Accelerated Coordinate Descent with Arbitrary Sampling and Best Rates for Minibatches

Filip Hanzely[1] Peter Richtárik[1, 2, 3]

King Abdullah University of Science and Technology[1] University of Edinburgh[2] MIPT[3]

PROBLEM

smooth & strongly convex

$$\min_{x\in\mathbb{R}^n} f(x)$$

very large

ASSUMPTIONS

σ strong convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x-y\rangle + \frac{\sigma}{2}\|x-y\|^2$$

positive definite matrix

$\mathbf{M}$ smoothness

$$f(x) \leq f(y) + \langle \nabla f(y), x-y\rangle + \frac{1}{2}(x-y)^\top \mathbf{M}(x-y)$$

Recovers L smoothness for $\mathbf{M} = L\mathbf{I}$ identity matrix

(Lipschitz continuity of gradients)

EXAMPLES

Lemma [3]

$$f(x) = \sum_{i=1}^m \phi_i(\mathbf{A}_i x) \Rightarrow f \text{ is } \mathbf{M} \text{ smooth}$$

ϕ_i is L_i smooth

$$\mathbf{M} = \sum_{i=1}^m L_i \mathbf{A}_i^\top \mathbf{A}$$

Logistic regression $f(x) = \frac{1}{m}\sum_{i=1}^m \log\left(1+\exp\left(\mathbf{A}_{i:}x\cdot b\right)\right) + \frac{\lambda}{2}\|x\|^2$

Dual of SVM with squared hinge loss

indicator function

$$f(x) = \frac{1}{\lambda n^2}\sum_{j=1}^m\left(\sum_{i=1}^n b_i \mathbf{A}_{ji} x_i\right)^2 - \frac{1}{n}\sum_{i=1}^n x_i + \frac{1}{4n}\sum_{i=1}^n x_i^2 + \mathcal{I}_{[0,\infty]}(x)$$

$n > m \Rightarrow$ coordinate descent (CD) is the state of the art

KNOWN CONVERGENCE RATES

METHOD	PROBABILITY OF CHOOSING COORDINATE i	ERROR AFTER k STEPS
CD [1, 2] Coordinate Descent	$p_i = \frac{\mathbf{M}_{ii}}{\sum_j \mathbf{M}_{jj}}$	$\left(1-\frac{\sigma}{\mathrm{Trace}(\mathbf{M})}\right)^k$
ACD [4] Accelerated Coordinate Descent	$p_i = \frac{\sqrt{\mathbf{M}_{ii}}}{\sum_j \sqrt{\mathbf{M}_{jj}}}$	$\left(1-\frac{\sqrt{\sigma}}{\sum_{i=1}^n \sqrt{\mathbf{M}_{ii}}}\right)^k$

CONTRIBUTIONS

1. First ACD method with arbitrary sampling

Sampling: set-valued random variable $S\sim\mathcal{D}$ with values being subsets of $\{1,2,\dots,n\}$ $\quad p_i = \mathrm{Prob}(i\in S)$

Arbitrary sampling = any sampling with $p_i > 0$ for all i

2. Importance sampling for minibatches for CD and ACD (can be much better than uniform)

ESO: EXPECTED SEPARABLE OVERAPPROXIMATION

any $x\in\mathbb{R}^n$ any $h\in\mathbb{R}^n$ $p_i = \mathrm{Prob}(i\in S)$

$$\mathbb{E}_{S\sim\mathcal{D}}\left[f\left(x+\sum_{i\in S} h_i e_i\right)\right] \leq f(x) + \sum_{i=1}^n p_i \nabla_i f(x) h_i + \frac{1}{2}\sum_{i=1}^n p_i v_i h_i^2$$

sampling $\quad$ i[th] standard unit basis vector ("coordinate vector") $\quad$ ESO parameters (defined by this inequality)

Lemma [3] Hadamard product

$$\mathbf{P}\circ\mathbf{M} \preceq \mathrm{Diag}(p_1v_1,\dots,p_nv_n) \Rightarrow \text{ESO inequality holds}$$

"probability matrix" defined by $\mathbf{P}_{ij} = \mathrm{Prob}(i\in S\wedge j\in S)$

identity: $\mathbb{E}[|S|] = \mathrm{Trace}(\mathbf{P})$

Corollary $\mathrm{Prob}(|S|=1)=1$ $\Rightarrow$ ESO inequality holds with $v_i = \mathbf{M}_{ii}$

ACCELERATED COORDINATE DESCENT (ACD)

Algorithm 1 ACD (Accelerated coordinate descent with arbitrary sampling)

