Variable metric approaches for large-scale constrained least squares problems

Vanna Lisa Coli[(a)], Valeria Ruggiero[(b)] and Luca Zanni[(a)]

(a) Department of Physics, Computer Science and Mathematics, Modena, Italy
(b) Department of Mathematics and Computer Science, Ferrara, Italy

Statement of the problem

Constrained least squares problem

$$\min_{\boldsymbol{x}\in\Omega} \quad f(\boldsymbol{x}) = \frac{1}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2$$

- ▷ $A\in\mathbb{R}^{m\times n}, \boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{b}\in\mathbb{R}^m$
- ▷ $m \le n$, n very large, $A_{ij}\ge 0$, A not in memory (only matrix-vector operators involving A and A^T available)

▷ $\Omega\subset\mathbb{R}^n$, nonempty closed convex set defined by simple constraints

$\updownarrow$

non-expensive $\mathcal{O}(n)$ projection onto Ω

simple constraints

- ➡ $\boldsymbol{l}\le\boldsymbol{x}\le\boldsymbol{u}$
- ➡ $\boldsymbol{l}\le\boldsymbol{x}\le\boldsymbol{u}\wedge\boldsymbol{w}^T\boldsymbol{x}=c$
- ➡ $\boldsymbol{l}\le\boldsymbol{x}\le\boldsymbol{u}\wedge\boldsymbol{w}^T\boldsymbol{x}\le\rho$

Target imaging applications:

- fiber orientation from diffusion MRI data;
- tomography

Non-scaled state of the art approaches

Gradient Projection (GP)

[Birgin-Martínez-Raydan, *SIAM J.Optim.*, 2000]

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \lambda_k\left(P_\Omega(\boldsymbol{x}^{(k)}-\alpha_k\nabla f(\boldsymbol{x}^{(k)}))-\boldsymbol{x}^{(k)}\right)$$

Step-length selection α_k: recent rules based on Ritz-like values overcome popular Barzilai-Borwein rules [Fletcher, *Math. Program.*, 2012], [Porta-Prato-Zanni, *J. Sci. Comp.*, 2015]

Gradient Projection with extrapolation (GP_Ex)

$$\boldsymbol{x}^{(k+1)} = P_\Omega(\boldsymbol{y}^{(k)}-\alpha\nabla f(\boldsymbol{y}^{(k)})),$$

$$\boldsymbol{y}^{(k)} = \boldsymbol{x}^{(k)} + \beta_k(\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)})$$

Suitable choices for the β_k parameter: [Bertsekas, *Convex Opt.Theory*, 2009], [Beck-Teboulle, *SIAM J. Imaging Sci.*, 2009]

Acceleration strategy

Exploit scaled gradient directions

$$\mathrm{d}_k = -D_k\nabla f(\boldsymbol{x}^{(k)}), \quad D_k\in\mathcal{D}_{\rho_k}$$

where $\mathcal{D}_{\rho_k}$ is the set of symmetric positive definite matrices with eigenvalues $\tau_j^{(k)}$ s. t. $0<\frac{1}{\rho_k}\le\tau_j^{(k)}\le\rho_k$

Diagonal scaling matrix: the updating rule

$$D_k = \mathrm{diag}\left(\max\left\{\frac{1}{\rho_k}, \min\left\{\rho_k, \frac{\boldsymbol{x}^{(k)}}{A^TA\boldsymbol{x}^{(k)}},\right\}\right\}\right)$$

where $\rho_k=\sqrt{1+\frac{\gamma}{k^p}}$ with $\gamma>0$ and $p>2$

Scaled GP (SGP) algorithm

Initialize: $\boldsymbol{x}^{(0)}\in\Omega, \epsilon>0, \delta,\sigma\in(0,1), \alpha_0>0, D_0\in\mathcal{D}_{\rho_0}$

for $k=0,1,\dots,$

$\quad \boldsymbol{d}^{(k)} = P_{\Omega,D_k^{-1}}\left(\boldsymbol{x}^{(k)}-\alpha_kD_k\nabla f(\boldsymbol{x}^{(k)})\right)-\boldsymbol{x}^{(k)}$ (scaled gradient projection step)

$\quad \lambda_k = 1$

$\quad$ while $f(\boldsymbol{x}^{(k)}+\lambda_k\boldsymbol{d}^{(k)}) > f(\boldsymbol{x}^{(k)})+\sigma\lambda_k\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)}$

$\quad\quad \lambda_k=\delta\lambda_k$ (backtracking step)

$\quad$ end

$\quad \boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\lambda_k\boldsymbol{d}^{(k)}$

$\quad$ if $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|\le\epsilon\|\boldsymbol{x}^{(k)}\|$, break; end

$\quad$ define the diagonal scaling matrix $D_{k+1}\in\mathcal{D}_{\rho_{k+1}}$ (scaling updating rule)

