Parallel Majorization Minimization with Dynamically Restricted Domains for Nonconvex Optimization: Supplementary Material

Yan Kaganovsky*  Ikenna Odinaka*  David Carlson  Lawrence Carin
Duke University  Duke University  Columbia University  Duke University

Abstract

We provide proofs for the theorems presented in the main paper and additional numerical examples.

1 Proofs of Theorems and Lemmas

1.1 Proof of Lemma 2.1

Proof First, we need to verify that the conditions in (3)–(7) are satisfied. It follows directly from (14) that \( F(\hat{\theta}; \theta) = F(\theta) \) and \( \nabla_\theta F|_{\theta = \hat{\theta}} = \nabla_\theta F|_{\theta = \hat{\theta}} \) so (4)–(5) are satisfied. Also, by construction the entries of \( \hat{\theta} \) satisfy (6)–(7). Lastly, we check that the majorization domain \( \hat{\Omega}_M \) is a (non-empty) convex polyhedron so the domain of the surrogate is convex. Also, \( \hat{\theta} \in \text{int}(\hat{\Omega}_M) \) and \( \hat{\theta} \in \Omega_F \) by assumption so (6)–(7) are satisfied. Lastly, we check that the function is convex on \( \hat{\Omega}_M \). The Hessian of \( F \) is \( \mathcal{H}(\hat{\theta}) = \text{XDiag}(\{f_m''(\theta_T x_m)\}_{m=1}^M) X^T + X DX^T \). From the definition of \( D \) in (15) and (11) it follows that \( \mathcal{H}(\hat{\theta}) \geq 0 \) for \( \hat{\theta} \in \hat{\Omega}_M \), thus proving the convexity of \( F \) on \( \hat{\Omega}_M \). 

1.2 Proof of Lemma 2.2

Proof Define the line \( L(\alpha) := \{ \lambda \theta_2 + (1 - \lambda) \theta_1 | \lambda \in (0, \alpha) \} \). Since \( \theta_1, \theta_2 \in \Omega_F \) and \( \Omega_F \) is convex, then \( L(1) \subseteq \Omega_F \) and since \( \theta_1 \in \text{int}(\Omega_M) \) there exits \( \alpha_0 > 0 \) such that \( \theta \neq L(\alpha_0) \subseteq \Omega_M \). For \( \alpha_0 \leq 1 \), we also have \( L(\alpha_0) \subseteq L(1) \subseteq \Omega_F \) and therefore \( L(\alpha_0) \subseteq \Omega_F \cap \Omega_M \). Let \( \theta^* := \lambda \theta_2 + (1 - \lambda) \theta_1 \) with \( \lambda \in (0, \alpha_0) \), then \( \theta^* \in L(\alpha_0) \subseteq \Omega_F \cap \Omega_M \). Since \( F \) is convex we have \( F(\theta^*) \leq \lambda F(\theta_2) + (1 - \lambda) F(\theta_1) < \lambda F(\theta_2) + (1 - \lambda) F(\theta_1) \), where in the last inequality we used \( F(\theta_2) < F(\theta_1) \). 

* Indicates equal contributions. To appear in the Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. Copyright 2016 by the authors.

