Supplement Materials for "An Improved GLMNET for L1-regularized Logistic Regression"

Guo-Xun Yuan R96042@CSIE.NTU.EDU.TW
Chia-Hua Ho B95082@CSIE.NTU.EDU.TW
Chih-Jen Lin CJLIN@CSIE.NTU.EDU.TW
Department of Computer Science
National Taiwan University
Taipei 106, Taiwan

Editor: S. Sathiya Keerthi

I. Introduction

This document presents some materials not included in the paper. In Section II, we show that the solution of subproblem (13) converges to zero. In Section III, we show that newGLMNET has quadratic convergence if the loss function $L(\cdot)$ is strictly convex and the exact Hessian is used as H in the quadratic sub-problem. In Section IV, we show that newGLMNET terminates in finite iterations even with our shrinking. In Section V, because of the sparsity of the model, we conduct preliminary experiments to show that some features can be removed before L1 training to make training faster. In Section VI, we investigate the sensitivity of newGLMNET's adaptive inner stopping condition. Results are similar under different ways to adjust the inner stopping condition. Section VII gives a comparison to another method FISTA (Beck and Teboulle, 2009). The results show that newGLMNET is more efficient than FISTA.

II. Convergence of the Solution of Subproblem (13)

For any subsequence $\{\boldsymbol{w}^{k_m}\}$ of $\{\boldsymbol{w}^k\}$, in this section, we show that if

$$\{\boldsymbol{w}^{k_m}\} \text{ converges to some point and } \lim_{m\to\infty} \lambda_{k_m}\|\boldsymbol{d}^{k_m}\| = 0, \tag{II.1}$$

then

$$\lim_{m\to\infty} \|\boldsymbol{d}^{k_m}\| = 0,$$

where λ_{k_m} is the step size of the k_mth iteration in (20). This property will be used in Sections III and IV. The proof follows the approach of Hsieh and Dhillon (2011, Lemma 7).

If $\{\|\boldsymbol{d}^{k_m}\|\}$ does not converge to zero, then there is a subsequence of $\{\boldsymbol{d}^{k_m}\}$, denoted by $\{\boldsymbol{d}^{k_s}\}$, such that $\|\boldsymbol{d}^{k_s}\| > \delta$ for some $\delta > 0$. Because $\{\|\lambda_{k_s}\boldsymbol{d}^{k_s}\|\}$ also converges to zero, the line search condition is not satisfied for step size λ_{k_s}/β.

$$f\left(\boldsymbol{w}^{k_s} + \frac{\lambda_{k_s}}{\beta}\boldsymbol{d}^{k_s}\right) - f(\boldsymbol{w}^{k_s}) > \sigma\frac{\lambda_{k_s}}{\beta}(\nabla L(\boldsymbol{w}^{k_s})^T\boldsymbol{d}^{k_s} + \|\boldsymbol{w}^{k_s} + \boldsymbol{d}^{k_s}\|_1 - \|\boldsymbol{w}^{k_s}\|_1).$$

It can be rewritten by

$$\begin{aligned}&\frac{L\left(\boldsymbol{w}^{k_s}+\frac{\lambda_{k_s}}{\beta}\boldsymbol{d}^{k_s}\right)-L(\boldsymbol{w}^{k_s})+\|\boldsymbol{w}^{k_s}+\frac{\lambda_{k_s}}{\beta}\boldsymbol{d}^{k_s}\|_1-\|\boldsymbol{w}^{k_s}\|_1}{\frac{\lambda_{k_s}}{\beta}}\\&>\sigma(\nabla L(\boldsymbol{w}^{k_s})^T\boldsymbol{d}^{k_s}+\|\boldsymbol{w}^{k_s}+\boldsymbol{d}^{k_s}\|_1-\|\boldsymbol{w}^{k_s}\|_1).\end{aligned} \tag{II.2}$$

By (II.2) and the convexity of $\|\cdot\|_1$, we have

$$\begin{aligned}&\frac{L\left(\boldsymbol{w}^{k_s}+\frac{\lambda_{k_s}}{\beta}\boldsymbol{d}^{k_s}\right)-L(\boldsymbol{w}^{k_s})}{\frac{\lambda_{k_s}}{\beta}}+\|\boldsymbol{w}^{k_s}+\boldsymbol{d}^{k_s}\|_1-\|\boldsymbol{w}^{k_s}\|_1\\&\geq\frac{L\left(\boldsymbol{w}^{k_s}+\frac{\lambda_{k_s}}{\beta}\boldsymbol{d}^{k_s}\right)-L(\boldsymbol{w}^{k_s})+\|\boldsymbol{w}^{k_s}+\frac{\lambda_{k_s}}{\beta}\boldsymbol{d}^{k_s}\|_1-\|\boldsymbol{w}^{k_s}\|_1}{\frac{\lambda_{k_s}}{\beta}}\\&>\sigma(\nabla L(\boldsymbol{w}^{k_s})^T\boldsymbol{d}^{k_s}+\|\boldsymbol{w}^{k_s}+\boldsymbol{d}^{k_s}\|_1-\|\boldsymbol{w}^{k_s}\|_1).\end{aligned} \tag{II.3}$$

From Theorem 1(a) of (Tseng and Yun, 2009),

$$\nabla L(\boldsymbol{w})^T\boldsymbol{d}^{k_s}+\|\boldsymbol{w}+\boldsymbol{d}^{k_s}\|_1-\|\boldsymbol{w}\|_1\leq-\lambda_{\min}\|\boldsymbol{d}^{k_s}\|^2, \tag{II.4}$$

where $\lambda_{\min}$ is defined in Eq. (42) of Appendix A of the paper. We can combine (II.3) and (II.4) to

$$\begin{aligned}&\frac{L\left(\boldsymbol{w}^{k_s}+\frac{\lambda_{k_s}}{\beta}\boldsymbol{d}^{k_s}\right)-L(\boldsymbol{w}^{k_s})}{\frac{\lambda_{k_s}}{\beta}}-\nabla L(\boldsymbol{w}^{k_s})^T\boldsymbol{d}^{k_s}\\&>\ -(1-\sigma)(\nabla L(\boldsymbol{w}^{k_s})^T\boldsymbol{d}+\|\boldsymbol{w}^{k_s}+\boldsymbol{d}\|-\|\boldsymbol{w}^{k_s}\|)\\&\geq(1-\sigma)\lambda_{\min}\|\boldsymbol{d}^{k_s}\|^2.\end{aligned}$$

By dividing both sides by $\|\boldsymbol{d}^{k_s}\|$,

$$\begin{aligned}&\frac{L\left(\boldsymbol{w}^{k_s}+\frac{\lambda_{k_s}}{\beta}\|\boldsymbol{d}^{k_s}\|\frac{\boldsymbol{d}^{k_s}}{\|\boldsymbol{d}^{k_s}\|}\right)-L(\boldsymbol{w}^{k_s})}{\frac{\lambda_{k_s}}{\beta}\|\boldsymbol{d}^{k_s}\|}-\nabla L(\boldsymbol{w}^{k_s})^T\frac{\boldsymbol{d}^{k_s}}{\|\boldsymbol{d}^{k_s}\|}\\&\geq(1-\sigma)\lambda_{\min}\|\boldsymbol{d}^{k_s}\|\\&\geq(1-\sigma)\lambda_{\min}\delta.\end{aligned}$$

However, there exists a subsequence of $\{\boldsymbol{d}^{k_s}\}$ such that $\boldsymbol{d}^{k_s}/\|\boldsymbol{d}^{k_s}\|$ converges to some point because $\boldsymbol{d}^{k_s}/\|\boldsymbol{d}^{k_s}\|\in\{\boldsymbol{d}\mid\|\boldsymbol{d}\|\leq 1\}$, which is compact. Besides, $\{\lambda_{k_s}\|\boldsymbol{d}^{k_s}\|\}$ also converges to zero. Then the left side converges to zero in the subsequence, so there is a contradiction. As a result, $\{\boldsymbol{d}^{k_m}\}$ converges to zero.

III. Quadratic Convergence of newGLMNET

In Appendix B of the paper, we showed that if the loss term $L(\boldsymbol{w})$ is strictly convex, the convergence rate of newGLMNET is at least linear. Note that we set $H=\nabla^2L(\boldsymbol{w})+\nu\mathcal{I}$ in

(19) of the paper because $\nabla^2 L(\boldsymbol{w})$ may not be positive definite. If it is known beforehand that the loss term $L(\boldsymbol{w})$ is strictly convex, then $\nabla^2 L(\boldsymbol{w})$ is positive definite without the $\nu\mathcal{I}$ term. In this section, we show that newGLMNET converges quadratically if the loss term $L(\boldsymbol{w})$ is strictly convex and $H = \nabla^2 L(\boldsymbol{w})$.[1] The proof follows the approach in Hsieh et al. (2011, Section 7.3), and we assume that

- each sub-problem is exactly solved,
- shrinking is not considered,
- $L(\boldsymbol{w})$ is strictly convex, and
- σ in the line search (Eq. (20) of the paper) is in the $(0, 0.5)$ interval.