Parameters: i.i.d. proper samplings $S^k\sim\mathcal{D}$; $v\in\mathbb{R}^n_{++}$; $\sigma>0$; stepsize parameters $\eta,\theta>0$.
Initial iterate $y^0 = z^0 \in \mathbb{R}^n$
for $k = 0,1\dots$ **do**
$\quad x^{k+1} = (1-\theta)y^k + \theta z^k$
$\quad$ Get $S^k\sim\mathcal{D}$
$\quad y^{k+1} = x^{k+1} - \sum_{i\in S^k}\frac{1}{v_i}\nabla_i f(x^{k+1})e_i$ $\quad$ i[th] partial derivative
$\quad z^{k+1} = \frac{1}{1+\eta\sigma}\left(z^k + \eta\sigma x^{k+1} - \sum_{i\in S^k}\frac{\eta}{p_i}\nabla_i f(x^{k+1})e_i\right)$
end for

CONVERGENCE RATE (ACD)

Theorem

$$\mathbb{E}\left[\Psi^k\right] \leq \left(1-0.618\sqrt{\sigma\min_i\frac{p_i^2}{v_i}}\right)^k \Psi^0$$

Lyapunov function

$$\Psi^k = \frac{1}{\theta^2}(f(y^k)-f(x^*)) + \frac{1}{2(1-\theta)}\|z^k - x^*\|^2$$

$$\theta = 0.618\sqrt{\sigma\min_i\frac{p_i^2}{v_i}}$$

constant depending on $\mathbf{M}$ and $\mathcal{D}$

Optimal v has the form: $v_i = cp_i^2$

Rate has the form $\left(1-0.618\sqrt{\frac{\sigma}{c}}\right)^k$

Smaller c ⇒ Better rate

IMPORTANCE SAMPLING FOR MINIBATCHES

Two samplings S satisfying $\mathbb{E}[|S|] = \tau$ minibatch size

Sampling 1
(standard; no importance sampling)
sample uniformly from all subsets of $\{1,2,\dots,n\}$ of size τ

Rate $(1-0.618\sqrt{\sigma/c_1})^k$

Sampling 2
(new; importance sampling)
sample each i independently with probability

$$p_i = \frac{2\mathbf{M}_{ii}}{\sqrt{\mathbf{M}_{ii}^2 + 2\mathbf{M}_{ii}\delta} + \mathbf{M}_{ii}}$$

δ s.t. $\sum p_i = \tau$

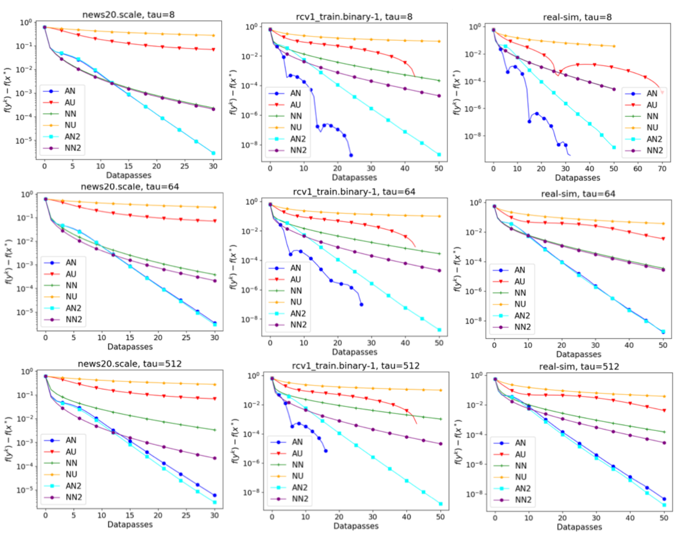
Rate $(1-0.618\sqrt{\sigma/c_2})^k$

Theorem (ACD) $\forall\,\mathbf{M}$: $c_2 \leq \mathcal{O}(\tau)\,c_1$ $\quad\exists\,\mathbf{M}$ such that $c_2 \leq \mathcal{O}\left(\frac{\tau^2}{n^2}\right)c_1$

In words: Sampling 2 can be at most $\mathcal{O}(\sqrt{\tau})$ times worse than Sampling 1
Sampling 2 can be $\mathcal{O}\left(\frac{n}{\tau}\right)$ times better than Sampling 1

Theorem (CD)

Importance sampling can be at most $\mathcal{O}(1)$ times worse than uniform
Importance sampling can be $\mathcal{O}(n)$ times better than uniform



REFERENCES

[1] Yurii Nesterov. **Efficiency of coordinate descent methods on huge-scale optimization problems.** *SIAM Journal on Optimization* 22(2), 341-362, 2012.

[2] Peter Richtárik and Martin Takáč. **On optimal probabilities in stochastic coordinate descent methods.** *Optimization Letters* 10(6), 1233-1243, 2016.

[3] Zheng Qu and Peter Richtárik. **Coordinate descent with arbitrary sampling II: expected separable overapproximation.** *Optimization Methods and Software* 31(5), 858-884, 2016.

[4] Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik and Yang Yuan. **Even faster accelerated coordinate descent using non-uniform sampling.** *ICML* 2016.

[5] Dominik Csiba and Peter Richtárik. **Importance sampling for minibatches.** *Journal of Machine Learning Research* 19(27), 1-21, 2018.