Regularized Nonlinear Acceleration

SparTaN — Sparse Representations and Compressed Sensing Training Network

Damien Scieur (damien.scieur@inria.fr), **Alexandre d'Aspremont and Francis Bach**

SIERRA Team - École Normale Supérieure/INRIA, Paris

Problem formulation

We want to minimize a (convex) function f:

$$\min_{x \in \mathbb{R}^d} f(x) \ , \quad \text{Solution: } x^*$$

If we use gradient method with step size $1/L$ (the constant of smoothness), we get the sequence

$$\{x_0,\ x_1,\ \dots,\ x_N\} \quad \text{where} \quad x_{i+1} = x_i - \tfrac{1}{L} f'(x_i)$$

Linearizing the equation around x^* gives

$$x_{i+1} = x_i - \tfrac{1}{L}\big(f''(x^*)(x_i - x^*) + \mathcal{O}(\|x_i - x^*\|^2)\big)$$

We have the **perturbed vector auto-regressive (VAR) process**

$$x_{i+1} - x^* = A(x_i - x^*) + \mathcal{O}(\|x_i - x^*\|^2)$$

where $A = I - \frac{1}{L} f''(x^*)$.

Two main questions

- How can we accelerate convergence of VAR processes?
- What is the impact of perturbations?

Structure of VAR processes

Assume we have the sequence $\{x_0,\ x_1,\ \dots,\ x_{N+1}\}$ produced by

$$x_{i+1} - x^* = A(x_i - x^*) = A^{i+1}(x_0 - x^*)$$

Averaging with coefficients c_i (with unitary sum) gives

$$\textstyle\sum_{i=0}^N c_i x_i = x^* + \sum_{i=0}^N c_i A^i (x_0 - x^*) = x^* + \underbrace{p(A)(x_0 - x^*)}_{=\text{Error term}}$$

We need to find c which **minimizes the norm of the matrix polynomial** (i.e. the error term).

Chebyshev acceleration

Idea: Similar to Nesterov's method. It uses coefficients c which minimize the worst case of $\|p(A)(x_0 - x^*)\|$, i.e.

$$c_{\text{Cheby}} = \arg\min_{c\,:\,\mathbf{1}^T c = 1} \left\{ \max_{A:0 \preceq A \preceq \sigma I} \left\| \textstyle\sum_{i=0}^N c_i A^i \right\| \right\}$$

where $\sigma = (1 - \mu/L) < 1$ (in the case of gradient method).
Advantage: Coefficients known in advance.
Drawbacks: Not adaptive, requires the knowledge of μ and L.
Rate of convergence: Optimal if applied on gradient method for minimizing quadratics, not generalizable for non-linear objective.

Acceleration of VAR processes

The mean of VAR processes follows

$$\textstyle\sum_{i=0}^N c_i x_i = x^* + p(A)(x_0 - x^*)$$

We need to minimize $p(A)(x_0 - x^*)$ **using only** x_i.
Main trick: The differences follow

$$x_{i+1} - x_i = (x_{i+1} - x^*) - (x_i - x^*) = (A - I)(x_i - x^*)$$

So their mean is

$$\textstyle\sum_{i=0}^N c_i (x_{i+1} - x_i) = (A - I)p(A)(x_0 - x^*)$$

We can minimize (over c) the combination of differences:

$$\left\|\textstyle\sum_{i=0}^N c_i (x_{i+1} - x_i)\right\| \approx 0 \quad \Rightarrow \quad \|p(A)(x_0 - x^*)\| \approx 0$$

Problem: We do not observe A!

Minimal Polynomial Extrapolation

Idea: Similar to **conjugate gradient** without knowing A. Instead of minimizing $\|p(A)(x_0 - x^*)\|$, MPE solves

$$\min_{c:\mathbf{1}^T c = 1} \left\|\textstyle\sum_{i=0}^N c_i (x_{i+1} - x_i)\right\| = \min_{c:\mathbf{1}^T c = 1} \|Uc\|, \ \text{Solution: } c = \frac{(U^T U)^{-1}\mathbf{1}}{\mathbf{1}^T (U^T U)^{-1}\mathbf{1}}$$

Advantages: No parameter, adaptive, complexity $O(d)$.
Drawback: **Extremely unstable**, and works rarely when applied on non-linear functions **because U is a Krylov matrix**.
Rate of convergence: Similar to Chebyshev's rate (with multiplicative constant).

Regularized MPE (main contribution)

Idea: Solves the regularized version of MPE:

$$\min_{c:\mathbf{1}^T c = 1} \|Uc\| + \lambda \|c\|^2 \quad \Rightarrow \quad c = \frac{(U^T U + \lambda I)^{-1}\mathbf{1}}{\mathbf{1}^T (U^T U + \lambda I)^{-1}\mathbf{1}}$$

If $U^T U$ is perturbed by matrix P, then $\lambda = O(\|P\|)$.

