Path following algorithms for ℓ_2-regularized M-estimation with approximation guarantee

Yunzhang Zhu[1] and Renxiong Liu[2]

[1]Department of Statistics
The Ohio State University

[2]Statistics and Data Science Team
Nokia Bell Labs

Background

- Modern machine learning algorithms are often formulated as regularized M-minimization problems:

$$\theta(\lambda) = \arg\min_{\theta} L_n(\theta) + \lambda p(\theta)\,,$$

  where $L_n(\theta)$ denotes an empirical loss function, $p(\theta)$ denotes a regularization function, and $\lambda > 0$ is a tuning parameter.
- Often $\theta(\lambda)$ can not be computed, and path-following algorithms are usually used to obtain a sequence of solutions at some selected grid points to produce an approximated solution path.
- There is a paucity of literature on how to choose these grid points and how accurately one should solve the optimization problem at the selected grid points.

ℓ_2-regularized M-estimation problem

- We consider the solution path of an ℓ_2-regularized M-estimation problem. Our goal is to approximate the solution path

$$\theta(t) = \arg\min_{\theta \in \mathbb{R}^p} \left\{ (e^t - 1) \cdot L_n(\theta) + (1/2) \cdot \|\theta\|_2^2 \right\} \qquad (1)$$

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 approximated solutions $\{\theta_k\}_{k=1}^N$ at these grid points, we construct an approximated solution path over $[0, t_{\max})$ through linear interpolation. More specifically, we define a piecewise linear solution path $\tilde{\theta}(t)$ as follows

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

where $t_0 = 0$ and $\theta_0 = \mathbf{0}$.

Approximation Errors

- To assess how well the linear interpolation $\tilde{\theta}(t)$ approximates $\theta(t)$, we use the function-value suboptimality of the solution paths defined by

$$\sup_{0 \le t \le t_{\max}} \left\{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \right\}, \tag{2}$$

  where $f_t(\theta) := (1 - e^{-t}) L_n(\theta) + e^{-t}(\|\theta\|_2^2/2)$ is a scaled version of the objective function in (1).
- The global approximate errors (2) can be bounded by a set of local approximation errors $\sup_{t_k \le t \le t_{k+1}} \{ f_t(\tilde{\theta}(t)) - f_t(\theta(t)) \}$, where we show that:

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

Approximation Errors

- The **interpolation error** is irreducible once the grid points are chosen, while the **optimization error** does depend on the algorithm and can be pushed to be arbitrarily small if we run the algorithm long enough at each grid point.
- It is therefore natural to consider a stopping criterion scheme that balances the two types of errors:

$$\underbrace{e^{t_{k+1}} \max\left\{\left(\frac{1-e^{-t_{k+1}}}{1-e^{-t_k}}\right)^2 \|\nabla f_{t_k}(\theta_k)\|_2^2,\ \|\nabla f_{t_k}(\theta_{k+1})\|_2^2\right\}}_{\text{optimization error}}$$

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

A general path following algorithm

- **Input**: $\epsilon > 0$, $C_0 \leq 1/4$, $c_1 \geq 1$, $c_2 > 0$, $0 < \alpha_{\max} \leq 5^{-1}$ and $t_{\max} \in (0, \infty]$.
- **Output**: grid points $\{t_k\}_{k=1}^N$ and an approximated solution path $\tilde{\theta}(t)$.
- Initialize $k = 1$. Compute

$$\alpha_1 = \min\{\alpha_{\max}, \ln(1 + \frac{\sqrt{\epsilon}}{\|\nabla L_n(\mathbf{0})\|_2})\}.$$

Starting from $\mathbf{0}$, iteratively calculate θ_1 by minimizing $f_{t_1}(\theta)$ until

$$\|\nabla f_{t_k}(\theta_k)\|_2 \leq C_0 \frac{(e^{\alpha_k} - 1)}{(e^{t_k} - 1)} \|\theta_k\|_2 \tag{3}$$

is satisfied for $k = 1$.

A general path following algorithm

- While
$$\frac{c_2\left(1-e^{-\max(\alpha_k,\, t_{\max}-t_k)}\right)}{e^{t_k}-1} \le \epsilon \text{ or } t_k \ge t_{\max}$$
is not satisfied, compute
$$\alpha_{k+1} = \min\left\{\ln(1+\frac{c_1(e^{\alpha_1}-1)\|\nabla L_n(\mathbf{0})\|_2 e^{t_k/2}(1-e^{-t_k})}{\|\theta_k\|_2}), \alpha_{\max},\, 2\alpha_k\right\},$$
update $t_{k+1} = t_k + \alpha_{k+1}$. Starting from θ_k, iteratively compute θ_{k+1} by minimizing $f_{t_{k+1}}(\theta)$ until (3) is satisfied. Meanwhile, update $k = k + 1$.
- Construct a solution path $\tilde{\theta}(t)$ through linear interpolation of $\{\theta_k\}_{k=1}^N$.

