Logistic Regression Confined by Cardinality-Constrained Sample and Feature Selection (Supplementary Material)

Ehsan Adeli*, *Member, IEEE,* Xiaorui Li*, Dongjin Kwon, Yong Zhang, and Kilian M. Pohl

S1 Theoretical Discussions Quoted from the Literature

Note that all the citations and reference numbers in this Section are the same as the main paper.

Theorem 4.1 of [66]:

> *"The point of this restriction is that a strictly convex function with bounded level sets has a unique minimizer x_{min}.*
>
> ***Theorem 4.1.*** *(Convergence of one-dimensional Nelder-Mead method.) Let f be a strictly convex function on $\mathcal{R}^1$ with bounded level sets. Assume that the Nelder-Mead algorithm is applied to f with parameters satisfying $\rho > 0, \chi > 1, \chi > \rho, \rho\chi \geq 1$, and $0 < \gamma < 1$, beginning with a nondegenerate initial simplex Δ_0. Then both endpoints of the Nelder-Mead interval converge to x_{min}."*

To prove this theorem, the authors introduced several intermediate lemmas. For detailed analysis please refer to [66].

Theorems 2.1 and 4.3 and Proposition 3.1 of [65]:

> *"Mathematically, all these applications can be formulated into the following ℓ_0 minimization problems:*
>
> $$\min_{x \in \mathcal{X}} \{f(x) : g(x) \leq 0, h(x) = 0, \|x_J\|_0 \leq r\}, \quad (1.1)$$
>
> $$\min_{x \in \mathcal{X}} \{f(x) + \nu \|x_J\|_0 : g(x) \leq 0, h(x) = 0\} \quad (1.2)$$
>
> *for some integer $r \geq 0$ and $\nu \geq 0$ controlling the sparsity (or cardinality) of the solution, where $\mathcal{X}$ is a close convex set in the n-dimensional Euclidean space $\mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ are continuously differentiable functions, and $\|x_J\|_0$ denotes the cardinality of the subvector formed by the entries of x indexed by J.*
>
> ***Theorem 2.1.*** *Assume that x^* is a local minimizer of problem (1.1). Let $J^* \subseteq J$ be an index set with $|J^*| = r$ such that $x^*_j = 0$ for all $j \in \bar{J}^*$, where $\bar{J}^* = J \setminus J^*$. Suppose that the following Robinson condition*
>
> $$\left\{ \begin{bmatrix} g'(x^*)d - v \\ h'(x^*)d \\ (I_{\bar{J}^*})^\top d \end{bmatrix} : \begin{array}{c} d \in \mathcal{T}_{\mathcal{X}}(x^*), v \in \mathbb{R}^m, \\ v_i \leq 0, i \in \mathcal{A}(x^*) \end{array} \right\} = \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^{|J|-r} \quad (2.1)$$
>
> *holds, where $g'(x^*)$ and $h'(x^*)$ denote the Jacobian of the functions $g = (g_1, \ldots, g_m)$ and $h = (h_1, \ldots, h_p)$ at x^*, respectively, and*
>
> $$\mathcal{A}(x^*) = \{1 \leq i \leq m : g_i(x^*) = 0\}. \quad (2.2)$$
>
> *Then, there exists $(\lambda^*, \mu^*, z^*) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n$ together with x^* satisfying*
>
> $$\begin{aligned} & -\nabla f(x^*) - \nabla g(x^*)\lambda^* - \nabla h(x^*)\mu^* - z^* \in \mathcal{N}_{\mathcal{X}}(x^*), \\ & \lambda^*_i \geq 0, \lambda^*_i g_i(x^*) = 0, i = 1, \cdots, m, \\ & z^*_j = 0, j \in \bar{J} \cup J^*, \end{aligned} \quad (2.3)$$
>
> *where $\bar{J}$ is the complement of J in $\{1, \ldots, n\}$.*
>
> ***Proposition 3.1.*** *Let $\mathcal{X}_i \subseteq \mathbb{R}$ and $\phi_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, \ldots, n$ be given. Suppose that r is a positive integer and $0 \in \mathcal{X}_i$ for all i. Consider the following ℓ_0 minimization problem:*
>
> $$\min \left\{ \sum_{i=1}^{n} \phi_i(x_i) : \|x\|_0 \leq r, x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \right\}. \quad (3.1)$$
>
> *Let $\tilde{x}^*_i \in \operatorname{Argmin}\{\phi_i(x_i) : x_i \in \mathcal{X}_i\}$, and let $I^* \subseteq \{1, \ldots, n\}$ be the index set corresponding to the r largest values of $\{v^*_i\}_{i=1}^n$, where $v^*_i = \phi_i(0) - \phi_i(\tilde{x}^*_i)$ for $i = 1, \ldots, n$. Then x^* is an optimal solution of problem (3.1), where x^* is defined as follows:*
>
> $$x^*_i = \begin{cases} \tilde{x}^*_i & \text{if } i \in I^*; \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n.$$
>
> ***Theorem 4.3.*** *Assume that $\epsilon_k \to 0$. Let $\{(x^k, y^k)\}$ be the sequence generated by the above PD method, $I_k = \{i^k_1, \cdots, i^k_r\}$ be a set of r distinct indices in $\{1, \ldots, |J|\}$ such that $(y^k)_i = 0$ for any $i \notin I_k$, and let $J_k = \{J(i) : i \in I_k\}$. Suppose that the level set $\mathcal{X}_\Upsilon := \{x \in \mathcal{X} : f(x) \leq \Upsilon\}$ is compact. Then, the following statements hold:*
>
> (a) *The sequence $\{(x^k, y^k)\}$ is bounded.*
>
> (b) *Suppose (x^*, y^*) is an accumulation point of $\{(x^k, y^k)\}$. Then, $x^* = y^*$ and x^* is a feasible point of problem (1.1). Moreover, there exists a subsequence K such that $\{(x^k, y^k)\}_{k \in K} \to (x^*, y^*)$, $I_k = I^*$ and $J_k = J^* := \{J(i) : i \in$*