To begin, we show that newGLMNET with $H = \nabla^2 L(\boldsymbol{w})$ still converges to the optimal solution. We only need to show

$$\exists \lambda_{\min} > 0 \text{ s.t. } H^k \succeq \lambda_{\min}\mathcal{I}, \forall k$$

still holds (Eq. (42) of the paper). Because $L(\boldsymbol{w})$ is strictly convex and $\nabla^2 L(\boldsymbol{w})$ is continuous,

$$\lambda_{\min} \equiv \min_{\boldsymbol{w} \in \{\boldsymbol{w} | f(\boldsymbol{w}) \leq f(\boldsymbol{w}^1)\}} \lambda(\nabla^2 L(\boldsymbol{w})) > 0 \tag{III.1}$$

exists, where $\lambda(\nabla^2 L(\boldsymbol{w}))$ is the smallest eigenvalue of $\nabla^2 L(\boldsymbol{w})$. Therefore, similar to Appendix A of the paper, newGLMNET with $H = \nabla^2 L(\boldsymbol{w})$ converges to the optimal solution.

Next, we divide the indices into three groups.

$$\begin{aligned} P &= \{j \mid \nabla_j L(\boldsymbol{w}^*) = -1, \} \\ Z &= \{j \mid -1 < \nabla_j L(\boldsymbol{w}^*) < 1, \} \\ N &= \{j \mid \nabla_j L(\boldsymbol{w}^*) = 1, \}, \end{aligned}$$

where $\boldsymbol{w}^*$ is the unique optimal solution. From the optimality condition,

$$\begin{cases} w_j \geq 0 & \text{if } j \in P, \\ w_j = 0 & \text{if } j \in Z, \\ w_j \leq 0 & \text{if } j \in N. \end{cases}$$

Therefore, it is equivalent to solving the following problem.

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & F(\boldsymbol{w}) \equiv L(\boldsymbol{w}) + \sum_{j \in P} w_j - \sum_{j \in N} w_j \\ \text{subject to} \quad & w_j \geq 0 \text{ if } j \in P \\ & w_j = 0 \text{ if } j \in Z \\ & w_j \leq 0 \text{ if } j \in N \end{aligned} \tag{III.2}$$

If all constraints in (III.2) are satisfied during the optimization process, then (III.2) becomes an unconstrained problem.

$$\min_{\boldsymbol{w}} \quad F(\boldsymbol{w}).$$

[1] Note that even if $\nabla^2 L(\boldsymbol{w})$ is positive definite, newGLMNET only guarantees to converge at least linearly if we use $H = \nabla^2 L(\boldsymbol{w}) + \nu\mathcal{I}$; see Appendix B of the paper.

Previous works (e.g., Dunn, 1980) have shown that if $F(\boldsymbol{w})$ is strictly convex, $\nabla^2 F(\boldsymbol{w})$ is Lipschitz continuous, and $\boldsymbol{w}^k$ is updated by

$$\boldsymbol{w}^{k+1} = \arg\min_{\boldsymbol{w}} \nabla F(\boldsymbol{w}^k)^T(\boldsymbol{w} - \boldsymbol{w}^k) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^k)\nabla^2 F(\boldsymbol{w}^k)(\boldsymbol{w} - \boldsymbol{w}^k), \tag{III.3}$$

then $\{\boldsymbol{w}^k\}$ converges quadratically to $\boldsymbol{w}^*$ for all $\boldsymbol{w}^1$ sufficient close to $\boldsymbol{w}^*$. Because the sub-problem of newGLMNET is exactly the same as (III.3) (if the sign of $\boldsymbol{w}$ is fixed), to have quadratic convergence, we only need to show that newGLMNET with L1-regularized logistic regression has the following properties.

1. $F(\boldsymbol{w})$ is strictly convex and $\nabla^2 F(\boldsymbol{w})$ is Lipschitz continuous,
2. when k is large enough, the step size is always one, and
3. when k is large enough, all constraints in (III.2) are satisfied.

In the following three subsections, we will prove that the three properties hold.

III.1 Strict Convexity and Lipschitz Continuity

Because $L(\boldsymbol{w})$ is strictly convex, $F(\boldsymbol{w})$ is also strictly convex. Besides, for all $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$,

$$\begin{aligned}
\|\nabla^2 F(\boldsymbol{w}_1) - \nabla^2 F(\boldsymbol{w}_2)\| &= \|XD(\boldsymbol{w}_1)X^T - XD(\boldsymbol{w}_1)X^T\| \\
&\leq \|X\|\|X^T\|\|D(\boldsymbol{w}_1) - D(\boldsymbol{w}_2)\| \\
&\leq \|X\|\|X^T\|(\max_i |D_{ii}(\boldsymbol{w}_1) - D_{ii}(\boldsymbol{w}_2)|) \qquad \text{(III.4)} \\
&= \|X\|\|X^T\| \left(\max_i |\nabla D_{ii}(\tilde{\boldsymbol{w}})^T(\boldsymbol{w}_1 - \boldsymbol{w}_2)|\right) \\
&= \|X\|\|X^T\| \left(\max_i |(\tilde{\tau}(1-\tilde{\tau})^2 - \tilde{\tau}^2(1-\tilde{\tau}))y_i\tilde{\boldsymbol{w}}^T(\boldsymbol{w}_1 - \boldsymbol{w}_2)|\right) \\
&= \|X\|\|X^T\| \left(\max_i |(\tilde{\tau}(1-\tilde{\tau})(1-2\tilde{\tau})y_i\boldsymbol{x}_i)^T(\boldsymbol{w}_1 - \boldsymbol{w}_2)|\right) \\
&\leq \|X\|\|X^T\| \left(\max_i \|\tilde{\tau}(1-\tilde{\tau})(1-2\tilde{\tau})y_i\boldsymbol{x}_i\|\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|\right),
\end{aligned}$$

where $\tilde{\tau} = \tau(y_i\tilde{\boldsymbol{w}}^T\boldsymbol{x})$ for some $\tilde{\boldsymbol{w}}$ between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, and (III.4) is from that $D(\boldsymbol{w})$ is diagonal. Because $\tilde{\tau}$ is between zero and one, $\tilde{\tau}(1-\tilde{\tau})(1-2\tilde{\tau}) \in (-1, 1)$. Then,

$$\|\nabla^2 F(\boldsymbol{w}_1) - \nabla^2 F(\boldsymbol{w}_2)\| \leq \left(\|X\|\|X^T\| \max_i \|\boldsymbol{x}_i\|\right) \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|.$$

Hence, $\nabla^2 F(\boldsymbol{w})$ is Lipschitz continuous.

III.2 Step Size is Always One After k Large Enough

To simplify the notation, we define

$$\begin{aligned}
\boldsymbol{d} &\equiv \arg\min_{\boldsymbol{d}} \nabla L(\boldsymbol{w})^T\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T\nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \|\boldsymbol{w} + \boldsymbol{d}\| - \|\boldsymbol{w}\|, \\
\tilde{L}(t) &\equiv L(\boldsymbol{w} + t\boldsymbol{d}),
\end{aligned}$$

and the right-hand side of the sufficient decrease condition (Eq. (20) of the paper) with $\gamma = 0$ as

$$\Delta \equiv \nabla L(\boldsymbol{w})^T \boldsymbol{d} + \|\boldsymbol{w} + \boldsymbol{d}\| - \|\boldsymbol{w}\|.$$

Because $\nabla^2 L(\boldsymbol{w})$ is Lipschitz continuous, we have

$$\tilde{L}''(t) - \tilde{L}''(0) \leq |\tilde{L}''(t) - \tilde{L}''(0)| = |\boldsymbol{d}^T(\nabla^2 L(\boldsymbol{w} + t\boldsymbol{d}) - \nabla^2 L(\boldsymbol{w}))\boldsymbol{d}| \leq t\Lambda\|\boldsymbol{d}\|^3, \qquad \text{(III.5)}$$

where Λ is the Lipschitz constant. Because $\tilde{L}''(0) = \boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d}$, (III.5) can be rewritten as

$$\tilde{L}''(t) \leq \boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + t\Lambda\|\boldsymbol{d}\|^3.$$

