Accelerated, Parallel and PROXimal coordinate descent

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

THE UNIVERSITY of EDINBURGH

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

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 least squares, support vector machines, non-negative factorization…
- Convergence in $O(1/k)$: we bring it to $O(1/k^2)$ and we update several coordinates in parallel

APPROX

Pick $z_0 \in \mathrm{dom}\psi$, set $\tau = \mathbf{E}[|\hat{S}|]$, $\theta_0 = \frac{\tau}{n}$, $u_0 = 0$
for $k \geq 0$ **do**
  Generate a random set of coordinates $S_k \sim \hat{S}$
  for $i \in S_k$ **do**

$$t_k^{(i)} = \arg\min_{t\in\mathbb{R}^{N_i}} \langle \nabla_i f(\theta_k^2 u_k + z_k), t\rangle + \frac{n\theta_k \beta L_i}{2\tau}\|t\|_{(i)}^2 + \psi_i(z_k^{(i)} + t)$$

$$z_{k+1}^{(i)} \leftarrow z_k^{(i)} + t_k^{(i)}$$

$$u_{k+1}^{(i)} \leftarrow u_k^{(i)} - \frac{1 - \frac{n}{\tau}\theta_k}{\theta_k^2} t_k^{(i)}$$

  end for

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$$

end for
OUTPUT: $\theta_k^2 u_{k+1} + z_{k+1}$

Parallel

Assume:

- f is partially separable of degree ω:

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

  f_j depends on at most ω coordinates

- ∇f coordinatewise Lipschitz: $\forall x \in \mathbb{R}^N, t \in \mathbb{R}^{N_i}$,

$$\|\nabla_i f(x + U_i t) - \nabla_i f(x)\|_{(i)}^* \leq L_i \|t\|_{(i)}$$

- $\hat{S}$ is a τ-nice sampling: If $|S| = \tau$, then $\mathbf{P}(\hat{S} = S) = \frac{1}{\binom{n}{\tau}}$

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

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

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

Proximal

Lemma. *For all $k \geq 0$*

$$\theta_k^2 u_{k+1} + z_{k+1} = \sum_{l=0}^k \gamma_k^l z_l$$

where $\gamma_k^0, \gamma_k^1, \dots, \gamma_k^k \geq 0$ and $\sum_{l=0}^k \gamma_k^l = 1$

$\Rightarrow \theta_k^2 u_{k+1} + z_{k+1} \in \mathrm{dom}\psi$, $\theta_k^2 u_k + z_k \in \mathrm{dom}\psi$

Accelerated

Theorem. *Suppose that $(f, \hat{S}) \sim$ ESO(β, w). Denote $\tau = \mathbf{E}[|\hat{S}|] > 0$:*

$$\mathbf{E}[F(x_k) - F(x_*)] \leq \frac{4n^2}{(k\tau + n)^2} C$$

where $C = (1 - \frac{\tau}{n})(F(x_0) - F(x_*)) + \frac{\beta}{2}\|x_0 - x_*\|_w^2$

