A Dual-Augmented Block Minimization Framework for Learning with Limited Memory
Ian E.H. Yen¹, Shan-Wei Lin², Shou-De Lin²
¹University of Texas at Austin. ²National Taiwan University

Abstract
In this work, we consider Empirical Risk Minimization (ERM) when data size is larger than the memory capacity of machines.
State-of-the-art batch algorithms become slow due to I/O.
Online algorithms converge slowly (especially for non-smooth regularizer), while existing distributed approach requires data to fit into memory of several machines.
We propose a Block Minimization framework that generalizes (Yu et al. 2010) for SVM to that for any convex ERM, which can be integrated with any convex optimization solver to achieve global fast convergence in limited-memory condition.

Regularized Empirical Risk Minimization (ERM)
Given a data set \[ D = \{ (\phi_n, y_n) \}_{n=1}^N, \] the ERM estimates model through

\[
\min_{\mathbf{w}} F(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \sum_{n=1}^{N} L_n(\phi_n, \mathbf{w})
\]

where \[ \mathbf{w} \in \mathbb{R}^d \] are parameters to be estimated, \[ \phi_n \] is \[ d \times p \] feature matrix of \[ n \]-th sample, and \[ \mathcal{L}(\cdot), \mathcal{L}_n(\cdot) \] are loss function and regularizer.

Examples
- **Multiclass Classification**: \( (p = 3^y) \), where \( y \) is label set.
- Logistic loss: \( \mathcal{L}(\mathbf{w}) = \log \sum_{y=1}^{Y} e^{\langle \mathbf{w}, \phi_n \rangle} - \langle \mathbf{w}, \phi_n \rangle y_n \).
- Hinge loss: \( \mathcal{L}(\mathbf{w}) = \max\{ 1 - \langle \mathbf{w}, \phi_n \rangle y_n, 0 \} \).
- **Multitask Regression**: \( (p = K, \text{ where } K \text{ is tasks}) \)
- Square loss: \( \mathcal{L}(\mathbf{w}) = \frac{1}{2} \| \mathbf{w} \|_2^2 \).
- Others: Ranking, Matrix Completion, Structured Learning, Clustering etc.,
- **Regularizers**: L2 norm \( \| \mathbf{w} \|_2 \), L1 norm \( \| \mathbf{w} \|_1 \), Group norm \( \| \mathbf{W} \|_2 \), Nuclear norm \( \| \mathbf{W} \|_F \), etc.

Strong Convexity & Smoothness
- A function \( f(x) \) is strongly convex iff it is lower bounded by a simple quadratic function
  \[
f(y) \geq f(x) + \nabla f(x)^{T}(y-x) + \frac{1}{2}M \| x-y \|_2^2
\]
  for some constant \( M > 0 \) and \( x, y \in \text{dom}(f) \).
- A function \( f(x) \) is smooth iff it is upper bounded by a simple quadratic function
  \[
f(y) \leq f(x) + \nabla f(x)^{T}(y-x) + \frac{1}{2}M \| x-y \|_2^2
\]
  for some constant \( 0 < M < \infty \) and \( x, y \in \text{dom}(f) \).

**Theorem 1**: A convex function \( f(\cdot) \) is smooth with parameter \( M \) if and only if its conjugate function \( f^*(\cdot) \) is strongly convex with parameter \( m = 1/M \).

Dual Form
The dual of ERM problem (1) is of the form

\[
\min_{\alpha} G(\mathbf{\alpha}) = R^{*}(-\sum_{n=1}^{N} \Phi_n^{T} \mathbf{\alpha}_n) + \sum_{n=1}^{N} L_n(\mathbf{\alpha}_n).
\]

**Block Coordinate Descent on (4) guarantees convergence only for smooth \( R^{*}(\cdot) \) (strongly convex \( R(\cdot) \)), which does not hold for most of regularizers.

Use Proximal Minimization to ensure convergence for any convex ERM.

Duality Formulation
The dual of ERM problem (1) is of the form

\[
\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \sum_{n=1}^{N} L_n(\phi_n, \mathbf{w})
\]

which, by Theorem 1, has a dual problem of smooth \( \mathcal{R}(\cdot) \) since the augmented regularizer \( \mathcal{R}(\cdot) \) is strongly convex.

Let \( \mathcal{L}(\mathbf{w}, \mathbf{a}) \) be the Lagrangian of (5). Our algorithm performs Block-Coordinate Descent on dual of (5), which minimizes block of variables \( \alpha_n \) via

\[
\mu_n \left( \sum_{i \neq n} \Phi_i^{T} \mathbf{\alpha}_i - \Phi_n^{T} \mathbf{\alpha}_n \right) = \mu_n (\mathbf{1} - \Phi_n) \mathbf{\alpha}_n
\]

which requires only data in block \( B \) and can be solved via any solver designed for (1), where vector \( \mu_n \) memorizes historical gradient given by data not in \( B \): (7)

- **Algorithmic Framework**

Disk
Memory

\[
\left( w^{t+1}, \mu^{t+1} \right) \leftarrow \text{Solver} (\Phi_B, \mu^t)
\]

\[
\mu^{t+1} = w^{t+1} - \mathbf{\alpha}^{t+1}
\]

Save \( w^{t+1} \)
Load \( w^{t+1} \)

**Convergence of Block Minimization**

- The dual of (5) takes the form

\[
\min_{\mathbf{w}} \mathcal{R}(\mathbf{w}) = \sum_{n=1}^{N} \frac{1}{2} \| \mathbf{w} - \mathbf{\alpha}_n \|^2
\]

where \( \mathcal{R}(\cdot) \) is the convex conjugate of \( \mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w}) + \frac{1}{2} \| \mathbf{w} - \mathbf{\alpha}_n \|^2 \).

- Since \( \mathcal{R}(\mathbf{w}) \) is strongly convex with parameter \( m = 1/\mathcal{L}_2(\cdot) \), the convex conjugate \( \mathcal{R}^{*}(\cdot) \) is smooth with parameter \( M = \mathcal{L}_2(\cdot) \) according to Theorem 1.

The augmented dual (9) is composite of a convex, smooth function plus a convex, block-separable function, for which BCD has guaranteed convergence to optimum. In particular, with probability \( 1 - p \)

\[
\mathcal{F}(\alpha) - F^* \leq c \mathcal{K}(\alpha) \log \frac{1}{\epsilon} - F^*
\]

for some constant \( c > 0 \) and probability \( 1 - p \).

- Due to non-expansiveness of proximal operator, we show that solving sub-problem (5) with tolerance \( \epsilon \) suffices for convergence to \( \epsilon \), overall precision where \( t \) is the number of outer iterations required by (12), (13).

- The overall procedure requires \( O(K \log(1/\epsilon) \log(1/t)) \) block minimization steps if \( \mathcal{R}_L(\cdot), \mathcal{R}_R(\cdot) \) are strictly convex and smooth or polynomial. Otherwise, we need \( O(K \log(1/\epsilon) \log(1/t)) \) block minimization steps as long as \( \mathcal{R}_L(\cdot) \) is smooth.

**Experiments**

- **Data**: Year-Pred, WebSpam, RCV1
- **Accuracy**: 315,000, 31,500, 681,740
- **Error**: 4,673.51, 51,630, 8,088,636
- **Memory (GB)**: 18.7, 12.0, 0.80