Accelerated, Parallel and PROXimal coordinate descent

Olivier Fercoq, olivier.fercoq@ed.ac.uk and Peter Richtárik, peter.richtarik@ed.ac.uk

Problem

Minimize for \( x \in \mathbb{R}^N \) the composite function \( F \)

\[ \min_{x \in \mathbb{R}^N} \{ F(x) = f(x) + \psi(x) \} \]

- \( f : \mathbb{R}^N \to \mathbb{R}, \) convex, differentiable, not strongly convex
- \( \psi : \mathbb{R}^N \to \mathbb{R} \cup \{ +\infty \}, \) convex, separable

\( \psi(x) = \sum_{i=1}^n \psi_i(x^{(i)}) \)

Examples: \( \lambda \| x \|_1, I_{[0,1]}(x), \lambda \sum_{i=1}^n \| x^{(i)} \|_2 \)

Coordinate descent

- At each iteration, one solves a 1-dimensional optimization problem
- Very cheap iterations: for sparse problems, less than the cost of summing two vectors
- Many iterations are required
- Famous in machine learning: \( L_1 \)-regularised linear least squares, support vector machines, non-negative factorisation,
- Convergence in \( O(1/k) \): we bring it to \( O(1/k^2) \) and we update several coordinates in parallel

APPROX

Pick \( z_0 \in \text{dom} \psi \), set \( \tau = \frac{1}{\sqrt{N}} \), \( u_0 = 0 \)

for \( k \geq 0 \) do

Generate a random set of coordinates \( S_k \sim S \)

for \( i \in S_k \) do

\[ t_k^{(i)} = \arg \min_{t \in [0,1]} \{ \nabla f(\theta_k^{(i)} u_k + z_k, t) \} \]

\[ z_k^{(i)} = z_k^{(i)} + t_k^{(i)} \]

\[ u_k^{(i)} + 1 \leftarrow u_k^{(i)} - 1 - \frac{2 \theta_k}{\theta_k^{(i)}} \]

end for

\[ \theta_k = \sqrt{2 + \sqrt{4 - 2 \frac{\theta_k^{(i)}}{\theta_k^{(i)}}}} \]

end for

OUTPUT: \( \theta_k u_k + z_k \)

Parallel

Assume:

- \( f \) is partially separable of degree \( \omega \):

\[ f(x) = \sum_{j=1}^m f_j(x) \]

\( f_j \) depends on at most \( \omega \) coordinates

\( \nabla f \) coordinatewise Lipschitz: \( \forall x \in \mathbb{R}^N, t \in \mathbb{R}^N, \)

\[ \| \nabla f(x + U(t)) - \nabla f(x) \|_i \leq L_i \| t \|_i \]

\( \hat{S} \) is a \( \tau \)-nice sampling:

If \( |S| = \tau \), then \( P(S = S) = \frac{1}{\binom{n}{\tau}} \)

\( \beta = 1 + \frac{(\omega - 1)(\tau - 1)}{\max(1, n - 1)} \)

Then for all \( x, h \in \mathbb{R}^N, (f, \hat{S}) \sim \text{ESO}(\beta, L) \)[4]:

\[ E \left[ f(x + h|\hat{S}) \right] \leq f(x) + \frac{\tau}{n} \| \nabla f(x, h) + \frac{\beta}{2} \| h \|_L^2 \]

Proximal

Lemma. For all \( k \geq 0 \)

\[ \theta_k^2 u_{k+1} + z_{k+1} = \sum_{i=0}^k \gamma_i z_i \]

where \( \gamma_0, \gamma_1, \ldots, \gamma_k \geq 0 \) and \( \sum_{i=0}^k \gamma_i = 1 \)

\[ \Rightarrow \theta_k^2 u_{k+1} + z_{k+1} \in \text{dom} \psi, \quad \theta_k^2 u_k + z_k \in \text{dom} \psi \]

L1-regularised L1 regression

Dorothea dataset: \( m = 800, N = 100,000, \omega = 6.061, \tau = 4, \epsilon = 0.1 \) (smoothing \( \beta \))

\[ F(x) = \| Ax - b \|_1 + \| x \|_1 \]

Support Vector Machines

Malicious URL dataset:

\( m = 2,306,120, N = 3,231,961, \tau = 1 \)

\[ F(x) = \frac{1}{2N} \sum_{j=1}^m \left( \sum_{i=1}^N y_{ij}^2 \right) - \frac{1}{N} \sum_{i=1}^N x_i + I_{[0,1]}(x) \]

Accelerated

Theorem. Suppose that \( (f, \hat{S}) \sim \text{ESO}(\beta, L) \).

Denote \( \tau = E[|\hat{S}|] > 0 \):

\[ E[F(x_k) - F(x)] \leq \frac{4 \tau}{(k+1) \tau^2} \]

where \( C = (1 - \frac{1}{\tau})(F(x_0) - F(x_1)) + \frac{\tau}{\epsilon} \| x_0 - x_1 \|_L^2 \)

For \( \epsilon > 0 \), we obtain an \( \epsilon \)-solution in expectation after at most

\[ k \geq \frac{2n}{\tau} \left( \frac{1 - \frac{1}{\tau}}{\epsilon} \right) \frac{F(x_0) - F(x_1)}{2} + \frac{\tau}{\epsilon} \| x_0 - x_1 \|_L^2 \]

Conclusion

- First accelerated, parallel and proximal coordinate descent method
- Needs to be able to compute \( \nabla f(\theta_k^2 u_k + z_k) \) without actually summing the 2 vectors: this includes quadratics, smoothed \( L_1 \) norm and logistic regression
- Very promising numerical experiments on machine learning problems: several times faster than the state of the art

Perspectives:
- Nonuniform samplings
- Line search
- Universal algorithms
- AdaBoost

Acknowledgement

EPSRC grant EP/1017127/1 (Mathematics for Fast Digital Resources), EPSRC grant EP/G036136/1 and the Scottish Funding Council (NAIS), EPSRC grant EP/K02325X/1 (Accelerated Coordinate Descent Methods for Big Data Problems) and Simons Institute for the Theory of Computing.

References


