A Fast algorithm for Sparse Reduced-Rank Regression

École des Ponts ParisTech

{ Benjamin.Dubois, Jean-François.Delmas and Guillaume.Obozinski }@enpc.fr

Problem

Learning a **row-sparse** and **low-rank** matrix for a multitask regression problem.

$$\min_{B\in\mathbb{R}^{p,k}:\,\mathrm{rk}(B)\leq r} \frac{1}{2}\|Y-XB\|_F^2+\lambda\|B\|_{1,2}$$

1. Nondifferentiability due to $\|B\|_{1,2}$
2. Nonconvexity due to $\mathrm{rk}(B)\leq r$

Related Work

1. The analytical solution when $\lambda=0$ requires inverting a large matrix.
2. The biconvex formulation [2] when $\lambda=0$ requires proper initialization.
3. Replacing the rank constraint with the trace-norm $\|B\|_*$ makes each iteration expensive and introduces a bias.

Formulation

Setting $B=UV^T$ with $U\in\mathbb{R}^{p,r}$, $V\in\mathbb{R}^{k,r}$ and $V^TV=I_r$, the problem becomes

$$\min F_\lambda(U) \qquad \text{(SRRR)}$$

where $F_\lambda(U)=f(U)+\lambda\|U\|_{1,2}$ and,

$$f(U):=\frac{1}{2}\|XU\|_F^2-\|Y^TXU\|_*.$$

Studying the geometry of a latent problem lets us show that f is strongly convex in certain directions around the optima and therefore satisfies a Polyak-Łojasiewicz inequality. The latter leads to linear convergence.

The Latent Problem

Let PSQ^T be the SVD of $(X^TX)^{-\frac{1}{2}}X^TY$ and ℓ its rank. Let $P^\perp$ be s.t. $(P,P^\perp)$ is orthonormal. With the invertible change of variables

$$U=(X^TX)^{-\frac{1}{2}}(PSA+P^\perp C), A\in\mathbb{R}^{\ell,r}, C\in\mathbb{R}^{p-\ell,r}$$

we show that $\min f(U)$ is equivalent to

$$\min f_a(A)+\frac{1}{2}\|C\|_F^2$$

$$\text{where } f_a(A):=\frac{1}{2}\|SA\|_F^2-\|S^2A\|_*$$

Proposition 1. *The minima of f_a are*

$$\left\{\tilde{I}R \mid R\in\mathcal{O}_r\right\}$$

where $\tilde{I}=(1_{\{i=j\}})_{1\leq i\leq\ell,\,1\leq j\leq r}$ and $\mathcal{O}_r=\left\{R\in\mathbb{R}^{r,r},\ R^TR=I_r\right\}$.

References

[1] Csiba, D. and Richtárik, P. (2017). Global convergence of arbitrary-block gradient methods for generalized Polyak-Łojasiewicz functions. *arXiv preprint arXiv:1709.03014*.

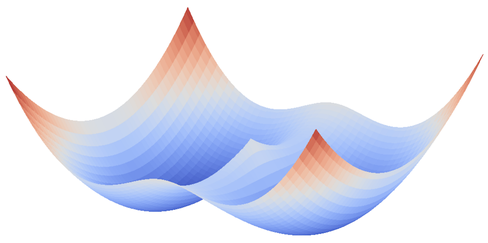
[2] Park, D., Kyrillidis, A., Caramanis, C., and Sanghavi, S. (2016). Finding low-rank solutions via non-convex matrix factorization, efficiently and provably. *arXiv preprint arXiv:1606.03168*.

Geometry : Local strong convexity on cones

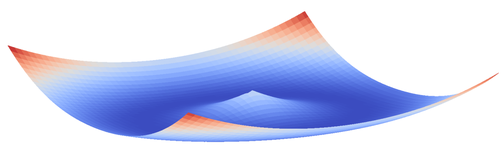
The function f_a is a difference of orthogonal-invariant convex functions. Around the optima, it can be thought of, when $r\geq 2$, as the bottom of a glass bottle with a punt.

Proposition 2. *There exists $\bar{\lambda}>0$ s.t. for any $\lambda<\bar{\lambda}$, there exists $L>\mu>0$ and a sublevel set $\mathcal{V}_\lambda$ of F_λ that can be partitioned into $\{\mathcal{C}_R\cap\mathcal{V}_\lambda\}_{R\in\mathcal{O}_r}$ so that f is L-smooth on $\mathcal{V}_\lambda$ and F_λ is μ-strongly convex on every $\mathcal{C}_R\cap\mathcal{V}_\lambda$, where*

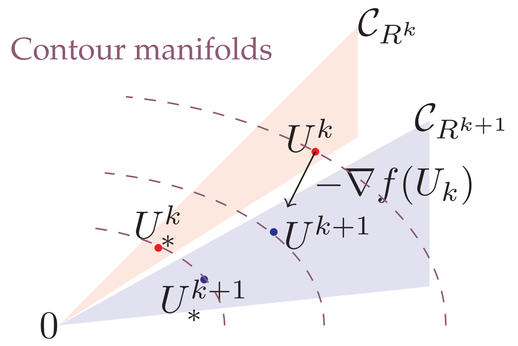
$$\mathcal{C}_R:=\left\{(X^TX)^{-\frac{1}{2}}(PS\begin{pmatrix}A_1\\A_2\end{pmatrix}+P^\perp C)R,\ (A_1,A_2,C)\in\mathbb{R}^{r+(\ell-r)+(p-\ell),r},\ A_1\succ 0\right\}$$



Possible landscape for f in the special case $r=1,p=2$



Representation of f around one of the two connected components of the optima when $r\geq 2$



Representation of one iteration when $\lambda=0$

Proposition 2, which is proved by computing the Hessian of f_A, implies that the restriction of F_λ on a cone satisfies a Polyak-Łojasiewicz inequality. Given the orthogonal invariance, this implies that F_λ satisfies such an inequality in the whole neighborhood.