$I^*\}$ for some index set $I^* \subseteq \{1, \ldots, |J|\}$ when $k \in K$ is sufficiently large.

(c) *Let x^*, K and J^* be defined above, and let $\bar{J}^* = J \setminus J^*$. Suppose that the Robinson condition (2.1) holds at x^* for such $\bar{J}^*$. Then, $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ is bounded, where $\lambda^k = \rho_k[g(x^k)]^+$, $\mu^k = \rho_k h(x^k)$, $\varpi^k = \rho_k(x^k_J - y^k)$. Moreover, each accumulation point $(\lambda^*, \mu^*, \varpi^*)$ of $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ together with x^* satisfies the first-order optimality conditions (2.3) with $z^*_j = \varpi^*_i$ for all $j = J(i) \in \bar{J}^*$. Further, if $\|x^*_J\|_0 = r$, h's are affine functions, and f and g's are convex functions, then x^* is a local minimizer of problem (1.1).*

"

Refer to [65] for the proofs of the above theorems.

Definition 4, Assumptions 1 and 4, Theorems 5.2 and 5.4 of [48]:

"In particular, we propose a feasible augmented Lagrangian (FAL) method for solving them, which solves a sequence of partially regularized unconstrained optimization problems in the form of

$$\min_{x\in\mathbb{R}^n} \left\{ F(x) := f(x) + \lambda \sum_{i=r+1}^{n} \phi(|x|_{[i]}) \right\}. \qquad (22)$$

In addition, for convenience of presentation, let $\Phi(x) := \sum_{i=r+1}^{n} \phi(|x|_{[i]})$.
Definition 4. *(first-order stationary point) $x^* \in \mathbb{R}^n$ is a first-order stationary point of (22) if*

$$0 \in \nabla f(x^*) + \lambda\, \partial\Phi(x^*).$$

Assumption 1. *ϕ is lower semi-continuous and increasing in $[0, \infty)$. Moreover, $\phi(0) = 0$.*
Assumption 4. *(i) f is continuously differentiable in $\mathcal{U}(x^0; \Delta)$ for some $x^0 \in \mathbb{R}^n$ and $\Delta > 0$, and moreover, there exists some $L_f > 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\| \le L_f \|x - y\|, \forall x, y \in \mathcal{U}(x^0; \Delta),$$

where

$$\mathcal{U}(x^0; \Delta) := \{x : \|x - z\| \le \Delta \text{ for some } z \in \mathcal{S}(x^0)\}.$$

$$\mathcal{S}(x^0) := \{x \in \mathbb{R}^n : F(x) \le F(x^0)\}.$$

(ii) F is bounded below and uniformly continuous in $\mathcal{S}(x^0)$.
(iii) The quantities A and B defined below are finite:

$$A := \sup_{x\in\mathcal{S}(x^0)} \|\nabla f(x)\|, B := \sup_{x\in\mathcal{S}(x^0)} \sum_{i=r+1}^{n} \phi(|x|_{[j]}).$$

