Stochastic Proximal Gradient Algorithm

Eric Moulines

Institut Mines-Télécom / Telecom ParisTech / Laboratoire Traitement et Communication de l'Information

Joint work with: Y. Atchade, *Ann Arbor, USA*,
G. Fort *LTCI/Télécom Paristech*
and the kind help of A. Juditsky *LJK, Grenoble*

Problem Statement

$$\textbf{(P)} \quad \mathrm{Argmin}_{\theta \in \Theta} \left\{ -\ell(\theta) + g(\theta) \right\} ,$$

where ℓ is a smooth log-likelihood function or some other smooth statistical learning function, and g is a possibly non-smooth convex penalty term.

This problem has attracted a lot of attention with the growing need to address high-dimensional statistical problems

This work focuses on the case where the function ℓ and its gradient $\nabla \ell$ are both intractable, and where $\nabla \ell$ is given by

$$\nabla \ell(\theta) = \int H_\theta(x) \pi_\theta(\mathrm{d}x),$$

for some probability measure π_θ.

Network structure

- Problem: Estimate sparse network structures from measurements on the nodes.
- For discrete measurement: amounts to estimate a Gibbs measure with pair-wise interactions

$$f_\theta(x_1, \ldots, x_p) = \frac{1}{Z_\theta} \exp\left\{ \sum_{i=1}^{p} \theta_{ii} B_0(x_i) + \sum_{1 \le j < i \le p} \theta_{ij} B(x_i, x_j) \right\},$$

for a function $B_0 : \mathsf{X} \to \mathbb{R}$, and a symmetric function $B : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$, where X is a finite set.

- The absence of an edge encodes conditional independence.

Network structure

- Each graph represents a model class of graphical models; learning a graph then is a model class selection problem.
- Constraint-based approaches: test conditional independence from the data and then determine a graph that most closely represents those independencies.
- Score-based approaches combine a metric for the complexity of the graph with a measure of the goodness of fit of the graph to the data... but the number of graph structures grows super-exponentially, and the problem is in general NP-hard.

Network structure

- For $x \in \mathsf{X}^p$, define $\bar{B}(x) \stackrel{\text{def}}{=} (B_{jk}(x_j, x_k))_{1 \le j,k \le p} \in \mathbb{R}^{p \times p}$.
- The ℓ^1-penalized maximum likelihood estimate of θ is obtained by solving an optimization problem of the form (P) where ℓ and g are given by

$$\ell(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left\langle \theta, \bar{B}(x^{(i)}) \right\rangle - \log Z_\theta, \quad g(\theta) = \lambda \sum_{1 \le k < j \le p} |\theta_{jk}| \, .$$

Fisher identity

- Fact 1: Z_θ is the normalization constant is given by

$$Z_\theta = \sum_x \exp(\langle \theta, \bar{B}(x) \rangle$$

where the sum is over all the possible configurations.

- Fact 2: the gradient $\nabla \log Z_\theta$ is the expectation of the sufficient statistics:

$$\nabla \log Z_\theta = \sum_x \bar{B}(x) f_\theta(x)$$

- Problem: None of these quantities can be computed explicitly... Nevertheless, they can be estimated using Monte Carlo integration.

General framework

$$\text{Argmin}_{\theta \in \Theta} \left\{ -\ell(\theta) + g(\theta) \right\},$$

where

- ℓ is a smooth log-likelihood function or some other smooth statistical learning function,
- g is a non-smooth convex sparsity-inducing penalty.

re

- the function ℓ and its gradient $\nabla \ell$ are intractable,
- The score function $\nabla \ell$ is given by

$$\nabla \ell(\theta) = \int H_\theta(x) \pi_\theta(\mathrm{d}x),$$

for some probability measure π_θ on some measurable space $(\mathsf{X}, \mathcal{B})$, and some function $H_\theta : \mathsf{X} \to \Theta$.

Definition

- Definition: Proximal mapping associated with closed convex function g and stepsize γ

$$\text{prox}_\gamma(\theta) = \text{Argmin}_{\vartheta \in \Theta} \left(g(\vartheta) + (2\gamma)^{-1} \|\vartheta - \theta\|_2^2 \right)$$

- If $g = \mathbb{I}_{\mathcal{K}}$, where $\mathcal{K}$ is a closed convex set ($\mathbb{I}_{\mathcal{K}}(x) = 0$, $x \in \mathcal{K}$, $\mathbb{I}_{\mathcal{K}}(x) = \infty$ otherwise), then prox_γ is the Euclidean projection on $\mathcal{K}$

$$\text{prox}_\gamma(\theta) = \text{Argmin}_{\vartheta \in \mathcal{K}} \|\vartheta - \theta\|_2^2 = P_{\mathcal{K}}(\theta)$$

