Optimization Problem
We study the optimization problem
\[
\min_{x \in \mathbb{R}^n} f(x) + \lambda \|x\|_1
\]

- \( f : \mathbb{R}^n \to \mathbb{R} \) is twice continuously differentiable and convex
- logistic loss (logistic regression)
- log-linear model (log likelihood)
- mean-square error (linear regression)
- \( \| \cdot \| \) promotes sparsity
- \( \lambda > 0 \) is a weighting parameter

Applications
- Machine Learning: Classification/regression problems, e.g. predict customer-product relation for recommendation system
- Natural Language Processing: Given corpus and word embeddings, train language model with log-likelihood function
- Machine Translation: Comparing the performance of different translation systems using a chosen score function

State-of-the-Art
- LIBLINEAR: Proximal-Newton, coordinate descent, supports popular machine learning libraries (e.g., Sklearn and EML)
- OBA: Orthant-wise strategy, suitable for problems with non-diagonal dominant Hessian, flexibility in subproblem solver
- ASA-CG: Active-set algorithm for box-constrained optimization problems, general \( f \) allowed, only uses first derivatives

FaRSA: Fast Reduced Space Algorithm
Basic Idea
- Each iteration, “intelligently” choose whether to
  - \(( \omega \)-step\) optimize over nonzero variables
  - \(( \beta \)-step\) allow zero variables to become nonzero

Optimality measure over zero and nonzero variables
- \[ \| \hat{\omega}(x) \| \text{ is an optimality measure in the nonzero variables:} \]
- \[ \beta(x) \text{ is an optimality measure in the zero variables:} \]
- \[ \| \omega(x) \| = 0 \iff \| \hat{\omega}(x) \| = \| \beta(x) \| = 0 \]
- \( x^* \) solves problem \( (1) \iff \| \omega(x^*) \| = \| \hat{\omega}(x^*) \| = 0 \)

Pseudocode
1. Initialize: \( x_0 \) and weighting parameter \( \lambda \geq 0 \)
2. for \( k = 0, 1, 2, \ldots \) do
3. if \( \| \omega(x_k) \| \geq \| \hat{\omega}(x_k) \| \) then
4. Obtain Newton-CG step in nonzero variables: \( S_k \)
5. Perform an orthant-restricted linesearch: \( \alpha_k \)
6. else
7. Obtain reduced-ISTA step in zero variables: \( S_k \)
8. Perform standard linesearch: \( \alpha_k \)
9. end if
10. \( x_{k+1} = x_k + \alpha_k S_k \)
11. end for

Key Properties
- global convergence: limit points of \( \{ x_k \} \) are minimizers of \( f(x) + \lambda \|x\|_1 \)
- active-set identification: \( x_k \) in same orthant as \( x^* \) for \( k \) large
- local convergence: \( \{ x_k \} \to x^* \) at a superlinear rate
- scalability: Newton-CG step only uses Hessian-vector products

Numerical Experiments
- Test datasets from LIBSVM repository with \( f \) as logistic loss
- Global comparison and convergence to state-of-the-art

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Time (seconds)</th>
<th>Final Objective Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>breast-cancer</td>
<td>0.30009</td>
<td>0.30011</td>
</tr>
<tr>
<td>liver-disorders</td>
<td>6.9871</td>
<td>13.9159</td>
</tr>
<tr>
<td>leukemia</td>
<td>0.1140</td>
<td>0.0621</td>
</tr>
<tr>
<td>skin-nonskin</td>
<td>0.48005</td>
<td>0.48005</td>
</tr>
<tr>
<td>colon-cancer</td>
<td>0.51685</td>
<td>0.51685</td>
</tr>
<tr>
<td>breast-cancer</td>
<td>0.00123</td>
<td>0.00123</td>
</tr>
<tr>
<td>heart</td>
<td>0.0704</td>
<td>0.0351</td>
</tr>
<tr>
<td>german numer</td>
<td>0.0704</td>
<td>0.0351</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.0222</td>
<td>0.0119</td>
</tr>
<tr>
<td>australian</td>
<td>21.1252</td>
<td>509.6747</td>
</tr>
<tr>
<td>australian</td>
<td>0.1514</td>
<td>0.1029</td>
</tr>
<tr>
<td>australian</td>
<td>3.2802</td>
<td>20.8732</td>
</tr>
<tr>
<td>australian</td>
<td>0.46300</td>
<td>0.46300</td>
</tr>
<tr>
<td>australian</td>
<td>0.14125</td>
<td>0.14126</td>
</tr>
<tr>
<td>australian</td>
<td>0.0072</td>
<td>0.0037</td>
</tr>
<tr>
<td>australian</td>
<td>1.4968</td>
<td>0.9301</td>
</tr>
<tr>
<td>australian</td>
<td>0.12206</td>
<td>0.12206</td>
</tr>
</tbody>
</table>

References

daniel.p.robinson@jhu.edu, tchen59@jhu.edu, frank.e.curtis@gmail.com