Theorem 5.1. *Let $\{x^k\}$ and $\bar{L}_k$ be generated in Algorithm 1, and let*

$$\bar{L} := \max\{L_{\max}, \tau \underline{L}, \tau(L_f + c)\}, \underline{L} := \frac{2(A\Delta + B)}{\Delta^2},$$

where A, B, L_f and Δ are given in Assumption 4. Under Assumption 4, the following statements hold:

(i) *For each $k \ge 0$, the inner termination criterion (24) is satisfied after at most*

$$\left\lfloor \frac{\log \bar{L} - \log L_{\min}}{\log \tau} + 1 \right\rfloor$$

inner iterations;

(ii) *$F(x^k) \le F(x^0)$ and $\bar{L}_k \le \bar{L}$ for all $k \ge 0$.*

Theorem 5.2. *Let the sequence $\{x^k\}$ be generated by Algorithm 1. There holds:*

(i) *$\|x^{k+1} - x^k\| \to 0$ as $k \to \infty$;*

(ii) *Any accumulation point of $\{x^k\}$ is a first-order stationary point of (22);*

(iii) *For any $\epsilon > 0$, x^k is a first-order ϵ-stationary point of problem (22) when k is sufficiently large.*

We are now ready to discuss how to solve efficiently the subproblem (23) of Algorithm 1. Clearly, (23) is equivalent to

$$\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2} \left\| x - \left(x^k - \frac{1}{L_k} \nabla f(x^k) \right) \right\|_2^2 + \frac{\lambda}{L_k} \sum_{i=r+1}^{n} \phi(|x|_{[i]}) \right\},$$

which is a special case of a more general problem

$$\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2} \|x - a\|_2^2 + \tilde{\lambda} \sum_{i=r+1}^{n} \phi(|x|_{[i]}) \right\}, \qquad (31)$$

for some $a \in \mathbb{R}^n$ and $\tilde{\lambda} > 0$. In what follows, we show that problem (31) can be solved as $n - r$ number of one-dimensional problems in the form of (26).
Theorem 5.4. Suppose that ϕ satisfies Assumption 1. Let I^* be the index set corresponding to the $n - r$ smallest entries of $|a|$ and $x^* \in \mathbb{R}^n$ be defined as follows:

$$x^*_i \in \begin{cases} \operatorname*{Argmin}_{u\in\Re} \left\{ \frac{1}{2}(u - a_i)^2 + \tilde{\lambda}\phi(|u|) \right\} & \text{if } i \in I^*, \\ \{a_i\} & \text{otherwise} \end{cases}$$

$i = 1, \ldots, n$. Then x^* is an optimal solution of problem (31).
"

For more detailed discussions, please refer to [48].

S2 MRI Data Preprocessing and Feature Extraction

Preprocessing of the T1-weighted (T1w) MR images involves noise removal [Coupé et al., 2008] and correcting field inhomogeneity via N4ITK (Version 2.1.0) [Tustison et al., 2010]. Next, the brain mask is segmented by majority voting [Rohlfing et al., 2004] across maps extracted by FSL BET (Version 5.0.6) [Smith, 2002], AFNI 3dSkullStrip (Version AFNI_2011_12_21_1014) [Cox, 1996], FreeSurfer mri-gcut (Version 5.3.0) [Sadananthan et al., 2010], and the Robust Brain Extraction (ROBEX) method (Version 1.2) [Iglesias et al., 2011]. We further apply the cross-sectional approach of FreeSurfer (Version 5.3.0) software [Dale et al., 1999], [Reuter et al., 2012] to the skull-stripped T1w MRI of each subject in order to measure the *mean curvature (MeanCurv)*, *surface area (SurfArea)*, *gray matter volume (GrayVol)*, and *average thickness (ThickAvg)* of 34 bilateral cortical Regions Of Interest (ROIs) [2 hemispheres × 4 measurement types × 34 ROIs = 272], the volumes of 8 bilateral sub-cortical ROIs (*i.e.*, thalamus, caudate, putamen, pallidum, hippocampus, amygdala, accumbens, cerebellar cortex) [$2 \times 8 = 16$], the volumes of 5 subregions of the corpus callosum (posterior, mid-posterior, central, mid-central and anterior), and the combined volume of all white matter hypointensities [$5 + 1 = 6$]. Additionally, supratentorial volume (svol) of the left and right lateral ventricles, and the third ventricle [$1 + 2 \times 2 = 5$] are measured by non-rigidly aligning the SRI24 atlas [Rohlfing et al., 2010] to the T1w MRI of the subject via ANTS (Version: 2.1.0) [Avants et al., 2008]. In addition to svol, each subject is thus represented by the z-scores of 298 morphometric measures (*i.e.*, features).