- if $g(\theta) = \sum_{i=1}^p \lambda_i |\theta_i|$ then prox_g is shrinkage (soft threshold) operation

$$[S_{\lambda,\gamma}(\theta)]_i = \begin{cases} \theta_i - \gamma\lambda_i & \theta_i \geq \gamma\lambda_i \\ 0 & |\theta_i| \leq \gamma\lambda_i \\ \theta_i + \gamma\lambda_i & \theta_i \leq -\gamma\lambda_i \end{cases}$$

Proximal gradient method

Unconstrained problem with cost function split in two components

$$\text{Minimize } f(\theta) = -\ell(\theta) + g(\theta)$$

- $-\ell$ convex, differentiable with $\mathrm{dom}(g) = \mathbb{R}^n$
- g closed, convex, possibly non differentiable... but prox_g is inexpensive !

Proximal gradient algorithm

$$\theta^{(k)} = \mathrm{prox}_{\gamma_k g}(\theta^{(k-1)} + \gamma_k \nabla \ell(\theta^{(k-1)}))$$

where $\{\gamma_k,\ k \in \mathbb{N}\}$ is a sequence stepsizes, which either be constant, decreasing or determined by line search

Interpretation

- Denote

$$\theta^+ = \text{prox}_\gamma(\theta + \gamma \nabla \ell(\theta))$$

- from definition of proximal operator:

$$\begin{aligned}\theta^+ &= \text{Argmin}_\vartheta (g(\vartheta) + (2\gamma)^{-1} \|\vartheta - \theta - \gamma \nabla \ell(\theta)\|_2^2) \\ &= \text{Argmin}_\vartheta (g(\vartheta) - \ell(\theta) - \nabla \ell(\theta)^T (\vartheta - \theta) + (2\gamma)^{-1} \|\vartheta - \theta\|_2^2) \ .\end{aligned}$$

- θ^+ minimizes $g(\vartheta)$ plus a simple quadratic local model of $-\ell(\vartheta)$ around θ
- If $\gamma \leq 1/L$, the surrogate function on the RHS majorizes the target function, and the algorithm might be seen as a specific instance of the Majorization-Minimization algorithm.

Some specific examples

- if $g(\theta) = 0$ then proximal gradient $=$ gradient method.

$$\theta^{(k)} = \theta^{(k-1)} + \gamma_k \nabla \ell(\theta^{(k-1)})$$

- if $g(\theta) = I_{\mathcal{K}}(\theta)$, then proximal gradient $=$ projected gradient

$$\theta^{(k)} = P_{\mathcal{K}}(\theta^{(k-1)} + \gamma_k \nabla \ell(\theta^{(k-1)})) \; .$$

- if $g(\theta) = \sum_i \lambda_i |\theta_i|$ then proximal gradient $=$ soft-thresholded gradient

$$\theta^{(k)} = S_{\lambda,\gamma_k}(\theta^{(k-1)} + \gamma_k \nabla \ell(\theta^{(k-1)}))$$

Gradient map

The proximal gradient may be equivalently rewritten as

$$\theta^{(k)} = \theta^{(k-1)} - \gamma_k G_{\gamma_k}(\theta^{(k-1)})$$

where the function G_γ is given by

$$G_\gamma(\theta) = \frac{1}{\gamma}(\theta - \mathsf{prox}_\gamma(\theta + \gamma \nabla \ell(\theta)))$$

The subgradient characterization of the proximal map implies

$$G_\gamma(\theta) \in -\nabla \ell(\theta) + \partial g(\theta - \gamma G_\gamma(\theta))$$

Therefore, $G_\gamma(\theta) = 0$ if and only if θ minimizes $f(\theta) = -\ell(\theta) + g(\theta)$

Convergence of the proximal gradient

Assumptions: $f(\theta) = -\ell(\theta) + g(\theta)$

- $\nabla \ell$ is Lipschitz continuous with constant $L > 0$

$$\|\nabla \ell(\theta) - \nabla \ell(\vartheta)\|_2 \leq L\|\theta - \vartheta\|_2 \ \forall \theta, \vartheta \in \Theta$$

- optimal value $f^\star$ is finite and attained at $\theta^\star$ (not necessarily unique)

Theorem

$f(\theta^{(k)}) - f^\star$ decreases at least as fast as $1/k$

- *if fixed step size $\gamma_k \leq 1/L$ is used*
- *if backtracking line search is used*

Back to the original problem !

