Stochastic FISTA algorithms: so fast ?

G. Fort[1], L. Risser[1], Y. Atchadé[2], E. Moulines[3]

[1]IMT & CNRS, Univ. de Toulouse, France (work partially supported by ANR-11-LABX-0040-CIMI)
[2]Univ. Michigan, USA [3]Ecole Polytechnique, France

Problem

Solve

$$\text{argmin}_{\theta\in\Theta}\left(f(\theta)+g(\theta)\right)$$

where **A1)** $g:\mathbb{R}^d\to[0,+\infty]$ is convex, not identically $+\infty$ and lower semi-continuous; $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is continuously differentiable on $\Theta:=\{\theta\in\mathbb{R}^d: g(\theta)+|f(\theta)|<\infty\}$ and its gradient is L-lipschitz on Θ.
The function f and its gradient are numerically intractable
A2) for any $\theta\in\mathbb{R}^d$,

$$\nabla f(\theta)=\int_{\mathsf{X}} H(\theta,x)\,\pi_\theta(\mathrm{d}x)$$

where X is a topological space endowed with its Borel σ-field, π_θ is a probability measure on X and $H:\mathbb{R}^d\times\mathsf{X}\to\mathbb{R}^d$ is measurable. In addition, $x\mapsto H(\theta,x)$ is π_θ-integrable for any $\theta\in\mathbb{R}^d$.

Perturbed FISTA

Choose positive sequences: $\{\gamma_n\}_{n\geq1}$, $\{t_n\}_{n\geq1}$ s.t.

$$t_n\geq1,\qquad \tau_n:=\gamma_n t_{n-1}^2-\gamma_{n+1}t_n(t_n-1)\geq0.$$

Define iteratively, given θ_0 and $t_0=1$

$$\vartheta_n=\theta_n+\frac{t_{n-1}-1}{t_n}\left(\theta_n-\theta_{n-1}\right),$$
$$\theta_{n+1}=\text{Prox}_{\gamma_{n+1},g}\left(\vartheta_n-\gamma_{n+1}\left(\nabla f(\vartheta_n)+\eta_{n+1}\right)\right),$$

where η_{n+1} is a perturbation (since non explicit gradient) and

$$\text{Prox}_{\gamma,g}(u):=\text{argmin}_{\theta\in\Theta}\left(g(\theta)+\frac{1}{2\gamma}(\theta-u)^2\right)$$

Penalized ML with intractable likelihood

▷ ***Latent Variable Model.*** (the dependence upon the observations is omitted)

$$f(\theta)=-\ell_N(\theta)=-\frac{1}{N}\log\int_{\mathsf{X}}p(x,\theta)\;\mu(\mathrm{d}x)$$

where for any θ, $p(\cdot,\theta)$ is the likelihood of the complete data model; x collects the missing variables.
The gradient of f is an expectation w.r.t. the a posteriori distribution - known up to a constant

$$\nabla f(\theta)=-\frac{1}{N}\int_{\mathsf{X}}\partial_\theta\log p(x,\theta)\;\pi_\theta(\mathrm{d}x)\qquad\text{with}\qquad\pi_\theta(\mathrm{d}x):=\frac{p(x,\theta)\mu(\mathrm{d}x)}{\int p(u,\theta)\mu(\mathrm{d}u)}.$$

▷ ***Binary Graphical Model.*** N i.i.d. observations $Y^{(n)}$ on $\{0,1\}^p$ from the distribution

$$\pi_\theta(\mathrm{d}y_{1:p})=\frac{1}{Z_\theta}\exp\left(\sum_{i=1}^p\theta_i y_i+\sum_{1\leq i<j\leq p}\theta_{ij}1_{y_i=y_j}\right)$$

with an untractable normalizing constant Z_θ. The normalized negative log-likelihood f is intractable and its gradient too

$$\partial_{\theta_i}f(\theta)=\frac{1}{N}\sum_{n=1}^N\left(Y_i^{(n)}-\int_{\mathsf{X}}x_i\pi_\theta(\mathrm{d}x)\right),\qquad\partial_{\theta_{ij}}f(\theta)=\frac{1}{N}\sum_{n=1}^N\left(1_{Y_i^{(n)}=Y_j^{(n)}}-\int_{\mathsf{X}}1_{x_i=x_j}\pi_\theta(\mathrm{d}x)\right).$$

The gradient of f is an expectation w.r.t. π_θ, intractable (sum over 2^p terms)

Monte Carlo perturbation

Approximation of the gradient by a **biased** Monte Carlo sum: given a Markov chain $\{X_{\ell,n+1}\}_{\ell\geq0}$, from a Markov Chain Monte Carlo sampler with invariant distribution $\pi_{\vartheta_n}(\mathrm{d}x)$, set

$$H_{n+1}:=\frac{1}{m_{n+1}}\sum_{\ell=1}^{m_{n+1}}H(X_{\ell,n+1},\vartheta_n)\qquad\Longrightarrow\qquad\eta_{n+1}=H_{n+1}-\nabla f(\vartheta_n),\ \text{and}\ \mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right]\neq0$$