$\quad$ define the step-length α_{k+1} (step-length updating rule)

end

Scaled GP_Ex algorithm

Initialize: $\boldsymbol{x}^{(0)}\in\mathbb{R}^n, \epsilon>0, \boldsymbol{y}^{(0)}=\boldsymbol{x}^{(0)}, 0<\delta<1, a>2,$ $\alpha_0>0, D_0\in\mathcal{D}_{\rho_0}$

for $k=0,1,\dots,$

1. $\boldsymbol{x}^{(k+1)}=P_{\Omega,D_k^{-1}}\left(\boldsymbol{y}^{(k)}-\alpha_kD_k\nabla f(\boldsymbol{y}^{(k)})\right)$ (scaled gradient proj. step)

$\quad$ if $f(\boldsymbol{x}^{(k+1)}) > f(\boldsymbol{y}^{(k)})+\nabla f(\boldsymbol{y}^{(k)})^T(\boldsymbol{x}^{(k+1)}-\boldsymbol{y}^{(k)}) + \frac{1}{2\alpha_k}\|\boldsymbol{x}^{(k+1)}-\boldsymbol{y}^{(k)}\|^2_{D_k^{-1}}$

$\quad\quad \alpha_k=\delta\alpha_k$; goto **1.**; (backtracking step)

$\quad$ end

$\quad$ if $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|\le\epsilon\|\boldsymbol{x}^{(k)}\|$, break; end

$\quad \beta_{k+1}=\frac{k}{k+1+a}$; $\quad \alpha_{k+1}=\alpha_k$;

$\quad \boldsymbol{y}^{(k+1)}=\boldsymbol{x}^{(k+1)}+\beta_{k+1}\left(\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\right)$ (extrapolation step)

$\quad$ define the diagonal scaling matrix $D_{k+1}\in\mathcal{D}_{\rho_{k+1}}$ (scaling updating rule)

end

Convergence analysis

SGP method

- ▷ Let f be convex
- ▷ Let the solution set X^* be not empty
- ▷ ∇f globally Lipschitz on Ω
- ▷ $\alpha_k\in[\alpha_{min},\alpha_{max}]$, $\quad 0<\alpha_{min}\le\alpha_{max}$

⇨

$$\boldsymbol{x}^{(k)}\xrightarrow{k\to\infty}\boldsymbol{x}^*, \quad \boldsymbol{x}^*\in X^*$$

$$f^{(k)}-f^*=\mathcal{O}\left(\frac{1}{k}\right)$$

[Bonettini-Prato, *Inverse Problems*, 2015]

Scaled GP_Ex method

- ▷ Let f be convex
- ▷ Let the solution set X^* be not empty
- ▷ ∇f Lipschitz continuous on Ω

⇨

$$\boldsymbol{x}^{(k)}\xrightarrow{k\to\infty}\boldsymbol{x}^*, \quad \boldsymbol{x}^*\in X^*$$

$$f^{(k)}-f^*\le\frac{C}{(k-1+a)^2}$$

[Bonettini-Porta-Ruggiero, *SIAM J. Sci. Comput.*, 2016]

Numerical results

Estimating the fibre orientation from diffusion MRI data

Let V be the number of voxels into a brain region

- Voxel-by-voxel approach: solve V minimization problems

$$\min_{\boldsymbol{x}^{(v)}}\frac{1}{2}\|\Phi_{m\times n}\boldsymbol{x}^{(v)}-\boldsymbol{b}^{(v)}\|_2^2 \quad \text{sub. to} \quad \boldsymbol{x}^{(v)}\ge\boldsymbol{0}, \ \|\boldsymbol{x}^{(v)}\|_0\le k$$

$(m=15,\ n=201,\ k=3,\ V=16\times16\times5)$

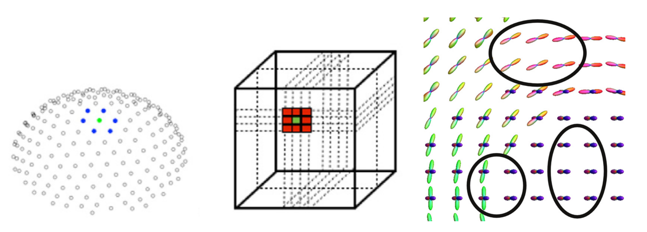
- Whole region approach [Auría-Daducci-Thiran-Wiaux, *NeuroIm.*, 2015]