- The score function $\nabla \ell$ is given by

$$\nabla \ell(\theta) = \int H_\theta(x) \pi_\theta(\mathrm{d}x) \ .$$

  Therefore, at each iteration, the score function should be approximated.
- The case where $\pi_\theta = \pi$ and and random variables $\{X_n,\ n \in \mathbb{N}\}$ each marginally distributed according to $\pi =$ online learning (Juditsky, Nemirovski, 2010, Duchi et al, 2011).

Back to the original problems !

$$\nabla \ell(\theta) = \int H_\theta(x) \pi_\theta(\mathrm{d}x) \ .$$

- π_θ depends on the unknown parameter θ...
- Sampling directly from π_θ is often not directly feasible. But one may construct a Markov chain, with Markov kernel P_θ, such that $\pi_\theta P_\theta = \pi_\theta$
- The Metropolis-Hastings algorithm or Gibbs sampling provides a natural framework to handle such problems.

Stochastic Approximation / Mini-batches $g \equiv 0$

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H_{n+1}$$

where H_{n+1} approximates $\nabla \ell(\theta_n)$.

- Stochastic Approximation: $\gamma_n \downarrow 0$ and $H_{n+1} = H_{\theta_n}(X_{n+1})$ and $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$.
- Mini-batches setting: $\gamma_n \equiv \gamma$ and

$$H_{n+1} = m_{n+1}^{-1} \sum_{j=0}^{m_{n+1}-1} H(\theta_n, X_{n+1,j}) \, ,$$

where $m_n \uparrow \infty$ and $\{X_{n+1,j}\}_{j=1}^{m_{n+1}}$ is a run of the length m_{n+1} of a Markov chain with transition kernel P_{θ_n}.

- Beware ! For SA, n iterations $= n$ simulations. For minibatches, n iterations $= \sum_{j=1}^{n} m_j$ simulations.

Averaging

$$\bar{\theta}_n \stackrel{\text{def}}{=} \frac{\sum_{k=1}^n a_k \theta_k}{\sum_{k=1}^n a_k} = \left(1 - \frac{a_n}{\sum_{k=1}^n a_k}\right) \bar{\theta}_{n-1} + \frac{a_n}{\sum_{k=1}^n a_k} \theta_n \ .$$

- Stochastic approximation: take $a_n \equiv 1$, $\gamma_n = C n^{-\alpha}$ with $\alpha \in (1/2, 1)$, then

$$\sqrt{n} \left(\bar{\theta}_n - \theta_*\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \sigma^2)$$

- Mini-batch SA: take $a_n \equiv m_n$, $\gamma_n \equiv \gamma \leq 1/(2L)$ and $m_n \to \infty$ sufficiently fast, then

$$\sqrt{n} \left(\bar{\theta}_{N_n} - \theta_*\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \sigma^2)$$

where N_n is the number of iterations for n simulations: $\sum_{k=1}^{N_n} m_k \leq n < \sum_{k=1}^{N_n+1} m_k$.

Stochastic Approximation

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\nabla\ell(\theta_n) + \gamma_{n+1}\eta_{n+1}$$
$$\eta_{n+1} = H_{\theta_n}(X_{n+1}) - \nabla\ell(\theta_n) = H_{\theta_n}(X_{n+1}) - \pi_{\theta_n}(H_{\theta_n}) \ .$$

- Idea Split the error into a martingale increment + remainder term
- Key tool Poisson equation

$$\hat{H}_\theta - P_\theta\hat{H}_\theta = H_\theta - \pi_\theta(H_\theta) \ .$$

Decomposition of the error

$$\begin{aligned}\eta_{n+1} &= H_{\theta_n}(X_{n+1}) - \pi_{\theta_n}(H_{\theta_n}) \\ &= \hat{H}_{\theta_n}(X_{n+1}) - P_{\theta_n}\hat{H}_{\theta_n}(X_{n+1}) \\ &= \hat{H}_{\theta_n}(X_{n+1}) - P_{\theta_n}\hat{H}_{\theta_n}(X_n) + P_{\theta_n}\hat{H}_{\theta_n}(X_{n+1}) - P_{\theta_n}\hat{H}_{\theta_n}(X_n)\end{aligned}$$

We further split the error

$$\begin{aligned}&P_{\theta_n}\hat{H}_{\theta_n}(X_{n+1}) - P_{\theta_n}\hat{H}_{\theta_n}(X_n) \\ &= P_{\theta_{n+1}}\hat{H}_{\theta_{n+1}}(X_{n+1}) - P_{\theta_n}\hat{H}_{\theta_n}(X_n) + P_{\theta_n}\hat{H}_{\theta_n}(X_{n+1}) - P_{\theta_{n+1}}\hat{H}_{\theta_{n+1}}(X_{n+1}) \, .\end{aligned}$$

To prove that the remainder term goes to zero, it is required to prove the regularity of the Poisson solution with respect to θ, to prove that $\theta \mapsto \hat{H}_\theta$ and $\theta \mapsto P_\theta \hat{H}_\theta$ is smooth in some sense.... this is not always a trivial issue !