Convergence results assuming, almost-surely, for some $p>1$

$$\sup_n m_{n+1}^2\mathbb{E}\left[\|\mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right]\|^2\right]<\infty,\qquad m_{n+1}\mathbb{E}\left[\|\eta_{n+1}-\mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right]\|^2|\mathcal{F}_n\right]\leq C_n\qquad\sup_n\mathbb{E}[C_n^p]<\infty.$$

Convergence Results

A3) The function f is convex and the set $\mathcal{L}:=\text{argmin}_\theta(f+g)$ is non empty. The following sum exists for some $\theta_\star\in\mathcal{L}$

$$\sum_n\gamma_{n+1}t_n\left\langle z_n-\theta_\star;\eta_{n+1}\right\rangle,\qquad z_n:=\theta_n+t_n\left(\text{Prox}_{\gamma_{n+1},g}(\vartheta_n-\gamma_{n+1}\nabla f(\vartheta_n))-\theta_n\right)$$

Theorem Under A1)-A3), then $\lim_n\gamma_{n+1}t_n^2F(\theta_{n+1})$ exists, and $F:=f+g$

$$\gamma_{n+1}t_n^2F(\theta_{n+1})+\frac{1}{2}\|z_n-\theta_\star\|^2+\sum_{k=1}^n\tau_kF(\theta_k)\leq\gamma_1F(\theta_1)+\frac{1}{2}\|\theta_1-\theta_\star\|^2+\sum_{k=1}^n\gamma_{k+1}^2t_k^2\|\eta_{k+1}\|^2-\sum_{k=1}^n\gamma_{k+1}t_k\left\langle z_k-\theta_\star;\eta_{k+1}\right\rangle$$

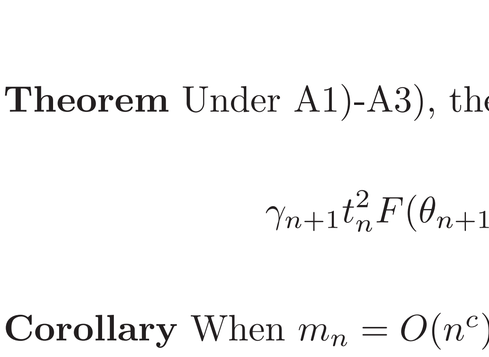
Corollary When $m_n=O(n^c)$, $\gamma_n=O(n^{-a})$ and $t_n=O(n)$: choose $a\in[1,2[$ and $c=2-a$ OR $a\in[0,1]$ and $c=3-2a$. Then

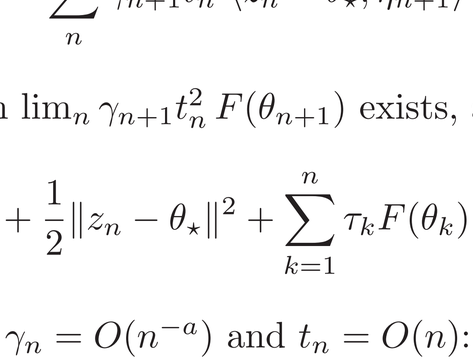
$$\lim_n n^{2-a}F(\theta_n)\ \text{exists a.s.}\qquad\sup_n n^{2-a}\mathbb{E}\left[F(\theta_n)\right]<\infty\qquad\sup_n\sum_{k=1}^n kF(\theta_k)<\infty\ \text{a.s.}\qquad\sup_n\sum_{k=1}^n k\;\mathbb{E}\left[F(\theta_k)\right]<\infty.$$

Rate for Monte Carlo-FISTA $O(n^{-2})$ with $a=0$ and $c=3$. For a precision δ: $O(\delta^{-1/2})$ iterations and thus a MC computational cost δ^{-2}.
Rate for Monte Carlo-ISTA $O(n^{-1})$ for the averaged estimator $n^{-1}\sum_{k=1}^n\theta_k$, with $a=0$ and $c=1$. For a precision δ: MC comput. cost δ^{-2}.

Numerical Illustration: binary graphical model

(left) # of non zero components when $n\in\{50,100,1000,1500,2000\}$ (right) Probability to be non null, at convergence.





$$\nabla f(\tau)=\phi(\tau)+\left\langle\int S(x)\pi_\tau(\mathrm{d}x);\psi(\theta)\right\rangle$$

$$g(\theta)=\lambda\sum_{1\leq i<j\leq p}|\theta_{ij}|+\mu\sum_{i=1}^p\theta_i^2$$

Alg.1: (P-ISTA) $t_n=1$, $m_n=O(\sqrt{n})$, $\gamma_n=O(1/\sqrt{n})$

Alg.2 to Alg.4: $\gamma_n=O(1)$, $m_n=O(n^3)$, $t_n=n^b$ with $b\in\{1,0.5,\epsilon\}$.

Alg.5 $S_{n+1}=\gamma_{n+1}m_{n+1}^{-1}\sum_{\ell=1}^{m_{n+1}}S(X_{\ell,n+1})+(1-\delta_{n+1})S_n$

References

- Y. Atchadé, G. Fort, E. Moulines. *On perturbed proximal gradient algorithms* J. Mach. Learn. Res. Vol. 18, 2017.
- G. Fort, E. Ollier and A. Samson *Stochastic Proximal Gradient for Penalized Mixed Models.* Statistics and Computing, 2018.