1.3 Proof of Lemma 2.3

Proof Let \( g(w) : \mathbb{R} \to \mathbb{R} \) be a convex function. From Jensen’s inequality \( g(\sum_k w_k) = g(\sum_k r_k w_k / r_k) \leq \sum_k r_k g(w_k / r_k) \) for any \( r = [r^1, r^2, \ldots, r^K] \in \mathbb{R}^K \) with \( 1 \leq K \leq p, \text{ s.t. } r \geq 0 \) and \( \|r\|_1 = 1 \). Now set \( g_m(v) = \hat{f}_m(\theta_T x_m + v) \) (recall that \( \hat{f} \) in (16) is globally convex) and rewrite \( \hat{f}_m(\theta_T x_m) = g_m((\theta - \hat{\theta})^T x_m) \) and then apply the above inequality with \( w_k = (\theta - \hat{\theta})^T x_m \) for each \( m \) separately which leads to \( \hat{f}_m(\theta_T x_m) \leq \sum_k r_k \hat{f}_m(\theta_T x_m + (\theta - \hat{\theta})^T x_m / r_k) \). For each \( m \) we choose the \( r_k \) given in (20) which satisfies the conditions of Jensen’s inequality. From (16), (12), and (11) it follows that \( \hat{f}_m(\theta_T x_m) \leq \hat{f}_m(\theta_T x_m) \leq \sum_k r_k \hat{f}_m(\theta_T x_m + (\theta - \hat{\theta})^T x_m / r_k) \) for any \( m \) and \( \theta \in \hat{\Omega}_M \). Summing over \( m \) we obtain that \( F \leq \sum_m S_m \) with \( S_m \) defined in (19) which proves that (3) holds for \( \hat{\Omega}_M \). By using (12) and (16), it is simple to check directly that (4)–(7) are satisfied.

1.4 Proof of Lemma 3.2

Proof We have \( x^* \in A(x^*) \) and the constraints as specified by \( \Omega_F \) in (1) are qualified. Then there exit Lagrange multipliers \( \{\eta_i\}_{i=1}^I \subseteq \mathbb{R} \) and \( \{\mu_j\}_{j=1}^J \subseteq \mathbb{R} \) such that the following KKT conditions hold

\[
\nabla C(x^*) + \sum_{i=1}^I \eta_i \nabla g_i(x^*) + \sum_{j=1}^J \mu_j \nabla h_j(x^*) = 0 \quad (A1)
\]

\[
g_i(x^*) \leq 0, \quad \eta_i \geq 0, \quad g_i(x^*) \eta_i = 0, \quad \forall i \in [I] \quad (A2)
\]

\[
h_j(x^*) = 0, \quad \mu_j \in \mathbb{R}, \quad \forall j \in [J], \quad (A3)
\]

where \( [I] = \{1, 2, \ldots, I\} \) and \( [J] = \{1, 2, \ldots, J\} \), and we used (5) so that \( \nabla S(x^*) = \nabla C(x^*) \). Equations (A1)–(A3) are exactly the KKT conditions for the program in (1) which are satisfied by \( \{x_i^*, \eta_i^\star_{i=1}^I, \mu_j^\star_{j=1}^J\} \), and therefore \( x^* \) is a stationary point of (1). 

1.5 Proof of Theorem 3.3

Proof \( \Omega_F \subseteq \mathbb{R}^p \) is assumed closed and bounded, and it is therefore compact. \( \theta^{(t)} \in \Omega_F \) so \( \theta^{(t)} \) lies in a
2 Sigmoid-Loss SVM

Figure 1 shows a comparison between the loss functions considered in this work

\[
\begin{align*}
0-1: & \quad f(z) = (-\text{sign}(z) + 1)/2, & (A4) \\
\text{hinge:} & \quad f(z) = \max(0, 1 - z), & (A5) \\
\text{logistic:} & \quad f(z) = \log(1 + \exp(-z)), & (A6) \\
\text{sigmoid:} & \quad f(z) = 1 - \tanh(z). & (A7)
\end{align*}
\]

![Figure 1: A comparison between the 0-1, logistic, Hinge, and Sigmoid loss functions.](image)

3 Example for Choosing the Majorization Domain

To illustrate the majorization-minimization procedure proposed in the paper, a simple 1D example is shown in Fig. 2, where the blue curve is the original objective, the green curve is the global surrogate when \([a, b] = (-\infty, \infty)\), and the red curve is the local surrogate when \([a, b]\) are chosen according to Algorithm 4.1 and Algorithm 4.2 (“shallow region” case). It can be seen in Fig. 2 that using the local surrogate with lower curvature leads to taking a larger step than when using a global surrogate. Note that at the iteration shown, each surrogate leads to taking a step from the expansion point (marked by an asterisk) to the minimum of the surrogate (marked by circles). It should be noted however, that the minimum for the high-dimensional problem in (2) does not necessarily occur at the minimum points of \(f_m\). Also note that all surrogates are convex but neither of them are quadratic.