Minibatch case

Assume that the Markov kernel is nice ...

Bias

$$\mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right] = m_{n+1}^{-1} \sum_{j=0}^{m_{n+1}-1} \left(\nu_{\theta_n} P_{\theta_n}^j H_{\theta_n} - \pi_{\theta_n} H_{\theta_n}\right) = O(m_{n+1}^{-1})$$

Fluctuation

$$m_{n+1}^{-1} \sum_{j=0}^{m_{n+1}-1} H_{\theta_n}(X_j) - \pi_{\theta_n}(H_{\theta_n})$$

$$= m_{n+1}^{-1} \sum_{j=1}^{m_{n+1}-1} \hat{H}_{\theta_n}(X_j) - P_{\theta_n}\hat{H}_{\theta_n}(X_{j-1}) + \text{remainders}$$

Minibatches case

- Contrary to SA, the noise η_{n+1} in the recursion

$$\theta_{n+1} = \theta_n + \gamma \nabla \ell(\theta_n) + \gamma \eta_{n+1}$$

converges to zero a.s. and the stepsize is kept constant $\gamma_n = \gamma$.

- Idea: perturbation of a discrete time dynamic system

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k + \gamma \nabla \ell(\tilde{\theta}_k)$$

having a unique fixed point and a Lyapunov function ℓ: $\ell(\theta_{k+1}) \geq \ell(\theta_k)$ in presence of vanishingly small perturbation η_{n+1}.

- a.s convergence of perturbed dynamical system with a Lyapunov function can be established under very weak assumptions...

Stochastic proximal gradient

The stochastic proximal gradient sequence $\{\theta_n,\ n \in \mathbb{N}\}$ can be rewritten as

$$\theta_{n+1} = \text{prox}_{\gamma_{n+1}}(\theta_n + \gamma_{n+1}\nabla\ell(\theta_n) + \gamma_{n+1}\eta_{n+1}) \ ,$$

where $\eta_{n+1} \stackrel{\text{def}}{=} H_{n+1} - \nabla\ell(\theta_n)$ is the *approximation error*.

Questions:

- Convergence and rate of convergence in the SA and mini-batch settings ?
- Stochastic Approximation / Minibatch: which one should I prefer ?
- Tuning of the parameters (stepsize for SA, size of minibatches, averaging weights, etc...)
- Acceleration (*à la* Nesterov) ?

Main result

Lemma

Suppose that $\{\gamma_n,\ n \in \mathbb{N}\}$ *is decreasing and* $0 < L\gamma_n \leq 1$ *for all* $n \geq 1$. *For any* $\theta_* \in \Theta$, *and any nonnegative sequence* $\{a_n,\ n \in \mathbb{N}\}$,

$$\left(\sum_{j=1}^{n} a_j\right)\{f(\bar{\theta}_n) - f(\theta_*)\}$$

$$\leq \frac{1}{2}\sum_{j=1}^{n}\left(\frac{a_j}{\gamma_j} - \frac{a_{j-1}}{\gamma_{j-1}}\right)\|\theta_{j-1} - \theta_\star\|^2 + \frac{a_0}{2\gamma_0}\|\theta_0 - \theta_*\|^2$$

$$+ \sum_{j=1}^{n} a_j \left\langle T_{\gamma_j}(\theta_{j-1}) - \theta_\star, \eta_j \right\rangle + \sum_{j=1}^{n} a_j \gamma_j \|\eta_j\|^2 ,$$

where $T_\gamma(\theta) \stackrel{\text{def}}{=} \mathrm{prox}_\gamma\left(\theta + \gamma\nabla\ell(\theta)\right)$ *is the gradient proximal map,*

Stochastic Approximation setting

Take $a_j = \gamma_j$ and decompose, using the Poisson equation, $\eta_j = \xi_j + r_j$, where ξ_j is a martingale term and r_j is a remainder term.

$$\left(\sum_{j=1}^{n} \gamma_j\right) \{f(\bar{\theta}_n) - f(\theta_*)\} \leq \frac{a_0}{2\gamma_0} \|\theta_0 - \theta_*\|^2$$

$$+ \sum_{j=1}^{n} \gamma_j \left\langle T_{\gamma_j}(\theta_{j-1}) - \theta_\star, \xi_j \right\rangle + \sum_{j=1}^{n} \gamma_j^2 \|\eta_j\|^2 + \text{remainders} \,,$$