By integrating both sides with t, we have

$$\tilde{L}'(t) \leq \tilde{L}'(0) + t\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + t^2\frac{1}{2}\Lambda\|\boldsymbol{d}\|^3 = \nabla L(\boldsymbol{w})^T\boldsymbol{d} + t\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + t^2\frac{1}{2}\Lambda\|\boldsymbol{d}\|^3.$$

By integrating both sides again, we have

$$\tilde{L}(t) \leq \tilde{L}(0) + t\nabla L(\boldsymbol{w})^T\boldsymbol{d} + \frac{1}{2}t^2\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \frac{1}{6}t^3\Lambda\|\boldsymbol{d}\|^3.$$

Replace t with one and add the one norm term. Then we have

$$\begin{aligned}
f(\boldsymbol{w} + \boldsymbol{d}) &= \|\boldsymbol{w} + \boldsymbol{d}\|_1 + L(\boldsymbol{w} + \boldsymbol{d}) \\
&\leq \tilde{L}(0) + \|\boldsymbol{w}\|_1 + \nabla L(\boldsymbol{w})^T\boldsymbol{d} + \|\boldsymbol{w} + \boldsymbol{d}\|_1 - \|\boldsymbol{w}\|_1 + \frac{1}{2}\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \frac{1}{6}\Lambda\|\boldsymbol{d}\|^3 \\
&= f(\boldsymbol{w}) + \Delta + \frac{1}{2}\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \frac{1}{6}\Lambda\|\boldsymbol{d}\|^3.
\end{aligned}$$

Because $\boldsymbol{d}$ minimizes the quadratic sub-problem, for any $\alpha \in (0, 1)$,

$$\begin{aligned}
&0 \geq q(\boldsymbol{d}) - q(\alpha\boldsymbol{d}) \\
&= \left(\nabla L(\boldsymbol{w})^T\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \|\boldsymbol{w} + \boldsymbol{d}\|_1\right) - \left(\alpha\nabla L(\boldsymbol{w})^T\boldsymbol{d} + \alpha^2\frac{1}{2}\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \|\boldsymbol{w} + \alpha\boldsymbol{d}\|_1\right) \\
&\geq (1 - \alpha)\nabla L(\boldsymbol{w})^T\boldsymbol{d} + (1 - \alpha^2)\frac{1}{2}\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \|\boldsymbol{w} + \boldsymbol{d}\|_1 - (\alpha\|\boldsymbol{w} + \boldsymbol{d}\|_1 + (1 - \alpha)\|\boldsymbol{w}\|_1) \\
&= (1 - \alpha)\left(\nabla L(\boldsymbol{w})^T\boldsymbol{d} + (1 + \alpha)\frac{1}{2}\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \|\boldsymbol{w} + \boldsymbol{d}\|_1 - \|\boldsymbol{w}\|_1\right).
\end{aligned}$$

By taking $\alpha \to 0$, we have

$$0 \geq \nabla L(\boldsymbol{w})^T\boldsymbol{d} + \boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} + \|\boldsymbol{w} + \boldsymbol{d}\| - \|\boldsymbol{w}\|.$$

Therefore,

$$\lambda_{\min}\|\boldsymbol{d}\|^2 \leq \boldsymbol{d}^T \nabla^2 L(\boldsymbol{w})\boldsymbol{d} \leq -\nabla L(\boldsymbol{w})^T\boldsymbol{d} - \|\boldsymbol{w} + \boldsymbol{d}\| + \|\boldsymbol{w}\| = -\Delta, \text{ and}$$

$$f(\boldsymbol{w} + \boldsymbol{d}) - f(\boldsymbol{w}) \leq \left(\frac{1}{2} - \frac{\Lambda\|\boldsymbol{d}\|}{6\lambda_{\min}}\right)\Delta.$$

Because $\boldsymbol{d}$ converges to zero by Section II, when k is large enough, the step size is one for all $\sigma < 1/2$.

III.3 Constraints can be Ignored After k Large Enough

The cases for $j \in P$, $j \in N$, and $j \in Z$ are similar, so we only show the case for $j \in Z$. Please see Hsieh et al. (2011, Lemma 11) for the case $j \in P$.

Assume there exists an infinite subsequence $\{\boldsymbol{w}^{s_t}\}$ such that $w_j^{s_t} \neq 0, \forall t$. When t is large enough, then $\boldsymbol{w}^{s_t} = \boldsymbol{w}^{s_t-1} + \boldsymbol{d}$ by Section III.2, where

$$\boldsymbol{d} = \arg\min_{\boldsymbol{d}} \nabla L(\boldsymbol{w}^{s_t-1})^T \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T \nabla^2 L(\boldsymbol{w}^{s_t-1})\boldsymbol{d} + \|\boldsymbol{w}^{s_t-1} + \boldsymbol{d}\| - \|\boldsymbol{w}^{s_t-1}\|. \tag{III.6}$$

Because $w_j^{s_t} = w_j^{s_t-1} + d_j \neq 0$ and $\boldsymbol{d}$ is the optimal solution of (III.6), we have

$$\nabla_j L(\boldsymbol{w}^{s_t-1}) + (\nabla^2 L(\boldsymbol{w}^{s_t-1})\boldsymbol{d})_j = +1 \text{ or } -1.$$

Besides, by the convergence of $\{\boldsymbol{w}^k\}$ and Section II, $\boldsymbol{d}$ converges to zero, so $\nabla_j L(\boldsymbol{w}^{s_t-1})$ should converge to $+1$ or -1. This is a contradiction to the assumption that $\nabla_j L(\boldsymbol{w}^*) \in (-1, 1)$. Therefore, $w_j^k = 0$ if $j \in Z$ for all k large enough.

IV. Finite Termination and Asymptotic Convergence of Algorithm 3 with Shrinking

In Appendix A of the paper, without considering shrinking, we showed that if the subproblem (13) is exactly solved, then any limit point of $\{\boldsymbol{w}^k\}$ generated by newGLMNET is an optimal point. In this section, we show the following two results if shrinking is applied and the subproblem is exactly solved.

1. For any given stopping tolerance ϵ_{out}, newGLMNET in Algorithm 3 terminates in finite iterations.
2. Any limit point of $\{\boldsymbol{w}^k\}$ is optimal.

We denote J^k as the working set in the kth iteration. The followings are the steps of our proof.

1. Proof by contradiction. Assume there is a infinite sequence $\{\boldsymbol{w}^k\}$ generated by Algorithm 3.
2. $\{f(\boldsymbol{w}^k)\}$ strictly decreases and converges to a point $\tilde{f}$.
3. $\{\boldsymbol{w}^{k_m}\}$, a subsequence of $\{\boldsymbol{w}^k\}$, converges to a point $\tilde{\boldsymbol{w}}$, and so do $\{\boldsymbol{w}^{k_m-1}\}$ and $\{\boldsymbol{w}^{k_m+1}\}$.
4. Then $\|\nabla^S f(\boldsymbol{w}^{k_m})\|_1$ converges to zero, and therefore Algorithm 3 terminates in finite iterations.
5. We can assume that $\boldsymbol{w}^{k_m}$ for all m have the same working set J.
6. $\tilde{\boldsymbol{w}}$ is an optimal solution in terms of J.
7. $\nabla f(\tilde{\boldsymbol{w}}) = \boldsymbol{0}$, so $\tilde{\boldsymbol{w}}$ is actually an optimal solution of the whole problem.
8. Any limit point of $\{\boldsymbol{w}^k\}$ is an optimal point.