PL Inequalities

The Polyak-Łojasiewicz inequality generalizes the fact for a μ-strongly convex and differentiable function,

$$F(U)-F^*\leq\frac{1}{2\mu}\|\nabla F(U)\|_F^2, \qquad \text{(PL)}$$

For a non-smooth function $F=f+g$, f being L-smooth, the proximal variant is $F(U)-F^*\leq\lambda(U)$ where

$$\lambda(U):=-L\min_{U'}[\langle\nabla f(U),U'-U\rangle+\frac{L}{2}\|U'-U\|_F^2+g(U')]$$

In a unifying framework, [1] show how the PL inequalities lead to linear convergence for first-order (randomized block) descent methods.

Convergence

We apply Moreau-Yosida regularization to the nondifferentiable trace-norm and consider a **proximal gradient algorithm involving a backtracking line search** with parameter β. Let $(U_k)_{k\geq 0}$ be a generated sequence.

Global convergence to a critical point

Proposition 3. *The sequence $(U^k)_{k\geq 0}$ has finite length $\sum_k\|U^{k+1}-U^k\|_F<\infty$, and F-converges to a critical point $\tilde{U}\in\mathbb{R}^{p,r}$ such that $0\in\partial F(\tilde{U})$,*

$$\lim_{k\to\infty}F(U^k)=F(\tilde{U}).$$

Asymptotic linear convergence

Extending the results in [1] we obtain

Proposition 4. *Let U^{k+1} be generated from $U^k\in\mathcal{V}_\lambda$. Then $U^{k+1}\in\mathcal{V}_\lambda$ and denoting $\rho=\min(\frac{1}{2},\beta\frac{\mu}{L})$, we have*

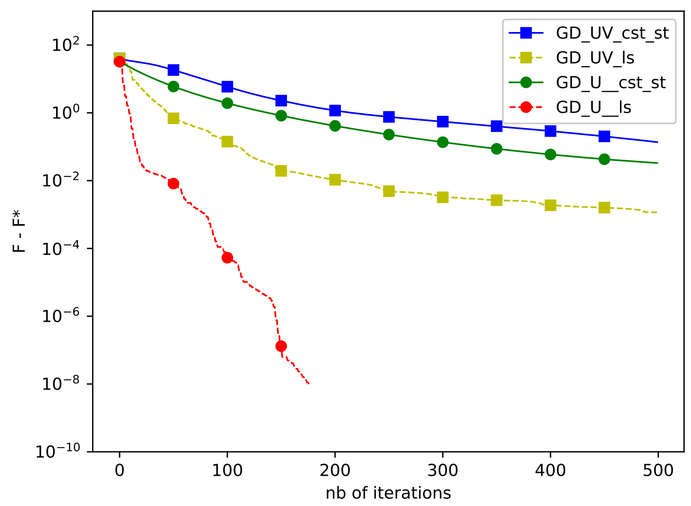
$$F(U^{k+1})-F^*\leq(1-\rho)\,(F(U^k)-F^*).$$

Experiments

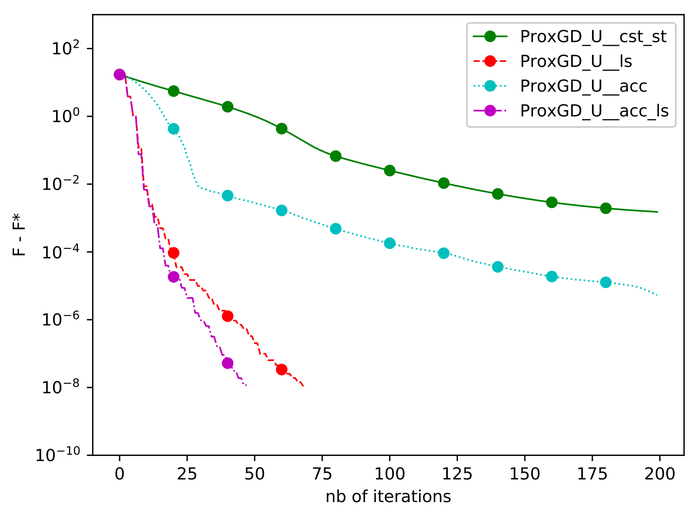
When $\lambda=0$, we compare our formulation SRRR (GD_U) with the biconvex one [2] (GD_UV)

$$\min_{U,V}\frac{1}{2}\|Y-XUV^T\|_F^2+\frac{\mu}{4}\|U^TU-V^TV\|_F^2$$

and consider different variants : constant step (cst_st), line search (ls) and accelerated versions (acc, acc_ls).



Case $\lambda=0$.



Case $\lambda=0.1$.

The line search and the acceleration significantly increase the speed of convergence.