S3 Additional Experiments

In addition to the experiments in the main paper, to compare our method to published findings, we also apply the methods to the real data provided by two benchmark datasets of the UCI machine learning repository [Lichman, 2013]: the Lymphography Domain Dataset and the SPECTF Heart Dataset. Both datasets are characterized by redundant and uninformative features with respect to group separation. In addition, the Lymphography Domain Dataset contains samples that are outliers, *i.e.*, they do not belong to either of the two cohorts under investigation.

S3.1 UCI Benchmark Datasets

The Lymphography Domain dataset contains 148 samples, each being represented by 19 features[1]. The samples are divided into four classes: 2 are labeled as 'normal', 81 as 'metastases', 61 as 'malignant lymphoma', and 4 as 'fibrosis'. The first and the last classes are quite small so that we view them as outliers in the experiment of distinguishing the metastases from the malignant lymphoma samples. For simplicity, the normal cases are assigned to the metastases group ($N = 83$) and the fibrosis cases to the malignant lymphoma group ($N = 65$). Thus, the two groups of interest are unequal in size and contain outliers.

The SPECTF Heart Dataset contains measurements of cardiac Single Proton Emission Computed Tomography (SPECT) images of 267 subjects. 212 participants are labelled as abnormal and 55 as normal. Each SPECT image is summarized by 44 continuous measurements, which are generated by counting the number of 'rested' and 'stressed' voxels in 22 regions of interest (ROIs). Given this type of evaluation, we expect high redundancy between features of this imbalanced dataset.

S3.2 Results of Comparison

Table S1 summaries the accuracy scores of all implementations on the two benchmark datasets. As on the synthetic dataset, our proposed sample-feature selection scheme with hybrid regularization (*i.e.*, SFS$_H$) achieved a higher BAcc score on both datasets than the other proposed implementations (*i.e.*, SFS$_P$, SFS$_C$), which were again higher than all alternative approaches. The same observation was true for the F1-score. Of those methods, SFS$_C$ reported the lowest precision and recall, which were at least as high as those of the alternative approaches. Furthermore, SFS$_H$ never selected outliers to be included in the cost function during training.

Method	**Lymphography**				**Heart**			
	BAcc	Pre	Rec	F_1	BAcc	Pre	Rec	F_1
SFS$_H$	**87.9**	0.87	**0.89**	**0.88**	**82.9**	0.78	0.82	**0.80**
SFS$_P$	85.5	**0.88**	0.82	0.85	81.5	**0.79**	0.80	0.79
SFS$_C$	81.8	0.83	0.79	0.81	81.7	**0.79**	0.80	0.79
JFSS$_{\ell_1}$-W	77.8	0.79	0.72	0.75	76.8	0.75	0.79	0.77
JFSS$_{\ell_1}$	69.5	0.75	0.62	0.68	55.4	0.56	**0.96**	0.71
LR$_{\ell_0}$-W	76.8	0.79	0.78	0.78	78.5	0.75	0.80	0.77
SFS+SVM-W	75.9	0.75	0.73	0.74	79.1	0.74	0.80	0.77
SVM-W	74.6	0.75	0.72	0.73	78.8	0.69	0.80	0.74
SVM	66.7	0.78	0.55	0.65	58.0	0.57	0.84	0.68

TABLE S1: Comparison of the results on benchmark datasets. In each column the best result is typeset in bold typeface.