The red term is a martingale with a bracket bounded by

$$\sum_{j=1}^{n} \gamma_j^2 \|\theta_{j-1} - \theta_\star\|^2 \mathbb{E}\left[\|\xi_j\|^2 \,\middle|\, \mathcal{F}_{j-1}\right]$$

If $\sum_{j=1}^{\infty} \gamma_j = \infty$ and $\sum_{j=1}^{\infty} \gamma_j^2 < \infty$, $\{\bar{\theta}_n,\ n \in \mathbb{N}\}$ converges. rate of convergence $\ln(n)/\sqrt{n}$ by taking $\gamma_j = j^{-1/2}$.

Minibatch setting

Theorem

Let $\{\bar{\theta}_n,\, n \geq 0\}$ *be the average estimator. Then for all* $n \geq 1$,

$$\left(\sum_{j=1}^{n} a_j\right) \mathbb{E}\left[f(\bar{\theta}_n) - f(\theta_*)\right]$$

$$\leq \frac{1}{2}\sum_{j=1}^{n}\left(\frac{a_j}{\gamma_j} - \frac{a_{j-1}}{\gamma_{j-1}}\right)\mathbb{E}\left[\|\theta_{j-1} - \theta_\star\|^2\right] + \frac{a_0}{2\gamma_0}\mathbb{E}[\|\theta_0 - \theta_*\|^2]$$

$$+ \sum_{j=1}^{n} a_j \mathbb{E}\left[\|\theta_{j-1} - \theta_\star\|\, \epsilon_{j-1}^{(1)}\right] + \sum_{j=1}^{n} a_j \gamma_j \mathbb{E}\left[\epsilon_{j-1}^{(2)}\right] \,.$$

where

$$\epsilon_n^{(1)} \stackrel{\text{def}}{=} \|\mathbb{E}\left[\eta_{n+1} \,|\, \mathcal{F}_n\right]\|\,, \quad \epsilon_n^{(2)} \stackrel{\text{def}}{=} \mathbb{E}\left[\|\eta_{n+1}\|^2 \,\middle|\, \mathcal{F}_n\right] \,.$$

Convergence analysis

Corollary

Suppose that $\gamma_n \in (0, 1/L]$, *and there exist constants* $C_1, C_2, B < \infty$ *such that, for* $n \geq 1$,

$$\mathbb{E}[\epsilon_n^{(1)}] \leq C_1 m_{n+1}^{-1},\ \mathbb{E}[\epsilon_n^{(2)}] \leq C_2 m_{n+1}^{-1}, \quad \textit{and} \quad \sup_{n \in \mathbb{N}} \|\theta_n - \theta_\star\| \leq B\ , \mathbb{P}-\textit{a.s.}$$

Then, setting $\gamma_n = \gamma$, $m_n = n$ *and* $a_n \equiv 1$,

$$\mathbb{E}\left[f(\theta_n) - f(\theta_\star)\right] \leq C/n \quad \textit{and} \quad \mathbb{E}\left[f(\theta_{N_n}) - f(\theta_\star)\right] \leq C/\sqrt{n}$$

where N_n *is the number of iterations for* n *simulations. This is the same rate than for the SA (without the logarithmic term).*

Potts model

We focus on the particular case where $\mathsf{X} = \{1, \ldots, M\}$, and $B(x, y) = \mathbb{1}_{\{x=y\}}$, which corresponds to the well known Potts model

$$f_\theta(x_1, \ldots, x_p) = \frac{1}{Z_\theta} \exp\left\{ \sum_{i=1}^{p} \theta_{ii} B_0(x_i) + \sum_{1 \leq j < i \leq p} \theta_{ij} \mathbb{1}_{\{x_i = x_j\}} \right\}.$$

- The term $\sum_{i=1}^{p} \theta_{ii} B_0(x_i)$ is sometimes referred to as the external field and defines the distribution in the absence of interaction.
- We focus on the case where the interactions terms θ_{ij} for $i \neq j$ are nonnegative. This corresponds to networks with there is either no interaction, or collaborative interactions between the nodes.