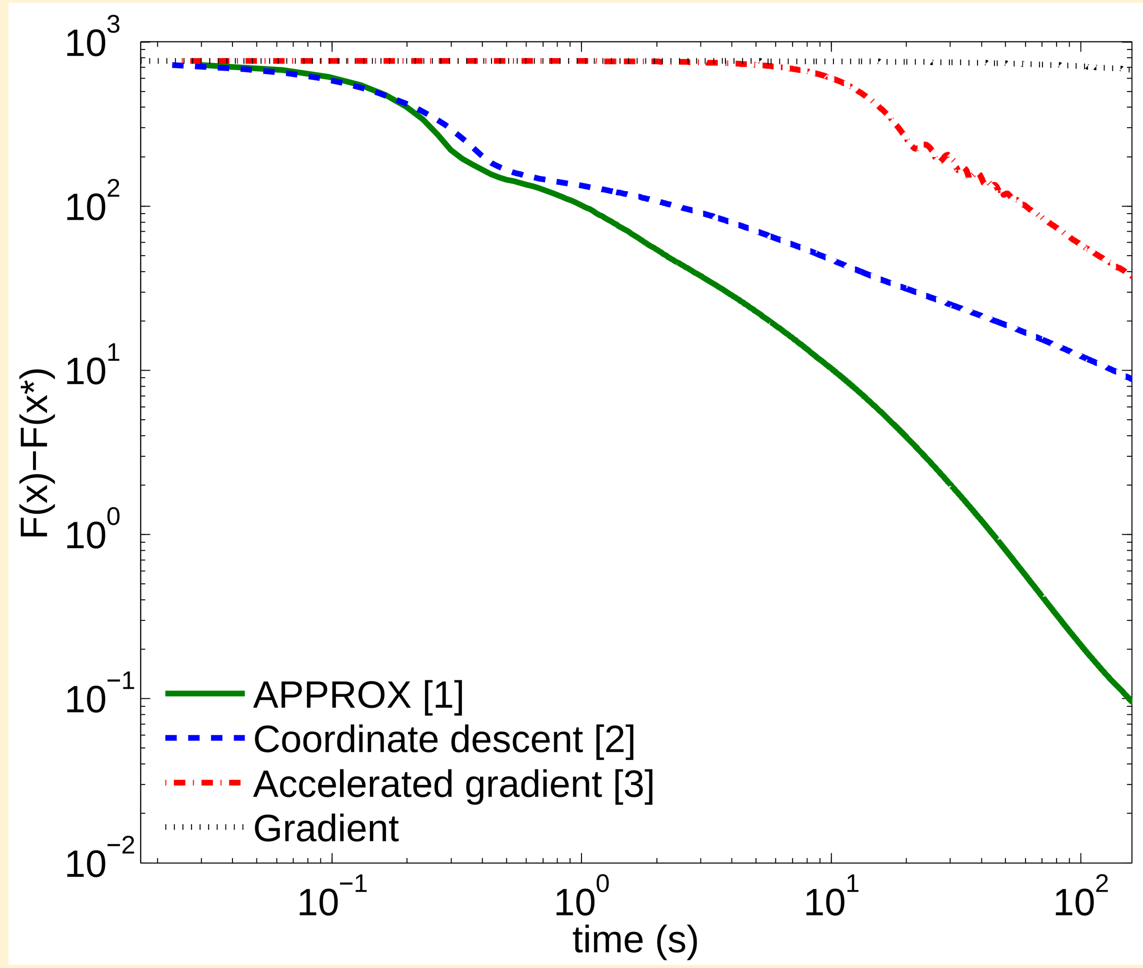
For $\varepsilon > 0$, we obtain an ε-solution in expectation after at most

$$k \geq \frac{2n}{\tau}\sqrt{\frac{(1 - \frac{\tau}{n})(F(x_0) - F(x_*)) + \frac{\beta}{2}\|x_0 - x_*\|_L^2}{\varepsilon}}$$

L_1-regularised L_1 regression

Dorothea dataset: m=800, N=100,000, ω=6,061, τ=4, $\epsilon = 0.1$ (smoothing [3])