4 Additional Details Regarding the Numerical Experiments

Experiments performed on the MNIST dataset utilize all the available examples for digit “3” (6131 for training, and 1010 for testing) and for digit “5” (5421 for training, and 892 for testing). For the 20Newsgroups dataset, we also used all available examples for newsgroup 1 (480 for training, and 318 for testing) and for newsgroup 20 (376 for training, and 251 for testing). For the TB dataset we split the data into 80% training and 20% testing examples, which amounts to 260 training and 70 testing examples for HIV negative, and 133 training and 28 test examples for HIV positive. The feature vectors from the 20 Newsgroups dataset were transformed using the transformation \(\log(1 + x)\), which led to an improvement in the performance of L-BFGS and gradient descent for logistic-regression. All algorithms were run till one of the following stopping criteria was met: (1) the relative change in the objective between two consecutive iterations was less than \(10^{-6}\); (2) the magnitude of the gradient was less than \(10^{-8}\); (3) the relative change in the norm of \(\theta\) between two consecutive iterations was less than \(10^{-2}\).

5 Additional Results

Table 1 shows the classification accuracy (%) on test set using Logistic regression.
Figure 2: Top: an example of the proposed local (red curve) and global (green curve) surrogate for a 1D function (blue curve) given by \( f(x) = I \exp(-x) + r - y \log(I \exp(-x) + r) \) with \( I = 10^5 \), \( y = 10^4 \), \( r = 10 \). Bottom: second derivative of \( f \). The expansion point (marked by an asterisk) is located at a “shallow region”. The majorization domain for the local surrogate is \([a, b] = (-\infty, 7]\), which is computed by Algorithm 4.1 and Algorithm 4.2. The right boundary \( b \) is chosen between the point of minimum curvature (marked by a square) and the expansion point. Here we chose \( \alpha = 0.5 \) and \( \beta = 0.3 \) for the parameters of Algorithm 4.1. The convex extension of the local surrogate beyond \( b \) is not shown.

Table 1: Classification accuracy (%) on test set using Logistic regression. LIBLIN uses an L1 penalty and the rest of the methods use a nonconvex log-penalty. For the latter, 10 different random initializations were used and the mean and standard deviation are presented. GD=Gradient Descent, RProp=RMSProp, AGrad=AdaGrad, PMM=Parallel Majorization-Minimization, DRD=Dynamically Restricted Domain.

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST</th>
<th>20 News</th>
<th>TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIBLIN</td>
<td>96.69</td>
<td>79.61</td>
<td>89.69</td>
</tr>
<tr>
<td>LBFGS</td>
<td>96.13 ± 0.11</td>
<td>76.2 ± 1.06</td>
<td>88.45 ± 1.34</td>
</tr>
<tr>
<td>CG</td>
<td>96.4 ± 0.13</td>
<td>77.93 ± 0.85</td>
<td>85.57 ± 2.06</td>
</tr>
<tr>
<td>GD</td>
<td>95.6 ± 0.68</td>
<td>77.21 ± 3</td>
<td>87.63 ± 0.73</td>
</tr>
<tr>
<td>PSCA</td>
<td>89.54 ± 0</td>
<td>68.7 ± 0.36</td>
<td>46.19 ± 0.46</td>
</tr>
<tr>
<td>RProp</td>
<td>96.34 ± 0.11</td>
<td>83.18 ± 0.63</td>
<td>55.26 ± 25.05</td>
</tr>
<tr>
<td>AGrad</td>
<td>95.17 ± 0.08</td>
<td>81.3 ± 0.34</td>
<td>54.84 ± 24.67</td>
</tr>
<tr>
<td>PMM</td>
<td>96.27 ± 0</td>
<td>75.89 ± 0.32</td>
<td>90.72 ± 0</td>
</tr>
<tr>
<td>PMM-DRD</td>
<td>96.49 ± 0.17</td>
<td>76.68 ± 0.17</td>
<td>90.72 ± 0</td>
</tr>
</tbody>
</table>

References