Algorithm

At the k-th iteration, and given $\mathcal{F}_k = \sigma(\theta_1, \ldots, \theta_k)$:

1. generate the X^p-valued Markov sequence $\{X_{k+1,j}\}_{j=0}^{m_{k+1}}$ with transition P_{θ_k} and initial distribution ν_{θ_k}, and compute the approximate gradient

$$H_{k+1} = \frac{1}{n} \sum_{i=1}^{n} \bar{B}(x^{(i)}) - \frac{1}{m_{k+1}} \sum_{j=1}^{m_{k+1}} \bar{B}(X_{k+1,j}) \ ,$$

2. Compute

$$\theta_{k+1} = \Pi_{\mathcal{K}_a} \left(s_{\gamma_{k+1},\lambda} \left(\theta_k + \gamma_{k+1} H_{k+1} \right) \right) \ ,$$

the operation $s_{\gamma,\lambda}(M)$ soft-thresholds each entry of the matrix M, and the operation $\Pi_{\mathcal{K}_a}(M)$ projects each entry of M on $[0, a]$.

MCMC scheme

For $j \neq i$, we set $b_{ij} = \mathrm{e}^{\theta_{ij}}$. Notice that $b_{ij} \geq 1$. For $x = (x_1, \ldots, x_p)$,

$$f_\theta(x) = \frac{1}{Z_\theta} \exp\left(\sum_{i=1}^{p} \theta_{ii} B_0(x_i)\right) \prod_{1 \leq j < i \leq p} \left(b_{ij} \mathbb{1}_{\{x_i = x_j\}} + \mathbb{1}_{\{x_i \neq x_j\}}\right).$$

Augment the likelihood with auxiliary variables $\{\delta_{ij},\ 1 \leq j < i \leq p\}$, $\delta_{ij} \in \{0, 1\}$ such that the joint distribution of (x, δ) is given by

$$\bar{f}_\theta(x, \delta) \propto \exp\left(\sum_{i=1}^{p} \theta_{ii} B_0(x_i)\right) \times \prod_{j<i} \left(\mathbb{1}_{\{x_i = x_j\}} b_{ij} \left(1 - b_{ij}^{-1}\right)^{\delta_{ij}} b_{ij}^{\delta_{ij}-1} + \mathbb{1}_{\{x_i \neq x_j\}} 0^{\delta_{ij}} 1^{1-\delta_{ij}}\right).$$

The marginal distribution of x in this joint distribution is the same f_θ given above.

MCMC scheme

- The auxiliary variables $\{\delta_{ij},\ 1 \leq j < i \leq p\}$ are conditionally independent given $x = (x_1, \ldots, x_p)$; if $x_i \neq x_j$, then $\delta_{ij} = 0$ with probability 1. If $x_i = x_j$, then $\delta_{ij} = 1$ with probability $1 - b_{ij}^{-1}$, and $\delta_{ij} = 0$ with probability b_{ij}^{-1}.
- The auxiliary variables $\{\delta_{ij},\ 1 \leq j < i \leq p\}$ defines an undirected graph with nodes $\{1, \ldots, p\}$ where there is an edge between $i \neq j$ if $\delta_{ij} = 1$, and there is no edge otherwise.
- This graph partitions the nodes $\{1, \ldots, p\}$ into maximal clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ (a set of nodes where there is a path joining any two of them).
- Notice that $\delta_{ij} = 1$ implies $x_i = x_j$. Hence all the nodes in a given cluster holds the same value of x.

$$\bar{f}_\theta(x|\delta) \propto \prod_{k=1}^{K} \left[\exp\left(\sum_{i \in \mathcal{C}_k} \theta_{ii} B_0(x_i) \right) \prod_{j<i,(i,j)\in\mathcal{C}_k} \mathbb{1}_{\{x_i = x_j\}} \right].$$

Wolff algorithm

Given $X = (X_1, \ldots, X_p)$

1. Randomly select a node $i \in \{1, \ldots, p\}$, and set $\mathcal{C}_0 = \{i\}$.
2. Do until $\mathcal{C}_0$ can no longer grow. For each new addition j to $\mathcal{C}_0$, and for each $j' \notin \mathcal{C}_0$ such that $\theta_{jj'} > 0$, starting with $\delta_{jj'} = 0$, do the following. If $X_j = X_{j'}$, set $\delta_{jj'} = 1$ with probability $1 - \mathrm{e}^{-\theta_{jj'}}$. If $\delta_{jj'} = 1$, add j' to $\mathcal{C}_0$.
3. If $X_i = v$, randomly select $v' \in \{1, \ldots, M\} \setminus \{v\}$, and propose a new vector $\tilde{X} \in \mathsf{X}^p$, where $\tilde{X}_j = v'$ for $j \in \mathcal{C}_0$ and $\tilde{X}_j = X_j$ for $j \notin \mathcal{C}_0$. Accept $\tilde{X}$ with probability

$$1 \wedge \exp\left((B_0(v') - B_0(v)) \sum_{j \in \mathcal{C}_0} \theta_{jj} \right).$$