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

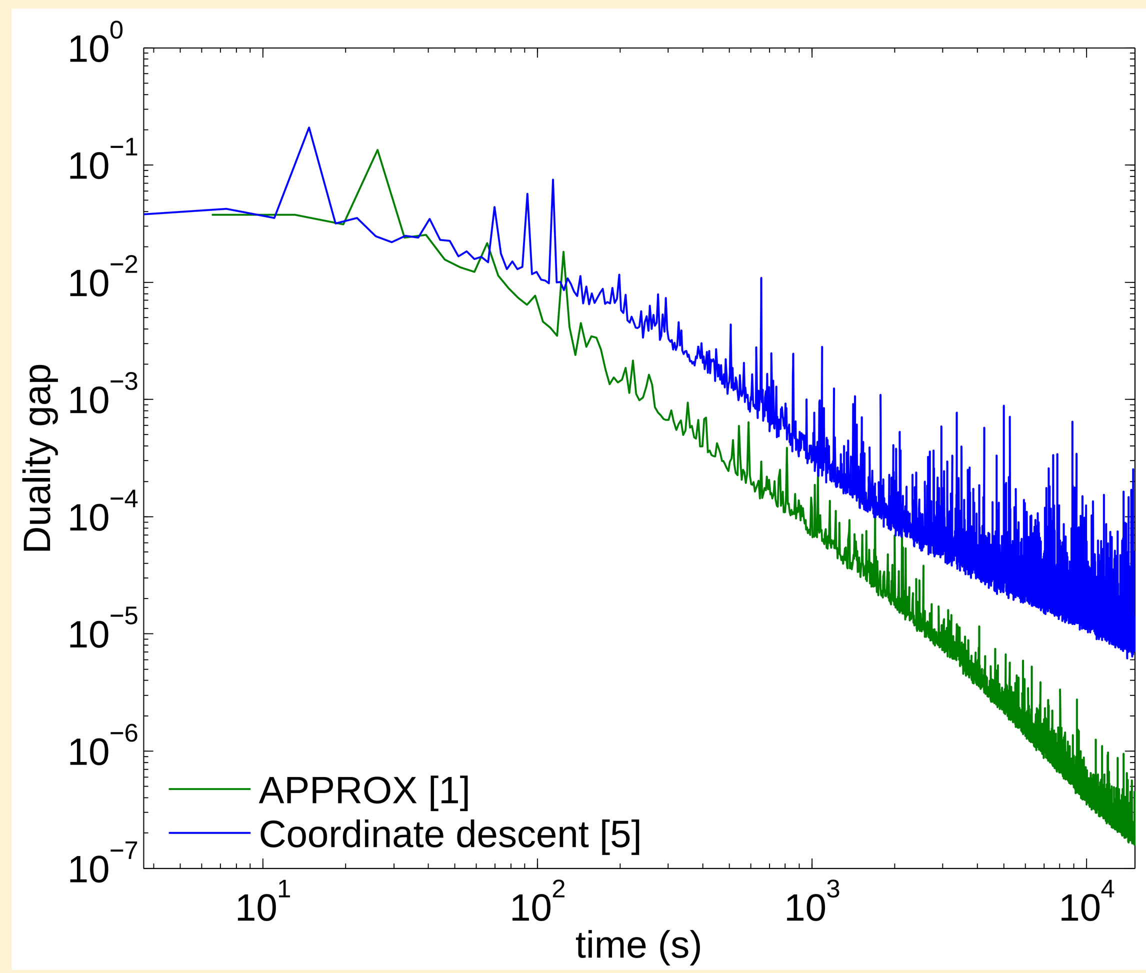


Comparison of algorithms: $\|Ax - b\|_1 + \|x\|_1$

Support Vector Machines

Malicious URL dataset: m = 2,396,130, N = 3,231,961, τ = 1

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



One order of magnitude faster for the dual SVM

Conclusion

- First accelerated, parallel and proximal coordinate descent method
- Needs to be able to compute $\nabla_i f(\theta_k^2 u_k + z_k)$ 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/I017127/1 (Mathematics for Vast 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

[1] Olivier Fercoq and Peter Richtárik. Accelerated, parallel and proximal coordinate descent. *Preprint arXiv:1312.5799*, 2013.

[2] Olivier Fercoq and Peter Richtárik. Smooth minimization of nonsmooth functions by parallel coordinate descent. *Preprint arXiv:1309.5885*, 2013.

[3] Yurii Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[4] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization problems. *Preprint arXiv:1212.0873*, 2012.

[5] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599.