Theoretical guarantees

Theorem 1

Suppose that $L_n(\theta)$ is differentiable and convex. For any $\epsilon > 0$ and $t_{\max} \in (0, \infty]$, assume that either $t_{\max} < \infty$ or $\|\theta(t_{\max})\|_2 < \infty$. Then, our proposed path following algorithm terminates after finite number of iterations, and when terminated, the solution path $\tilde{\theta}(t)$ satisfies

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

Computational complexity

Theorem 2

Suppose that $L_n(\theta)$ is differentiable and convex. For any $\epsilon > 0$ and $t_{\max} \in (0, \infty]$, assume that either $t_{\max} < \infty$ or $\|\theta(t_{\max})\|_2 < \infty$. The total number of grid points required for our proposed path following algorithm to achieve an ϵ-suboptimality:

$$\sup_{0 \le t \le t_{\max}} \{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\} \lesssim \epsilon$$

is at most $\mathcal{O}(\epsilon^{-1/2})$.

Numerical studies

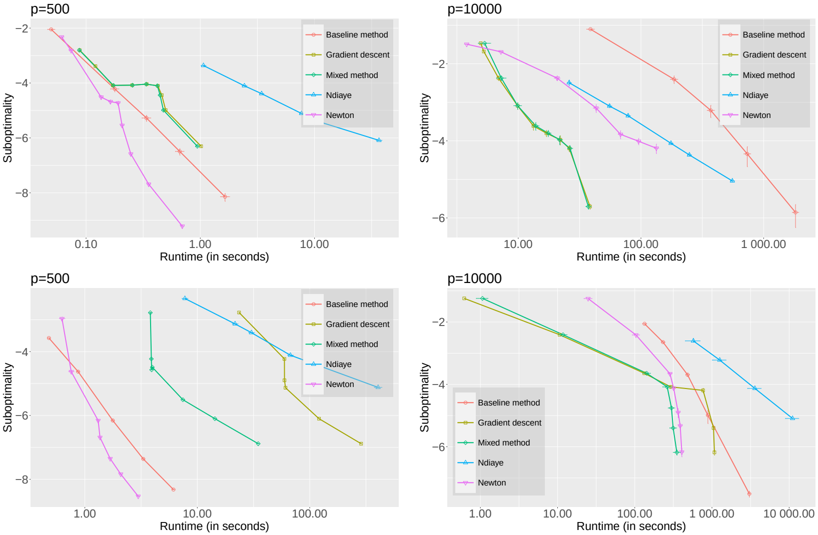

Figure: Runtime v.s. suboptimality when applied to ridge regression (upper panels) and ℓ_2-regularized logistic regression (lower panels).

Numerical studies

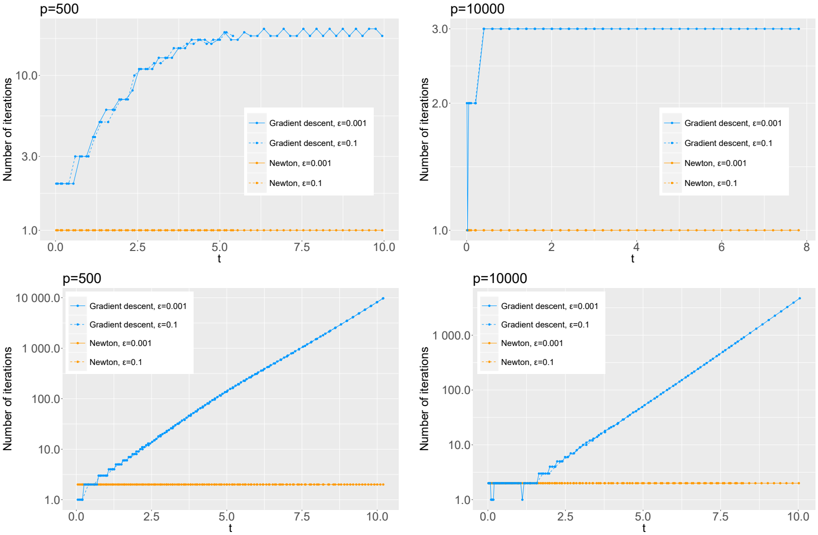

Figure: Number of iterations at each grid point for the Newton and gradient descent methods applying to ridge regression (upper panels) and ℓ_2-regularized logistic regression (lower panels).

Future works

- Extension to nonconvex loss function (see Section 3 of our paper).
- Extension to the case where the loss function or the regularizer are not differentiable.