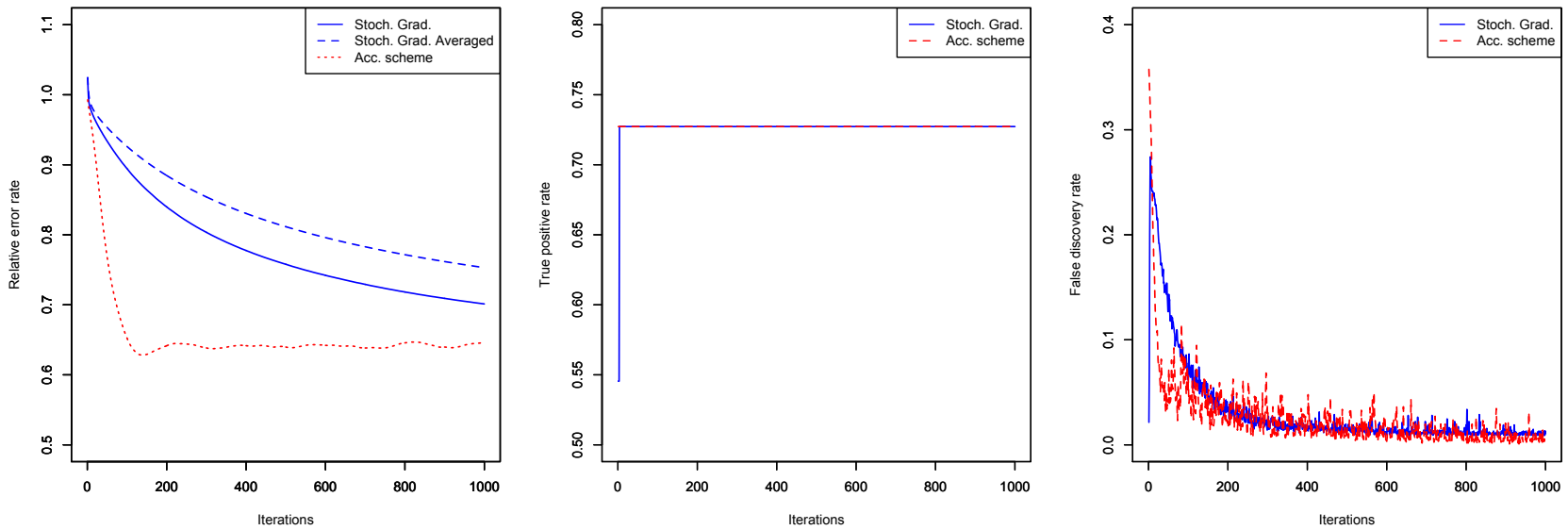

Figure : Simulation results for $p = 50$, $n = 500$ observations, 1% of off-diagonal terms, minibatch, $m_n = 100 + n$