Under a given ϵ_{out}, if newGLMNET by Algorithm 3 does not stop in finite iterations, then $\{f(\boldsymbol{w}^k)\}$ is an infinite sequence and

$$\|\nabla^S_{J^k} f(\boldsymbol{w}^k)\|_1 > \epsilon_{\text{out}} \tag{IV.1}$$

for all k. From Theorem 1(a) of Tseng and Yun (2009), we have

$$f(\boldsymbol{w}^{k+1}) - f(\boldsymbol{w}^k) \leq -\sigma\lambda_k\lambda_{\min}\|\boldsymbol{d}\|^2 \leq -\sigma\lambda_{\min}\|\lambda_k\boldsymbol{d}\|^2, \forall k, \tag{IV.2}$$

where σ and λ_k are defined in (20), and $\lambda_{\min} = \nu$ as mentioned in Eq. (42) of Appendix A of the paper. Note that $d_j = 0$ for $j \notin J^k$ because we consider shrinking now and $\lambda_k > 0$ because $\nabla L(\boldsymbol{w})$ is Lipschitz continuous; see Appendix A of the paper and Lemma 5(b) of Tseng and Yun (2009). Eq. (IV.2) guarantees that $\{f(\boldsymbol{w}^k)\}$ strictly decreases. Because $\{f(\boldsymbol{w}^k)\}$ is lower bounded, it converges to a value $\tilde{f}$.

$$\lim_{k\to\infty} f(\boldsymbol{w}^k) = \tilde{f}. \tag{IV.3}$$

Because the level set $\{\boldsymbol{w} \mid f(\boldsymbol{w}) \leq f(\boldsymbol{w}^1)\}$ is bounded and closed, we can find a subsequence of $\{\boldsymbol{w}^k\}$ such that the subsequence $\{\boldsymbol{w}^{k_m}\}$ converges to $\tilde{\boldsymbol{w}}$. Besides, with $\lim_{k\to\infty} f(\boldsymbol{w}^{k+1}) - f(\boldsymbol{w}^k) = 0$, by taking the limit of (IV.2), we have

$$\lim_{m\to\infty} \|\lambda_{k_m-1}\boldsymbol{d}^{k_m-1}\| = 0. \tag{IV.4}$$

Further,

$$\lim_{m\to\infty} \boldsymbol{w}^{k_m-1} = \lim_{m\to\infty} \boldsymbol{w}^{k_m} = \lim_{m\to\infty} \boldsymbol{w}^{k_m+1} = \tilde{\boldsymbol{w}} \tag{IV.5}$$

because $\lambda_{k_m}\boldsymbol{d}^{k_m} = \boldsymbol{w}^{k_m+1} - \boldsymbol{w}^{k_m}$. From the result in Section II, (IV.4) and (IV.5) then imply

$$\lim_{m\to\infty} \|\boldsymbol{d}^{k_m-1}\| = 0. \tag{IV.6}$$

Then we show that $\|\nabla^S f(\boldsymbol{w}^{k_m})\|_1$ converges to zero. If $\boldsymbol{d}^{k_m-1}_{J^{k_m-1}} \in \mathbf{R}^{|J^{k_m-1}|}$ is the solution of the subproblem based on $\boldsymbol{w}^{k_m-1}$, by the optimality of the subproblem, we have for $j \in J^{k_m-1}$,

$$\nabla_j L(\boldsymbol{w}^{k_m-1}) + (H^{k_m-1}\boldsymbol{d}^{k_m-1})_j \in \begin{cases} \{-1\} & \text{if } w_j^{k_m} > 0, \\ \{1\} & \text{if } w_j^{k_m} < 0, \\ [-1,1] & \text{if } w_j^{k_m} = 0, \end{cases}$$

where $d_{j'}^{k_m-1} = 0$ for $j' \notin J^{k_m-1}$. By changing the signs of both sides and adding $\nabla_j L(\boldsymbol{w}^{k_m})$, we have for $j \in J^{k_m-1}$,

$$\begin{aligned} &\nabla_j L(\boldsymbol{w}^{k_m}) - \nabla_j L(\boldsymbol{w}^{k_m-1}) - (H^{k_m-1}\boldsymbol{d}^{k_m-1})_j \\ &\quad\in \begin{cases} \{\nabla_j L(\boldsymbol{w}^{k_m}) + 1\} & \text{if } w_j^{k_m} > 0, \\ \{\nabla_j L(\boldsymbol{w}^{k_m}) - 1\} & \text{if } w_j^{k_m} < 0, \\ [\nabla_j L(\boldsymbol{w}^{k_m}) - 1, \nabla_j L(\boldsymbol{w}^{k_m}) + 1] & \text{if } w_j^{k_m} = 0. \end{cases} \end{aligned} \tag{IV.7}$$

From the definition of the minimum-norm subgradient, (IV.7) leads

$$|\nabla_j^S f(\boldsymbol{w}^{k_m})| \leq |\nabla_j L(\boldsymbol{w}^{k_m}) - \nabla_j L(\boldsymbol{w}^{k_m-1}) - (H^{k_m-1}\boldsymbol{d}^{k_m-1})_j|. \tag{IV.8}$$

The right-hand side converges to to zero because of (IV.6) and the continuity of $\nabla L(\boldsymbol{w})$. Therefore,

$$\lim_{m\to\infty} |\nabla_j^S f(\boldsymbol{w}^{k_m})| = 0, \forall j \in J^{k_m-1}. \tag{IV.9}$$

Further, for $j \notin J^{k_m-1}$, by the way we choose the working set,

$$\nabla_j L(\boldsymbol{w}^{k_m-1}) \in [-1,1] \text{ and } w_j^{k_m-1} = 0.$$

Because $w_j^{k_m-1} = w_j^{k_m}$, we have for $j \notin J^{k_m-1}$

$$\nabla_j L(\boldsymbol{w}^{k_m}) - \nabla_j L(\boldsymbol{w}^{k_m-1}) \in [\nabla_j L(\boldsymbol{w}^{k_m}) - 1, \nabla_j L(\boldsymbol{w}^{k_m}) + 1] \text{ and } w_j^{k_m} = 0,$$

and similar to (IV.8),

$$|\nabla_j^S f(\boldsymbol{w}^{k_m})| \leq |\nabla_j L(\boldsymbol{w}^{k_m}) - \nabla_j L(\boldsymbol{w}^{k_m-1})|.$$

The right-hand side also converges to zero. With (IV.9), we have $\|\nabla^S f(\boldsymbol{w}^{k_m})\|$ converges to zero, a contradiction to (IV.1). As the result, the stopping condition (29) can be satisfied in finite iterations.

Next, we show that $\tilde{\boldsymbol{w}}$ is actually an optimal solution. Because of the finite (i.e., 2^n) possible working sets, one set must appear infinite times. As a result, we can assume all $\{\boldsymbol{w}^{k_m}\}$ have the same working set J. If not, we can take a subsequence of $\{\boldsymbol{w}^{k_m}\}$ again to remove the elements with different working sets. By the way of finding the working set J, we have

$$w_j^{k_m} = d_j = 0 \text{ and } \nabla_j^S f(\boldsymbol{w}^{k_m}) = 0, \text{ for all } j \notin J. \tag{IV.10}$$

Assume $\tilde{\boldsymbol{w}}$ is not an optimal solution (in terms of J). If we construct a subproblem based on $\tilde{\boldsymbol{w}}$, Tseng and Yun (2009, Lemma 2) show that the optimal solution $\|\tilde{\boldsymbol{d}}\| \neq 0$. Further, from (IV.2), we get

$$f(\tilde{\boldsymbol{w}} + \tilde{\lambda}\tilde{\boldsymbol{d}}) < f(\tilde{\boldsymbol{w}}), \tag{IV.11}$$

where $\tilde{\lambda}$ is the step size obtained by linear search. Besides, for any $\boldsymbol{w}^{k_m}$, we have

$$f(\boldsymbol{w}^{k_m} + \tilde{\lambda}\tilde{\boldsymbol{d}}) \geq f(\boldsymbol{w}^{k_m+1}). \tag{IV.12}$$

Because $f(\boldsymbol{w})$ is continuous, taking the limit of (IV.12) implies

$$f(\tilde{\boldsymbol{w}} + \tilde{\lambda}\tilde{\boldsymbol{d}}) = \lim_{m\to\infty} f(\boldsymbol{w}^{k_m} + \tilde{\lambda}\tilde{\boldsymbol{d}}) \geq \lim_{m\to\infty} f(\boldsymbol{w}^{k_m+1}) = f(\tilde{\boldsymbol{w}}), \tag{IV.13}$$

which is a contradiction to (IV.11). Therefore, $\tilde{\boldsymbol{w}}$ is an optimal solution (in terms of J).

We have shown that $\tilde{\boldsymbol{w}}$ is optimal in terms of the working set J, so

$$\nabla_J^S f(\tilde{\boldsymbol{w}}) = \boldsymbol{0}. \tag{IV.14}$$

For $j \notin J$, because $w_j^{k_m} = 0$ are constant for all m,

$$|\nabla_j^S f(\boldsymbol{w}^{k_m})| = \max(0, \nabla_j L(\boldsymbol{w}^{k_m}) - 1).$$

Besides, $\nabla_j L(\boldsymbol{w}^{k_m})$ is continuous, so $\max(0, \nabla_j L(\boldsymbol{w}^{k_m}) - 1)$ is continuous. Therefore, by (IV.10), for $j \notin J$,

$$\begin{aligned} 0 &= \lim_{m\to\infty} |\nabla_j^S f(\boldsymbol{w}^{k_m})| \\ &= \lim_{m\to\infty} \max(0, \nabla_j L(\boldsymbol{w}^{k_m}) - 1) \\ &= \max(0, \nabla_j L(\tilde{\boldsymbol{w}}) - 1) = |\nabla_j^S f(\tilde{\boldsymbol{w}})|. \end{aligned} \tag{IV.15}$$

By (IV.14) and (IV.15), we obtain

$$\|\nabla^S f(\tilde{\boldsymbol{w}})\| = 0.$$

Therefore, $\tilde{\boldsymbol{w}}$ is an optimal solution of the whole problem (1).