Advantages: **Adaptive**, stable, complexity $O(d)$.
Drawback: Parameter λ (can be found by line-search).
Rate of convergence: **Asymptotically optimal:**

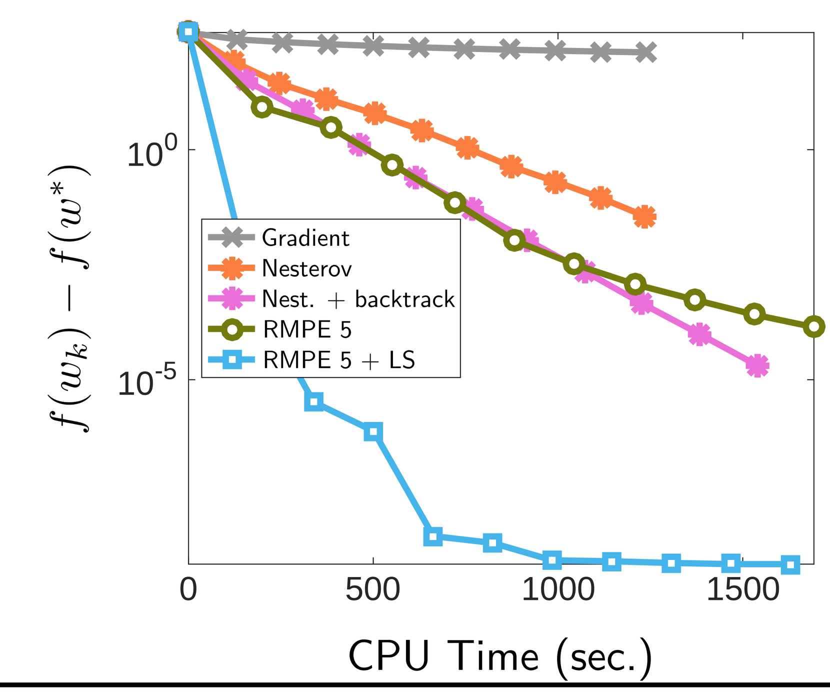
when $\|x_0 - x^*\| \to 0$, then $\|\sum_{i=0}^N c_i x_i - x^*\| = O(1 - \sqrt{\mu/L})^N \|x_0 - x^*\|$

The **global bound** depends of "Regularized Chebyshev Polynomials".

Numerical experiments

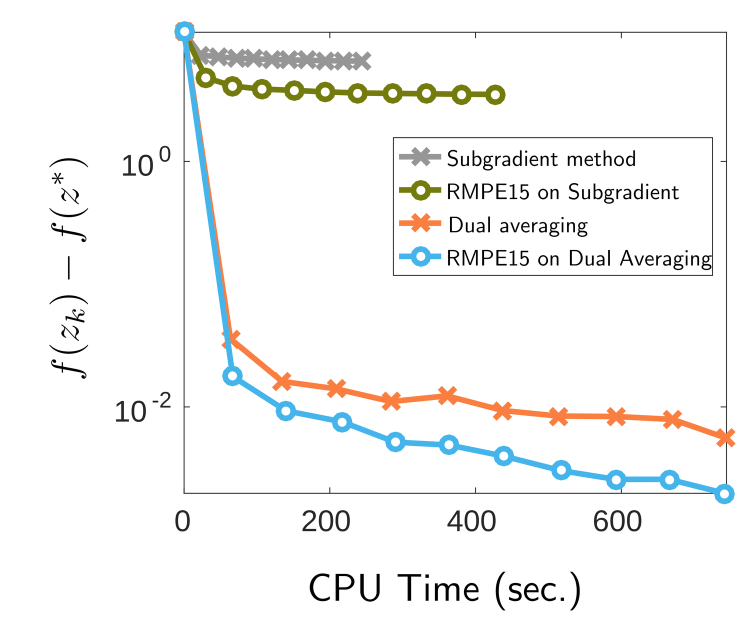
Logistic regression

$$\min_w \sum_{i=1}^m \log\left(1 + \exp(-y_i X_i^T w)\right) + \frac{\tau}{2}\|w\|^2$$



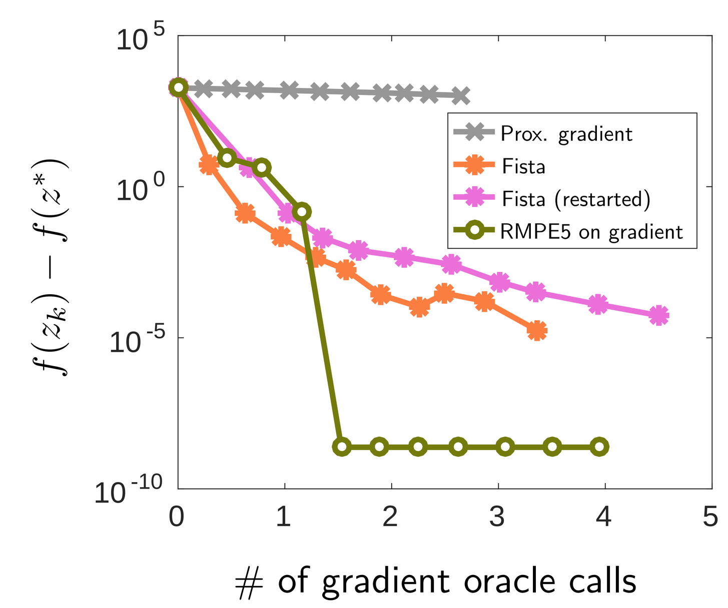
Max Cut (Dual)

$$\min_z \lambda_{\max}\big(\mathrm{Lap}(G) + \mathbf{diag}(z)\big) - \mathbf{1}^T z$$



SVM (Dual)

$$\min_{z \in [0,1]} \frac{1}{2}\|X\,\mathbf{diag}(y) z\|_2^2 - \mathbf{1}^T z$$



Acknowledgements

We received fundings from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under grant agreement n^o 607290 SpaRTaN, as well as support from ERC SIPA and the chaire *Économie des nouvelles données* with the *data science* joint research initiative with the *fonds AXA pour la recherche*.

Main References

- Cabay S. and Jackson LW. A polynomial extrapolation method for finding limits and antilimits of vector sequences.
- Mešina M. Convergence acceleration for the iterative solution of the equations x=ax+f.
- Smith D., Ford W. and Sidi A. Extrapolation methods for vector sequences.
- Tyrtyshnikov E. How bad are Hankel matrices?