As mentioned, we turned the Lymphography dataset into a binary classification problem by interpreting the small classes as outliers that were merged with the large ones. Published accuracy scores are not directly comparable to our findings as they use different validation schemes to solve the multi-class classification problem and report on the accuracy without normalizing for group size. For example, the knowledge-based approach by Centnik *et al.* [Cestnik et al., 1987] reports 76% accuracy, the rule-induction methods by Clark and Niblett [Clark and Niblett, 1987] achieves 83% accuracy, and the probability series expansion classifiers by Agarwal *et al.* [Agarwal and Hudson, 2017] measures 86.4% accuracy. In comparison, the 'un-normalized' accuracy score of SFS$_H$is 88.1%, which would drop down to 87.4% if we viewed all 'outliers' as misclassified (*i.e.*, the original multi-class problem). We conclude that on this dataset our findings are highly competitive to existing publications.

With respect to the Heart dataset, the proposed techniques are superior to all other methods not only in terms of the balanced accuracy but also for the F_1-score and the balance between the precision and recall. Note, only the methods that perform sample selection (*i.e.*, the proposed methods and JFSS$_{\ell_1}$-W) report balanced precision and recall scores (*i.e.*, the difference is less than 5%). Also, it is important to note that JFSS$_{\ell_1}$, which is based on ℓ_1 regularization, was not designed for highly imbalanced cases. The reimplemented version of this method using a weighted loss function performs relatively well on this dataset, but still inferior to our methods. Unlike the rigerous 10-fold cross-validation scheme proposed here, the experimental design for the Heart dataset in previous works [Pant et al., 2017], [Cios et al., 1997] split the dataset into a balanced training set (containing 40 samples for each class) and use the remaining samples for testing (172 samples for one class and 12 for the other). They report the accuracy instead of the balanced accuracy score. The CLIP3 [Cios et al., 1997] algorithm achieved a 77.0% accuracy. The authors in [Pant et al., 2017] also obtained accuracies of 78.4% using twin SVM and 83.3% using neural network methods. Our proposed implementations, especially SFS$_H$, show comparable results even through they are trained on imbalanced data.

[1] This lymphography domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data.

References

[Agarwal and Hudson, 2017] Agarwal, S. and Hudson, C. M. (2017). Probability series expansion classifier that is interpretable by design. In *Interpretable ML Symposium, NIPS*.

[Avants et al., 2008] Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41.

[Cestnik et al., 1987] Cestnik, B., Kononenko, I., and Bratko, I. (1987). Assistant 86: A knowledge-elicitation tool for sophisticated users. In *European Conference on European Working Session on Learning*, pages 31–45. Sigma Press.

[Cios et al., 1997] Cios, K. J., Wedding, D. K., and Liu, N. (1997). Clip3: Cover learning using integer programming. *Kybernetes*, 26(5):513–536.

[Clark and Niblett, 1987] Clark, P. and Niblett, T. (1987). Induction in noisy domains. In *EWSL*, pages 11–30.

[Coupé et al., 2008] Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., and Barillot, C. (2008). An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. *IEEE transactions on medical imaging*, 27(4):425–441.

[Cox, 1996] Cox, R. W. (1996). Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173.

[Dale et al., 1999] Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194.

[Iglesias et al., 2011] Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans. on Medical Imaging*, 30(9):1617–1634.

[Lichman, 2013] Lichman, M. (2013). UCI machine learning repository.

[Pant et al., 2017] Pant, H., Soman, S., Sharma, M., et al. (2017). Scalable twin neural networks for classification of unbalanced data. *arXiv:1705.00347*.

[Reuter et al., 2012] Reuter, M., Schmansky, N. J., Rosas, H. D., and Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418.

[Rohlfing et al., 2004] Rohlfing, T., Russakoff, D. B., and Maurer, C. R. (2004). Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE transactions on medical imaging*, 23(8):983–994.

[Rohlfing et al., 2010] Rohlfing, T., Zahr, N. M., Sullivan, E. V., and Pfefferbaum, A. (2010). The SRI24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31(5):798–819.

[Sadananthan et al., 2010] Sadananthan, S. A., Zheng, W., Chee, M. W., and Zagorodnov, V. (2010). Skull stripping using graph cuts. *NeuroImage*, 49(1):225–239.

[Smith, 2002] Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155.

[Tustison et al., 2010] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320.