Accelerated Coordinate Descent with Arbitrary Sampling and Best Rates for Minibatches

Filip Hanzely1  Peter Richtárik1, 2, 3
King Abdullah University of Science and Technology1  University of Edinburgh2  MIPT3

ACCELERATED COORDINATE DESCENT (ACD)

Algorithm 1 ACD (Accelerated coordinate descent with arbitrary sampling)

1. Parameters: i.i.d. proper samplings S^k \sim \mathcal{D} \subseteq [n]; \sigma > 0; stepsize parameters \eta_i > 0.
2. Initial iterate \mathbf{x}^0 = \mathbf{x}^* \in \mathbb{R}^n.
3. for \ k = 0, 1, \ldots do
4. \quad \mathbf{x}^{k+1} = (1 - \eta_k y_k) \mathbf{x}^k + \eta_k y_k \mathbf{M} \mathbf{x}^k.
5. Get \mathbf{S}^k = \mathcal{D}.
6. \quad y^{k+1} = \mathbf{M}^{k+1}\mathbf{M}^{-1}(\mathbf{x}^{k+1} - \mathbf{x}^k) + \frac{1}{\eta_k} \nabla f(\mathbf{x}^{k+1})^T y^{k+1}.
7. \quad \mathbf{x}^{k+1} = \frac{1}{\eta_k} \nabla f(\mathbf{x}^{k+1}) + \frac{1}{\eta_k} \mathbf{M} \mathbf{x}^{k+1}.
8. end for

CONVERGENCE RATE (ACD)

Theorem

\[ E[\mathbf{S}^k] \leq \left(1 - \frac{0.618}{\mathbf{M}^{k+1}}\mathbf{M}^{-1}\mathbf{M} \right) \frac{\sigma}{\eta_k} \mathbf{M}^{k+1} \]

Optimal \( \varepsilon \) has the form: \( \mathbf{v}_1 = \mathbf{c}_1 \mathbf{M}^{k+1} \)

Rate has the form (1 - 0.618 \( \sqrt{\mathbf{M}} \))

Smaller \( \varepsilon \) \quad Better rate

IMPORTANCE SAMPLING FOR MINIBATCHES

Two samplings \( S \) satisfying \( E[|S|] = \) \( \tau \) - minibatch size

Sampling 1

(standard; no importance sampling)

sample uniformly from all subsets of \( \{1, 2, \ldots, n\} \) of size \( \tau \)

Rate \( (1 - 0.618\sqrt{\mathbf{M}}) \)

Sampling 2

(new; importance sampling)

sample each \( i \) independently with probability \( \mathbf{p}_i \)

Rate \( (1 - 0.618\sqrt{\mathbf{M}}) \)

THEOREM (ACD)

\[ \forall \mathbf{M} : c_2 \leq O(\mathbf{M} \mathbf{M}^{-1}) \mathbf{M} \quad \exists \mathbf{M} \text{ such that } c_2 \leq O(\mathbf{M} \mathbf{M}^{-1}) \mathbf{M} \]

In words: Sampling 2 can be at most \( O(\mathbf{M} \mathbf{M}^{-1}) \mathbf{M} \) times worse than Sampling 1

Sampling 2 can be \( O(\mathbf{M} \mathbf{M}^{-1}) \mathbf{M} \) times better than Sampling 1

THEOREM (CD)

Importance sampling can be at most \( O(\mathbf{M}) \) times better than uniform

Importance sampling can be \( O(\mathbf{M}) \) times better than uniform

REFERENCES