Finally, we prove that any limit point of $\{\boldsymbol{w}^k\}$ is optimal. From (IV.3) and the result that $\{\boldsymbol{w}^{k_m}\}$ converges to an optimal solution, $\{f(\boldsymbol{w}^{k_m})\}$ globally converges to the optimal function value f^*. Because $f(\boldsymbol{w})$ is continuous, any limit point of $\{\boldsymbol{w}^k\}$ has the function value f^*. Thus, the proof is complete.

V. Rules for Feature Elimination

In Section 5.3, we describe our shrinking scheme for removing features in solving L1-regularized problems. Some recent works such as El Ghaoui et al. (2010); Tibshirani et al. (2011) proposed rules to cheaply eliminate features prior to the L1 training. We refer to these rules as screening methods. Different from shrinking, which couples with a learning algorithm, screening can be viewed as a data preprocessing step. The screening methods identify some features corresponding to $w_j^* = 0$ in the final model. By (safely) pruning those features, the size of a training data set can be reduced. Further, these screening methods are independent of optimization methods to solve (1).

Because screening methods aim to quickly identify some zero elements without solving the optimization problem, the effectiveness may be limited. In the following experiments, we investigate how GLMNET (or newGLMNET) may benefit from a screening method. Instead of applying a specific screening method, we consider the best situation where all indices with $w_j^* = 0$ can be identified. Because such an ideal screening method is not available, we run an L1 solver first on the original data set to identify all zero elements of $\boldsymbol{w}^*$.

We conduct the running time comparison between GLMNET and newGLMNET on "the screened data sets." All other experimental settings are the same as those described in Section 6. Results in Figure 1 indicate that the running time of GLMNET and newGLMNET on the screen sets of news20 and gisette is significantly reduced. However, on the two larger sets rcv1 and yahoo-korea, the running time is similar with/without the screening procedure. The reason might be that the shrinking procedure in GLMNET and newGLMNET is very effective to remove some features.

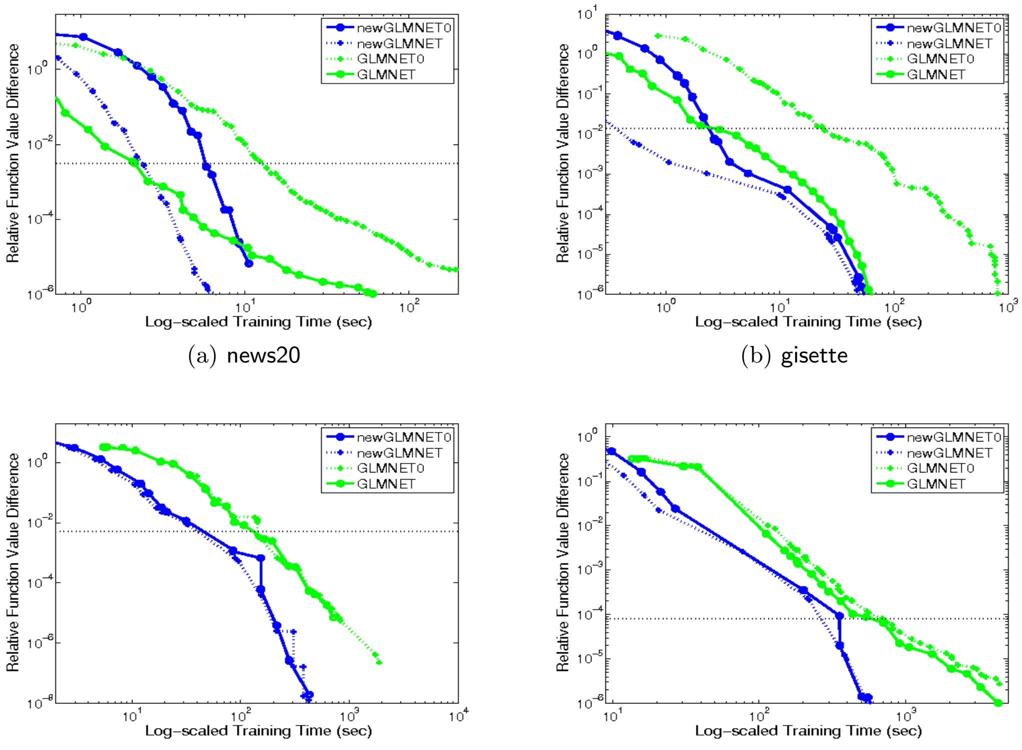
The preliminary experiments here show that the pre-processing step of pruning some zero elements can possibly improve the training speed of GLMNET and newGLMNET on some data sets.

VI. Update Rules for the Inner Stopping Tolerance ϵ_{in}

In the paper, the inner stopping condition (28) is updated by

$$\epsilon_{\text{in}} \leftarrow \epsilon_{\text{in}}/4$$

if the CD procedure exits after one iteration. To see if newGLMNET is sensitive to the reduction ratio, in Figures 2 and 3 we respectively show function-value reduction and accuracy of using $\epsilon_{\text{in}}/2, \epsilon_{\text{in}}/4,$ and $\epsilon_{\text{in}}/10$. Results are very similar, although the update by $\epsilon_{\text{in}}/2$ is slightly worse. One possible reason is that the sub-problem is loosely solved and more



(a) news20

(b) gisette

(c) rcv1

(d) yahoo-korea

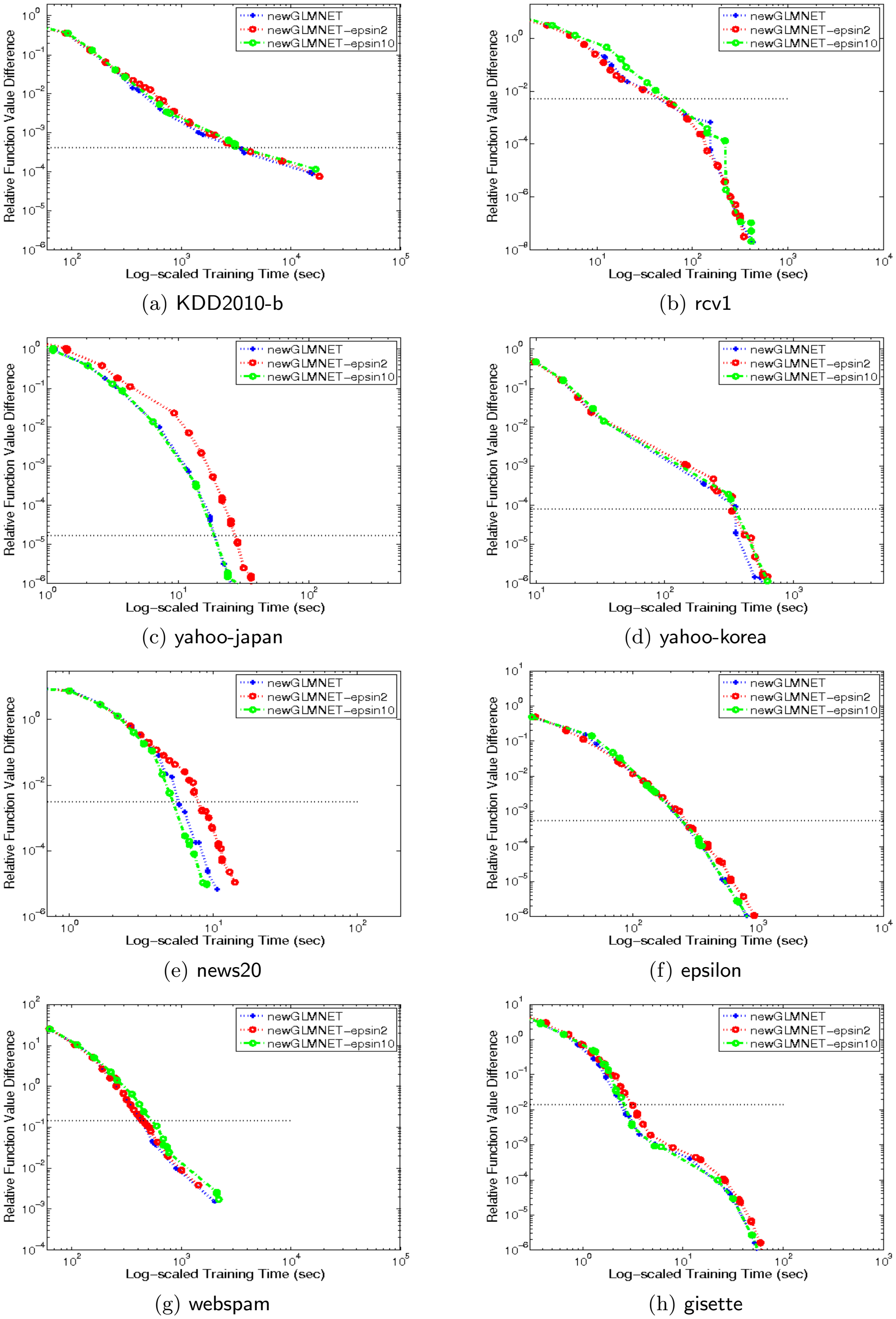
Figure 1: Running time comparisons between GLMNET and newGLMNET on the ideally screened data sets (i.e., features with $w_j^* = 0$ are removed before training). The legend GLMNET0 and newGLMNET0 are used, respectively, to indicate the curves of running GLMNET and newGLMNET on the original data for reference.