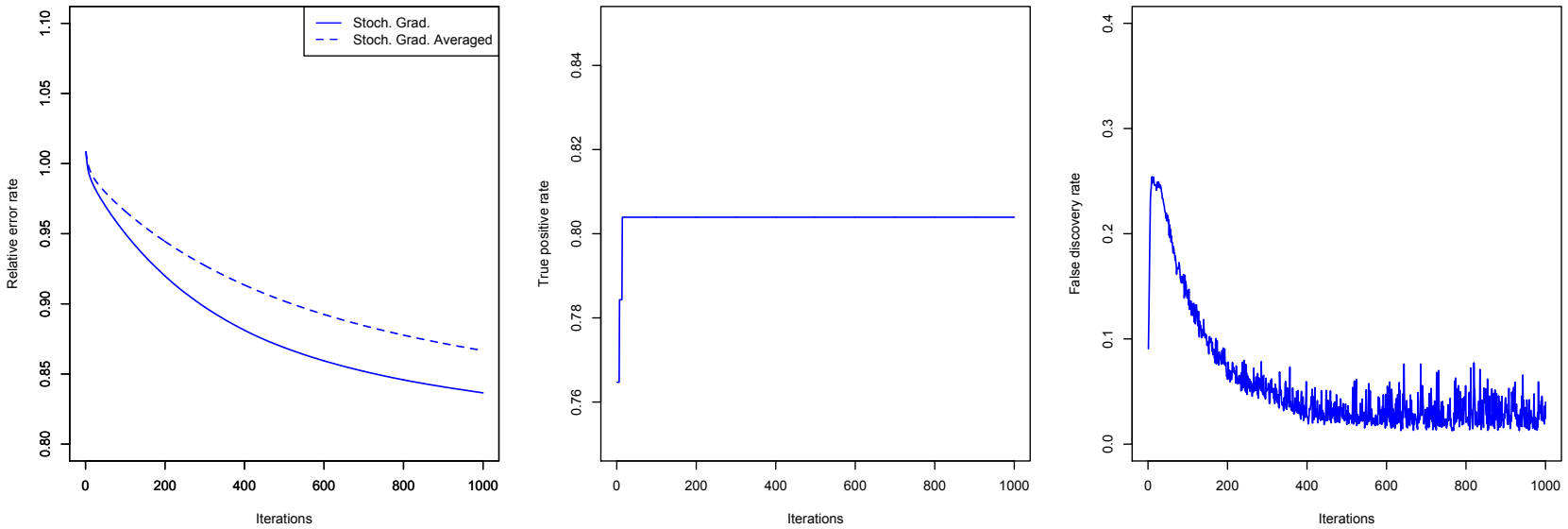

Figure : Simulation results for $p = 100$, $n = 500$ observations, 1% of off-diagonal terms, $n = 500$ observations, 1% of off-diagonal terms, minibatch, $m_n = 100 + n$

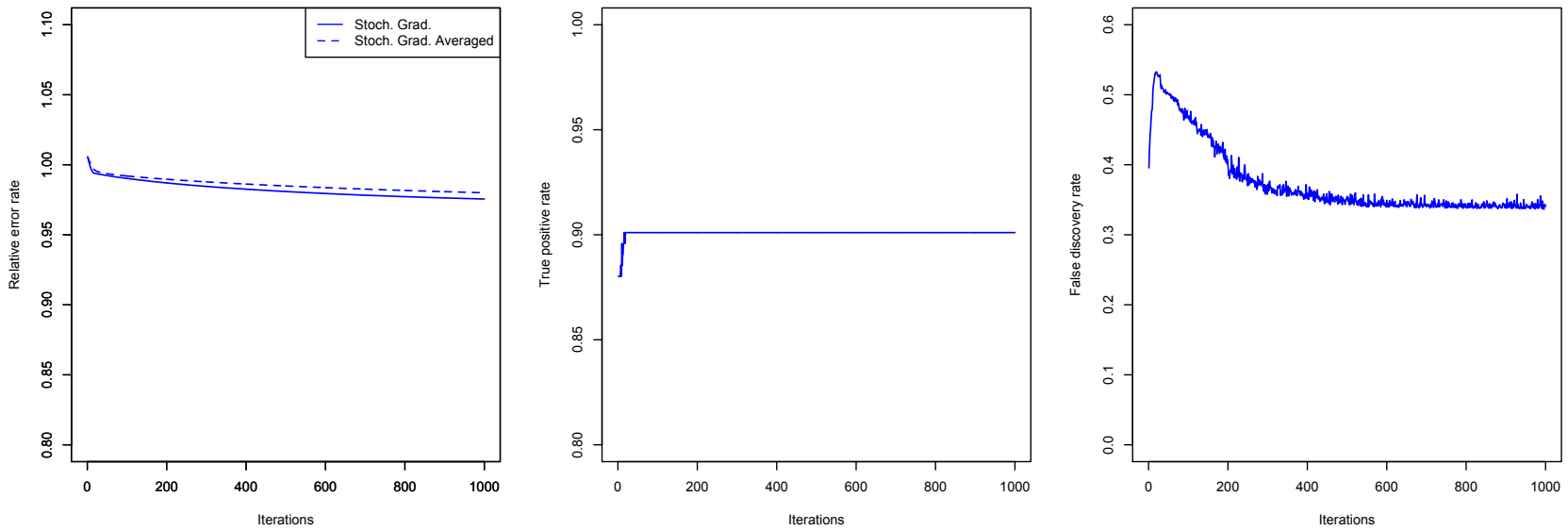

Figure : Simulation results for $p = 200$, $n = 500$ observations, 1% of off-diagonal terms, 1% of off-diagonal terms, minibatch, $m_n = 100 + n$

Take-home message

- Efficient and globally converging procedure for penalized likelihood inference in incomplete data models are available if the complete data likelihood is globally concave with convex sparsity-inducing penalty (provided that computing the proximal operator is easy)
- Stochastic Approximation and Minibatch algorithms achieve the same rate, which is $1/\sqrt{n}$ where n is the number of simulations. Minibatch algorithms are in general preferable if the computation of the proximal operator is complex.
- Thanks for your attention... and patience !