exploit voxelwise sparsity + spatial coherence in neighbour voxels

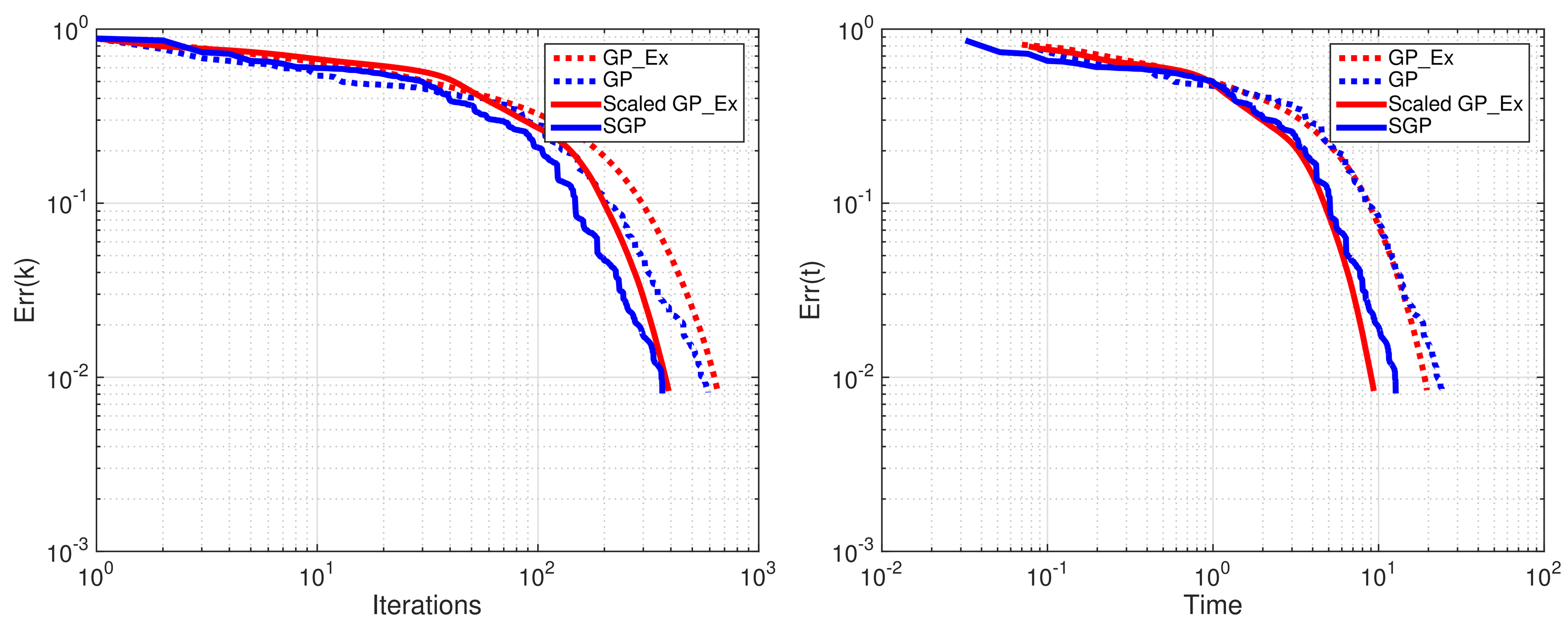
⇨ reweighted $\ell1$-minimization scheme: solve a sequence of pb. of type

$$\min_{\boldsymbol{X}}\frac{1}{2}\|\Phi_{M\times N}\boldsymbol{X}-\boldsymbol{B}\|_2^2 \quad \text{sub. to} \quad \boldsymbol{X}\ge\boldsymbol{0}, \ \|\boldsymbol{X}\|_{W,1}\le K$$

- $M=m\cdot V=19200$
- $N=n\cdot V=257280$
- $K=k\cdot V=3840$
- $\|\boldsymbol{X}\|_{\boldsymbol{W},1}=\sum_{i=1}^N W_i|X_i|, \quad W_i>0$



	GP_Ex	GP	Scaled GP_Ex	SGP
iterations	649	589	390	365
seconds	19.61	24.19	9.38	12.73



Err $:=\left\|\Phi_{M\times N}\boldsymbol{X}-\boldsymbol{B}\right\|_2-\left\|\Phi_{M\times N}\boldsymbol{X}^*-\boldsymbol{B}\right\|_2$, where $\boldsymbol{X}$ is the approximated solution and $\boldsymbol{X}^*$ is a ground–truth, obtained by executing GP_Ex with very high accuracy.

Works in progress

- ➤ application to Computed Tomography data [Jensen-Jørgensen-Hansen-Jensen, *BIT Numer. Math*, 2012]
- ➤ performance comparisons with different scaling rules in literature [Bonettini-Chiuso-Prato, *SIAM J. Sci. Comp.*, 2015], [Hager-Mair-Zhang, *Math. Program.*, 2009]
- ➤ derivation of new scaling strategies (extension to box constraints)

References

[1] Auría A., Daducci A., Thirana J.-P. and Wiaux, Y., *Structured sparsity for spatially coherent fibre orientation estimation in diffusion MRI*, Neuroimage, (2015)

[2] Bonettini S., Porta F. and Ruggiero V., *A variable metric inertial method for convex optimization*, SIAM Journal of Scientific Computing, (2016)

[3] Bonettini S. and Prato M., *New convergence results for the scaled gradient projection method*, Inverse Problems, (2015)

[4] Porta F., Prato M. and Zanni L., *A new steplength selection for scaled gradient methods with application to image deblurring*, Journal of Scientific Computing, (2015)