iterations are needed to achieve similar function values of using $\epsilon_{\text{in}}/4$ or $\epsilon_{\text{in}}/10$. Therefore, a large number of gradient evaluations of the loss function causes longer running time.

VII. Comparison between newGLMNET and FISTA

We conducted experiment to compare FISTA and newGLMNET for solving L1-regularized logistic regression (1). To have a simple comparison, we do not consider shrinking. The Lipschitz constant of logistic loss is $C\|X^T\|\|X\|$ as shown in Appendix A of the paper; see Eqs. (39) and (41). We do not count the time of evaluating the constant when comparing FISTA and newGLMNET.

At each iteration, FISTA finds the gradient direction, but newGLMNET finds the Newton direction. Therefore, FISTA needs less time in each iteration. However, because of the slow convergence speed, FISTA is overall slower than newGLMNET. Figure 4 shows the relative function value difference versus the training time. Each point corresponds to an iteration.



(a) KDD2010-b

(b) rcv1

(c) yahoo-japan

(d) yahoo-korea

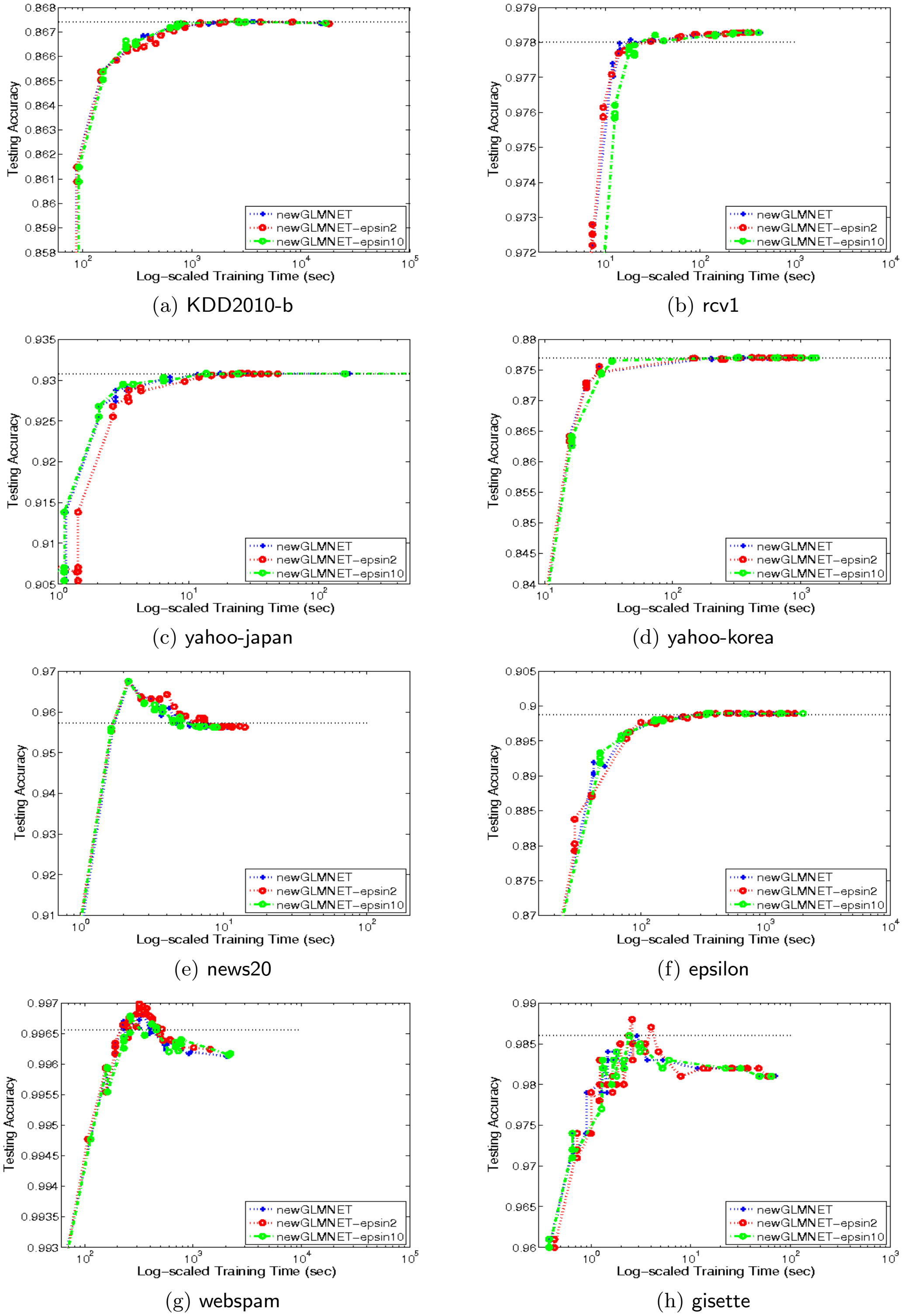
(e) news20

(f) epsilon

(g) webspam

(h) gisette

Figure 2: L1-regularized logistic regression: relative difference to the optimal function value versus training time. Both x-axis and y-axis are log-scaled. We compare newGLMNET with three update rules for inner stopping tolerance: $\epsilon_{\text{in}} \leftarrow \epsilon_{\text{in}}/2, \epsilon_{\text{in}}/4,$ and $\epsilon_{\text{in}}/10$.



