 5(3):735–743, September 1984. ISSN 0196-5204. doi: 10.1137/0905052.

Yichao Wu. An ordinary differential equation based solution path algorithm. *Journal of Nonparametric Statistics*, 23(1):185–199, 2011.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. doi: 10.1007/s00365-006-0663-2.

Hua Zhou and Yichao Wu. A generic path algorithm for regularized statistical estimation. *Journal of the American Statistical Association*, 109:686–699, 2014.