(a) KDD2010-b

(b) rcv1

(c) yahoo-japan

(d) yahoo-korea

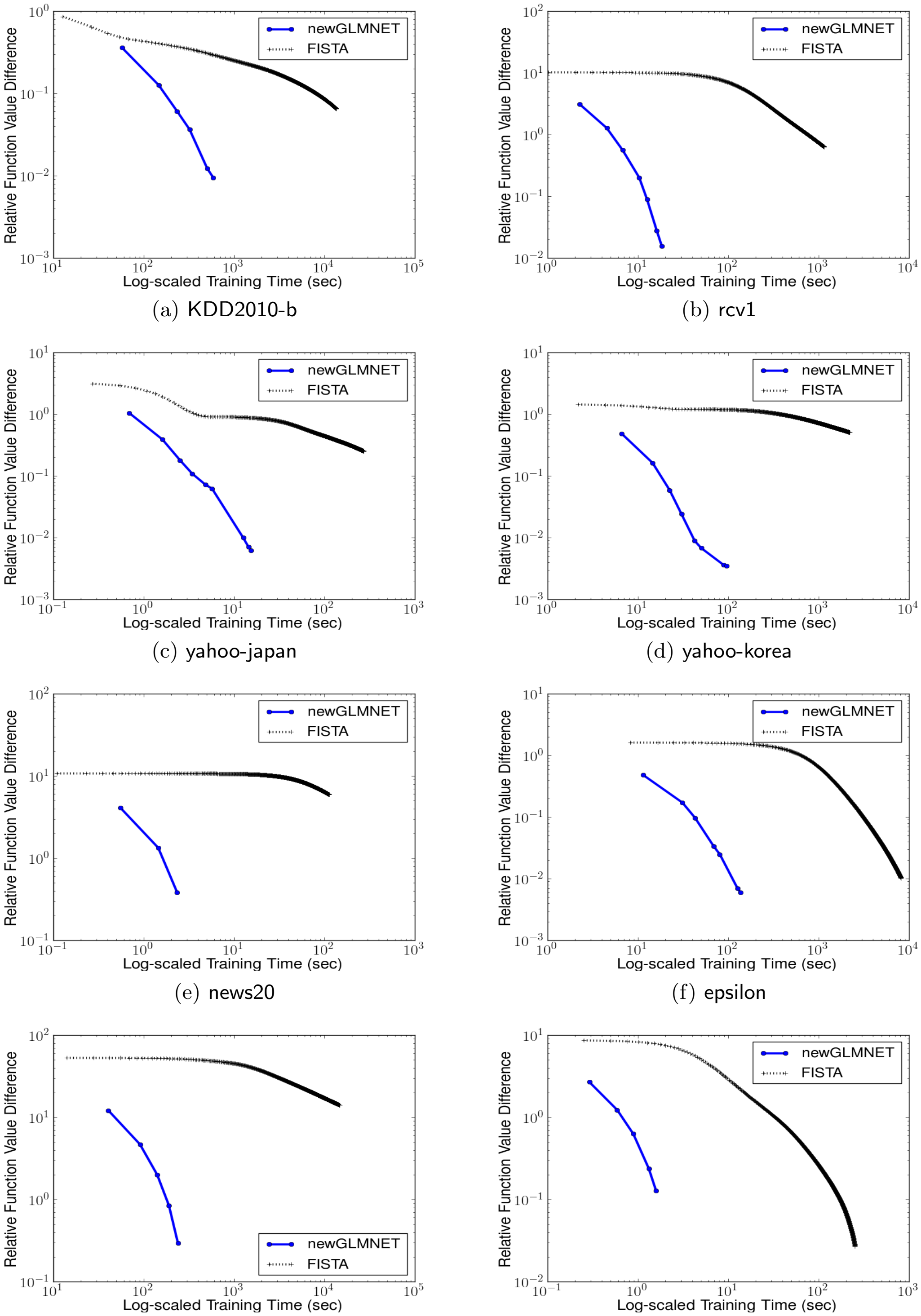
(e) news20

(f) epsilon

(g) webspam

(h) gisette

Figure 3: L1-regularized logistic regression: testing accuracy versus training time (log-scaled). We compare newGLMNET with three update rules for inner stopping tolerance: $\epsilon_{\text{in}} \leftarrow \epsilon_{\text{in}}/2, \epsilon_{\text{in}}/4,$ and $\epsilon_{\text{in}}/10$.



(a) KDD2010-b

(b) rcv1

(c) yahoo-japan

(d) yahoo-korea

(e) news20

(f) epsilon

(g) webspam

(h) gisette

Figure 4: Comparison of newGLMNET and FISTA for L1-regularized logistic regression (1). Relative difference to the optimal function value versus training time. Both x-axis and y-axis are log-scaled.

We also tried to use FISTA to solve the sub-problem (13) in GLMNET or newGLMNET. Sub-problem (13) can be considered as a L1-regularized least-square regression problem. We show details in Appendix A of this document. In Beck and Teboulle (2009), FISTA was applied to solve this type of problems. Because each run of newGLMNET involves several iterations, we present only results of solving the first and the last sub-problems in Figures 5 and 6, respectively. Each point in the figures corresponds to a FISTA iteration or a CD cycle of newGLMNET. We can see that CD is faster than FISTA. This result seems to be consistent with the theoretical convergence property. For CD, because (13) is strongly convex after adding $\nu\mathcal{I}$ to the Hessian in (7), we can apply Tseng and Yun's proof to have linear convergence, but for FISTA, Beck and Teboulle (2009) proved only a sub-linear convergence rate.

Appendix A. Solving Sub-problem (13) by FISTA

Sub-problem (13) minimizes $q_k(\boldsymbol{d})$, which is the second order approximation of problem (1).

$$q_k(\boldsymbol{d}) \equiv \overbrace{\underbrace{C\sum_{i=1}^{l}(\tau(y_i(\boldsymbol{w}^k)^T\boldsymbol{x}_i)-1)y_i\boldsymbol{x}_i)^T}_{\equiv\nabla L(\boldsymbol{w}^k)^T}\boldsymbol{d}+\frac{1}{2}\boldsymbol{d}^T\underbrace{(CX^TDX+\nu\mathcal{I})}_{\equiv H^k}\boldsymbol{d}}^{\equiv\bar{q}_k(\boldsymbol{d})}+\|\boldsymbol{w}^k+\boldsymbol{d}\|_1-\|\boldsymbol{w}^k\|_1. \tag{A.1}$$

Eq. (16) is the gradient of $\bar{q}_k(\boldsymbol{w}^k)$ in each direction, and the they can be rewritten as

$$\nabla\bar{q}_k(\boldsymbol{d}) \equiv L(\boldsymbol{w}^k)+H^k\boldsymbol{d}. \tag{A.2}$$

Minimizing Eq. (A.1) is actually a least squares problem because H is symmetric positive definite matrix and therefore can be decomposed to A^TA for some nonsingular matrix A. So Eq. (A.1) is equal to

$$\frac{1}{2}(A\boldsymbol{d}+A^{-T}\nabla L(\boldsymbol{w}^k))^2+\|\boldsymbol{w}^k+\boldsymbol{d}\|_1+\text{constant}.$$

Given any point $\hat{\boldsymbol{d}}$, FISTA minimizes the second order approximation of $\bar{q}_k(\boldsymbol{d})$ plus the regularization term

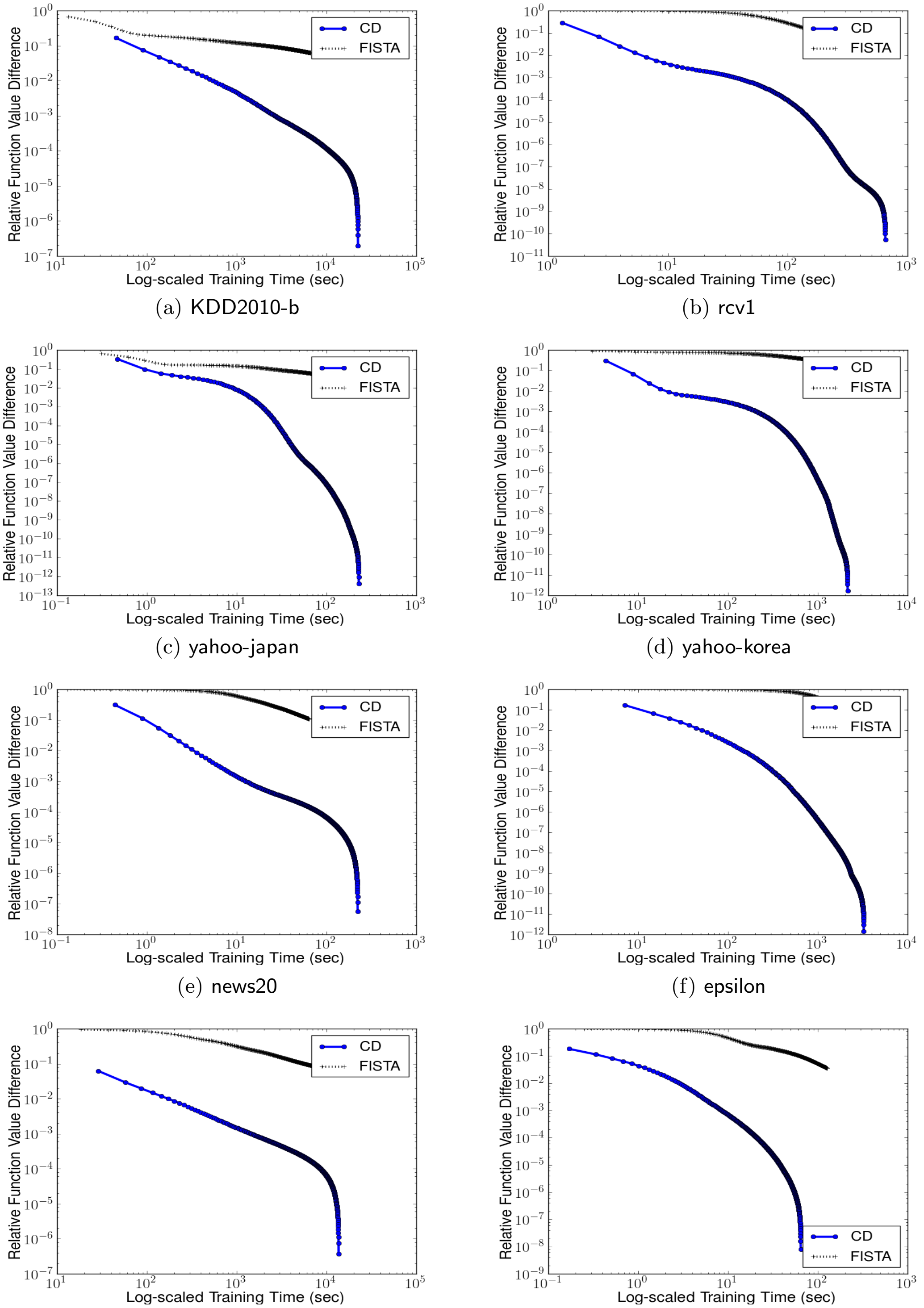
$$Q(\hat{\boldsymbol{d}},\boldsymbol{d}) \equiv \bar{q}_k(\hat{\boldsymbol{d}})+\nabla\bar{q}_k(\hat{\boldsymbol{d}})^T(\boldsymbol{d}-\hat{\boldsymbol{d}})+\frac{\Lambda}{2}\|\boldsymbol{d}-\hat{\boldsymbol{d}}\|_2^2+\|\boldsymbol{d}+\boldsymbol{w}^k\|_1, \tag{A.3}$$

where Λ is the Lipschitz constant of $\nabla\bar{q}_k(\boldsymbol{d})$. By Eq. (A.2), $\|H\| = \|CX^TDX+\nu I\|$ is a Lipschitz constant. It can be further simplified to $C\|X^T\|\|X\|+\nu$ by triangle inequality and Appendix A.

As a separable problem, minimizing problem (A.3) has a closed-form solution similar to Eq. (9)

$$d_j = \begin{cases} \hat{d}_j - \frac{\nabla_j\bar{q}_k(\hat{\boldsymbol{d}})+1}{\Lambda} & \text{if } \nabla_j\bar{q}_k(\hat{\boldsymbol{d}})+1 \le \Lambda(w_j^k+\hat{d}_j), \\ \hat{d}_j - \frac{\nabla_j\bar{q}_k(\hat{\boldsymbol{d}})-1}{\Lambda} & \text{if } \nabla_j\bar{q}_k(\hat{\boldsymbol{d}})-1 \ge \Lambda(w_j^k+\hat{d}_j), \\ -w_j^k & \text{otherwise.} \end{cases} \tag{A.4}$$

The procedure of sub-problem (13) is Algorithm 1.



(a) KDD2010-b

(b) rcv1

(c) yahoo-japan

(d) yahoo-korea

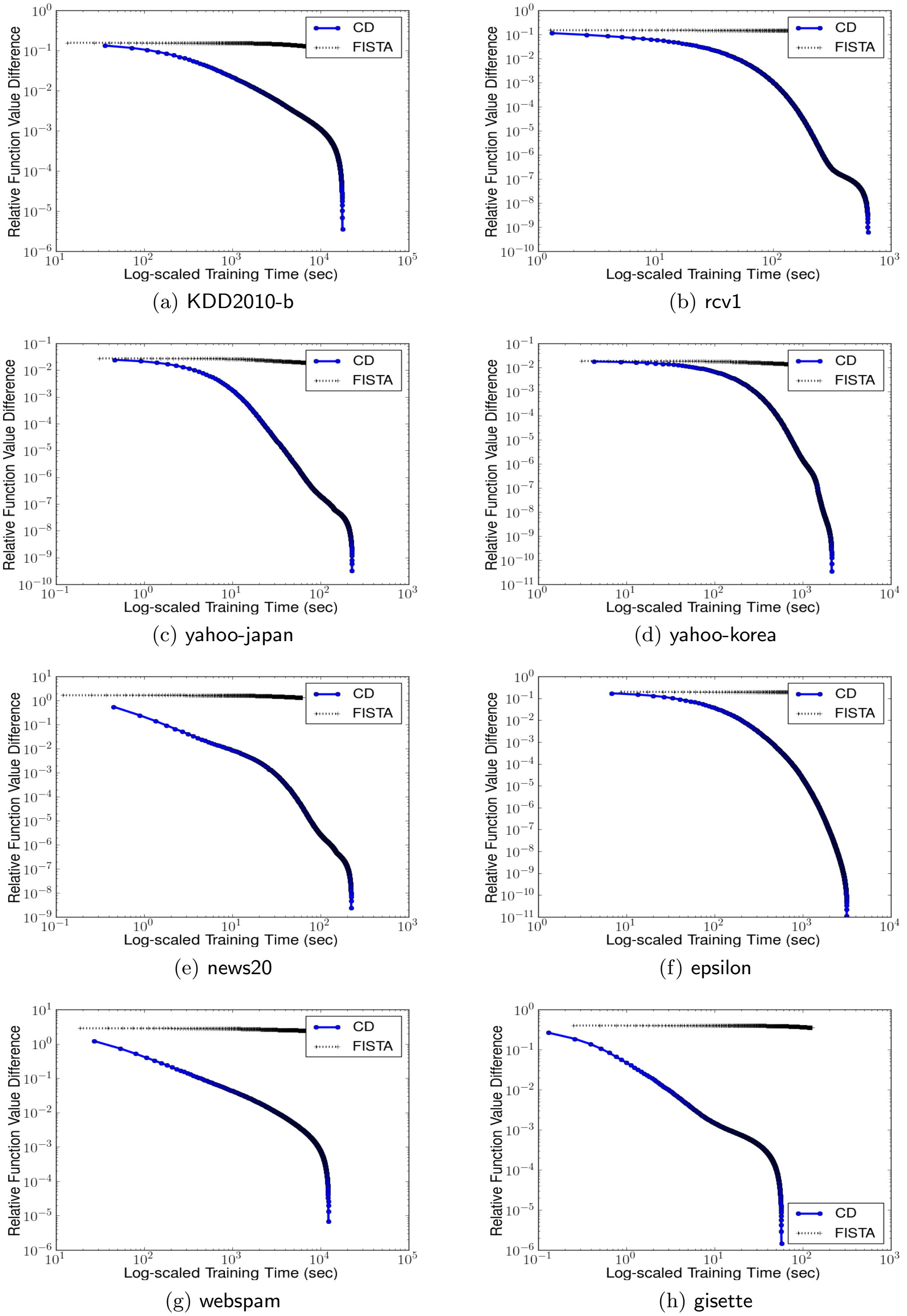
(e) news20

(f) epsilon

(g) webspam

(h) gisette

Figure 5: Comparison of CD and FISTA for solving the first sub-problem (13). We present relative difference to the optimal function value of the sub-problem versus training time. Both x-axis and y-axis are log-scaled.



(a) KDD2010-b

(b) rcv1

(c) yahoo-japan

(d) yahoo-korea

(e) news20

(f) epsilon

(g) webspam

(h) gisette

Figure 6: Comparison of CD and FISTA for solving the last sub-problem (13). We present relative difference to the optimal function value of the sub-problem versus training time. Both x-axis and y-axis are log-scaled.

Algorithm 1 FISTA algorithm for solving the inner problem.

```
• Given the Lipschitz constant $\Lambda = C\|X^T\|\|X\| + \nu$, $t^1 = 1$, $\boldsymbol{d}^0 = \mathbf{0}$, and $\hat{\boldsymbol{d}}^1 = \boldsymbol{d}^0$.
• for $p = 1, 2, \ldots, 1000$
    1. $\boldsymbol{d}^p \leftarrow \arg\min_{\boldsymbol{d}} Q(\hat{\boldsymbol{d}}^p, \boldsymbol{d})$ by (A.4).
    2. if $\|\nabla \bar{q}_k(\boldsymbol{d}^p)\|_1 \leq \epsilon_{\text{inner}}$
          break.
    3. $t^{p+1} \leftarrow (1 + \sqrt{1 + 4(t^p)^2})/2$.
    4. $\hat{\boldsymbol{d}}^{p+1} \leftarrow \boldsymbol{d}^p + (\boldsymbol{d}^p - \boldsymbol{d}^{p-1})(t^p - 1)/t^{p+1}$
```

References

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Joseph C. Dunn. Newton's method and the Goldstein step length rule for constrained minimization. *SIAM Journal on Control and Optimization*, 18:659–674, 1980.

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. Technical report, EECS Department, University of California, Berkeley, 2010.

Cho-Jui Hsieh and Inderjit S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the Seventeenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011.

Cho-Jui Hsieh, Matyas A. Sustik, Pradeep Ravikumar, and Inderjit S. Dhillon. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems (NIPS) 24*, 2011.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonatha Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of Royal Statistical Society: Series B*, 2011.

Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, 2009.